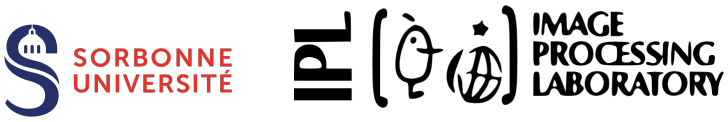

Sorbonne Université - Image Processing Lab (Universitat de Valencia)

Master thesis

Main Confounder Analysis and Application to Geosciences

Author:
H. Durand
ID: 3672141

Supervisors:
G. Camp-Valls
G. Varhando

November 2022

First of all, I would like to thank the whole team that supervised me, Gustau and Gherardo, for the kindness they showed me and the precious advice they gave me. I would particularly like to thank Gherardo who, in spite the fact we mainly worked remotely, always knew how to steer me in relevant directions while giving me great freedom in my work.

Abstract

Contents

1 Introduction

Climate models or more generally Earth System Models (ESMs) have become central to the study of climate evolution, both for the assessment of past climates and for projections of future climate. These models were among the first applications of numerical computation in the 1950s (see Platzman [1]) when the use of the "super-computers" of the time enabled the field of weather and climate prediction to experience a real boom. The structure of these models has become more complex over the last few decades, first including ocean circulation (see Manabe and Bryan [2]) in 1969, then the contribution of the radiation balance modified by human forcing linked to CO2 emissions (see Manabe and Wetherald [3]) in 1975, leading finally to the creation of the Intergovernmental Panel for Climate Change (IPCC) in 1988, whose mission is "[...] to assess, in a systematic, clear and objective manner, the scientific, technical and socio-economic information needed to improve our understanding of the risks associated with human-induced global warming [...]". The various components of the ESMs are generally modelled by systems of partial differential equations (PDEs) describing various processes such as fluid mechanics (described by the Navier-Stockes equations) or thermodynamics for modelling the ocean and atmosphere or biological and chemical processes describing marine and terrestrial ecosystems. These processes encompass spatial and temporal scales of different order, ranging from the collision between cloud particles of the order of a micron in size to the deep circulation of ocean, of the order of 1000 to 10000 km. The limited computing power of today's supercomputers does not allow the creation of models representing the entire Earth system at a sufficiently small scale to model small-scale processes such as cloud formation or the formation and circulation of plankton. Furthermore, human contributions to climate change are now widely accepted and their uncertain evolution complicates the modellers' projections.

At the same time, we have observed a great development of statistical tools allowing the analysis of increasingly complex and high dimmensional data. Although these tools have quickly become standard in some fields of scientific research (health sciences, economics, etc.), the climate sciences have for a long time remained relatively closed to them. They were mainly relying on statitical methods of the beginning of the 20th century such as Principal Components Analysis (also known as Empirical Orthogonal Functions in climate sciences), correlation analysis and linear regression. Recently, the abundance of climate data from model simulations, Earth-orbiting satellites, and in situ observations coupled with the recent advances in developping field such as Machine Learning have allowed many advances in our understanding of the earth system, putting statistical analysis back at the center of a lot of research in this field.

Of particular interest in this work, the field of causal analysis, at the crossroads of statistics and computer science, is struggling to make its entry into the climate sciences for several understandable reasons. First the chaotic and non-stationary nature of the climate makes invalidates a large number of assumptions that are often made in causal inference or discovery (e.g. stationarity, i.i.d distributed data, etc...). Second, the concept of intervention, central to Pearl's causal framework (one of the widely considered framework in causality), seems at odds with the physical

description of the climate system since it requires that it be possible to change one of the variables without affecting the rest of the system which seems to contradict the laws of conservation of energy.

That being said, it seems that causal analysis has the potential to give us a better understanding of the climate system and its interaction with ecosystems and human activity.This framework would seem, for example, to be particularly suitable for analyzing the impact of policy aiming to tackle climate change.

Thus, the nature of the studied system brings many theoretical and practical challenges. Here we focus in a specific problem at the junction of causal discovery, causal representation learning and causal inference which is the discovery of unobserved confounders from high dimensional proxy variables for unbiased estimation of causal effect.

As this work is still in its infancy, we report here the main challenges of this problem, study the main methods used in the literature in similar contexts and propose a research path to address their main shortcomings.

2 Scope

The first part of this report aims to give the reader a general understanding of what causality is and what it can contribute to our understanding of climate. In subsection 3.1 we briefly describe what are the main ideas behind causality and what it brings in addition to classical statistics. We then formalize it with a brief description of the main mathematical settings and assumptions in subsection 3.2. As this work is at the crossroad of causal discovery, causal inference and causal representation learning we will try to disambiguate these terms in **??** and finally give some examples of applications in geoscience in 3.3.

The problem of variable adjustement for causal effect estimation has been widely studied in a lot of different fields, each considering differents settings. Specifically, this is of primer importance in health science and economics to assess the effect of a treatment on an outcome (e.g. the effect of drug medication on healing or the effect of a public policy of the average level of education of a population). We first give a clear description of what we expect from variable adjustement for causal effect estimation in climate sciences in subsection 4.1 and then describe the different methodologies that, to the best of our knowledge, are the most used to tackle this problem in subsection 4.3. This leads to domain specific difficulties that we will describe in subsection **??**. We describe in section 5 our approach adress this problematic and the different difficulties that it brings. We present our main results applied to real application (see 6.1) and finally we will discuss our results and give futher path of research in section 7.

3 Context

The discovery of the causes at the source of the various phenomena observed, whether physical, chemical, biological, psychological or even social, has been central to scientific research for several centuries. In spite of this, causal language remained

neglected (or even prohibited) in mathematics throughout the first half of the 20th century with the mathematisation of Statistics and is still largely so today. We learn in school that *"Correlation does not imply causation"* and rightly so, but we do not learn to distinguish a spurious correlation due to a common cause from a direct causal link. Thus, it should be noted that geneticist Sewall Wright – geneticist from the early 20th century – developp, concomitantly with modern statists, a framework known as *Path Analysis* which is still widely used in a lot of scientific field such as biology, sociology or econometrics. This can certainly be considered as one of the first formalization of causality.

The lack of formalisation of the concept of causality makes it difficult to make causal statements such as event A causes event B. In the course of the 20th century, we have seen situations that seem aberrant today, such as the industry questioning tobacco consumption as a causal factor in lung cancer on the pretext of a simple correlation that could be due to external factors. The theoretical framework of causality being unclear, it was necessary to wait for

But in recent decades we assisted to what Judea Pearl calls the Causal Revolution (see Pearl and Mackenzie [4]) with a multiplication of theoretical developpement and formalization of what we understand by causation. Many frameworks exists but one of the most widely used in many domains is the graphical causal model framework (see Pearl [5]) where a directed graph express the causal relation between different variables. In the following section we try to give some insights on the usefullness of the causal framework.

3.1 Causality

Although time is the most obvious clue when one seeks to reason causally (causes should precedes consequences), the end of the twentieth century, with the work of Pearl, Rebane or Spirtes, saw the emergence of a formalization of the concept of causality that does not require considering time. It seems obvious that time alone cannot distinguish real causal link from spurious associations caused by some external factors. One could for example look at his barometer falling a few minutes before it starts to rain and yet, common sense will not make him think that this caused the rain. Let us, as an intoduction example, consider observations of precipitation in different regions of Europe during summer period, say in Denmark and the Mediterranean to rely on the work of Kretschmer et al. [6]. One could analyse the correlation between spatially averaged observation in this those regions and find a significant association ($r = -0.24$). A climatologist will quickly conclude that this association is spurious and that it is actually caused by a third phenomenon, namely the North Atlantic Oscillation, which in its positive phase causes particularly rainy periods in the Mediterranean and dry anomalies in northern Europe during the summer. He is actually having a causal reasoning about the phenomena he observes (which is often the case in experimental science). Reasoning that could be translated, for example, by the graph [5].

Indeed, an apparent correlation between two phenomena A and B does not necessarily imply a causal link between them, but if this is not the case, it generally implies that a third phenomenon C is a common cause of A and B (see figure

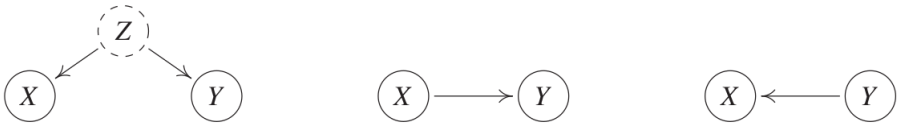

Figure 1: Causal graphs summarizing the Reinchenbach's common cause principle. If X and Y are dependent one from an other, then we are in one of this three scenario. (Left) Z is a common cause of X and Y. (Middle) X causes Y. (Right) Y causes X. From Peters et al. [7]

[1](Left)). This is entailed in the famous common cause principle of Reichenbach which can be formulated as follow

Principle 3.1 [1] *(Reichenbach's common cause principle) If two random variables X and Y are statistically dependent ($X \perp\!\!\!\perp Y$), then there exists a third variable Z that causally influences both. (As a special case, Z may coincide with either X or Y .) Furthermore, this variable Z screens X and Y from each other in the sense that given Z, they become independent, $X \perp\!\!\!\perp Y|Z$.*

It should be noted that a few exception applies to this principle, Peters et al. [7] raise three of them

1. The random variables we observe are conditioned on others
2. The random variables only appear to be dependent. For example, they may be the result of a search procedure over a large number of pairs of random variables that was run without a multiple testing correction.
3. Similarly, both random variables may inherit a time dependence and follow a simple physical law, such as exponential growth. The variables then look as if they depend on each other, but because the i.i.d. assumption is violated, there is no justification of applying a standard independence test.

Another central notion in the causal analysis of phenomena is that of independent mechanisms. Let's say consider we are given a sample of city observations with their average annual temperature and altitude from which we can estimate a joint distribution $p(a, t)$ of the altitude A and the temperature T. Here the common sense tells us that the altitude should cause the temperature ($A \to T$) and not the opposit. This idea seems to come from the fact that by changing the altitude of the city (say by make it fly on a flying platform) then the temperature of the city would drop. On the other hand, if the average temperature of the city would increase (say due to global warming) this would not affect the altitude of the city. This is actually based on the concept of **intervention**: What would happen if I intervened on this variable A? Would it affect this variable B?

[1]From Peters et al. [7]

One could wonder if this could be expressed in a probabilistic manner. For this, let's consider the two following factorization of the joint distribution

$$\begin{aligned} p(a,t) &= p(a|t)p(t) \\ &= p(t|a)p(a) \end{aligned}$$

We could argue that the second factorization seems more relevant as it is possible to imagine a meteorological mechanism $p(t|a)$ describing how temperature is affected by altitude (threw pressure levels, winds and other meteorological phenomena) which is independant of the distribution of the altitude $p(a)$. In contrast, it is way more complicated to think about a mechanism $p(a|t)$ describing how altitude is related to temperature independently of the distribution of the temeprature $p(t)$. One could reformulate it in the following way : if $A \to T$ then the distribution $p(a)$ and the mechanism $p(t|a)$, describing how T is affected by A, should be independent. This is entailed in the Independent Mechanism Principle.

Principle 3.2 [2] *(Independent mechanisms) The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other. In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other conditional distributions. In case we have only two variables, this reduces to an independence between the cause distribution and the mechanism producing the effect distribution.*

Causal analysis is therefore a tool that makes it possible, with the help of hypotheses concerning the various mechanisms studied, to answer questions that statistics alone cannot address. We formalise the main hypotheses that allow us to reason in a causal manner in the following section.

3.2 Formalization of causality

3.2.1 Structural Causal Models and interventions

As mentioned earlier, causal analysis is based on probabilistic modelling but adds additional information about the relationship between the different variables which is usually represented by a directed graph where the arrows represent causal associations. In this work we will mainly consider Structural Causal Models (SCMs), defined as follow

Definition 3.1 [3] *(Structural causal models)*
A structural causal model (SCM) $\mathcal{C} := (S, P_N)$ consists of a collection S of d (structural) assignments

$$X_i := f_j(PA_j, N_j), \qquad j = 1, ..., d \tag{1}$$

[2] From Peters et al. [7]
[3] From Peters et al. [7]

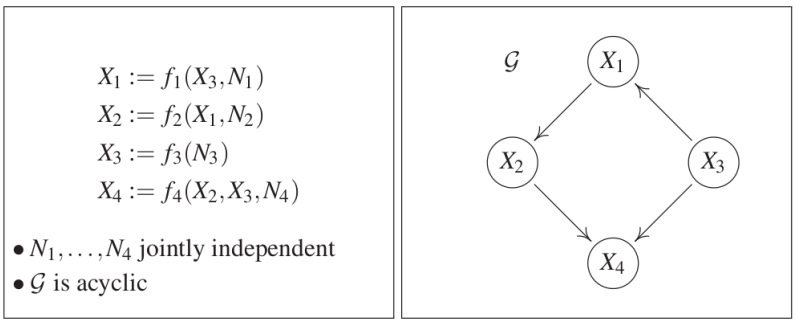

Figure 2: Example of an SCM (left) with corresponding graph (right). From Peters et al. [7]

where $PA_j \subset X_1, ..., X_d$ X_j are called parents of X_j ; and a joint distribution $P_N = P_{N_1,...,N_d}$ over the noise variables, which we require to be jointly independent; that is, P_N is a product distribution. The graph $\mathcal{G}$ of an SCM is obtained by creating one vertex for each X_j and drawing directed edges from each parent in PA_j to X_j , that is, from each variable X_k occurring on the right-hand side of equation (1) to X_j (see figure [2]). We henceforth assume this graph to be acyclic. We sometimes call the elements of PA_j not only parents but also direct causes of X_j , and we call X_j a direct effect of each of its direct causes. SCMs are also called (nonlinear) Structural Equation Models (SEMs).

One of the main advantages of SCMs over classical probabilistic models is that they entail **intervention** distributions in addition of the observational distribution.

Proposition 3.1 *(Entailed distributions)*
An SCM $\mathcal{C}$ defines a unique distribution over the variables $X = (X_1, ..., X_d)$ such that $X_j = f_j(PA_j, N_j)$, in distribution, for $j = 1, ..., d$. We refer to it as the entailed distribution $P_X^{\mathcal{C}}$ and sometimes write P_X.

Definition 3.2 *(Atomic intervention) Consider an SCM $\mathcal{C} := (S, P_N)$ and its entailed distribution $P_{\boldsymbol{X}}^{\mathcal{C}}$. We replace one (or several) of the structural assignments to obtain a new SCM $\tilde{\mathcal{C}}$. Assume that we replace the assignment for X_k by a real value a we then call the entailed distribution of the new SCM an atomic intervention distribution and say that the variables whose structural assignment we have replaced have been intervened on. We denote the new distribution by $P_{\boldsymbol{X}}^{\tilde{\mathcal{C}}} = P_{\boldsymbol{X}}^{\mathcal{C};do(X_k):=a}$.*

We refer the reader to Peters et al. [7] (Definition 6.8) for a more complete definition of intervention distribution generalized to any assignement such that $X_k := \tilde{f}(\tilde{PA}_k, \tilde{N}_k)$.

The markov property is central in causality in most of the methodologies relies on it. However, as it is show in Peters et al. [7] (Proposition 6.31), assuming that the observed distribution comes from an underlying SCM is sufficient to proove that Markov property is fullfilled.

Definition 3.3 [4] *(Markov property) Given a DAG $\mathcal{G}$ and a joint distribution $P_{\boldsymbol{X}}$, this distribution is said to satisfy*

[4] From [7]

1. *the global Markov property with respect to the DAG $\mathcal{G}$ if*

$$\boldsymbol{A} \perp\!\!\!\perp_{\mathcal{G}} \boldsymbol{B} | \boldsymbol{C} \Rightarrow \boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B}$$

   for all disjoint vertex sets A, B, C.

2. *the local Markov property with respect to the DAG $\mathcal{G}$if each variable is independent of its non-descendants given its parents, and*

3. *the Markov factorization property with respect to the DAG $\mathcal{G}$ if*

$$p(x) = p(x_1, ..., x_d) = \prod_{j=1}^{d} (x_j | pa_j^{\mathcal{G}}).$$

   For this last property, we have to assume that $P_{\boldsymbol{X}}$ has a density p; the factors in the product are referred to as causal Markov kernels describing the conditional distributions $P_{X_j|PA^{\mathcal{G}}}$.

Theorem 1 [5] *(Equivalence of Markov properties) If $P_{\boldsymbol{X}}$ has a density p, then all Markov properties in Definition [3.3] are equivalent.*

We refer the reader to Theorem 3.27 from Lauritzen [8] for a detailed proof.

3.2.2 Confounding

As will be detailed in section [4], the notion of confounder is central to our problem. Comming back to our example of the North Atlantic Oscillation driving precipitation in Europe, one could look at the distribution (see [3.2]) of the precipitation in Denmark when intervening on the precipitation in the Medieteranean. It seems obvious that the intervened distribution would be the same as the observed distribution. However, one would observed that modifying the values of the precipitation in the Mediteranean would affect the conditional ditribution of the precipitation in Denmark. This is known to be a confounding effect. And in this case the confounder is known to be the North Atlantic Oscillation.

Definition 3.4 *(Confounding) Consider an SCM $\mathcal{C}$ over nodes V with a directed path from X to Y , $X, Y \in V$. The causal effect from X to Y is called confounded if $p^{\mathcal{C};do(X:=x)}(y) \neq p^{\mathcal{C}}(y|x)$. Otherwise, the causal effect is called "unconfounded."*

When aiming to estimate the causal effect of one variable on another in the presence of potential confounders, it is important to adjust one's estimate to these so as not to obtain a biased estimate. The notion of Valid Adjustement Set consider the idea that a set of variables is sufficient when adjusting to get an unbiased estimate.

Definition 3.5 *(Valid Adjustement Set) Consider an SCM $\mathcal{C}$ over nodes V and let $Y \notin PA_X$ (otherwise we have $p^{\mathcal{C};do(X:=x)}(y) = p^{\mathcal{C}}(y)$). We call a set $Z \subset V\{X, Y\}$ a valid adjustment set for the ordered pair (X, Y) if*

$$p^{\mathcal{C};do(X:=x)} = \int_z p^{\mathcal{C}}(y|x, z) p^{\mathcal{C}}(z) dz$$

If Z is in $\mathbb{Z}$, we have $p^{\mathcal{C};do(X:=x)} = \sum_z p^{\mathcal{C}}(y|x, z) p^{\mathcal{C}}(z)$

[5] From Peters et al. [7]

Theorem 2 *(Valid adjustment sets) Consider an SCM over variables $\boldsymbol{X}$ with $X, Y \in \boldsymbol{X}$ and $Y \notin PA_{\boldsymbol{X}}$. Then, the following three statements are true.*

1. *"parent adjustment":*

$$Z := PA_{\boldsymbol{X}}$$

   is a valid adjustement set for (X, Y).

2. *"backdoor criterion": Any $Z \subset \boldsymbol{X}$ $\{X, Y\}$ with*

   - *Z contains no descendant of X and*
   - *Z blocks all paths from X to Y entering X through the backdoor ($X \leftarrow ...$)*

   is a valid adjustment set for (X, Y).

3. *"toward necessity": Any $Z \subset \boldsymbol{X}$ $\{X, Y\}$ with*

   - *Z contains no descendant of any node on a directed path from X to Y (except for descendants of X that are not on a directed path from X to Y) and*
   - *Z blocks all non-directed paths from X to Y*

   is a valid adjustment set for (X, Y).

We refer the reader to Peters et al. [7] (Proposition 6.41) for detailed explanations about *backdoor path* and proofs.

need description of faithfulness assumption

As previously stated, there is different frameworks to tackle causal questions

3.2.3 Granger causality

Until now we have not had recourse to the notion of time to arrive at causal conclusions and this is one of the strengths of the theoretical framework developed by Pearl, Spirtes and Glymour at the end of the last century. Nevertheless, it seems obvious that temporal information plays a role in the notion of causality and it would be a pity to deprive ourselves of it. A straightforward methodology, considering time series, relying on the fact that causes should precede effects (also known as the time order assumption) as been developped by Clive Granger in its early work Granger [9]. Granger's idea was that a variable X should be considered as a cause of an other variable Y if X contains unique information about the future of Y. This framework is known as Granger Causality (GC) and most of the time we reformulate this idea of unique information contained X on the future of Y by the predictive power of the past of X on the future of Y. Let's consider the very simple bivariate linear case.

We consider two time series $(X_t)_{t\in\mathbb{Z}}$ and $(Y_t)_{t\in\mathbb{Z}}$ and the two following autoregressive models for $(Y_t)_{t\in\mathbb{Z}}$, named *restricted* and *full* to highlight the fact that the first

is modeling $(Y_t)_{t\in\mathbb{Z}}$ only considering it's own past and that the second also consider the past of $(X_t)_{t\in\mathbb{Z}}$. This gives us

$$Y_t^{res} = \sum_{i=1}^{\tau} a_i^{res} Y_{t-i} + \epsilon_{res} \tag{2}$$

$$Y_t^{full} = \sum_{i=1}^{\tau} a_i^{full} Y_{t-i} + \sum_{i=1}^{\tau} b_i^{full} X_{t-i} + \epsilon_{full} \tag{3}$$

Where a^{res}, a^{full} and b^{full} are the regression coefficient of the models, ϵ_{res} and ϵ_{full} are random noise considered as independent of X_t and Y_t and τ is the considered time lag. We can now use a statistical test to compare both residuals of the *restricted* and *full* models to assess wether or not $(X_t)_{t\in\mathbb{Z}}$ is Granger causing $(Y_t)_{t\in\mathbb{Z}}$. We typically use a F-test

$$F = \frac{RSS_{res} - RSS_{full}/(r-s)}{RSS_{full}/(T-r)} \tag{4}$$

where RSS_{res} and RSS_{full} are the residual sum of squares for the *restricted* and *full* models. Using this test, we reject the null hypothesis stating that $(Y_t)_{t\in\mathbb{Z}}$ is not Granger caused by $(X_t)_{t\in\mathbb{Z}}$ if the observed test statistic F exceeds the $(1-\alpha)\%$ quantile of an F-distribution with $r-s$ and $T-r$ degrees of freedom.

It is important to raise that, restricted to the bivariate case, granger causality can be severly misleading due for example to potential confounders. Let's for example consider the case where a third time serie $(Z_t)_{t\in\mathbb{Z}}$ is granger causing $(Y_t)_{t\in\mathbb{Z}}$ with a lag of two and granger causing $(X_t)_{t\in\mathbb{Z}}$ with a lag of one

$$\begin{aligned} Z_t &= a_1 Z_{t-1} + \epsilon_Z \\ X_t &= a_2 Z_{t-1} + \epsilon_X \\ Y_t &= a_3 Z_{t-2} + \epsilon_Y \end{aligned}$$

Then using bivariate models considering only $(X_t)_{t\in\mathbb{Z}}$ and $(Y_t)_{t\in\mathbb{Z}}$ for testing the null hypothesis $\mathcal{H}_0 : (X_t)_{t\in\mathbb{Z}} \not\to (Y_t)_{t\in\mathbb{Z}}$ will lead to uncorrect result by rejecting the null hypothesis and thus inferring that $(X_t)_{t\in\mathbb{Z}}$ causes $(Y_t)_{t\in\mathbb{Z}}$ when it is not. In fact, in that case, considering only $(X_t)_{t\in\mathbb{Z}}$, $(Y_t)_{t\in\mathbb{Z}}$, the past of the serie $(X_t)_{t\in\mathbb{Z}}$ is containing unique information about the future of $(Y_t)_{t\in\mathbb{Z}}$. Thus it is important to conditioned our regression on potential (and potentially multivariate) confounders $(Z_t)_{t\in\mathbb{Z}}$. We then consider the following *restricted* and *full* models

$$Y_t^{res} = \sum_{i=1}^{\tau} a_i^{res} Y_{t-i} + \sum_{i=1}^{\tau} B_i^{res} Z_{t-i} + \epsilon_{res} \tag{5}$$

$$Y_t^{full} = \sum_{i=1}^{\tau} a_i^{full} Y_{t-i} + \sum_{i=1}^{\tau} B_i^{full} Z_{t-i} + \sum_{i=1}^{\tau} c_i^{full} X_{t-i} + \epsilon_{full} \tag{6}$$

Where B^{res} and B^{full} are matrixes of the regression parameters of $(Z_t)_{t\in\mathbb{Z}}$ on $(Y_t)_{t\in\mathbb{Z}}$.

The Granger Causality framework can be widely extended for example by considering non-linear relationships between the variables with for example kernel regression or neural networks regressors, we refer the reader to Shojaie and Fox [10] for futher developpement of nonlinear GC.

It seems also important to raise the fact that GC does not directly mesure causality relations but assess wether or not a serie X is predictive of another serie Y.

That being said, an important theorem directly relate Granger Causality to causality.

Theorem 3 *(Granger causality justification) Consider an SCM without instantaneous effects for the time series $(X_t)_{t\in\mathbb{Z}}$ such that the induced joint distribution is faithful with respect to the corresponding full time graph. Then the summary graph has an arrow from X^j to X^k if and only if there exists a $t \in Z$ such that*

$$X_t^k \not\perp\!\!\!\perp X_{past(t)}^j | X_{past(t)}^{-j} \tag{7}$$

We refer the reader to the book Peters et al. [7] Appendix C.14 for the proof of this theorem.

This imply that if there is no instantaneous effect in the series $(X_t)_{t\in\mathbb{Z}}$ and that the SCM if faithfull then X^j is a direct cause of X^k if and only if X_t^k is not independent of the past X_t^j knowing the past of all other variables contained in $(X_t)_{t\in\mathbb{Z}}$.

It has also be shown that GC is closely related to Conditional Mutual Information (CMI) and thus not only considering predictability power of the considered variables but using Information Theory to asses wether or not the past of $(X_t)_{t\in\mathbb{Z}}$ is containing unique information about the future of $(Y_t)_{t\in\mathbb{Z}}$. We refer the reader to Amblard and Michel [11] for futher developpement of this idea.

3.3 Causality in geosciences

With the important theoretical developments of the last few years, causal methods have become an essential part of some scientific disciplines, especially in health sciences and in econometrics. These methods are particularly suitable for assessing the effect of drug treatments on recovery or the effect of a public policy on the economic health of a country. However, they have taken longer to emerge in the climate sciences and geosciences in general, for both methodological and philosophical reasons.

Indeed, the climate, due to its chaotic and non-stationary nature, is outside the theoretical framework of the first methods developed. However, during the last decade, large number of researchers have been interested in applying causal methods to climate science, leading to the emergence of new methods particularly suited to this framework.

The PC algorithm and its extensions (PCMCI, LPCMCI, etc...) is particularly well suited to build graphs from time series. In Runge et al. [12], authors decompose

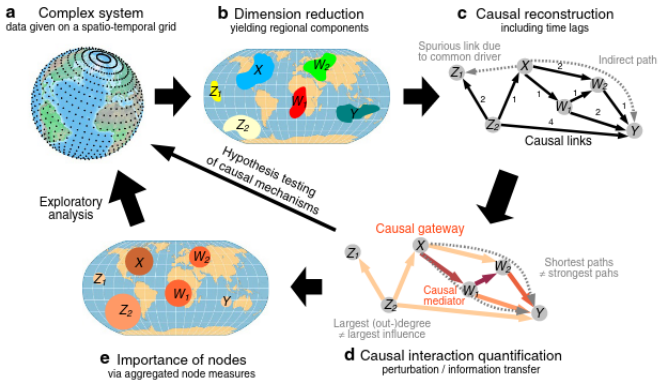

Figure 3: Schematic representation of our proposed approach illustrated for the complex system Earth.
(a) Climatological variables such as sea-level pressure or surface temperature are typically provided as time series at locations on a regular grid. (b) In the first step of our approach, a Varimax-dimension reduction yields a small set of regional components (here denoted by X, Y, y) representing climatological subprocesses with corresponding time series. (c) In the second step of our approach, this smaller set of variables allows for a reconstruction of the causal network (black links, labels denote time lags). An important pitfall in non-causal networks (as constructed from pairwise association measures such as cross-correlation) is that links can be spurious due to a common driving by another process or due to transitivity effects leading to indirect paths (grey dashed arrows in c). Perturbations cannot propagate along common-driver links (for example, between X and Z1) affecting network measures like the degree. Further, indirect paths such as from X to Y affect shortest path lengths in non-causal networks. (d) In the third step of our approach, the aim is to directly quantify the causal effect between pairs of components based on the causal network (Tigramite approach) and detect through which components and how much the causal effect is mediated. In the linear framework studied here this can be achieved by causal effect measures based on suitable link weights, where the weight of a link, for example X-W1, indicates the causal effect of a 1 s.d. perturbation in X on W1 (see Fig. 3b for a formula relating link weights to causal effects). Binary causal networks do not properly account for different link strengths which affects classical network measures (grey highlights in d). This analysis can be used to test specific hypotheses, but also to construct aggregate node measures (e) to identify components with high cumulative causal effect either as sources (causal gateways) or as intermediate nodes on causal pathways (causal mediators).

the climate systems in differents modes of variability using a Varimax rotation on the extracted principal components (see 4.3.2) and reconstruct a causal graph using the PC algorithm leading to teleconnection[6] pattern discoveries. Applied to a data set of atmospheric dynamic, the method yelds to insights into the Pacific–Indian Ocean interaction relevant for monsoonal dynamics.

In their paper, aithors of Runge et al. [13] highlights the potential applications of discovery and causal inference methods in the geosciences but also the specific challenges in this field. One of them being the discovery of features causaly relevant. In fact the causal discovery methodologies can not be apply to low scale observation for computational reason and for interpretability issues. But at the same time, climate variabilities extracted by tradictional statistical tools (as PCA) are not necessarily the most suitable for causal discovery as their are agnostic of the causal problematic. It is with this in mind that we develop in the following an approach that seeks to discover relevant climate variabilities for causal effect estimation.

[6]Teleconnections are statistical association between different climates variabilities. Authors of Runge et al. [12] claim for a need to go over statistical association within a causal framework leading allowing potential intervention in the model

4 Problematic

4.1 Problem description

So far, we have assumed that all the variables of interest are observed (except for noise). In practice, this is rarely the case and it is therefore necessary to have methods leading to unbiased estimates of the parameters of interest in our causal model. As described in subsection 3.2, a bias may appear in our estimation of the causal effect of the variable X on the variable Y if we are not controlling for the potential confounders. This idea is clearly entailed in the famous Simpson's paradox (see Simpson [14]). This *"refers to the phenomenon whereby an event C increases the probability of E in a given population p and, at the same time, decreases the probability of E in every subpopulation of p"* (Pearl [5]).

An interesting example of this "paradox" is the well known kidney stone example from medical study. We are interested in determining which treatment between treatment A (medication) and treatment B (surgical intervention) is the most effective to treat kidney stones. Let's first have a look at some observational data

	Treatment A	Treatment B
Recovery rate	78% (273/350)	83% (289/350)

From those data it may appear that treatment B is preferable. But as treatment A is clearly less invasive as treament B, medication is in general prefered to treat patient with small kidney stone when treatment A is prefered for those having large kidney stones. Splitting the patient in those two groups, *Small kidney stiones* and *Large kidney stones* patients we get the following recovery rates

	Treatment A	Treatment B
Small kidney stones	93% (81/87)	87% (234/270)
Large kidney stones	73% (192/263)	69% (55/80)
Both	93% (81/87)	87% (234/270)

When looking at the recovery rates of treatment A and B for small and large stones individually, we get the opposite conclusion, in both cases treatment B should be prefered to maximize the recovery rate. This "paradox" comes from the fact that kidney stones sizes have also have a big influence on the recovery rate and not only the decision of Treatment. It should be thus considered as a confounding factor and we should control for this variable estimate the treatment effect of medication or surgical intervention on the recovery of kidney stones. The same effect have, for example, been observed in the fatality rate of covid in different countries (see [15]) where it has for example been observed that Italy had a higher survival rate but when controlling on on demographic informations such as age the opposite conclusion can be made (comming from the fact that chinese population is older and that age a a strong positive correlation with fatality rate).

This phenomenon is not restricted to categorical variables and can therefore be found in the statistical analysis of climate. An interesting example comming from the climate science community is the confounding bias of the North Atlantic Oscillation

Figure 4: Considered causal network showing the causal association between the North Atlantic Oscillation (NAO), precipitation in Denmark (DK) and in the Mediterranean region (MED). *From Kretschmer et al. [6]*

(NAO) on precipitation in Europe. The NAO is a well studied climate phenomenon over the North Atlantic Ocean of fluctuations in the difference of atmospheric pressure at sea level between the Icelandinc Low [7] and the Azores High[8]. It is one of the most important climate flucatuations in North Atlantic and is thus strongly related with climate in western Europe and Eastern America. Authors of Kretschmer et al. [6] proposed for example to use a causal framework to study the confounding bias that NAO imply in the estimation of the association between precipitation in mediterranean region and Denmark during summer. Summer precipitation in these two regions is negatively correlated (r=-0.24)[9], but as the authors point out, climate scientists will generally agree that this correlation does not actually imply a direct causal link and should therefore not be considered a teleconnection. In fact, it has been widely studied that the positive phases of the summer NAO (SNAO) are characterized by drought anomalies in northern Europe but concurrently by particularly wet conditions in the Mediterranean region. Thus, SNAO should be considered as a common driver (or confounder) between these two features and should be taken into account when modelling their interaction.

Considering this *expert knowlegde* (summarized in figure [4]) about the association between those three features, knowing the Summer North Atlantic Oscillation ($SNAO$) and precipitation in Denmark (DK) and in the Mediterranean region (MED), and assuming linear dependence and gaussian noises we can test the hypopthesis of no direct causal link between DK and MED. As it as be shown in Baba et al. [16], considering multivariate normal distribution, testing for conditional independence is equivalent as testing for null partial correlation. We can thus esti-

[7]Icelandic Low is a semi-permanent centre of low atmospheric pressure found between Iceland and southern Greenland

[8]Azores High is a large subtropical semi-permanent centre of high atmospheric pressure typically found south of the Azores in the Atlantic Ocean

[9]Authors considered summer-mean data (June–August) of precipitation in Denmark (DK; 50°–60°N, 2°–15°E) and the Mediterranean (MED; 36–41°N, 10°–30°E) provided by the NCEP reanalyses (https://psl.noaa.gov/data/gridded/data.ncep.reanalysis.html), and an index of the NAO provided by NOAA (https://psl.noaa.gov/data/climateindices/list).

mate the partial correlation between DK and MED by estimating the correlation between the residuals of the linear regression of MED knowing NAO and DK knowing $SNAO$, where the regression coefficient $\hat{\beta}_{DK}$ and $\hat{\beta}_{MED}$ can be estimated with the Least Squares method, giving

$$\begin{aligned} \hat{\epsilon}_{MED} &= \|MED - NAO\hat{\beta}_{MED}\|_2^2 \\ \hat{\epsilon}_{DK} &= \|DK - NAO\hat{\beta}_{DK}\|_2^2 \\ \rho_{MED\,DK.NAO} &= \text{corr}(\hat{\epsilon}_{MED}, \hat{\epsilon}_{DK}) \end{aligned}$$

where $\hat{\epsilon}_{DK}$ and $\hat{\epsilon}_{MED}$ the respective residuals of both regressions.

This leads to an estimated partial correlation $\rho_{MED\,DK.NAO}$ of 0.01. It appears that our hypothesized causal network seems to be correct and that DK and MED have no direct causal association, which seems consistent with our climatic understanding of these two phenomena.

It is easy in this context to assess the hypothesis that there is no causal link between those two features since we have a relatively clear understanding of the main climate mechanisms that generate them. Unfortunately, in many contexts we do not have this knowledge or it involves complex models that are much more difficult to interpret. For example, let us imagine that we do not have the expert knowledge that NAO is the main common driver of MED and DK. This makes it much more complicated to evaluate our hypothesis and raises many questions about the method to be used and the data that one should potentially include in the modelling. The climate system involves complex interrelationships between a large number of phenomena at a variety of spatial and temporal scales and controlling all of these features poses different difficulties both from a statistical (high dimensionality) and computational (algorhythmic complexity) perspective.

One could imagine that field knowledge may gives us insights about the features that we should includes in our modelling in order to reduce the bias of our causal estimate. Returning to our initial problem of estimating the causal association between DK and MED, we can first consider that climate knowledge allows us to know that sea level pressure is likely to induce a bias in our estimation without additional spatial information. We therefore need to condition our estimate on all sea level pressure observations with the highest possible spatial resolution. Since we are considering three-monthly time averages only over the summer period we are in a typical high-dimensional case (the spatial resolution of our data is much higher than the temporal resolution). A common approach to tackle this problem would then be to use a regression method well-suited for high-dimensional problem such as Lasso or Ridge regression to estimate the residuals of DK and MED regressed on sea surface pressure observations (HGT[10]). We believe that this method, although effective, suffers from a lack of interpetability and hardly allows a better understanding of the teleconnections between the different studied phenomena. Indeed, one could argue that it could be of great interest to discover what are the main climate oscillations that confound our two signals when estimating their association. To make it more

[10]The North Atlantic Oscillation considered in this study is computed using a Varimax PCA to monthly mean standardized 500-mb height (known as *hgt*) anomalies obtained from the CDAS in the analysis region 20°N-90°N between January 1950 and December 2000.

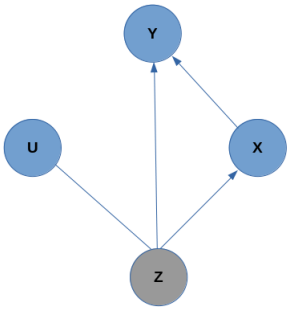

Figure 5: Causal Network representing our learning problem.

clear, in the precedent example, it might be interesting to be able to discover a signal strongly correlated with the NAO when estimating the partial correlation between DK and MED given HGT.

This brings us to our main research question : What are the main confounding signals of the two phenomena whose causal association we seek to estimate? In the following sections we will formalize it and give an overview of the related methods that we consider relevant and of interest to tackle this problem.

4.2 Problem formalization

Let us consider two time series $(X_t)_{t\in\mathbb{Z}} \in \mathbb{R}^{\mathbb{T}}$ and $(Y_t)_{t\in\mathbb{Z}} \in \mathbb{R}^{\mathbb{T}}$ and spatio-temporal proxy variables $(U_t)_{t\in\mathbb{Z}} \in \mathbb{R}^{\mathbb{T}\times}$ that potentially confound the causal relation between X and Y and are potentially high-dimensional ($d \gg T$). We aim to discover a mapping function w such that the variable $Z_t = w(U_t) \in \mathbb{R}^{T\times p}$ (with $p \ll d$) are a set of confounding variables which lead to an unbaised estimate of the causal effect of X on Y when controlling for them. This formulation is entailed in the causal graph in figure [5] where directed arrows represent causal association and undirected arrows for statistical association that are not necessarly causal.

We decide not to choose a causal direction for the relation between latent variables $(Z_t)_{t\in\mathbb{Z}}$ and $(U_t)_{t\in\mathbb{Z}}$ because it is a rather philosophical question wether the low-scale processes give rise to large-scale phenomena (which could be considered as a physical approach) or the other way around (more of a probabilistic approach) and we would like to keep the latitude to consider both. Also, while the causal links between Z, X and Y can potentially involve time, the w mapping is only in the spatial domain as we seek to approximate existing approaches in teleconnection discovery where climate oscillations generally involve spatial mapping only (e.g. component analysis in the spatial domain for the ENSO and NAO indexes).

Therefore, we formulate our learning problem as follow : Considering the causal graph in figure [5] we are interested in estimating the *Average Causal Effect* of X on Y, knowing $\mathbb{E}^{\mathcal{C};do(X):=x}[Y]$, by learning a *valid adjustment set* Z as a mapping w of proxy variables U.

4.3 Related work

4.3.1 Lasso

Proposed in 1996 by Robert Tibshirani in [17], the LASSO regression is a well studied estension of the Linear Least Squares regression that deals with high-dimensional predictors with sparse regression coeffcients. This a widely used methodology in health sciences and econometrics for covariates selection when one is interested in estamating the average treatment effect of, for example, medication on recovery rate or a policy on an outcome of interest in the presence of potential confounders. As stated in Koch et al. [18], in the presence of a large number of covariates relative to the number of samples, linear least squares regression may lead to strong overfitting when estimating the causal effect adjusted on all the potential covariates. Then, one could use the a LASSO regressions on both the treatment and the outcome from the covariates in order to select the potential confounders (the confounders would then be the intersection of the covariates selected from both regressions).

Our main concern regarding this methodology is that it doesn't seems appropriate to only select a few covariates when seeking to discover climate oscillations, we would rather like to aggregate them as it is done in a Principal Component Analysis (we may impose some sort of sparsity in the aggregation like it is done with rotated Varimax PCA). Also, it has be shown that for estimation purpose, Lasso regression is not the more appropriate approach and one should rather use an adaptative approach as it hs been done in Shortreed and Ertefaie [19].

4.3.2 Principal Component Analysis

Proposed by Karl Pearson in 1901 as a statistical formulation of the Pricipal Axes theorem, Principal Component Analysis (PCA) is a widely studied methods. It can briefly be defined as an orthogonal linear transformation of the data where the variance of the data is ordered from the greates on the projection axes. This methods, known as Empirical Orthogonal Function (EOF), have been extensively used in the climate community. This is a really powerfull tool to analyses climate variability as it is able to condense the information of high dimensional data sets and thus allows to extract the main climate variabilities.

A commonly used extension of PCA in climate sciences is the rotated PCA (RPCA) where an additional rotation is applied on the principal axes first extracted with PCA aiming to get a more interpretable of our data. The Varimax rotation developped in 1958 by psychologist Henry Felix Kaizer is certainly the most widely used rotation by climate scientist. It aims at finding a transformation for which *"each factor has a small number of large loadings and a large number of zero (or small) loadings"*[11]. This tends to generate more interpretable components as each extracted component represents a small number of original variable and can thus be considered as a sparser version of the PCA. Formally, it find the linear combination

[11] From Abdi [20]

of the original variables that maximized the variance of the loadings

$$R = \text{argmax}_R \left\{ \frac{1}{p} \sum_{j=1}^{k} \sum_{i=1}^{p} (\Lambda R)_{ij}^4 - \frac{1}{p} \sum_{j=1}^{k} \left(\sum_{i=1}^{p} (\Lambda R)_{ij}^2 \right)^2 \right\} \tag{8}$$

where Λ are the original components and R the rotation matrix.

The main drawback of PCA and its extension is that it is *outcome agnostic*, meaning that we are not necessarly extracting components (in our case climate oscillations) that are predictive or even associated with the phenomena of interest.

4.3.3 Othogonal Partial Least Squares Regression

A rather interesting approach when aiming to discover a latent representation $Z \in \mathbb{R}^{p \times N}$ highly predictive of a certain outcome $Y \in \mathbb{R}^{m \times N}$ from high dimensional data $X \in \mathbb{R}^{n \times N}$ is the one of Orthogonal Partial Least Squares. It aims to discover a projection matrix U that is highly predictive of Y by minimizing a Least Squares loss

$$\mathcal{L}(U, W) = \|Y - WU^T X\|_F^2 \tag{9}$$

where $\|.\|_F$ denotes the frobenius norm. Authors of Arenas-García and Gómez-Verdejo [21] show that it can be formulated as an eigen value decomposition problem.

Derivating the loss with regard to U we get

$$\frac{\partial \mathcal{L}(U, W)}{\partial U} = -2C_{XY}W + 2C_{XX}UW^TW = 0 \tag{10}$$

$$\Leftrightarrow U = C_{XX}^{-1} C_{XY} (W^T W)^{-1} \tag{11}$$

where C_{AB} is the covariance matrix of A and B up to a scalar mutliplication. Injecting [10] in our loss function and after some algebraic manipulation we get

$$\mathcal{L}(W) = Tr(C_{YY}) - Tr((W^TW)^{-1} W^T C_{XY}^T C_{XX}^{-1} C_{XY} W) \tag{12}$$

And as $Tr(C_{YY})$ is constant regarding W we have that our learning problem can be reformulated as follow

$$\begin{aligned} \max_U \quad & Tr(X^T C_{XY}^T C_{XX}^{-1} C_{XY} W \\ \text{s.t.} \quad & W^T W = I \end{aligned} \tag{13}$$

whose solution is obtained with a standard eigen value decompostion

$$C_{XY}^T C_{XX}^{-} 1 C_{XY} w = \lambda w \tag{14}$$

This method is a powwerfull tool to extract components that are relevant for predictive purpose. Also, as its formulation is rather simple, it can easily be extended to kernelized or sparse version as it is proposed in Arenas-García and Gómez-Verdejo [21].

Our main concern considering this approach is that (considering tha causal learning problem described in section 4) a non confounding variable that is strongly related to X (but not to Y) will be necessarly associated to Y through X (because $Z \to X \to Y$) thus PLS might extract this variable as a confounder when it is not.

4.3.4 Granger PCA

In subsection [4.3.2] we stated that one of the drawback of PCA was that it is *outcome agnostic* and in subsection [4.3.3] we described PLS and extensions that can be considered as *outcome aware* PCAs. An other approach recently presented in Varando et al. [22] is the Granger rotation which aims to find a rotation that extracts components which are ordered from the most to the least granger caused by an external phenomena. Considering a high dimensional time series $(X_t)_{t\in\mathbb{Z}} \in \mathbb{R}^{T\times d}$ caused by a one dimensional serie $(Y_t)_{t\in\mathbb{Z}} \in \mathbb{R}^T$, we search for a rotation for which the first components maximize the differences between the squared error of the $(X_t)_{t\in\mathbb{Z}}$ modeled by its past and the squared error of $(X_t)_{t\in\mathbb{Z}}$ modeled by its past and the past of $(Y_t)_{t\in\mathbb{Z}}$. Formally, considering linear autoregressive models we have want to find the linear combination U such as

$$\begin{aligned} U &= \text{argmax}_U \left\{RSS_0 - RSS1\right\} \\ RSS_0 &= \|X - (XU)_{past}W_0\|_2^2 \\ RSS_1 &= \|X - [(XU)\ Y]_{past}W_1\|_2^2 \end{aligned}$$

Where $(XU)_{past}$ are the past values[12] in the projected space, W_0 their repective regression coefficient in the linear autoregressive model, $[(XU)\ Y]_{past}$ the past values of the concatenation of the projection of X and Y andX_1 their respective regression coefficients.

Authors show that this problem can be solved as the following Eigen Value Decomosition (EVD) problem

$$\begin{aligned} \max_U \quad & U^T X^T (WW^T - VV^T) XU \\ \text{s.t.} \quad & U^T U = I \end{aligned} \tag{15}$$

with

$$\begin{aligned} W &= I - X_{past}(X_{past}^T X_{past})^{-1} X_{past}^T \\ V &= I - [XY]_{past}([XY]_{past}^T [XY]_{past})^{-1} [XY]_{past}^T \end{aligned}$$

where $[XY]_{past}$ is a simple column concatenation of X_{past} and Y_{past}.

[12]Here, when we consider the past of variable A_{past} we consider a simple column concatenation of the past values of A up to a specific time lag. For example consider a time lag of τ we have $A_{past} = [A_{t-1}|...|A_{t-\tau}]$.

5 Methodology

We describe in this section our approach to tackle the problem describe in section 4, knowing get an unbiased estimate of the causal effect of the serie $(X_t)_{t\in\mathbb{Z}} \in \mathbb{R}^T$ on $(Y_t)_{t\in\mathbb{Z}} \in \mathbb{R}^T$ potentially biased by unobserved confounders $(Z_t)_{t\in\mathbb{Z}} \in \mathbb{R}^{T\times p}$ that we would like to recover using high dimensional proxy variables $(U_t)_{t\in\mathbb{Z}} \in \mathbb{R}^{T\times d}$ (with $d \gg p$).

As a first approach, we decided to model our problem by a linear SCM $\mathcal{C}$ (of entailed distribution $\mathcal{P}^{\mathcal{C}}$) with gaussian random noises, knowing

$$Z_t \sim \mathcal{N}_p(0, \Sigma_z) \tag{16}$$
$$U_t = Z\tilde{W} + N_U \tag{17}$$
$$X_t = Z_t\beta_X + N_X \tag{18}$$
$$Y_t = Z_t\beta_Y + \alpha X + N_Y \tag{19}$$

with N_U, N_X and N_Y being random gaussian noises of respective variances $\Sigma_U \in \mathbb{R}^d$, $\sigma_X \in \mathbb{R}$, $\sigma_Y \in \mathbb{R}$. Here we consider the *"probabilistic"* approach and thus consider that high scale processes Z are observed threw a high-dimensional and therefore low scale proxy variables U.

Therefore, the parameters to be learn in the considered model are $\tilde{W}$, β_X, β_Y and α. As we consider a linear SCM with gaussian noise, the expectation of Y conditioned on X and Z is given by

$$\mathbb{E}[Y|X = x, Z = z] = ax + b^T z$$

and since Z is a *valid adjustement set* (see 3.5) we have that

$$\begin{aligned}\frac{\partial}{\partial x}\mathbb{E}^{\mathcal{C};do(X):=x}[Y] &= \frac{\partial}{\partial x}\mathbb{E}[Y|X = x, Z = z] \\ &= a\end{aligned}$$

Thus, model parameters can be estimated using simple regressions. The main problem here being that latent variables Z are not observed and thus we aim to estimate them or at least estimate a *valid adjustment set* for the estimation of parameter α.

5.1 Proposed methodology

5.1.1 Lasso selection

As a first approach, we seek to discover among a set of pre-extracted variabilities the one that are the the most confounding. For this purpose we first extract the first r principal components $(\boldsymbol{P}_i)_{i\in\{1,\dots,r\}}$ of the proxy variables U.

As raised in Shortreed and Ertefaie [19], inclusion of the proxy variables that only impact the cause X can inflate the standard errors without improving the bias but proxy that are only associated with the outcome and unrelated to the cause can improve precision.

Thus we select the s potential confounders $(\boldsymbol{C}_i)_{i\in\{1,...,r\}}$ as the principal components which have non null regression coefficient when regressing the outcome Y from X and $(\boldsymbol{P}_i)_{i\in\{1,...,r\}}$

$$\begin{aligned}\beta^{lasso} &= \text{argmin}_{\beta,\alpha}\|Y - \boldsymbol{P}\beta - \alpha X\|_2^2 + \lambda\|\beta\|_2^2 \\ C_i &= \boldsymbol{P_i}\mathbb{I}_{\beta_i^{lasso}\neq 0} \qquad \forall i \in \{1, ..., r\}\end{aligned}$$

When then order those components by confounding efect by performing an Ordinary Least Square regression from the selected components to both the cause X from $(\boldsymbol{C}_i)_{i\in\{1,...,s\}}$ and the outcome Y from $(\boldsymbol{C}_i)_{i\in\{1,...,s\}}$ and X.

$$\begin{aligned}\beta_X &= \text{argmin}_{\beta_X}\|X - R\beta_X\|_2^2 \\ \{\beta_Y, \alpha\} &= \text{argmin}_{\beta_Y,\alpha}\|Y - R\beta_Y - \alpha X\|_2^2\end{aligned}$$

The principal confounding component will then be considered as the one with highest confounding score, knowing $cs = \beta_X^2.\beta_X^2$.

Althought simple, this method tackle a few of the underlined difficulties of confounding adjustment. First, by first extracting the principal components we reduce the dimensionality of our considered data and we obtain an interpretable result (when just performing a Lasso regression on the all data would lead to a very sparse and thus uninterpretable solution). Second, as we first select the potential bias using a Lasso regression for Y we reduce the potential bias induced by the components that are only associated with the cause X or are independent from both the cause and the outcome.

5.1.2 Confounded Partial Least Squares

Considering the SCM described by [16] and given a valid adjustment set $\tilde{Z}$, we can estimate the parameters α, β_X and β_Y as the parameters of a Least Squares regression.

$$\mathcal{L} = \|Y - \alpha X - Z\beta_Y\|_2^2 + \|X - Z\beta_X\|_2^2 \tag{20}$$
$$\tag{21}$$

Estimating β_X it is rather straightforward as it is the regression coefficient of a multivariate linear regression.

$$\beta_X = (Z^T Z)^{-1} Z^T X \tag{22}$$

Derivating our loss with regard to α we have that

$$\frac{\partial \mathcal{L}}{\partial \alpha} = -2X^T(Y - \alpha X - Z\beta_Y) = 0 \tag{23}$$
$$\Leftrightarrow \alpha = X^{\ddagger}(Y - Z\beta_Y) \tag{24}$$

and with regard to β_Y we get

$$\frac{\partial \mathcal{L}}{\partial \beta_Y} = -2Z^T(Y - \alpha X - Z\beta_Y) = 0 \tag{25}$$

$$\Leftrightarrow \beta_Y = Z^{\ddagger}(Y - \alpha X) \tag{26}$$

Where $X^{\ddagger} = (X^T X)^{-1} X.T$ and $Z^{\ddagger} = (Z^T Z)^{-1} Z.T$ stands for the pseudo-inverse of X and Z. By injecting equation [24] in equation [26] we obtain

$$\beta_Y = Z^{\ddagger}(Y - X^{\ddagger} Y X)(1 - X Z^{\ddagger} X^{\ddagger} Z)^{-1} \tag{27}$$

and α is obtained by injecting equation [27] in equation [24]

$$\alpha = X^{\ddagger}(Y - Z Z^{\ddagger}(Y - X^{\ddagger} Y X)(1 - X Z^{\ddagger} X^{\ddagger} Z)^{-1}) \tag{28}$$

Similarly, assuming that weight matrix $\tilde{W}$ in equation 16 is invertible and denoting W its inverse we estimate W by considering fixed values of β_X, β_Y and α. Considering our loss function with $Z = UW$ and denoting $\tilde{Y} = Y - \alpha X$ we have

$$\mathcal{L} = \|\tilde{Y} - UW\beta_Y\|_2^2 + \|X - UW\beta_X\|_2^2 \tag{29}$$

$$\tag{30}$$

derivationg with regard to W gives us

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial W} &= -2U^T \tilde{Y} \beta_Y + 2U^T UW \beta_Y \beta_Y^T - 2U^T X \beta_X + 2U^T UW \beta_X \beta_X^T = 0 \\
&\Leftrightarrow U^T UW(\beta_Y \beta_Y^T + \beta_X \beta_X^T) = U^T(\tilde{Y}\beta_Y + X\beta_X)
\end{aligned}$$

And we finally get that

$$\Leftrightarrow W = (U^T U)^{-1} U^T (\tilde{Y}\beta_Y + X\beta_X)(\beta_Y \beta_Y^T + \beta_X \beta_X^T)^{-1} \tag{31}$$

As this problem have clearly many solutions, we propose to solve it in an iterative way, first optimizing W for fixed regression parameters $\hat{\beta}_X^{(k)}$, $\hat{\beta}_Y^{(k)}$ and $\hat{\alpha}^{(k)}$ and, find the optimal regression parameters $\hat{\beta}_X^{(k+1)}$, $\hat{\beta}_Y^{(k+1)}$ and $\hat{\alpha}^{(k+1)}$ given the induced latent variables $\hat{Z}^{(k)} = U\hat{W}^{(k)}$. During the optimization of W, we impose independence of the learn latent representation for interpretability purpose and problem simplification. This procedure is describe in the following algorythm.

As the implementation of the learning algorithm as not be done yet, we use in the result section the results given by an automated optimisation algorithm.

6 Experimental results

6.1 Ground experimentation

We now consider the causal problem presented in Kretschmer et al. [6] (example 3), knowing the confounding bias induced by the El-Nino-Souther Oscillation (ENSO)

Algorithm 1 Parameter estimation in the Confounded Partial Least Square problem

Require: Inital values of $\beta_X^{(1)}$, $\beta_Y^{(1)}$ and $\alpha^{(1)}$, observation U, X and Y, regularization parameter λ and

k= 1

while $|\alpha^{(k)} - \alpha^{(k-1)}| < \epsilon$ or $k = 1$ **do**

$W^{(k))} = (U^T U)^{-1} U^T (\tilde{Y}\beta_Y^{(k)} + X\beta_X^{(k)})(\beta_Y^{(k)}\beta_Y^{(k)T} + \beta_X^{(k)}\beta_X^{(k)T} + \lambda\mathbb{I})^{-1}$

$Z^{(k))} = UW^{(k))}$

$\beta_X^{(k+1)} = Z^{(k))T} X$

$\beta_Y^{(k+1)} = Z^{(k))T}(Y - (1 - X^{\ddagger} Z Z^T X)^{-1} X^{\ddagger}(Y - ZZ^T Y)$

$\alpha^{(k+1)} = X^{\ddagger}(Y - Z^{(k))}\beta_Y^{(k+1)}$

k = k+1

end while

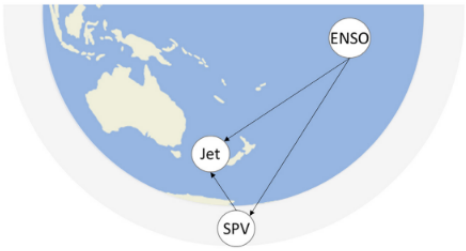

Figure 6: Causal network showing the hypothesized direct effect of ENSO on the position of the Southern Hemisphere jet stream (Jet), and its indirect causal influence mediated via the late-spring breakdown of the stratospheric polar vortex (SPV). From Kretschmer et al. [6]

on the South Pacific jet stream (JET) and the timing of the stratospheric polar vortex breakdown (SPV). The positive phase of ENSO (known as El Niño) generally induce high air surface pressure in the tropical western Pacific and it's negative phase (known as la Niña) with low air surface pressure in the same region. The two periods last several months each and typically occur every few years with varying intensity per period. But here the timing of the breakdown of the Southern Hemispheric Polar vortex also impacts the Jet Stream position. The causal graph describing the considered problem is given in figure [6].

Authors of Kretschmer et al. [6] uses indexes mesuring ENSO, SPV and JET and estimate the following regression coefficient

$$JET = -0.04ENSO + 0.39SPV + \epsilon \tag{32}$$

We now consider that the ENSO phenomenon in unknown but however, we would like to estimate the causal effect of SPV on JET. We propose to apply the two approaches described in section [5.1] considering low scale observations of the Sea Surface temperature. In fact the ENSO index considered by authors of Kretschmer et al. [6] is the second component of a Principal Component Analysis of SST in the region $5S - 5N$ and $170 - 120W$. In our case we consider SST on the entire

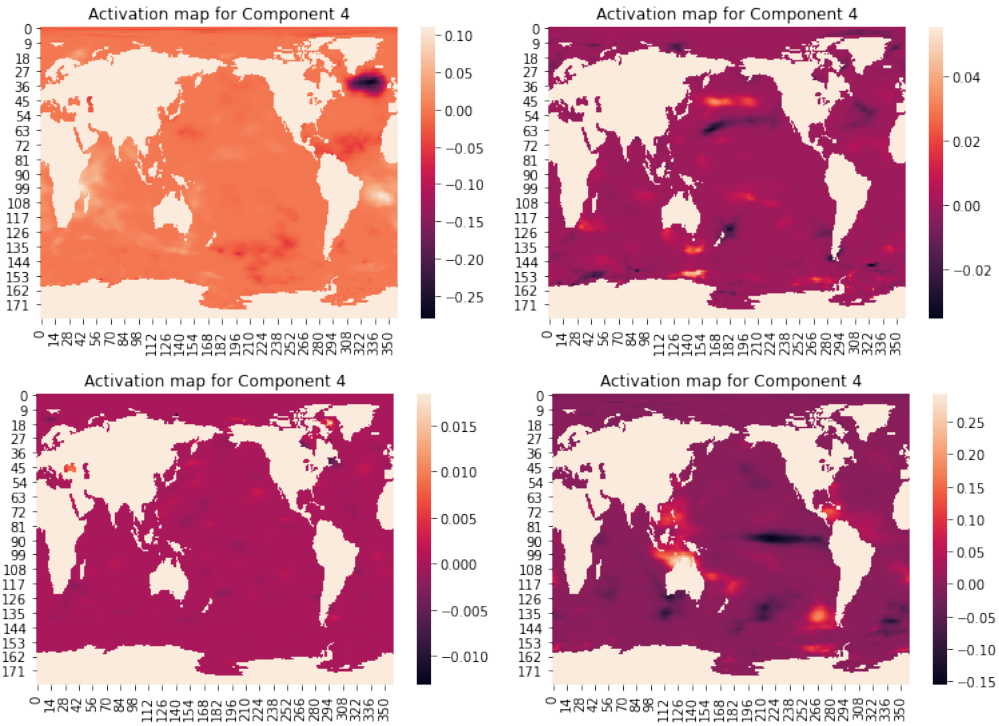

Figure 7: Activation map of the estmated weights W^3. (Top left) Component 1. (Top right) Component 2. (Bottom lleft) Component 3. (Bottom right) Component 4. We apply a cubic transformation to enhance our results.

globe, aiming to discover potentially new climate oscillation confounding the causal relation $SPC \rightarrow JET$ or confirming the importance to control on $ENSO$ when measuring this causal relation.

We first consider the Lasso selection approach. As it can be seen in figure [7], the Principal Component that is inducing the main confounding bias seems more related to SST activity in the North Atlantic which is confirmed by figure [8] where we can see that this first component is significatively correlated with the North Atlantic Oscillation (NAO). The second and third component do not seem to be correlated with any of the considered known oscillation as its activation is not very loaclized in a specific area. In contrast the Fourth component seems strongly associated with ENSO as we can see high weight activation in the region where ENSO index is computed.

The Confounding partial least square approach seems to give similar results. We see in figure [9] that the first extracted component seems more related to the North Atlantic Oscillation when the second and the third, rather similar (we can see that they are highly correlated in figure [10]) are more associated with strong activation on the West American coast, knwon to be mainly drived by ENSO variations. This is confirmed by the correlation analysis in figure [10]. The fourth component does not appear to be associated with any of the considered oscillations.

Our results seems to show the importance of adjusting the estimate of the causal effect of SPV on the southern hemispheric jet stream. This insight can be confirmed

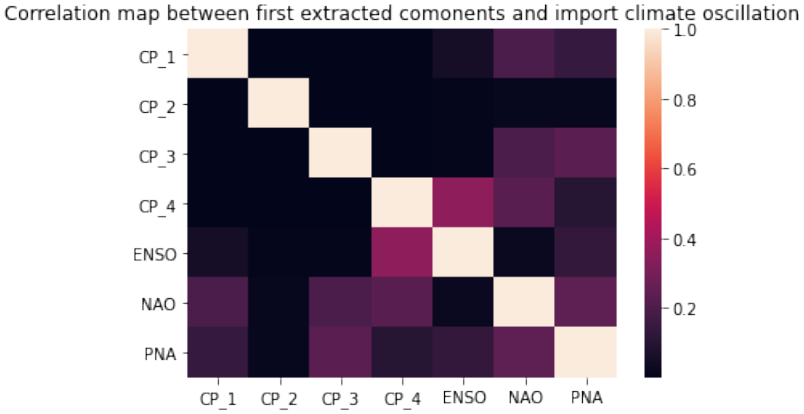

Figure 8: Correlation between extracted components and known climate oscillation

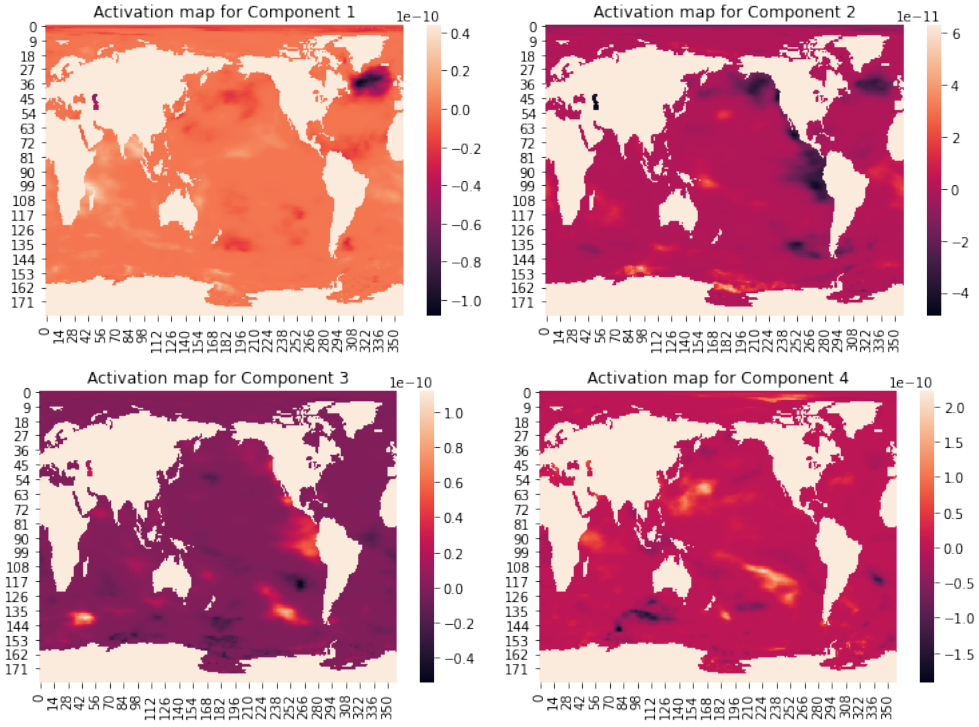

Figure 9: Activation map of the estmated weights W^3. (Top left) Component 1. (Top right) Component 2. (Bottom lleft) Component 3. (Bottom right) Component 4. We apply a cubic transformation to enhance our results.

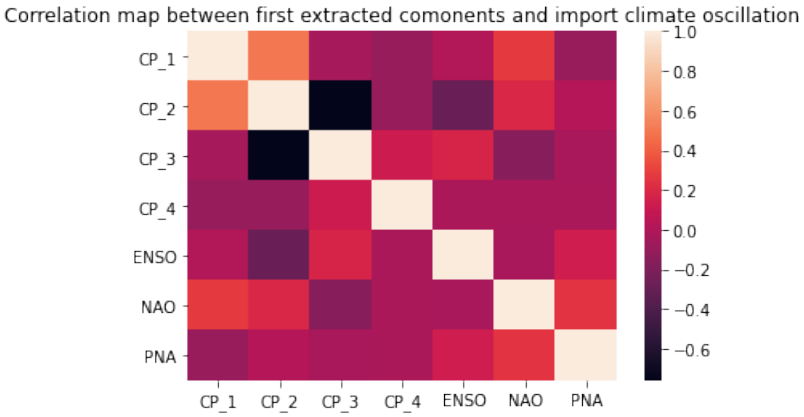

Figure 10: Correlation between extracted components and known climate oscillation

with simple regression which gives us array([-0.03376939, 0.21710294, 0.39332096])

$$JET = -0.03ENSO + 0.22NAO + 0.39SPV + \epsilon$$
$$SPV = 0.26ENSO - 0.02NAO$$

Thus ENSO and NAO have both similar confounding effect as the product of their respective regression coefficient in both regression are similar ($c_{ENSO} = -0.03 \times 0.26 = -0.0078$ and $c_{NAO} = 0.22 \times -0.02 = -0.0044$). However this hypothesis should futher explored.

7 Discussion

The results presented in this report are only preliminary and are by no means conclusive. We are currently working on theoretical and experimental results to confirm the relevance of our approach. Inspite of this fact, if these results turn out to be positive, we will consider extending our method to make it more appropriate to geoscientific problems.

First we envisage to consider the implication of time in our model extending the approach proposed in Varando et al. [22] with the Granger PCA.

It might also appear relevant to consider non linear dependencies between the variables in our causal learning problem which brings a wide range of question and problematics. A first approach could be to consider a kernelized version of our learning algorithm.

References

[1] George W Platzman. The ENIAC Computations of 1950 – Gateway to Numerical Weather Prediction. *Bulletin of the American Meteorological Society*, 60 (4):302–312, 1979.

[2] S. Manabe and K. Bryan. Climate calculations with a combined ocean-atmosphere model. *J. Atmos. Sci*, 26(4):786–789, 1969.

[3] Syukuro Manabe and Richard T Wetherald. The Effects of Doubling the CO2 Concentration on the climate of a General Circulation Model. *Journal of Atmospheric Sciences*, 32:3–15, 1975.

[4] Judea Pearl and Dana Mackenzie. *The Book of Why*. Basic Books, New York, 2018. ISBN 978-0-465-09760-9.

[5] Judea Pearl. *Causality*. Cambridge University Press, Cambridge, UK, 2 edition, 2009. ISBN 978-0-521-89560-6. doi: 10.1017/CBO9780511803161.

[6] Marlene Kretschmer, Samantha Adams, Rachel Prudden, Niall Robinson, Elena Saggioro, and Ted Shepherd. Quantifying causal pathways of teleconnections. *Bulletin of the American Meteorological Society*, 102, 07 2021. doi: 10.1175/BAMS-D-20-0117.1.

[7] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2017. ISBN 978-0-262-03731-0. URL https://mitpress.mit.edu/books/elements-causal-inference.

[8] Steffen L. Lauritzen. *Graphical Models*. Oxford University Press, 1996. ISBN 0-19-852219-3.

[9] Clive William John Granger. Investigating causal relations by econometric models and cross-spectral methods. 1969.

[10] Ali Shojaie and Emily B. Fox. Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application*, 9(1):289–319, 2022. doi: 10.1146/annurev-statistics-040120-010930. URL https://doi.org/10.1146/annurev-statistics-040120-010930.

[11] P.O. Amblard and Olivier Michel. The relation between granger causality and directed information theory: A review. *Entropy*, 15, 11 2012. doi: 10.3390/e15010113.

[12] Jakob Runge, V. Petoukhov, Jonathan Donges, Jaroslav Hlinka, Nikola Jajcay, Martin Vejmelka, David Hartman, , Milan Palus, and Juergen Kurths. Identifying causal gateways and mediators in complex spatio-temporal systems. *Nature Communications*, 6:8502, 10 2015. doi: 10.1038/ncomms9502.

[13] Jakob Runge, Sebastian Bathiany, Erik M. Bollt, G. Camps-Valls, Dim Coumou, Ethan R Deyle, Clark Glymour, Marlene Kretschmer, Miguel D. Mahecha, Jordi Muñoz-Marí, Egbert H. van Nes, J. Peters, Rick Quax, Markus Reichstein, Marten Scheffer, Bernhard Schölkopf, Peter L. Spirtes, George Sugihara, Jie Sun, Kun Zhang, and Jakob Zscheischler. Inferring causation from time series in earth system sciences. *Nature Communications*, 10, 2019.

[14] E. H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(2):238–241, 1951. ISSN 00359246. URL http://www.jstor.org/stable/2984065.

[15] Julius von Kügelgen, Luigi Gresele, and Bernhard Schölkopf. Simpson's paradox in covid-19 case fatality rates: a mediation analysis of age-related causal effects. 05 2020.

[16] Kunihiro Baba, Ritei Shibata, and Masaaki Sibuya. Partial correlation and conditional correlation as measure of conditional independence. *Australian New Zealand Journal of Statistics*, 46:657 – 664, 12 2004. doi: 10.1111/j.1467-842X.2004.00360.x.

[17] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346178.

[18] Brandon Koch, David Vock, and Julian Wolfson. Covariate selection with group lasso and doubly robust estimation of causal effects. *Biometrics*, 74, 06 2017. doi: 10.1111/biom.12736.

[19] Susan Shortreed and Ashkan Ertefaie. Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73, 03 2017. doi: 10.1111/biom.12679.

[20] Hervé Abdi. Factor rotations in factor analyses. 2003.

[21] Jerónimo Arenas-García and Vanessa Gómez-Verdejo. Sparse and kernel opls feature extraction based on eigenvalue problem solving. *Pattern Recognition*, 48, 05 2015. doi: 10.1016/j.patcog.2014.12.002.

[22] Gherardo Varando, Miguel-Ángel Fernández-Torres, Jordi Muñoz-Marí, and Gustau Camps-Valls. Learning causal representations with granger PCA. In *UAI 2022 Workshop on Causal Representation Learning*, 2022. URL https://openreview.net/forum?id=XsTEnaD_Lel.