On Spurious Causality, CO_2, and Global Temperature

Philippe Goulet Coulombe[1]

University of Pennsylvania

Maximilian Göbel

ISEG - Universidade de Lisboa

Abstract

Stips, Macias, Coughlan, Garcia-Gorriz, and Liang (2016) (*Nature Scientific Reports*) use information flows (Liang, 2008, 2014) to establish causality from various forcings to global temperature. We show that the formulas being used hinges on a simplifying assumption that is nearly always rejected by the data. We propose an adequate measure of information flow based on Vector Autoregressions, and find that most results in Stips et al. (2016) cannot be corroborated. Then, it is discussed which modeling choices (e.g., the choice of CO_2 series and assumptions about simultaneous relationships) may help in extracting credible estimates of causal flows and the transient climate response simply by looking at the joint dynamics of two climatic time series.

---

arXiv:2103.10605v1 [stat.AP] 19 Mar 2021

1   Introduction

Causality is fundamental to science. While causal statements can reasonably be made without hardship in controlled environments, things are far less straightforward when only observational data is available. Answering the question of what happens to Y if one intervenes on X is compromised by the simple fact that X and Y were not generated by exogenously modulating X, but by (often) endogenous interactions between the two variables.

Yet, for many scientific interrogations of capital importance – like the relationship between greenhouse gases and global temperature, only observational data is available. How does different forcings cause global mean surface temperature anomalies (GMTA)? In a popular article, Stips, Macias, Coughlan, Garcia-Gorriz, and Liang (2016) (henceforth, SMCGL) set up to answer the question using *information flow* (IF) from one variable to another for bivariate stochastic dynamical systems – a methodology developed in Liang (2008, 2014, 2015, 2016). Authors make grand claims about the technique being able to extract "true" and "rigorous" causality, which clash with the usually somber tone of causal analysis.

We show that the IF methodology will *not* extract causality for the near-universe of bivariate stochastic dynamic systems estimated on real data – with those of SMCGL included. In a nutshell, this occurs because Liang (2008) formulas being used throughout, assume that a certain statistical quantity is 0 when it is not. This can be tested, and it is rejected almost all the time. More precisely, Liang (2014) formula assumes that when conditioning on the past state of the system (X_{t-1} and Y_{t-1}), the remaining variation in X_t and Y_t (the innovations driving the system) are *uncorrelated*. In SMCGL, where the time unit t is one year, this implies that forcings and GMTA are uncorrelated *within the same year*, conditional on last year's values of both. That is unlikely to happen, and the data will testify to that. The assumption is presented as "linearity" (Liang, 2014), but it has little to do with it. Rather, it assumes that the system is identified, meaning that the joint dynamics of the time series data fall on the knife edge case where causality can indeed be claimed from the data without any external assumption (Sims, 1980). The problem is that for discretely sampled time series, the assumption is almost always rejected by lower-frequency data.[1] In short, SMCGL assume there to be no identification problem (in the sense of Sims 1980), and find no identification problem.

In this note, we detail the consequences of this high-standing omission on IF measures and SMCGL's results. First, in simulations (using data generating processes proposed by Liang (2014)) of which we know the true causality structure, it is shown that IF will often conclude that X_t is largely causing Y_t, when in fact the reverse is true. Second, we reconsider a key part of SMCGL's study where the authors investigate the causal structure between different forcings and GMTA. Using an appropriate methodology that accounts for the correlated innovations, it is found that in most instances, the data by itself cannot back SMCGL's claims. That is, unlike what the authors have put forward, it is not possible (within this framework) to claim that many

---

[1]In Tawia Hagan et al. (2019), IF are used on daily data, which can alleviate the problem *if there are no intra-day relationships*. This last condition is something that should be verified, not assumed.

forcings are causing GMTA's increase as a direct implication of the data. In other words, from the data *alone*, it is impossible to discriminate between certain forcings mostly causing GMTA or the reverse. External assumptions based on physical knowledge could remedy that. Or different data. We show that in the case of CO_2, results in accord with the scientific consensus can be obtained when using concentration directly rather than its radiative content.

The alternative methodology that we use are structural Vector Autoregressions (SVAR) which are simultaneous dynamic systems of equations. They can characterize a linear dynamic system in discrete time. The methodology was introduced to macroeconomics by Sims (1980) and is now so in many fields, ranging from neuroscience (Chen et al., 2011) to climate (Goulet Coulombe and Göbel, 2021). To document its reliability for this application, we report implied transient climate response (TCR) estimates of our alternative methodology. We find that those (i) largely depend on the necessary assumptions made about the simultaneous (with one year) impact of CO_2 forcing on GMTA but (ii) are in the range of recent estimates (Otto et al., 2013; Montamat and Stock, 2020) if one assumes simultaneous causality running from CO_2 forcing to GMTA.

Section 2 briefly review IFs and explain where the problematic assumption occurs. It also discusses relevant notions of Vector Autoregressions (VARs, Sims 1980) as a comprehensive framework to think about causality in multivariate time series systems. Section 3 display IFs possibly spurious behavior using simulated data where the true causality is known. Section 4 revisit the question of causality between different forcings and GMTA using appropriate tools. Section 5 concludes.

2 Information Flows, Non-Innocuous Assumptions and VARs

In this section, we review the basics of IFs as applied in empirical work, pin down the problematic assumption, explain why it is harmful through the lenses of a VAR, and propose an alternative measure of IF based on the VAR.

Liang (2008) considers the data generating process (DGP)

$$dX = F(X, t) dt + B(x, t) dW,$$
(1)

where $X = (X_1, X_2) \in \mathbb{R}^2$ are the state variables and $F = (F_1, F_2)$. $W = (W_1, W_2)$ is a standard Wiener process, with $\Delta W = \frac{dW}{dt}$ and $E(\Delta W_i) = 0$ and $E(\Delta W_i)^2 = \Delta t$. B is a 2×2 matrix and its entries b_{ij} govern how perturbations instantaneously impact the system. At this point, the only non-innocuous assumption is that of a bivariate system. The information flow from X_2 to X_1, $T_{2 \rightarrow 1}$, is then defined as

$$T_{2 \rightarrow 1} = \frac{dH_1}{dt} - \frac{dH_{1\cancel{2}}}{dt}$$
(2)

where $\frac{dH_1}{dt}$ is the evolution of the marginal entropy of X_1 and $\frac{dH_{1\cancel{2}}}{dt}$ denotes $\frac{dH_1}{dt}$ where the spillovers from X_2 are excluded. After lengthy derivations and reasonable assumptions (like

that W's, soon to be called structural shocks, are uncorrelated) it is obtained that

$$T_{j \to i} = -E\left[\frac{1}{\rho_i} \frac{\partial (F_i \rho_i)}{\partial X_i}\right] + \frac{1}{2}E\left[\frac{1}{\rho_i} \frac{\partial^2 (g_{ii}\rho_i)}{\partial X_i^2}\right], \tag{3}$$

where ρ is the joint probability density of variables X_i and X_j, and ρ_i denotes the marginal density of series X_i. Given that the model in (1) is not readily identified from the data, this is not yet operational. After further derivations and assuming $\mathbf{B} = \begin{bmatrix} b_{ii} & 0 \\ 0 & b_{jj} \end{bmatrix}$, Liang (2014) transforms (3) into a workable formula made of empirical moments

$$T_{j \to i} = \frac{\sigma_{i,i}\,\sigma_{i,j}\,\sigma_{j,\Delta i} - (\sigma_{i,j})^2\,\sigma_{i,\Delta i}}{(\sigma_{i,i})^2\,\sigma_{j,j} - \sigma_{i,i}\,(\sigma_{i,j})^2} \tag{4}$$

where $\sigma_{i,i}$ is the variance of i, $\sigma_{ij,}$ the covariance between i and j, and $\sigma_{i,\Delta i}$ is the covariance between i and the kth difference of i (we follow Liang (2015) and set $k = 1$). However, the validity of this appealing formula rests on the seemingly technical assumption of a diagonal $\mathbf{B}$. Our point is that this assumption is far from merely technical and very frequently wrong. Motivating the diagonal $\mathbf{B}$ among other assumptions, Liang (2014) states

> *Since the dynamics is unknown, we first need to choose a model. As always, a linear model is the natural choice, at least at the initial stage of development.*

The problem is that **(i)** of course, we must choose an empirical model, but we will try to avoid those that the data blatantly reject, **(ii)** $\mathbf{B}$ being diagonal has nothing to do with linearity, and **(iii)** there is no further potential "development" possible without this assumption, which, in effect, assumes the causal problem away. As a result, whenever the diagonal $\mathbf{B}$ is violated by the data, IFs – as currently used in empirical studies – provide spurious causality.

We now use a very popular framework, from macroeconomics, to think more clearly about $\mathbf{B}$. Liang (2008)'s flow of simplifications and assumptions makes his once sophisticated (1) collapse to that of a bivariate VAR with 1 lag

$$\begin{bmatrix} X_{1,t} \\ X_{2,t} \end{bmatrix} = \underbrace{\begin{bmatrix} c_1 \\ c_2 \end{bmatrix}}_{c} + \underbrace{\begin{bmatrix} a_{11}^1 & a_{12}^1 \\ a_{21}^1 & a_{22}^1 \end{bmatrix}}_{A} \begin{bmatrix} X_{1,t-1} \\ X_{2,t-1} \end{bmatrix} + \underbrace{\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}}_{B} \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}. \tag{5}$$

The seemingly innocuous assumption is much less so within a statistical framework relating assumptions directly to observable quantities. Indeed, the uncorrelatedness of W's, which here translates to that of ϵ, combined with $b_{21} = b_{12} = 0$ have a very stark implication. Let the number of endogenous variables be M and the number of lags P. Provided imposing $M = 2$ and $P = 1$ is reasonable (4) is valid if and only if regression residuals $\begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}$ are *not* cross-correlated. This is easily testable: one needs to estimate equations 1 and 2 separately by least squares, collect the residuals and calculate their correlation ρ_u. If the latter is different from

0 (and this could be formally tested with a t-test), then Liang (2008)'s simple formula does not apply. While $\rho_u = 0$ might be plausible in continuous time (or anything near it), this is a monumental stretch for data sampled at the yearly frequency (like in SMCGL). Fortunately, unlike true structural causality, the data can directly inform us on whether Liang (2014)'s formula is valid or not for specific pairs of time series.

2.1 Acknowledging the Identification Problem: Vector Autoregressions

This exposition follows closely Goulet Coulombe and Göbel (2021). In time series analysis, the "identification" problem originates from simultaneity in the data. We can learn whether

$$X_{t-1} \to Y_t \quad \text{or} \quad Y_{t-1} \to X_t$$

is more plausible. This is predictive causality in the sense of Granger (1969). However, the data itself cannot discriminate between

$$X_t \to Y_t \quad \text{and} \quad X_t \leftarrow Y_t.$$

In words, a correlation between X_t and Y_t can be generated by two different causal structures. Liang (2008)'s solution is to assume such relationships do not exist — yet, they do. Within a VAR, the problem boils down to the need for identifying C in

$$C\boldsymbol{y}_t = \Psi_0 + \sum_{p=1}^{P} \Psi_p \boldsymbol{y}_{t-p} + \boldsymbol{\varepsilon}_t, \tag{6}$$

where $\boldsymbol{y}_t$ is an M by 1 vector – meaning the dynamic system incorporates M variables. Ψ_p's parameterizes how each of these variables is predicted by its own lags and lags of the $M-1$ remaining variables. P is the number of lags being included. The matrix C characterizes how the M different variables interact contemporaneously — e.g., how total forcing affects GMTA within the *same* year (a time unit t in SMCGL). Finally, the *structural* disturbances are mutually *uncorrelated* with mean zero:

$$\boldsymbol{\varepsilon}_t = [\varepsilon_{1,t}, \ \dots \ , \varepsilon_{M,t}] \ \sim \ N(0, I_M).$$

Equation (6) is the so-called structural form of the VAR, which *cannot* be estimated because C is not identified by the data. SMCGL uses formula (4) which implicitly assumes a constrained version of (6) with $M = 2$, $P = 1$, and, most importantly, C being a diagonal matrix. The validity of their analysis hinges upon those constraints not being rejected by the data. In section 4, we find that the data disagrees with at least two of them.

Equation (6) is a structural model which can be used to answer causal questions directly. However, the elements of C are not plain regression coefficients, and cannot be estimated as such — they would be biased. It does not mean that they do not exist. The implications of their

existence can best be understood by looking at an estimable "reduced-form" VAR

$$\boldsymbol{y}_t = \boldsymbol{c} + \sum_{p=1}^{P} \Phi_p \boldsymbol{y}_{t-p} + \boldsymbol{u}_t, \tag{7}$$

where $\boldsymbol{c} = C^{-1}\Psi_0$ and $\Phi_p = C^{-1}\Psi_p$ are both regression coefficients obtained by running least squares separately on each equation. $\boldsymbol{u}_t$ are now regression residuals

$$\boldsymbol{u}_t = [u_{1,t}, \ ... \ , u_{M,t}] \ \sim \ N\left(0, \Sigma_u\right)$$

which will be cross-correlated if the true C is *not* diagonal. As mentioned earlier, Liang (2014)'s simplifying assumptions translate into $\Sigma_u = C^{-1'}C^{-1}$ being diagonal which is often at odds with the data. All the parameters of (7) can be estimated with traditional methods, but the model is not "structural" and cannot be used for causal inference. Structural VARs, which aim at uncovering "structural" causality (instead of predictive causality à la Granger 1969) acknowledge Σ_u being non-diagonal and provide ways to obtain C. As a byproduct, they can procure valid measures of information flow.

The raw material of causal measures are $\boldsymbol{\varepsilon}_t$, the structural disturbances entering the systems. However, those, like structural causality, are not directly extractable from the data: we only have $\boldsymbol{u}_t$ and translating those back to $\boldsymbol{\varepsilon}_t$ necessitates C. The latter is not directly attainable, but can be retrieved using the mapping $\Sigma_u = C^{-1'}C^{-1}$. In words, this means the covariance matrix of regression residuals from (7) can be used as raw material to retrieve the "structural" C. Mechanically, the identification problem emerges because there are many C's satisfying $\hat{\Sigma}_u = C^{-1'}C^{-1}$ — numerous causal structures deliver the same residuals' cross-correlations.

The strategy we opt for is the traditional Choleski decomposition of $\hat{\Sigma}_u$. This is one of many identification strategies for the VAR (Kilian and Lütkepohl, 2017). But among the catalog of methodologies, the Choleski decomposition is certainly popular (if not the most popular) in applied work and is simple to implement. Mechanically, it provides a lower-triangular matrix C, satisfying $\hat{\Sigma}_u = \mathbf{B}'\mathbf{B}$ where $\mathbf{B}$ is the same as from equation (5), but with dimensions $M \times M$. Its purpose is to transform cross-correlated regressions residuals u_t (equation (7)) into uncorrelated structural shocks $\boldsymbol{\varepsilon}_t$ (equation (6)). This is done by reversing the relationship $\boldsymbol{u}_t = C\boldsymbol{\varepsilon}_t$.

Uncorrelatedness is essential to study how GMTA responds to a given forcing, *keeping everything else constant*. Such a causal claim would be impossible when considering an impulse from correlated residuals u_t *as those always co-move*. A Choleski decompostion of Σ_u is *one way* to transform the observed u_t into the unobserved fundamental shocks $\boldsymbol{\varepsilon}_t$. The assumption underlying such an approach to orthogonalization is a causal ordering of shocks. The ordering restricts how variables interact with each other *within the same year*, conditional on the previous state of the system. In our bivariate setup, ordering forcing j after GMTA implies that forcing cannot impact GMTA within the same year. Ordering GMTA after a given forcing implies that GMTA cannot impact forcing j within the same year. SMCGL implicitly assumes both restric-

tions at the same time, which results in a rejected overidentified model. In contrast, when the model is just identified (when only one restriction is imposed), this choice cannot be validated by the data itself as it does not alter the likelihood.

When revisiting SMCGL's empirical work, we consider both orderings and document how sensitive conclusions are to that necessary choice.[2] There are cases where the sign of net IF (a qualitative notion) between j and GMTA depends on the ordering choice, and cases where it does not. For instance, we will find that whatever is assumed about C in a proper VAR system, total forcing appears to be causing GMTA much more than the reverse. In other cases, like CO_2-induced radiative forcing, this cannot be simply ruled out by the data.

2.2 An Adequate Measure of IF Based on the VAR

In a complete multivariate system like a VAR, the errors of the h-steps ahead forecast $y_{t+h,m}$ can be related back to structural shocks – that is, the anomalies driving the dynamics of the system. For instance, we can compute the share of the forecast error variance of GMTA 10 years from now that is attributable to CO_2 anomalies. Intuitively, if CO_2 is causing GMTA to increase, its exogenous impulses should be an important driver of GMTA's variance rather than GMTA anomalies themselves – a high information flow/transfer between the two variables. Accordingly, VAR forecast errors are

$$\boldsymbol{u}_{t+h} = \boldsymbol{y}_{t+h} - \hat{\boldsymbol{y}}_{t+h} = \sum_{h'=0}^{h-1} \Theta_{h'} \boldsymbol{\varepsilon}_{t+h-h'}$$

where Θ_h is a function of Φ_p's and C. See Kilian and Lütkepohl (2017) for further details. The forecast error variance decomposition (FEVD) of the whole system at horizon h can be analytically calculated using the entries of matrix $\Theta_{h'}$. Precisely,

$$\text{MSPE}_h = E(\boldsymbol{u}_{t+h} \boldsymbol{u} t +_h{}') = \sum_{h'=0}^{h-1} \Theta_{h'} \Theta_{h'}' .$$

An information flow "share" of i for j can be characterized by the share of forecast error variance of variable j attributable to structural shocks of i.[3] While those measures can be assessed for any horizon h (which can contain useful information), we focus on the cumulative sum since it is a measure of total flow. For non-stationary VARs (like those of the empirical section) we use a horizon of $h = 15$ years. In the case of stationary VAR process, FEVD measures quickly converge to their long-run value as h increases. Moreover, in that case, our FEVD-based IF measures can be more directly motivated from the Wold representation of a VAR process (Kilian and Lütkepohl, 2017). Finally, note that whenever the $b_{12} = b_{21} = 0$ assumption is approximately true, the FEVD approach and IFs give qualitatively similar answers.

---

[2]Of course, there are identification schemes outside of the family of "orderings" obtained by Choleski decomposition, but those are beyond the scope of this paper – and unnecessary to make our main point.

[3]For a discussion on how to think about "shocks" in a physical system, see Goulet Coulombe and Göbel (2021).

3 Simulations

This section showcases how IF can lead to a pretense of causal knowledge, with conclusions that are sometimes the exact opposite of the truth. We consider four data-generating processes (DGPs) where the true P is one. The first two correspond to what is proposed in Liang (2014). The last two are generic VAR positively autocorrelated processes with differing degrees of persistence. Following the notation above (equation (5)):

$$DGP(1) := \left\{ \boldsymbol{A} = \begin{bmatrix} 0.5 & 0.5 \\ 0 & 0.6 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 0.1 \\ 0.7 \end{bmatrix} \right. \tag{8a}$$

$$DGP(2) := \left\{ \boldsymbol{A} = \begin{bmatrix} -0.5 & 0.9 \\ -0.2 & 0.5 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right. \tag{8b}$$

$$DGP(3) := \left\{ \boldsymbol{A} = \begin{bmatrix} 0.5 & -0.2 \\ -0.5 & 0.25 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right. \tag{8c}$$

$$DGP(4) := \left\{ \boldsymbol{A} = \begin{bmatrix} 0.25 & -0.1 \\ -0.2 & 0.1 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right. \tag{8d}$$

As highlighted in the previous section, IFs are calculated assuming $b_{ij} = 0$. However, as will be reported and formally tested in section 4, this is frequently not the case for most time series, especially those sampled at low frequencies. Using a controlled simulation environment, we study how IFs behave for values of $\rho_u := corr(u_t^{X_1}, u_t^{X_2}) \in [-1,1]$.[4] Note that IFs are invariant to $\rho_u \neq 0$ emerging from $b_{12} = 0$ or $b_{21} = 0$, which is obviously problematic given the causal content of b_{ij}.[5]
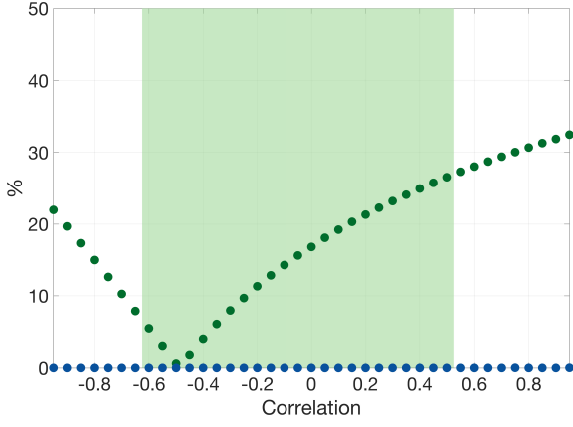
As our indicator of true underlying information flow, taking into account $\rho_u \neq 0$, we use the FEVD-based measure described earlier. Note that here, unlike the application to real data in section 4, *we know* what are the true b_{ji}'s. When $\rho_u \neq 0$, the correlation must be attributed to either b_{12} or b_{21}, or a combination of both. In a bivariate setup, this amounts to setting $b_{ij} = 0$, and attributing $\rho \neq 0$ to b_{ji}. Hence, it is possible to tell when standard IFs conclude falsehoods, because in simulations b_{12} and b_{21} are known.

In terms of notation, $Y_{i \to j}^{i,j}$ means the FEVD share at horizon $h = 10$ with $i = 1,2$ and $i \neq j$. The superscripts in $Y_{i \to j}^{i,j}$ determine the *true* ordering, hence i ordered before j. The subscripts indicate that we plot the contribution of variable i to the forecast error variance of variable j at horizon $h = 10$. The simulations have shown that $h = 10$ is sufficient for convergence.
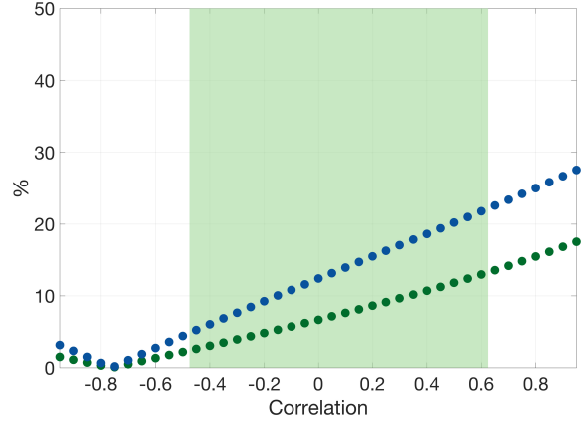
Figure 1 plots the absolute normalized information flows (NIF), $\tau_{i \to j}$ for DGP(1) through DGP(4), with varying ρ_u. A first observation is that $\tau_{i \to j}$ often varies significantly with ρ_u, even

---

[4]Variances of $u_t^{X_1}$ and $u_t^{X_2}$ are one.

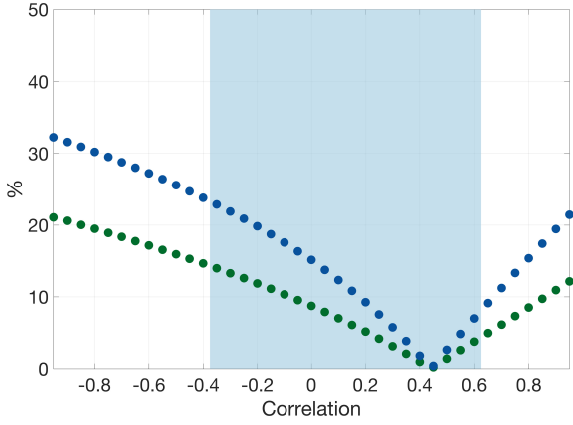[5]It is important to note that while we consider cases where either $b_{12} = 0$ or $b_{21} = 0$, there is a continuum of possibilities between those. We do so for simplicity of exposition (it makes the problem dichotomous). Moreover, setting either $b_{12} = 0$ or $b_{21} = 0$ to 0 corresponds to a causal ordering which is by far the most common identification scheme used in practice – and what we will be using in section 4.
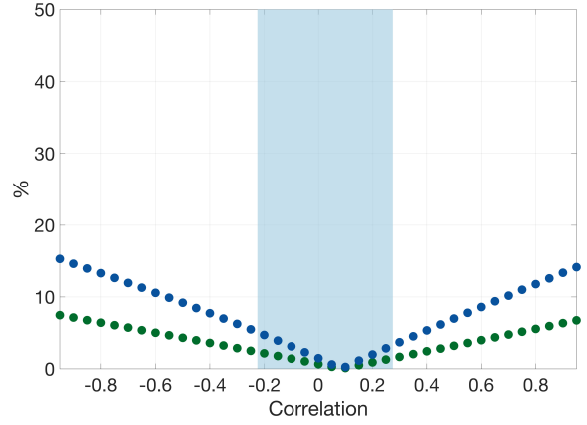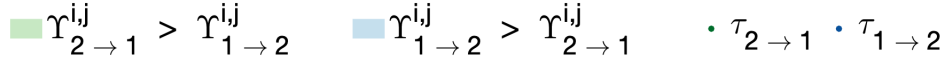
$$\square \, \Upsilon^{i,j}_{2 \to 1} > \Upsilon^{i,j}_{1 \to 2} \qquad \square \, \Upsilon^{i,j}_{1 \to 2} > \Upsilon^{i,j}_{2 \to 1} \qquad \cdot \, \tau_{2 \to 1} \quad \cdot \, \tau_{1 \to 2}$$

Notes: Color specification: the *light blue* area shows the region, in which FEVDs suggest $Y^{i,j}_{1 \to 2} > Y^{i,j}_{2 \to 1}$, regardless of the ordering. Similarly, the *light green* area shows the region, in which FEVDs suggest $Y^{i,j}_{2 \to 1} > Y^{i,j}_{1 \to 2}$ – independent of the ordering. The *white/non-colored* areas shows those regions, for which the ordering of $X1$ and $X2$ does matter. An unambiguous determination of $Y^{i,j}_{1 \to 2} > Y^{i,j}_{2 \to 1}$, respectively $Y^{i,j}_{2 \to 1} > Y^{i,j}_{1 \to 2}$, is not possible.

Figure 1: Ranking of NIFs ($\tau_{i \to j}$) vs Ranking of FEVDs ($Y^{i,j}_{i \to j}$)
Different Levels of Correlation; Horizon $h = 10$

when the very formula underlying it assumes $\rho_u = 0$. The relative strength of $\tau_{1 \to 2}$ and $\tau_{2 \to 1}$ can easily collapse to 0 or be much higher than what IFs report for $\rho_u = 0$.

The fact that IFs vary with ρ_u could give false hopes that they account for simultaneous relationships. We conduct a simple exercise to show it does not. An interesting empirical question is whether i is causing j more than j is causing i, which is the appealing promise of IFs. This moves beyond testing for Granger Causality and aims at quantifying the flow of information. The shaded green region corresponds to the values of ρ_u for which either the true underlying causality is that X_2 causes X_1 more than the reverse, *irrespective* of whether it emerges from either $b_{12} = 0$ or $b_{12} = 0$. For those values of ρ_u, when the true ordering is unknown, the *qualitative* conclusion about the sign of the net causality flow does not hinge on knowledge of b_{ij}.

Analogously, the blue shade represents values of ρ_u for which X_1 causing more X_2 is unanimous among b_{ij} configurations. White regions are values of ρ_u where conclusions about the sign of net causality flow cannot be determined from the data, i.e., the unknown b_{ij} is necessary to settle. In other words, given the data available to the modeler, X_2 causing more X_1 and vice versa are both equally likely, and sorting it out decisively implies making a successful guess on the true value of b_{ij}.

If IFs were correctly calibrated, the green dotted line should only be above the blue in the shaded green regions, and the dotted blue line above the green one in blue regions. Figure 1 clearly demonstrates this not to be the case. For every DGP, there is a substantial range of values (white spaces) for which IFs clearly conclude $\tau_{i \to j} > \tau_{j \to i}$ when there is one pair of (b_{ij}, b_{ji}) out of two *for which the reverse is true*. To address concerns about an horizon mismatch h between IFs and FEVD, Figure 4 in Appendix B shows results for $h = 2$. As it turns out, things worsen for IFs (the blue and green regions shrink) since the simultaneity problem is less diluted in dynamics at short horizons.

In sum, those simulations (largely inspired by Liang (2015) DGPs) show two things. First, the tractable IFs (from formula (4)) are functions of ρ_u even though they assume it to be zero. This compromises any statement on the strength of causal links. Second, for any DGP, there is an underlying pair (b_{ij}, b_{ji}) for which IFs' conclusion about the net causal flow is the opposite of reality. This is not only an artifact of large $|\rho_u|$, as exemplified by DGP(4).

4 GMTA and Radiative Forcing Revisited

In this section, we first revisit SMCGL's application of IFs to GMTA and forcings. Second, we look at what lies behind opaque FEVD measurements by reporting impulse response functions, and computing the transient climate response implied by simplistic bivariate VARs.

Global warming generated by man-made forcing is the prevalent generalization of the notion *climate change*. Numerous researchers have dedicated their works to this very relationship between anthropogenic forcing and constituents of global climate (Hansen et al., 2006; Andrews et al., 2010; Li et al., 2013; Notz and Stroeve, 2016) and the socio-economic consequences thereof (Nordhaus, 2014; Shukla et al., 2019). Despite overwhelming evidence for anthropogenic forcing being the main driver of global climate change (Masson-Delmotte et al., 2018), scientists have also observed that especially since the turn of the millennium, global temperature has plateaued despite ever-rising greenhouse gases and contrary to projections from key climate models.[6]

IFs hinge exactly on these bivariate relationships, which are attractive in their clarity, but are certainly a stark oversimplification of a complex system. Nevertheless, to give empirical content to our critique of the methodology, we study the relationship between GMTA and 7 forcings from SMCGL using both IFs and our FEVD-based remedy. The sample of annual means ranges

---

[6]Resolving this puzzle has led to reevaluate the role of oceans in the interplay of radiative forcing and the climatic response (Tollefson, 2014; Marotzke and Forster, 2015).

from 1850 - 2005.[7] We follow SMCGL and take our data from therein referenced data providers. Table 1 summarizes the results.[8]

In Table 1 we report estimated correlation between residuals of the bivariate VAR(1) implied by IFs. When forcing $P = 1$, in 5 cases out of 7, the null that $\hat{\rho}_u^{i,GMTA} = 0$ is rejected at least at the 10% level. When choosing P with Bayesian Information Criterion (BIC), only $\hat{\rho}_u^{Solar,GMTA} = 0$ cannot be rejected. Hence, as repeatedly mentioned in the text, IFs assume something that can be, and is rejected by the data. Moreover, in the light of simulations carried earlier, the qualitative and quantitative insights from IFs are often spurious under such conditions.

Naturally, we concentrate on FEVD results which correctly account for $\rho_u \neq 0$. However, setting $\rho_u = 0$ is only one of the empirical shortcomings of the empirical IF formula — it also sets P, the number of lags of each X_t, to be 1. Clearly, it is empirically plausible that $X_{i,t-2}$ or $X_{j,t-4}$ may have an impact on $X_{i,t}$ beyond what is channeled by single year lags ($X_{i,t-1}$ and $X_{j,t-1}$). In other words, $P = 1$ is extremely restrictive on climatic dynamics. When choosing P with BIC, results align better with prior scientific knowledge. Nonetheless, for the sake of completeness, we report both results ($P = 1$ and $P = P^*$, with P^* being BIC's choice).

With $P = P^*$, the fact that total forcing causes GMTA more than the reverse is without appeal. Nevertheless, the quantitative answer is, again, highly dependent on the ordering choice. After 15 years, total forcing anomalies are responsible for explaining between 28.0% and 47.4% of that of GMTA – depending on the preferred ordering. The net causal flow being higher from aerosol and solar to GMTA are also unanimous, but much smaller. Indecisive results are reported for Anthropogenic, Volcanic and PDO. Overall, Table 1 suggests that the data itself does not support the strong qualitative conclusions of SMCGL for their CO_2 measure and Anthropogenic.

When P is forced to one, as in IFs, inconclusive results are reported for total forcing, aerosol, volcanic and PDO. $P = 1$ specifications, irrespective of the ordering[9], conclude that GMTA is causing more CO_2 and Anthropogenic than the reverse. Only solar forcing results are unanimous, and in line with what climatic common wisdom suggests. Note that it also agrees with results from original IFs, which is not surprising given that $\hat{\rho}_{Solar,GMTA}$ is in the close vicinity of 0. However, choosing a proper P nearly doubles the share of forecast errors attributable to solar forcing.

---

[7]Data on the Pacific Decadal Oscillation (PDO) ranges from 1900 - 2005.

[8]The sample was restricted to 1850-2005 to match that of SMCGL. Table 5 reports results extending the sample to 2017.

[9]There are other identification schemes which cannot be cast as "orderings". That is, there are rotations of $\boldsymbol{u}_t$ (even when $\rho_u = 0$) giving different structural shocks. Hence, while the two orderings span a lot of possibilities (and those traditionally considered first in practice), they do represent the universe of rotations of $\boldsymbol{u}_t$ into $\boldsymbol{\epsilon}_t$.

Table 1: Empirical Results for the Bivariate Relationship
Between Various Forcings and GMTA

	Correlation	Normalized IF (IF × 100)		FEVD					
				Lags (P)	Correlation of Residuals (ρ_u)	Ordering: i, GMTA		Ordering: GMTA, i	
		$i \to$ GMTA	GMTA $\to i$			$i \to$ GMTA	GMTA $\to i$	$i \to$ GMTA	GMTA $\to i$
Total Forcing	0.73	**30.6**	20.8	4	0.23***	**47.4**	13.0	**28.0**	25.4
		(15.3)	(11.1)	1	0.29***	**51.4**	9.6	27.6	**28.7**
Anthropogenic	0.86	**39.8**	-20.0	4	-0.19**	**6.5**	3.7	3.8	**13.4**
		(35.7)	(-0.6)	1	-0.19**	5.0	**5.8**	2.2	**17.1**
CO$_2$ - ERF (W/m^2) SMCGL	0.86	**39.6**	-15.2	4	- 0.14*	6.5	**8.4**	5.6	**17.4**
		(35.1)	(-0.4)	1	-0.15*	2.8	**4.7**	1.1	**12.8**
Aerosol	-0.82	**35.9**	-24.5	4	-0.19**	**2.9**	0.6	**2.1**	1.6
		(24.3)	(-0.4)	1	-0.10	3.5	**4.0**	**1.8**	1.2
Solar	0.49	**13.5**	6.7	8	0.05	**8.5**	1.6	**6.6**	2.5
		(3.8)	(2.3)	1	0.08	**16.6**	4.2	**12.3**	6.8
Volcanic	0.09	**0.9**	-0.5	4	0.18**	**10.9**	0.8	3.1	**3.6**
		(0.2)	(-0.4)	1	0.20**	**7.1**	1.4	0.6	**3.7**
PDO	0.17	**-1.2**	-0.6	4	0.35***	**31.1**	0.9	6.3	**10.3**
		(-0.2)	(-0.5)	1	0.34***	**9.1**	0.5	0.2	**10.7**
CO$_2$ (Mt/yr)	0.82	**37.1**	-4.3	2	-0.10	8.9	2.1	**10.7**	0.6
		(27.0)	(-0.0)	1	-0.05	**4.2**	0.0	**4.4**	0.4
CO$_2$ (W/m^2)	0.86	**39.5**	-14.0	4	0.23***	5.2	**16.8**	2.9	**4.7**
		(34.5)	(-0.3)	1	0.07	1.6	**4.1**	0.9	**1.8**

Notes: i corresponds to the type of radiative forcing, listed in the left most column. The second column ("Correlation") documents the correlation between GMTA and variable i. FEVD values are taken at horizon $h = 15$, which translates into the contribution of variable i in the variance of the forecast error of variable j a decade and a half after the in-sample end date. Numbers in bold underline the highest absolute causal flow among a (i,GMTA) pair for a given measure. "*", "**", and "***" means that the null of the residuals cross-correlation of residuals is rejected at the 10%,5% and 1% level respectively.

4.1 What's up with CO$_2$?

Results for CO$_2$ are rather surprising. Irrespective of the ordering, GMTA is reported to cause CO$_2$ more than the reverse, a finding contradicting SMCGL's results and common wisdom.[10] However, SMCGL's CO$_2$ measure exhibit ill dynamic behavior that cannot possibly be that of a natural quantity. This becomes obvious when plotting their CO$_2$-ERF measure in first difference: it evolves according to a series of dichotomous jumps. This CO$_2$-ERF series, in which ERF stands for effective radiative forcing, is a concept presented in Myhre et al. (2013). The series itself is based on Etminan et al. (2016).

---

[10] Hansen et al. (2006) states that global warming started to accelerate not prior to the 1970s. Only after 1975 did the global temperature increase by approximately 0.2°C per decade.

Given the strange jumping behavior of the CO_2-ERF series, and to make our calculations comparable to other findings in the literature (Bruns et al., 2020; Montamat and Stock, 2020), we derive RF from CO_2 concentration as follows: we use the well established Meinshausen et al. (2017) data set on annual global means of CO_2 concentration, measured in parts per million (*ppm*). We follow Myhre et al. (2013) and transform the increase in CO_2 concentration in year t measured in *ppm* relative to the concentration in a given base year, $CO_{2,base}$, into radiative forcing, $RF_t^{CO_2}$ – measured in *W/m²* – as follows: $RF_t^{CO_2} = 5.35 \times \ln(CO_{2,t}/CO_{2,base})$. We use 1850 as base year following Bruns et al. (2020).

As it turns out, considering this less contentious CO_2 series does not resolve the apparently counterintuitive finding that GMTA explains a larger portion of the forecast error variance of CO_2 than vice versa. Such a finding has also been reported using a different methodology in Koutsoyiannis and Kundzewicz (2020). We explore a last avenue, that of using annual CO_2 emissions rather than $RF_t^{CO_2}$. This last attempt is successful in reconciling the FEVD approach with the traditional wisdom that CO_2 is causing GMTA "more" than the reverse. This finding is independent of the ordering choice.

In sum, *based on this particular time series evidence*, the causal link between certain forcings and GMTA remains disputable. What is less disputable are the effects of total forcing, CO_2 emissions, and solar forcing which all explain an important share of GMTA anomalies independently of arbitrary ordering preferences. Nonetheless, we see such analyses as rather primitive and potentially misleading. For instance, *X* causing "more" *Y* does not mean that the reverse causality is not quantitatively important, or climatologically relevant. We now turn to a more promising way to extract meaning out of selected bivariate VARs.

4.2   Impulse Response Functions

To open the black box of those rather abstract measurements, we report in Figure 3 impulse response functions (IRF) for the bivariate models of CO_2-GMTA and total forcing-GMTA. Since Sims (1980), the dominant approach for studying the properties of the VAR around its deterministic path has been IRFs to *structural shocks*. Their dynamic effect can be analyzed as that of a randomly assigned treatment because those have been transformed to be uncorrelated, which provides the "keeping everything else constant" interpretation.
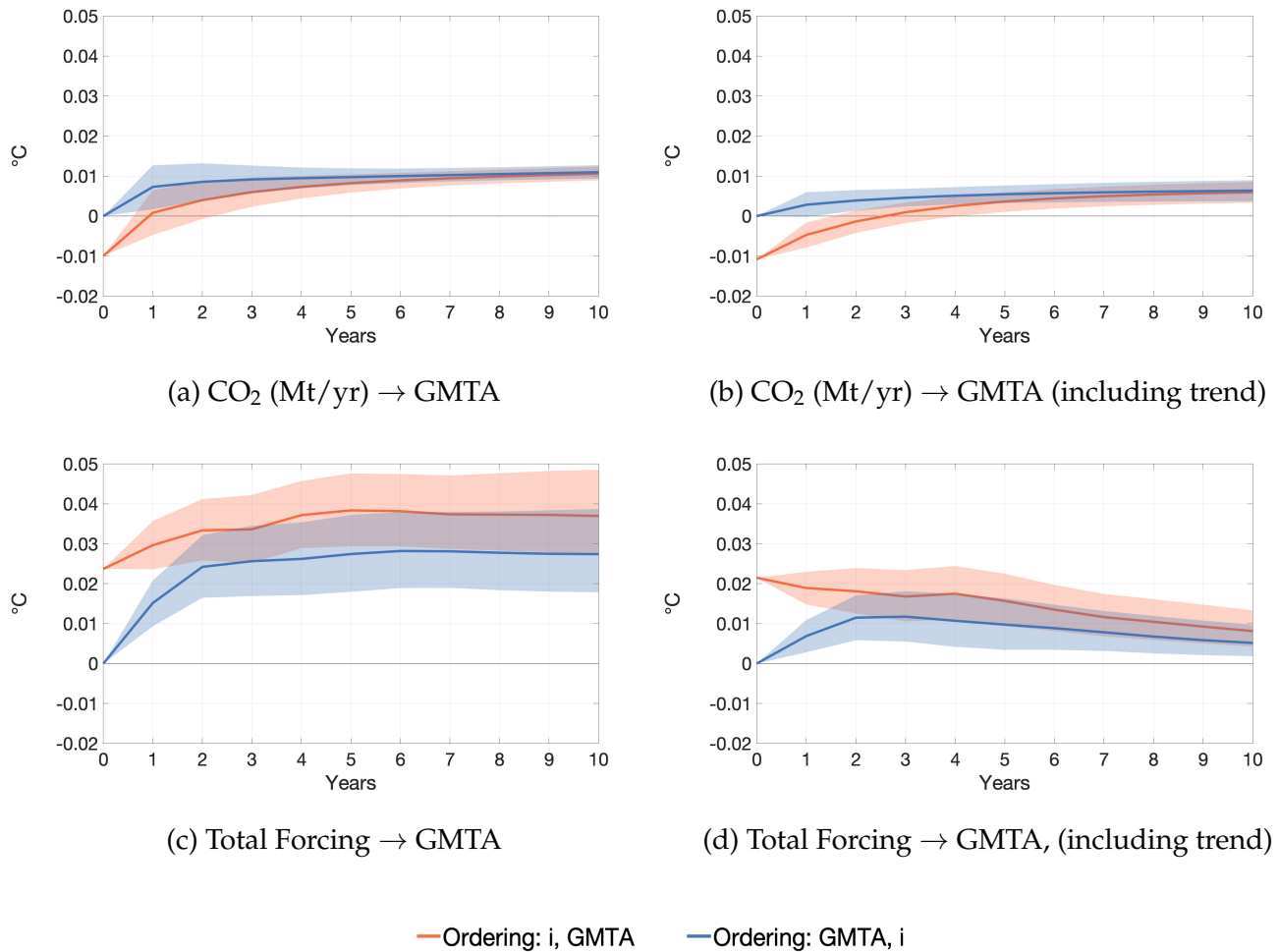
The IRF of a variable i to a one standard deviation shock of $\varepsilon_{j,t}$ is defined as

$$IRF(j \rightarrow i, h) = E(y_{i,t+h}|\boldsymbol{y}_t, \varepsilon_{t,j} = \sigma_{\varepsilon_j}) - E(y_{i,t+h}|\boldsymbol{y}_t, \varepsilon_{t,j} = 0). \tag{9}$$

Thus, it is the expected difference, h months after "impact", between a bivariate system that responded to an unexpected CO_2 increase, and the same system where no such increase occurred. In a linear VAR, the above takes a closed-form solution in terms of the matrices from (6). Models are now estimated with Bayesian methods, optimizing the hyperparameters of a standard Minnesota prior, and choosing the number of lags as reported in the gray-shaded rows of Table

1. The primary motivation is to obtain valid inference even in the presence of nonstationarity. Point estimates are nearly identical to that of OLS. For details on such choices in the context of climate data and a more thorough (yet introductory) treatment of IRFs, see Goulet Coulombe and Göbel (2021).
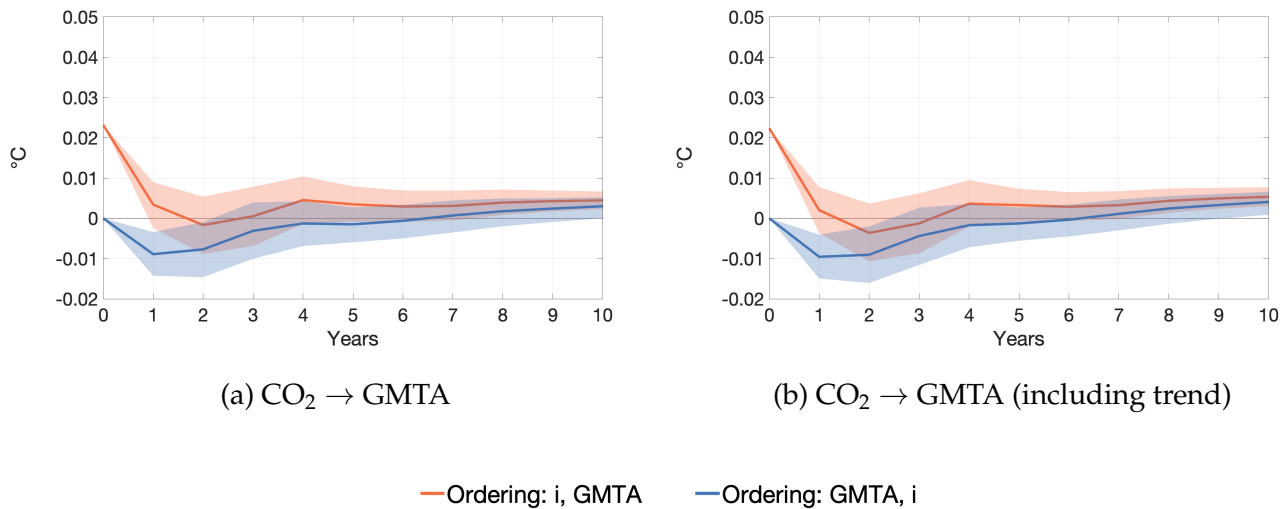
Figure 2: IRFs: Annual Emissions and Global Temperature

(a) CO_2 (Mt/yr) $\rightarrow$ GMTA

(b) CO_2 (Mt/yr) $\rightarrow$ GMTA (including trend)

(c) Total Forcing $\rightarrow$ GMTA

(d) Total Forcing $\rightarrow$ GMTA, (including trend)

—Ordering: i, GMTA   —Ordering: GMTA, i

Notes: We show impulse response functions from bivariate Vector Autoregressions of *GMTA* and annual CO_2 emissions. The right column includes a time trend as an additional exogenous regressor. The solid line is the median of 10,000 draws from the posterior distribution. Hyperparameters were optimized. The shaded area comprises the 68% credible region. Lags are those reported in Table 1 under the gray shaded rows.

In Figure 2 we show the effect of an unexpected increase in *annual* emissions on GMTA, while Figure 3 shows the response of GMTA radiative forcing generated by an unexpected rise in *cumulative* emissions.

Qualitatively, the impact of annual emissions and cumulative CO_2-induced forcing *shocks* on global temperature is vastly similar. In accord with findings in Goulet Coulombe and Göbel (2021) for the effect of CO_2 on Arctic sea ice extent, the impact of total forcing and CO_2 shocks is highly durable. This time, it is on GMTA rather than sea ice extent. As reflected in Figures 2 and 3, this result is qualitatively independent of the ordering choice. In both cases (and for both forcing variables), the effect of forcing takes about two years to completely settle in. However, it is clear that the reported short-run impact strongly depends on the identification assumptions,

Figure 3: IRFs: Cumulative Emissions and Global Temperature

(a) $CO_2 \rightarrow GMTA$ (b) $CO_2 \rightarrow GMTA$ (including trend)

——Ordering: i, GMTA   ——Ordering: GMTA, i

Notes: We show impulse response functions from bivariate Vector Autoregressions of *GMTA* and CO_2, and Total Forcing respectively. The right column includes a time trend as an additional exogenous regressor. The solid line is the median of 10,000 draws from the posterior distribution. Hyperparameters were optimized. The shaded area comprises the 68% credible region. Lags are those reported in Table 1 under the gray shaded rows.

which SMCGL completely abstract from.

Whether the slightly negative short-run response of GMTA in the first panel of Figure 2 favors the ordering $\{GMTA, CO_2\}$ over $\{CO_2, GMTA\}$ is debatable: Forster et al. (2020) find the reduction in global CO_2 emissions during the COVID-19 pandemic to have resulted in a short-run rise of global temperature. The key mechanism is a decline in the cooling-effect of aerosols as a result of less SO_2 emissions. The authors project a rise in global temperature over the first 24 months following the pandemic-induced reduction in global nitrogen oxide (NO_x) emissions.[11]

Lastly, a robustness check. Our variables are clearly nonstationary. While this does not pose a problem for Bayesian inference, it is nevertheless natural to wonder if results would be significantly altered by including a time trend. Accordingly, the second column reports the same IRFs, but for VAR specifications augmented with trends. Those show that adding such an exogenous explanatory variable does not change the dynamics of a GMTA response to an unexpected shock to CO_2. The addition of a trend to the bivariate model of GMTA and total forcing, allows GMTA to slowly revert to a lower impact – which is nevertheless highly persistent.

4.3   Are VAR Estimates Quantitatively Reasonable?

The transient climate response (TCR) is a frequently used metric to measure the impact of rising atmospheric CO_2 concentration on temperature. It is not only an indication of the trajectory of ongoing climate change, but also serves as a benchmark to evaluate the results of climate model projections (Phillips et al., 2020). The TCR is defined as the increase in temperature,

---

[11]Especially NO_2 is found to be well-correlated with CO_2 emissions (Forster et al., 2020).

between h_0 and h_T, under the assumption that CO_2 increases annually by 1%. h_T is defined as that point in time when – due to the steady annual increase of 1% – CO_2 concentration is twice as high as at date h_0 (Pretis, 2020; Montamat and Stock, 2020). Such a doubling of CO_2 would occur approximately after 70 years (Otto et al., 2013). Following the transformation of *ppm* to *W/m²* as suggested by Myhre et al. (2013) and Montamat and Stock (2020), a doubling of CO_2 under an annual increase in concentration of 1% would generate a radiative forcing of $5.35 \times \ln(2) \approx 3.7 \, \text{W/m}^2$.[12]

Typical estimates for TCR fall within a range of 1°C-2.5°C with a 66% probability, as summarized in the *IPCC 5th Assessment Report* (Bindoff et al., 2013). More recent estimates are well aligned with this range. Bruns et al. (2020) reports a point estimate of TCR ranging from 1.17°C to 1.85°C, depending on the type of data and model specification. Pretis (2020) embeds a two-component energy balance model into a cointegrated vector autoregressive model. His estimates vary across model specification and range from 1.24°C to 1.38°C. Phillips et al. (2020) report a global transient climate sensitivity of 2.05°C. The IV regression in Montamat and Stock (2020) allows for a differentiation of TCR measurements across different horizons, normalized to giving 70-year-horizon estimates. Their point estimates range in the neighborhood of 1.5°C within a 95% confidence interval of roughly 0.9°C to 2.1°C.

Despite being much more simplistic than the models of the aforementioned works on TCR, the bivariate VARs do also allow to estimate the impact of a doubling of CO_2 on temperature. Here we make use of the concept of IRFs, as presented in Equation (9). In particular, we estimate the impact on temperature when RF^{CO_2} increases by one standard deviation of its reduced-form residuals, σ_{ε_j}, where $j = RF^{CO_2}$, from a bivariate VAR of RF^{CO_2} and GMTA. Recalling the definition of TCR, a doubling of CO_2 concentration, which is achieved by an annual increase of 1% in atmospheric CO_2 concentration, generates an additional radiative forcing of approximately $5.35 \times \ln(2) \approx 3.7 \, \text{W/m}^2$. In our case, the shock σ_{ε_j} to CO_2 is a one-time event at horizon $h = 0$, but its effects are distributed over horizons $h = 1, 2, 3, ..., H$. This allows us to measure the *cumulative* increase, Ξ_i, in $i \in \left\{ RF^{CO_2}, GMTA \right\}$ generated by σ_{ε_j}, where $j = RF^{CO_2}$, at any horizon h:

$$\Xi_{j,h} = \sum_{s=0}^{h} IRF(j \to i, s) \ ,$$

where $IRF(j \to i, s)$ is defined as in Equation (9). Adapting the formula of Otto et al. (2013), we estimate TCR$_h$ as the increase in GMTA at horizon h as follows:

$$TCR_h = \Xi_{GMTA,h} \times \frac{5.35 \times \ln(2)}{\Xi_{RF^{CO_2},h}} \ , \tag{10}$$

where $\Xi_{GMTA,h}$ is the cumulative increase in global temperature at horizon h, resulting from the

---

[12] An annual increase of 1% in atmospheric CO_2 concentration results in a doubling of CO_2 after approximately 70 years, which is described more formally as: $h \times ln\left(\frac{1.01}{1}\right) = ln(2)$, for $h \approx 70$ (Montamat and Stock, 2020).

shock σ_{ε_j}, where $j = RF^{CO_2}$, at horizon $h = 0$. Likewise, $\Xi_{RF^{CO_2},h}$ is the cumulative increase in radiative forcing of CO_2 at horizon h, resulting from the shock σ_{ε_j}, where $j = RF^{CO_2}$, at horizon $h = 0$.

In Table 2 we present median point estimates of TCR_h for $h = 20$ and $h = 70$ (as in Montamat and Stock (2020)) from bivariate VARs of RF^{CO_2} and GMTA – with and without an exogenous time trend. Thus, Table 2 reports the TCR corresponding to the model specifications in Figure 3 (a) and (b). That is, we use Bayesian estimation techniques and deploy a Minnesota prior on our parameter estimates. Our VAR has four lags and the estimation is based on annual observations between 1850 and 2005.

Table 2: Transient Climate Response

Ordering	Without Trend		With Trend	
	TCR_{20}	TCR_{70}	TCR_{20}	TCR_{70}
CO_2, GMTA	1.99°C	2.06°C	2.17°C	2.58°C
GMTA, CO_2	0.57°C	1.82°C	0.85°C	2.39°C

The main message of Table 2 is twofold: first, the ordering of the variables heavily influences the final results for TCR_{20}, demonstrating the importance of respecting the possibility of cross-correlated residuals. Mechanically, the discordance brought up by the ordering choice vanishes at much longer horizons, and, as a result, TCR_{70} estimates are largely similar. However, at that horizon, it is the choice of whether or not to include a trend that can alter results significantly. Second, even though the TCR_h point estimates of the trend models are rather located at the upper bound of the IPCC range of 1°C-2.5°C, a simplistic bivariate VAR model including a constant and a time trend as additional exogenous regressors, is capable of providing a reasonable approximation of the rise in global mean temperature, triggered by a doubling of atmospheric CO_2 concentration.

TCR estimates can be helpful in choosing which ordering is most plausible. For instance, only by ordering CO_2 first do we get TCR_{20} to fall within the IPCC range. IRFs can also help sort things out. Ordering CO_2 second leads to a surprisingly lasting *negative* effect of CO_2 shocks on GMTA. An increasingly popular approach to VAR identification in macroeconomics is to use sign restrictions, where implausible IRF draws (based on economic theory) are tossed out (Uhlig, 2005). This dispenses the researcher from formulating a likely contentious causal ordering of variables. In a climate application, one could identify the VAR by rejecting specifications generating implausible $IRF(CO_2 \rightarrow GMTA, h)$ or TCRs. Applying this sort of reasoning leads us to favor – to nobody's surprise – the specification where simultaneous causality runs from CO_2 to GMTA. However, it is important to stress that this choice is obtained from prior knowledge on what is deemed reasonable and what is not, rather than our two time series.

5   Conclusion

This note is a cautionary tale about how seemingly innocuous simplifying assumptions can go wrong – especially when they are formulated without consulting the data. IFs, as proposed by SMCGL, is a concept that hinges on the assumption of zero correlation between the residuals of a bivariate VAR(1) process. In discrete time, especially with observations at lower frequencies, such an assumption is most often not justified. Both stylized simulations and an empirical application in the form of the transient climate response demonstrate that being negligent about cross-correlated residuals can lead to markedly different outcomes. Our results show that already in a bivariate system of CO_2 and GMTA, the resulting TCR depends on how one deals with correlated residuals.

FEVDs provide an alternative to IFs. Albeit being a decisive improvement over IFs, FEVDs are as good as their underlying statistical model. To avoid departing too much from the SM-CGL's framework, we only considered bivariate models. Climate systems obviously comprise numerous additional variables. Granville Tunnicliffe (2015) considers four in a VAR setup, but there could be many more. Additionally, the dynamics were assumed to be linear and time invariant, an approximation that should be eventually tested. Estrada et al. (2013) trend "breaks" model is one way to do it and they report results pointing in the same direction as ours. However, with machine learning tools becoming increasingly common use, more flexible alternatives could be used to yield further insights. Finally, we considered simplistic identification schemes which implied a causal ordering of variables. There is a plethora of more sophisticated schemes available (Kilian and Lütkepohl, 2017) and those could be used in future work, especially when moving beyond bivariate systems. Another – simpler – avenue is the use of data sampled at higher frequencies (like daily) which, by construction, makes the simultaneity problem much less of a Damocles sword.

References

Andrews, T., Forster, P., Boucher, O., Bellouin, N., and Jones, A. (2010). Precipitation, Radiative Forcing and Global Temperature Change. *Geophysical Research Letters*, 37(14).

Bindoff, N., Stott, P., AchutaRao, P., Allen, M., Gillett, N., Gutzler, D., Hansingo, K., Hegerl, G., Hu, Y., Jain, S., Mokhov, I., Overland, J., Perlwitz, J., Sebbari, R., and Zhang, X. (2013). Chapter 10 - Detection and Attribution of Climate Change: From Global to Regional. In *Climate Change 2013: The Physical Science Basis. IPCC Working Group I Contribution to AR5.* Cambridge University Press, Cambridge.

Bruns, S., Csereklyei, Z., and Stern, D. (2020). A Multicointegration Model of Global Climate Change. *Journal of Econometrics*, 214(1):175–197. Annals Issue: Econometric Models of Climate Change.

Chen, G., Glen, D. R., Saad, Z. S., Hamilton, J. P., Thomason, M. E., Gotlib, I. H., and Cox, R. W. (2011). Vector autoregression, structural equation modeling, and their synthesis in neuroimaging data analysis. *Computers in biology and medicine*, 41(12):1142–1155.

Estrada, F., Perron, P., and Martínez-López, B. (2013). Statistically derived contributions of diverse human influences to twentieth-century temperature changes. *Nature Geoscience*, 6(12):1050–1055.

Etminan, M., Myhre, G., Highwood, E., and Shine, K. (2016). Radiative Forcing of Carbon Dioxide, Methane, and Nitrous Oxide: A significant Revision of the Methane Radiative Forcing. *Geophysical Research Letters*, 43(24):12,614–12,623.

Forster, P., Forster, H., Evans, M., Gidden, M., Jones, C., Keller, C., Lamboll, R., Quéré, C., Rogelj, J., Rosen, D., Schleussner, C.-F., Richardson, T., Smith, C., and Turnock, S. (2020). Current and Future Global Climate Impacts Resulting from COVID-19. *Nature Climate Change*, 10(10):913–919.

Goulet Coulombe, P. and Göbel, M. (2021). Arctic amplification of anthropogenic forcing: A vector autoregressive analysis. *Journal of Climate*, Forthcoming.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438.

Granville Tunnicliffe, W. (2015). Atmospheric co2 and global temperatures: the strength and nature of their dependence. *Working Paper*.

Hansen, J., Sato, M., Ruedy, R., Lo, K., Lea, D., and Medina-Elizade, M. (2006). Global Temperature Change. *Proceedings of the National Academy of Sciences*, 103(39):14288–14293.

Kilian, L. and Lütkepohl, H. (2017). *Structural vector autoregressive analysis*. Cambridge University Press.

Koutsoyiannis, D. and Kundzewicz, Z. W. (2020). Atmospheric temperature and co2: Hen-or-egg causality? *Sci*, 2(4):83.

Li, C., Notz, D., Tietsche, S., and Marotzke, J. (2013). The Transient versus the Equilibrium Response of Sea Ice to Global Warming. *Journal of Climate*, 26(15):5624–5636.

Liang, X. (2008). Information Flow within Stochastic Dynamical System. *Phys. Rev. E*, 78(5):031113.

Liang, X. (2014). Unraveling the Cause-Effect Relation between Time Series. *Phys. Rev. E*, 90:052150.

Liang, X. (2015). Normalizing the Causality between Time Series. *Phys. Rev. E*, 92:022126.

Liang, X. (2016). Information Flow and Causality as rigorous Notions ab initio. *Phys. Rev. E*, 94:052201.

Marotzke, J. and Forster, P. (2015). Forcing, Feedback and Internal Variability in Global Temperature Trends. *Nature*, 517(7536):565–570.

Masson-Delmotte, V., Zhai, P., Pörtner, H.-O., Roberts, D., Skea, J., Shukla, P., Pirani, A., Moufouma-Okia, W., Péan, C., Pidcock, R., Connors, S., Matthews, J.B.R.and Chen, Y., Zhou, X., Gomis, M., Lonnoy, E., Maycock, T., Tignor, M., and Waterfield, T. (2018). IPCC, 2018: Global Warming of 1.5°C. An IPCC Special Report on the Impacts of Global Warming of 1.5°C above Pre-Industrial Levels and related Global Greenhouse Gas Emission Pathways, in the Context of Strengthening the Global Response to the Threat of Climate Change, Sustainable Development, and Efforts to Eradicate Poverty. *In press*.

Meinshausen, M., Vogel, E., Nauels, A., Lorbacher, K., Meinshausen, N., Etheridge, D., Fraser, P., Montzka, S., Rayner, P., Trudinger, C., Krummel, P., Beyerle, U., Canadell, J., Daniel, J., Enting, I. G., Law, R., Lunder, C., O'Doherty, S., Prinn, R., Reimann, S., Rubino, M., Velders, G., Vollmer, M., Wang, R., and Weiss, R. (2017). Historical greenhouse gas concentrations for climate modelling (cmip6). *Geoscientific Model Development*, 10(5):2057–2116.

Montamat, G. and Stock, J. (2020). Quasi-experimental estimates of the transient climate response using observational data. *Climatic Change*, 160(3):361–371.

Myhre, G., Shindell, D., Bréon, F.-M., Collins, W., Fuglestvedt, J., Huang, J., Koch, D., Lamarque, J.-F., Lee, D., Mendoza, B., Nakajima, T., Robock, A., Stephens, G., Takemura, T., and Zhang, H. (2013). Anthropogenic and Natural Radiative Forcing. In Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Doschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P., editors, *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 659–740. Cambridge University Press, Cambridge, UK.

Nordhaus, W. (2014). Estimates of the social cost of carbon: Concepts and results from the dice-2013r model and alternative approaches. *Journal of the Association of Environmental and Resource Economists*, 1(1/2):273–312.

Notz, D. and Stroeve, J. (2016). Observed arctic sea-ice loss directly follows anthropogenic co2 emission. *Science*, 354(6313):747–750.

Otto, A., Otto, F., Boucher, O., Church, J., Hegerl, G., Forster, P., Gillett, N. P., Gregory, J., Johnson, G., Knutti, R., Lewis, N., Lohmann, U., Marotzke, J., Myhre, G., Shindell, D., Stevens, B., and Allen, M. (2013). Energy Budget Constraints on Climate Response. *Nature Geoscience*, 6(6):415–416.

Phillips, P., Leirvik, T., and Storelvmo, T. (2020). Econometric Estimates of Earth's Transient Climate Sensitivity. *Journal of Econometrics*, 214(1):6–32. Annals Issue: Econometric Models of Climate Change.

Pretis, F. (2020). Econometric Modelling of Climate Systems: The Equivalence of Energy Balance Models and Cointegrated Vector Autoregressions. *Journal of Econometrics*, 214(1):256–273. Annals Issue: Econometric Models of Climate Change.

Shukla, P., Skea, J., Calvo Buendia, E., Masson-Delmotte, V., Pörtner, H.-O., Roberts, D., Zhai, P., Slade, R., Connors, S., van Diemen, R., Ferrat, M., Haughey, E., Luz, S., Neogi, S., Pathak, M., Petzold, J., Portugal Pereira, J., Vyas, P., Huntley, E., Kissick, K., Belkacemi, M., and Malley, J. (2019). Climate Change and Land: an IPCC Special Report on Climate Change, Desertification, Land Degradation, Sustainable Land Management, Food Security, and Greenhouse Gas Fluxes in Terrestrial Ecosystems. *In press*.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica: journal of the Econometric Society*, pages 1–48.

Stips, A., Macias, D., Coughlan, C., Garcia-Gorriz, E., and Liang, X. (2016). On the Causal Structure between CO2 and Global Temperature. *Scientific Reports*, 6:21691.

Tawia Hagan, D. F., Wang, G., San Liang, X., and Dolman, H. A. (2019). 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.

Tollefson, J. (2014). Climate Change: The Case of the Missing Heat. *Nature*, 505:276–278.

Uhlig, H. (2005). What are the effects of monetary policy on output? results from an agnostic identification procedure. *Journal of Monetary Economics*, 52(2):381 – 419.
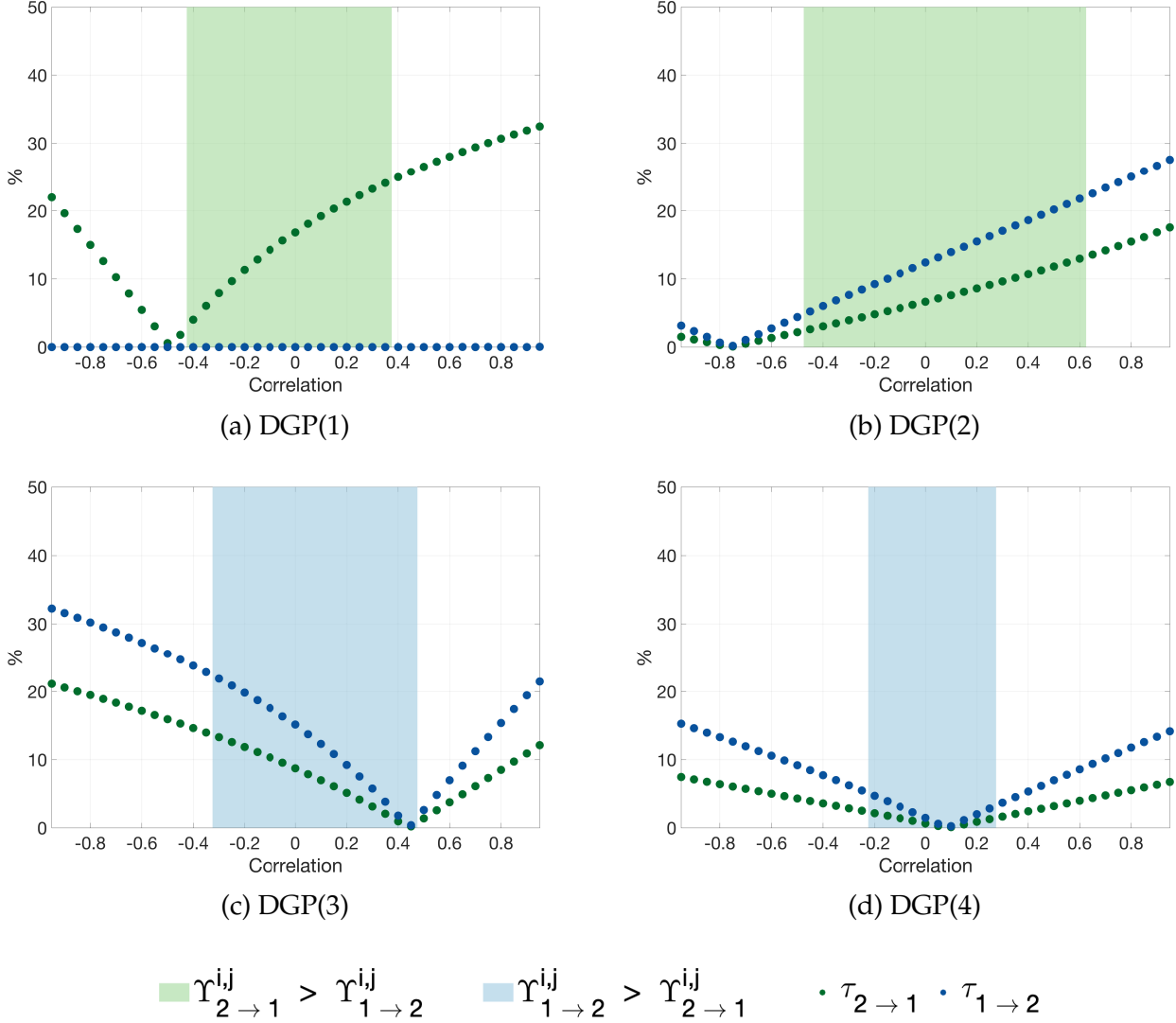
A Data Sources

For the empirical part of Section 4, we refer to the variables presented in Table 1 of SMCGL. Due to the sparse description of the data sources, our data set reduces to the annual global mean surface temperature anomalies (GMTA) and seven other representatives of radiative forcing. For five variables we could match the reported correlation coefficients as well as the IF values of SMCGL, Table 1. For *total forcing* and *solar* these measures do not coincide. For the analysis in Section 4.2 and 4.3, we use the CO_2 series based on Meinshausen et al. (2017). We transform *ppm* into W/m^2 as described in Section 4.

Table 3: List of Variables

Abbreviation	Description	Data Source
Total Forcing	annual; 1850-2005	KNMI Climate Explorer
Anthropogenic	annual; 1850-2005	KNMI Climate Explorer
CO_2-ERF (W/m^2)	annual; 1850-2005	KNMI Climate Explorer
Aerosol	annual; 1850-2005	KNMI Climate Explorer
Solar	annual; 1850-2005	KNMI Climate Explorer
Volcanic	annual; 1850-2005	KNMI Climate Explorer
PDO	annual averages of monthly observations; 1900-2005	KNMI Climate Explorer
GMTA	annual; global; 1900-2005	HadCRUT4
CO_2 (Million Tonnes / Year)	annual; global production-based emissions; 1850-2005	Our World in Data – not in use
CO_2 (ppm)	annual; global; 1850-2014 – based on Meinshausen et al. (2017)	1850-2014: IAC ETH Zürich
	annual; global; 2015-2017 – based on NOAA-ESRL	2015-2017: NOAA-ESRL

B Additional Simulation Results

In Figure 4 below, we compare NIF and $\text{FEVD}_{h=2}$. One might be concerned that results in Figure 1 suffer from a horizon mismatch between NIF and $\text{FEVD}_{h=10}$. Results suggest that reducing h worsens NIFs' problems by shrinking the "safe" regions. This is intuitive, the effect of assumptions on simultaneous relationships gets milder as we get further from $h = 0$.

(a) DGP(1)

(b) DGP(2)

(c) DGP(3)

(d) DGP(4)

$$\Upsilon^{i,j}_{2 \to 1} > \Upsilon^{i,j}_{1 \to 2} \qquad \Upsilon^{i,j}_{1 \to 2} > \Upsilon^{i,j}_{2 \to 1} \qquad \cdot\ \tau_{2 \to 1} \quad \cdot\ \tau_{1 \to 2}$$

Notes: Color specification: the *light blue* area shows the region, in which FEVDs suggest $Y^{i,j}_{1 \to 2} > Y^{i,j}_{2 \to 1}$, regardless of the ordering. Similarly, the *light green* area shows the region, in which FEVDs suggest $Y^{i,j}_{2 \to 1} > Y^{i,j}_{1 \to 2}$ – independent of the ordering. The *white/non-colored* areas shows those regions, for which the ordering of $X1$ and $X2$ does matter. An unambiguous determination of $Y^{i,j}_{1 \to 2} > Y^{i,j}_{2 \to 1}$, respectively $Y^{i,j}_{2 \to 1} > Y^{i,j}_{1 \to 2}$, is not possible.

Figure 4: Ranking of NIFs ($\tau_{i \to j}$) vs Ranking of FEVDs ($Y^{i,j}_{i \to j}$)
Different Levels of Correlation; Horizon $h = 2$

C   Additional Empirical Results

Table 4: Empirical Results for the Bivariate Relationship
Between Various Forcings and GMTA
Sample Period: 1850-**2017**

	Correlation	Normalized IF (IF × 100)		Lags (P)	Correlation of Residuals (ρ_u)	Ordering: i, GMTA		Ordering: GMTA, i	
		$i \to$ GMTA	GMTA $\to i$			$i \to$ GMTA	GMTA $\to i$	$i \to$ GMTA	GMTA $\to i$
Total Forcing	0.82	**36.8**	29.5	4	0.23***	**48.0**	15.6	28.3	**29.3**
		(17.2)	(15.3)	1	0.29***	**55.7**	14.4	30.0	**36.6**
Anthropogenic	0.91	**43.7**	-18.9	4	-0.16**	**6.5**	5.0	4.9	**13.5**
		(39.9)	(-0.6)	1	-0.15**	4.2	**5.1**	2.4	**13.7**
CO_2 - ERF (W/m^2) SMCGL	0.91	**43.5**	-13.0	4	-0.10	5.75	**9.2**	5.3	**15.4**
		(39.0)	(-0.3)	1	-0.11	2.1	**3.1**	1.2	**8.0**
Aerosol	-0.84	**37.9**	-45.7	4	-0.17**	**3.9**	2.9	**6.1**	0.0
		(19.4)	(-1.3)	1	-0.00	2.0	**33.3**	2.0	**33.1**
Solar	31.4	**7.0**	2.1	8	0.05	**4.6**	0.9	**3.1**	1.4
		(1.1)	(0.6)	1	0.06	**6.7**	1.0	**4.4**	2.0
Volcanic	0.11	**1.3**	-0.3	4	0.18**	**10.1**	0.5	**2.7**	2.4
		(0.2)	(-0.2)	1	0.20***	**7.1**	0.3	0.6	**3.7**
PDO	0.15	**-2.3**	-0.4	4	0.4***	**31.0**	0.8	3.9	**13.7**
		(-0.3)	(-0.3)	1	0.38***	**9.5**	0.3	0.7	**13.5**
CO_2 (Mt/yr)	0.89	**42.0**	1.0	2	-0.12	**7.9**	1.9	**8.8**	0.3
		(31.0)	(0.00)	1	-0.10	**5.4**	0.0	**5.4**	0.7
CO_2 (W/m^2)	0.91	**43.4**	-13.4	2	0.18**	5.0	**16.1**	6.1	**6.2**
		(38.3)	(-0.3)	1	0.06	1.5	**3.4**	1.0	**1.6**

Notes: i corresponds to the type of radiative forcing, listed in the left most column. The second column ("Correlation") documents the correlation between GMTA and variable i. FEVD values are taken at horizon $h = 15$, which translates into the contribution of variable i in the variance of the forecast error of variable j a decade and a half after the in-sample end date. Numbers in bold underline the highest absolute causal flow among a (i,GMTA) pair for a given measure. "*", "**", and "***" means that the null of the residuals cross-correlation of residuals is rejected at the 10%,5% and 1% level respectively. Sample period: 1850-2017.

Table 5: Transient Climate Response
Sample Period: 1850-**2017**

Ordering	Without Trend		With Trend	
	TCR_{20}	TCR_{70}	TCR_{20}	TCR_{70}
CO_2, GMTA	1.46°C	1.94°C	1.76°C	2.35°C
GMTA, CO_2	0.58°C	1.79°C	0.97°C	2.22°C