# scientific reports

OPEN

# Extracting causation from millennial-scale climate fluctuations in the last 800 kyr

Marco Baldovin[1,4✉], Fabio Cecconi[2], Antonello Provenzale[3] & Angelo Vulpiani[4]

The detection of cause-effect relationships from the analysis of paleoclimatic records is a crucial step to disentangle the main mechanisms at work in the climate system. Here, we show that the approach based on the generalized Fluctuation–Dissipation Relation, complemented by the analysis of the Transfer Entropy, allows the causal links to be identified between temperature, $CO_2$ concentration and astronomical forcing during the glacial cycles of the last 800 kyr based on Antarctic ice core records. When considering the whole spectrum of time scales, the results of the analysis suggest that temperature drives $CO_2$ concentration, or that are both driven by the common astronomical forcing. However, considering only millennial-scale fluctuations, the results reveal the presence of more complex causal links, indicating that $CO_2$ variations contribute to driving the changes of temperature on such time scales. The results also evidence a slow temporal variability in the strength of the millennial-scale causal links between temperature and $CO_2$ concentration.

Earth's climate is a complex nonlinear system in which multiple feedback mechanisms control the stability, variability and/or abrupt transitions between different climatic states (see e.g. Refs.[1,2]). Such feedbacks generate internal, intrinsic climatic oscillations and can amplify or damp the effects of external forcing factors, see e.g. Refs.[3–8]. In such a framework, identifying cause-effect relationships from an observed behavior is often difficult, and refined mathematical approaches and data analysis methods able to go beyond correlation estimates are needed to disentangle causation links, as discussed for example in Refs.[9–11].

One outstanding example of climate variability is the case of glacial-interglacial oscillations of the Pleistocene. In the last three million years, Earth's climate fluctuated between prolonged glacial periods, slowly developing through global temperature decrease and the build-up of extended ice sheets, and shorter interglacials with milder climate, generated by a relatively rapid (in geological sense) melting of the ice[12–15]. Such glacial cycles are believed to be a nonlinear reaction of the climate, or of some of its sub-systems, to the slow variation of the orbital forcing via amplifying feedbacks[16]. The Antarctic ice cores drilled at Vostok[17] and, more recently, by the EPICA project[18] have revealed the details of the glacial oscillations in the last few hundred thousand years[19,20]. Approximately synchronous variations of the reconstructed Antarctic temperature and of carbon dioxide concentration in the paleo-atmosphere are visible[21,22], although the precise lead-lag relationships between (Antarctic) temperature and $CO_2$ concentration are still a matter of debate. In particular, such lead-lag relations could vary on both the time scale and the specific period considered[23]. For example, a detailed correlation analysis of a high-resolution Antarctic ice core has indicated that during the last glacial period there is a lagged variation of $CO_2$ with respect to temperature on millennial time scales, which however becomes more complex at centennial time scales[24]. The interpretation of such time-scale dependence of the lead-lag relationships between $CO_2$ and $T$ can be offered in terms of the presence of multiple mechanisms at different time scales[23]. One possibility is the interplay of a slower process associated with the reorganization of the Southern Ocean carbon cycle, and faster (possibly abrupt) processes associated with Northern Hemisphere Dansgaard-Oeschger events[24,25]. Of course, whether the dynamics have been considered synchronous depends a lot on the quality of the age model and on whether lagged correlation or actual differences in specific change points are considered.

On the other hand, the correlation between two variables is not, in itself, a reliable measure of causation, as already pointed out in Ref.[23] for paleoclimate dynamics. A typical case in which correlation fails to catch the underlying causal structure is when two mutually independent variables $x_1(\tau)$ and $x_2(\tau)$ are driven by a common forcing $f(\tau)$, $\tau$ being the absolute time. In this case, a strong correlation can be easily misinterpreted as causation.

[1]CNRS, LPTMS, Université Paris-Saclay, 530 Rue André Rivière, 91405 Orsay, France. [2]CNR-Istituto dei Sistemi Complessi and INFN, Unità di Roma 1, 00185 Rome, Italy. [3]CNR-Istituto di Geoscienze e Georisorse, 56124 Pisa, Italy. [4]Dipartimento di Fisica, Università "Sapienza", 00185 Rome, Italy. ✉email: marco.baldovin@universite-paris-saclay.fr

This situation is frequently encountered in the climate system, as discussed in Ref.[10]. A relevant challenge is thus to disentangle the cause-effect relationships from the analysis of the two signals. In past works on glacial oscillations, the issue of causation has been addressed by the work in Ref.[26], which looked at the causal structure of the temperature-$CO_2$ concentration relations using an information flow approach[27]. However, confounding factors can be present and different causal relationships can exist on different time scales[23]. These issues should not be overlooked and they are further addressed in the present manuscript.

One reliable definition of cause-effect relationship, able to take into account different behaviors at different scales, is based on the observation of the average trajectory of $x_2(\tau)$ after an active perturbation of the variable $x_1(\tau)$ has been performed. This kind of "probing" resembles the idea behind the mathematical formal definition of causation by Judea Pearl[28]. In dynamical systems it can be characterized by looking at the linear response matrix function[29, 30]

$$R_{ij}(t; \tau) = \frac{\overline{\delta x_i(\tau + t)}}{\delta x_j(\tau)}. \tag{1}$$

Here $\delta x_j(\tau)$ is the value of an instantaneous, external perturbation operated on the variable $x_j$ at time $\tau$, while $\overline{\delta x_i(\tau + t)}$ is the average (over many repetitions of the experiment) of the difference between the perturbed trajectory of $x_i(\tau)$ and its unperturbed evolution. In what follows, we will assume that the above defined quantity does not depend on the absolute time $\tau$, but only on the lag $t$, so that $R_{ij}(t; \tau) \equiv R_{i,j}(t)$. From Eq. (1) it can be deduced that $R_{ij}(0) = \delta_{ij}$ by definition, where $\delta_{ij}$ is the Kronecker-delta. Thus the diagonal entries decay from the starting value $R_{ii}(0) = 1$, while the off-diagonal entries grow from the starting value $R_{i \neq j} = 0$. From a physical point of view, this simply means that the variables of a system cannot generate an immediate reaction on the others, as some (possibly very small) delay has to occur between a "cause" and its "effect".

Of course, Eq. (1) cannot be applied to any problem for which only time series referring to past events are available. However, a series of well known results of response theory show that $R_{ij}(t)$ can be written in terms of time correlations of suitable quantities[31, 32]. One of the possible formulations of this principle, sometimes called generalized Fluctuation–Dissipation Relation (generalized FDR), is:

$$R_{ij}(t) = -\left\langle x_i(t) \frac{\partial}{\partial x_j} \log P(\mathbf{x})\big|_{\mathbf{x}(0)} \right\rangle. \tag{2}$$

Here $P(\mathbf{x})$ is the stationary probability distribution of the whole phase-space vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ describing the system dynamics; the vector $\mathbf{x}$ is meant to include all the variables that determine the behavior of $x_i$ and $x_j$.

In particular, if the dynamics of a system of $n$ variables $x_1, \ldots, x_n$ is linear, the matrix of the response functions simplifies to (see Supplemental Material for a complete derivation)

$$\mathbb{R}(t) = \mathbb{C}(t)\mathbb{C}^{-1}(0), \tag{3}$$

that is easily determined from the elements of the correlation matrix, $C_{ij}(t) = \langle x_i(\tau + t)x_j(\tau) \rangle$, of the available data sets [where $\mathbb{C}^{-1}(0)$ is the inverse matrix of $C_{ij}(0) = \langle x_i(\tau)x_j(\tau) \rangle$]. Hereafter, we will denote with $\mathbb{M}$ a matrix and with $M_{ij}$ the scalar values of its entries. To appreciate the difference with a simple correlations analysis, it is useful to explicitly write the matrix (3) for a two-dimensional system $[x(t), y(t)]$, where the inversion of $\mathbb{C}(0)$ can be easily performed. We assume that $x(t)$ and $y(t)$ have zero average and unitary variance, as in the following we will always consider normalized signals of this sort. In this simple case, the 2×2 response matrix reads

$$\mathbb{R}(t) = \begin{pmatrix} \dfrac{C_{xx}(t) - C_{xy}(0) \, C_{xy}(t)}{1 - C_{xy}^2(0)} & \dfrac{C_{xy}(t) - C_{xy}(0) \, C_{xx}(t)}{1 - C_{xy}^2(0)} \\[3mm] \dfrac{C_{yx}(t) - C_{xy}(0) \, C_{yy}(t)}{1 - C_{xy}^2(0)} & \dfrac{C_{yy}(t) - C_{xy}(0) \, C_{yx}(t)}{1 - C_{xy}^2(0)} \end{pmatrix}. \tag{4}$$

The relation $C_{xx}(0) = C_{yy}(0) = 1$, following from the normalization of the data, has been used. In general, $C_{xy}(t)$ is different from $C_{yx}(t)$, the symmetry $C_{xy}(0) = C_{yx}(0)$ holding true only for $t = 0$. We remark that, even if the response can be computed as a combination of correlation functions, it provides information about the causal structure of the system which could not be deduced from cross-correlations alone[29]. Note that this result is quite robust: the presence of small nonlinearities in the dynamics is not expected to spoil the ability of Eq. (3) to detect causal links[30].

In principle, the generalized FDR solves the problem of inferring causal relations for any dynamical system, but its application strongly relies on the assumption that the dynamics of the chosen set of observables does not depend on any variable that is external to the system (i.e., in the language of stochastic processes, that the dynamics is Markovian). Moreover, the shape of the steady state distribution has to be known, at least approximately.

The application of this kind of analysis to the EPICA paleoclimate data is thus limited by two factors: (i) the lack of knowledge of a proper set of variables $\mathbf{x} = (x_1, \ldots, x_n)$ fully describing a Markovian dynamics and (ii) the relative shortage of data (about 1600 measurements, covering a range of 800 kyr), which would not allow a reliable estimate of the joint probability distribution $P(\mathbf{x})$, even if a valid set of observables $\mathbf{x}$ was known.

In this paper, we focus on the variations of the carbon dioxide concentration, $[CO_2]$, and the reconstructed temperature $T$, during glacial-interglacial oscillations. Assuming a sufficient time-scale separation between the astronomical forcing (with a typical time of the order of 20 kyr or more) and the internal climate variability on millennial or shorter time scales, we delineate a qualitative analysis of the causation relations between $[CO_2]$

and $T$ in the last 800 kyr. The strategy is based on the definition of a proper set of "fast" components whose dynamics is assumed to mainly reflect internal climate variability, associated with the interaction of the different climate sub-systems. Notice, however, that the intensity of such fast fluctuations could be modulated by the climate background state and thus by the astronomical forcing[33]. Here, the important hypothesis is that the fast oscillations are not completely "slaved" to the slow forcing. Within the reasonable assumption that the mutual interactions of the fast components can be approximately linearized, the response on fast time-scales of the order of one to two kyr can be inferred by exploiting Eq. (3), which, for these variables, can be evaluated with good accuracy even with the available quantity of data. Special attention should be given to the temporal resolution of the paleoclimatic data, which, in some portions of the record, may hamper the ability to safely detect millennial-scale oscillations, a point that is further addressed below and in the "Methods" section. Important messages of this work are that (a) FDR analysis provides relevant information on the causation relationships in (paleo)climate signals, (b) considering only unfiltered data including all time scales can lead to incomplete results, masking the possible emergence of more complex causal relationships on specific time scales, and (c) there is a clear long-term temporal variation in the strength of the millennial-scale causal relationships between temperature and $CO_2$ concentration.

## Results

As discussed in the Introduction, the generalized FDR can be used to unravel causal links between the variables of a physical system by analyzing their correlations, provided that (i) the dynamics is not subject to an external common driving which simultaneously forces several variables and (ii) the stationary distribution is known, at least approximately. The coupled dynamics of $T$ and $[CO_2]$ in the last 800 kyr does not fulfill any of these two conditions, as their behaviour is heavily conditioned by the external astronomical driving, and the relatively small amount of available data (about 1600 measurements spanning the whole record) does not allow to reconstruct a reliable joint probability distribution for the two quantities.

In the "Methods" section, we show that both the above difficulties can be circumvented as long as there is a sufficient scale separation between the (slow) typical time scale of the external driving and the (short) characteristic times of the interaction between the variables. This is indeed the case for Pleistocene climatic variability, where the time scale of the external astronomical forcing is of the order of 20 kyr or more (Milankovitch cycles[16]), while intense climatic fluctuations occur on a much faster time scale, of about 1 kyr.

The key ingredient for the analysis is thus a proper high-pass temporal filtering of the signals: before applying the FDR, we subtract from the time series of $T$ and $[CO_2]$ a running average (see "Methods" for details) over windows, $T_w$, of a few kyr. This basically allows to filter out any possible spurious correlation due for example to a common influence of the slow external forcing, while keeping the relevant information on the short-time mutual relationship. In addition, the distribution of the filtered variables is approximately Gaussian ("Methods" section, Fig. 4): this is consistent with the working hypothesis that the dynamics on the fast scales is approximately linear, and that the use of FDR in the form of Eq. (3) is justified.

The results of the analysis are shown in Fig. 1, where the response function computed with the generalized FDR is plotted as a function of time. The approach based on a straightforward application of Eq. (3) to the unfiltered data would suggest that the relative influence of temperature on $CO_2$ concentration, $T \rightarrow [CO_2]$, is much stronger than the reversed causal link $[CO_2] \rightarrow T$. When the formula is instead applied to the filtered data, a different scenario is observed. For the whole time series, the relative influence of the temperature on $[CO_2]$ is at most $\simeq 0.1$ (in a scale in which the self-response at time 0 is set equal to 1), and it almost vanishes for lags beyond 2 kyr; on the other hand the intensity of the $[CO_2] \rightarrow T$ causal relationship increases with respect to what is observed without high-pass filtering, doubling that of the reverse relation on the scale of 1 kyr. The causal link $[CO_2] \rightarrow T$ does now vanish at longer lags, displaying an oscillating behavior as shown also by the Transfer Entropy results.

As discussed in the "Methods" section, the width $T_w$ of the time window for the filter has been chosen to be 3 kyr. Figure 1 also shows that the method is quite robust with respect to the choice of $T_w$: indeed, we obtained an analogous behaviour of the response functions using $T_w = 2$ kyr and $T_w = 4$ kyr.

An important point concerns the temporal resolution of the two signals considered. Figure S4 of the Supplemental Material shows that the resolution of the $CO_2$ record is generally coarser than that of temperature, and both vary in time. To avoid possible spurious effects generated by the different temporal resolution and the different weight of the spline interpolation, we opted for degrading the resolution of the temperature signal to that of the carbon dioxide concentration (and vice-versa, in the few intervals where the temporal resolution of $[CO_2]$ is higher than that of $T$). Considering instead the results of the original (non degraded) data, as shown in Fig. S5 of the Supplemental Material, the main messages do not change.

A similar analysis, in which the transfer entropy (TE) between the two variables is computed before and after the filtering procedure (Fig. 1e,f), confirms that a high-pass filtering of the data is crucial to elucidate the causal relationships between temperature and $CO_2$ concentration in the last 800 kyr of the Pleistocene. A brief discussion on the concept of TE, which may be regarded as complementary to that of response, is given in the "Methods" section, where also further details about its application in this context are provided. Here it is worth noticing that the qualitative results are similar to those achieved with the generalized FDR approach: a direct application of the formula to the raw data would suggests a large influence of $T$ on $[CO_2]$; on the other hand, a proper temporal filtering leads to a more complex picture. A remarkable similarity between the behaviour of the two observables can be appreciated in Fig. 1d,f (recalling that TE, at variance with response, is a positive-definite quantity).
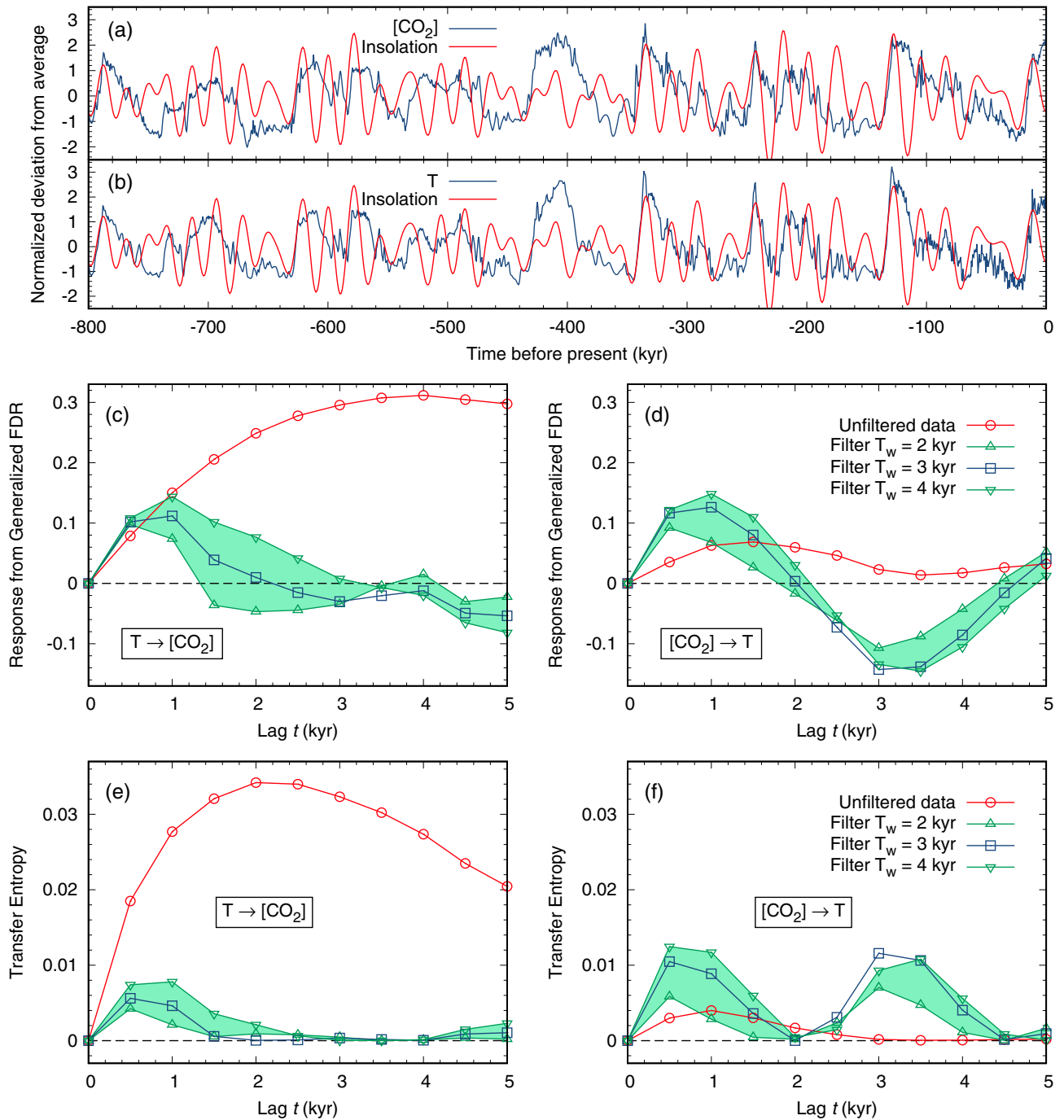
**Figure 1.** Analysis of the mutual influence between temperature $T$ and $CO_2$ concentration in paleoclimate data. Panels (**a**) and (**b**) show the deviation from average of the two signals as a function of time (blue curves), normalized by the standard deviation. The daily mean insolation at 65° N summer solstice, revealing the typical time scales of the external driving, is also reported (red curves; see "Methods" sections for details on the data sources). In panels (**c**) and (**d**) the response function, computed according to the Generalized FDR, is plotted. The analytical formula is given by the non-diagonal elements $R_{xy}(t)$ and $R_{yx}(t)$ of the linear response matrix (4), where $x$ and $y$ are the normalized $[CO_2]$ and $T$ signals shown in panels (**a**), (**b**) (and their high-pass filtered analogues). Panel (**c**) refers to the effect of $T$ on $[CO_2]$, while panel (d) shows the opposite relation. Red circles represent the results of a direct application of Eq. (3) on raw data, apparently suggesting that $T \to [CO_2]$ is stronger than $[CO_2] \to T$. The response on data filtered over $T_w = 3$ kyr window (blue squares) instead indicates that the impact of $[CO_2]$ on $T$ becomes larger. The result is robust with respect of $T_w$ variations by one kyr (green up/down triangles). A similar analysis, where TE are computed instead of generalised FDR, is shown in Panels (**e**) and (**f**). Here, we have considered the data for which the temporal resolution of the temperature record has been degraded to become similar to that of $CO_2$ concentration, as discussed in the "Methods" section. For the undegraded temperature data, the role of $CO_2$ driving is even larger, see Fig. S5 of the Supplemental Material.
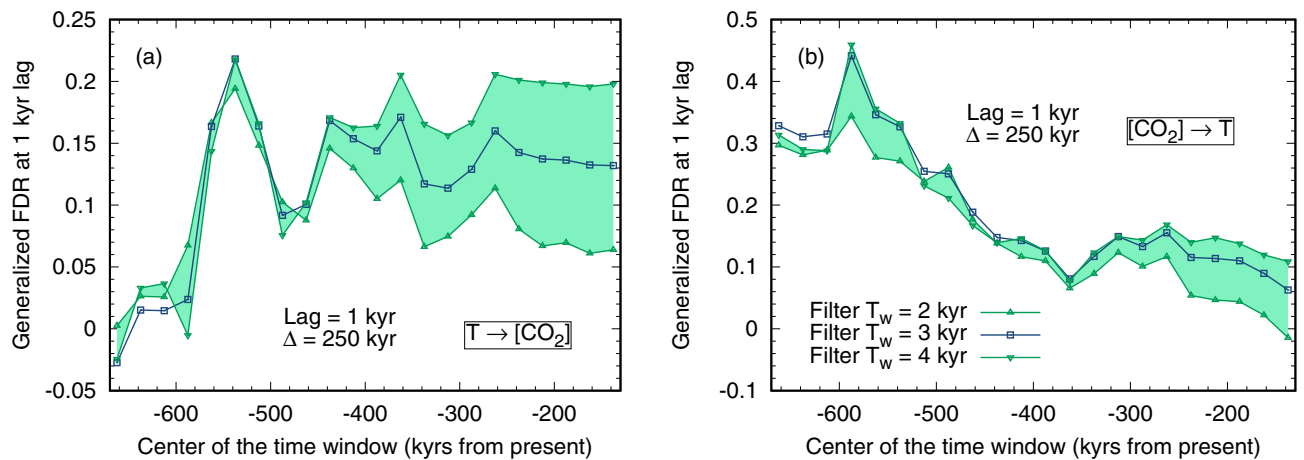
**Figure 2.** Values of the Generalized FDR at lag $t = 1$ kyr for the link $T \to [CO_2]$ (**a**) and $[CO_2] \to T$ (**b**), computed in a moving window with width $\Delta = 250$ kyr, slided along the whole record. All details as in panels (**c**) and (**d**) of Fig. 1.

## Discussion

The issue of which climatic signal drives which in the glacial-interglacial record is widely debated. For example, the analysis of Caillon et al.[34] indicated that $CO_2$ lagged Antarctic deglacial warming by 800 ± 200 years during a specific deglaciation event (Termination III 240,000 years ago). Subsequently, Parrenin et al.[22] found no asin-chronicity between Antarctic temperatures and $CO_2$ variations in the last deglaciation event (Termination TI), even though the situation is not always clear and it can vary with time[35, 36]. The work of Stips et al.[26], based on the use of the information flow to detect causal relations, revealed a complex pattern of cause-effect links, with a predominance of the Antarctic temperature driving $CO_2$ concentration when the whole record is considered. The analysis of millennial-scale fluctuations in the last glacial period showed that $CO_2$ seems to lag temperature by 500–1000 yr[24], while more complex relationships may exist on centennial time scales. Finally, the work of van Nes et al.[23] concluded that different relationships can exist on different time scales. Clearly, crucial to all these lag analyses is the availability of a safely calibrated time scale for both temperature and $CO_2$. In any case, here we found that the maximum value of the FDR for the effect of the $CO_2$ concentration on temperature for the high-pass filtered data is found between 500 and 1500 yr when considering the whole 800-kyr record. Thus, even an uncertainty of the age model of the order of 500 yr does not qualitatively change the results.

Here, we analysed causality links adopting the generalized Fluctuation–Dissipation Relation (see the "Methods" section for a detailed discussion of how this approach works), further confirming the results using the Transfer Entropy method. The main finding is that, using the data from Refs.[19, 20], we detected a causal link of temperature on $[CO_2]$ when considering the whole unfiltered record that includes both millennial-scale fluctuations and longer-term glacial-interglacial oscillations, in keeping with previous results[26]. On the scales of the astronomical forcing, albedo changes could drive temperature variations and consequently affect the whole cascade of climatic processes, including $CO_2$ changes. The causal link $T \to [CO_2]$ could thus be generated either by slow climatic processes, such as the global ocean's temperature-dependent ability to store $CO_2$, or simply reflect the fact that both climatic signals are controlled by a common driver, namely, the astronomical forcing with the related changes in summer solar insolation at high latitudes.

On the other hand, the significant novelty of our analysis is that the high-pass filtered paleoclimatic signal, including only fluctuations on scales of a few millennia, displays a more complex pattern of causal relationships, with mutual driving of $[CO_2]$ and temperature. This result is robust with respect to the precise value of the threshold used in the high-pass filter, which was varied between 2 and 4 kyr.

The results shown in Fig. 1 refer to the whole 800-kyr temporal period covered by the record, and they differ from the outcomes of some of previous analyses, performed on other signals spanning a more limited time range[24]. This supports the view that the strength of the causal links between temperature and $CO_2$, or more generally between the various components of the climate system, can vary with time, in line with the conclusions of Ref.[23]. Such view is confirmed by Figs. S6 and S7 of the Supplemental Material, where we analysed the first or the second half of the record. The results indicated that the strength of the causal links is different in the two periods. On millennial time scales, we always observed a mutual effect between $[CO_2]$ and $T$. In particular, the link $[CO_2] \to T$ is much stronger in the first 400 kyr of the record (Fig. S6), while in the last 400 kyr the reverse link $T \to [CO_2]$ becomes more relevant (Fig. S7).

To further explore this issue, in Fig. 2 we show the value of the Generalized FDR for $T \to [CO_2]$ and $[CO_2] \to T$ at lag 1 kyr (where the first maximum of the FDR is generally located), for a moving window of 250 ky slided along the whole time series. Interestingly, the effect of $[CO_2]$ on temperature is stronger in the earlier part of the record and it decreases in the course of time, while the reverse link $T \to [CO_2]$ grows in the early part of the record and then stabilizes at an approximately stationary value. Towards the end of the record, the two reverse links have comparable strength. At present, given the limited amount of data it can be difficult to disentangle between real variability and statistical fluctuations, but the results suggest a variability in the relative importance of the causal links between temperature and $CO_2$ concentration. As a word of caution, we also mention that

some of the inferred changes in causal relationships could be a result of changes in the data properties and the possibly non-stationary resolution properties of the two time series. In any case, these complex causal relationships would have been completely lost if we had considered only the unfiltered data.

A detailed analysis of the climatic processes inducing millennial-scale changes in $CO_2$ concentration is beyond the scope of this work. Here, we simply mention that the causes of the fluctuations of $CO_2$ concentration on such time scales are widely debated and still not fully clear, but most interpretations involve the role of $CO_2$ outgassing associated with changes in the ocean overturning circulation and/or marine ecosystem functioning[37, 38]. The negative value of the FDR, observed at a lag of about 3 kyr, can also point to a coupled oscillation in the climate system, although its nature remains currently undetermined. In any case, the results of our analysis support the view that internal climate mechanisms, rather than direct orbital forcing, are responsible for the main variability at millennial time scales in the last 800 kyr, in keeping with the conclusions of Ref.[23]. Further work using simple models such as that of Ref.[24] could help further addressing this issue; in this respect, it is worth noticing that the conclusion of Ref.[24] are consistent with our results for the last part of the analyzed time interval (see Fig. 2).

We emphasize that the results of the analysis presented here have to be regarded as qualitative. In fact, the relative scarcity of currently available data does not allow to claim the detection, within reasonable accuracy, of the detailed causal structure of glacial-interglacial dynamics on the whole spectrum of time scales. From a methodological point of view, our work clearly shows that the direct application of causality detection methods to unfiltered data may provide only a part of the story. The results reported here indicate that causality analysis can be a powerful approach to study paleoclimatic signals (such as multiple isotope records from ice cores, speleothems or sediments), provided that the data set is long enough and almost-linear interactions between the relevant climatic variables can be assumed on the temporal scales of interest.

## Methods

**Dataset.** The time series of reconstructed temperature and $CO_2$ concentration analyzed here are obtained from the EPICA Dome C ice drilling project in Antarctica, as described in Ref.[19]. Here we use the data described in Ref.[20], where the $CO_2$ concentration record was obtained by blending different ice cores and the chronology for the first 200 kyrs was revised and improved. In comparing the $CO_2$ concentration and temperature records, the issue of the gas-ice age difference (the so-called delta age) and its uncertainty should be kept in mind[39]. Here, we adopt the chronology indicated in the Ref.[20]. The time series of the insolation was calculated using the software provided at the site https://sites.google.com/site/geokotov/software and on the reconstructions of Ref.[40].

**Response function in the presence of slow external driving.** In this section we export the FDR formalism to cases in which slow time-dependent external driving is present. This is relevant in the climate context, where insolation drives the system on very slow time scales, compared to those for which experimental data are available.

First we will show that, for the considered class of models, the response function can be written in terms of correlation functions of suitably defined fast components of the dynamics. These components can be estimated, within reasonable approximations, from a proper filtering of the time series of the original variables. An example with a toy model is then discussed to illustrate the analytical results.

*Application of generalized FDR.* We will limit ourselves to the study of models in which the dynamics of the $n$ variables $x_i$, $i = 1, \ldots, n$ representing the state coordinates can be written as

$$\dot{\mathbf{x}} = -\mathbb{A}\mathbf{x} + \mathbf{c}f(t) + \boldsymbol{\xi}(t), \tag{5}$$

where $A$ is a $n \times n$ invertible, positive-definite and diagonalizable matrix; $\mathbf{c}$ is an $n$-dimensional vector of amplitudes, $\xi$ denotes a $\delta$-correlated diagonal noise $\langle \xi_i(t)\xi_j(s) \rangle = D_i \delta(t-s)\delta_{ij}$. We call $\tau_0$ the relaxation time of the free dynamics (i.e., without the forcing term), determined by the inverse of the spectral radius of $A$. The conditions on $A$ insure that the dynamics will not diverge in time. The diagonalizability requirement could actually be relaxed, as it is not essential to the proof, but it allows to simplify calculations: see Sec. 2 of the Supplemental Material for a brief discussion on this point. For our application to paleoclimate time series, it is reasonable to consider a slow forcing (e.g. periodic, or quasiperidic, with long periods)

$$f(t) = \sum_{i=1}^{l} a_i \cos(t/\tau_i + \phi_i), \tag{6}$$

where $\{a_i\}$ and $\{\phi_i\}$ are dimensionless constants $O(1)$. We assume that the $\{\tau_i\}$ are much larger than $\tau_0$, i.e.

$$\tau_l \geq \tau_{l-1} \geq \cdots \geq \tau_1 \gg \tau_0. \tag{7}$$

We are interested in the response of the system to an instantaneous perturbation $\mathbf{x}(0) \to \mathbf{x}(0) + \delta\mathbf{x}(0)$. In particular, we want to compute the response function (1), assuming that we ignore the details of the model, and that we only have access to the measured trajectories of the system. The generalized FDR (2) cannot be used *tout court* in this context; indeed, due to the presence of the external forcing, $f(t)$, the dynamics is not Markovian, because our set of variables $\mathbf{x}$ does not completely describe the state of the system.

It is then useful to decompose the full dynamics (5) into "slow" $\mathbf{x}_S$ and "fast" $\mathbf{x}_F = \mathbf{x} - \mathbf{x}_S$ components, evolving as

$$\dot{\mathbf{x}}_S = -\mathbb{A}\mathbf{x}_S + \mathbf{c}f(t) \tag{8}$$

$$\dot{\mathbf{x}}_F = -\mathbb{A}\mathbf{x}_F + \boldsymbol{\xi}(t) . \tag{9}$$

The above definition identifies a slow set of variables as those whose dynamics is only affected by the slow external forcing, while the uncorrelated noise is only present in the fast variables evolution.

Let us assume that the instantaneous change $\mathbf{x}(0) \to \mathbf{x}(0) + \delta\mathbf{x}(0)$, occurring at time $t = 0$, entirely affects the fast components $\mathbf{x}_F$. Of course we can always make such an assumption, as the only constraint imposed by the definition (9) is that the sum of $\mathbf{x}_F(0) + \mathbf{x}_S(0)$ is increased by $\delta\mathbf{x}(0)$. Denoting with "$\mathbf{x}^{(P)}$" the perturbed dynamics one has therefore

$$\mathbf{x}^{(P)}(t) - \mathbf{x}(t) = \mathbf{x}_F^{(P)}(t) - \mathbf{x}_F(t) \qquad \forall t > 0 , \tag{10}$$

which follows from the independence of the evolutions, $\mathbf{x}_F$ and $\mathbf{x}_S$. By definition, Eq. (1), the response function for the complete dynamics $\mathbf{x}$ is equal to that of the fast variables $\mathbf{x}_F$.

The physical meaning of this choice is easily understood in the context of paleoclimate, where the slow dynamics can be associated to the effect of the astronomical forcing, while the behaviour of the fast components is meant to be related to the internal climate dynamics. In this case our choice is equivalent to saying that the latter components are actually modified by an instantaneous perturbation (e.g. a large emission of $CO_2$ due to a volcano eruption), while the former, which only depend on astronomical motion, are not affected by this kind of events.

At this point, if the trajectories of the fast components, $\mathbf{x}_F$, were accessible, a plain employ of Eq. (2) would be possible, since the fast dynamics does not depend on the external forcing $f(t)$ and it is therefore Markovian. Moreover, since it is also described by a linear model, we could straightforwardly apply Eq. (3) and get:

$$\mathbb{R}(t) = \mathbb{C}_F(t)\mathbb{C}_F^{-1}(0) \tag{11}$$

with

$$\mathbb{C}_F(t) = \langle \mathbf{x}_F(t)\mathbf{x}_F^T(0) \rangle . \tag{12}$$

For the class of dynamics described by Eq. (5), this result provides an easy way to compute the response functions, once the dynamics of the fast variables is known.

*Evaluating of fast correlations from data filtering.* The computation of the response functions by means of Eq. (11) requires the evaluation of the correlation function matrix $C_F(t)$ appearing on the right hand side. The latter is usually not accessible from experiments and observations: if a large time-scale separation is present, however, such correlation functions can be estimated by considering a proper filtering of the dynamics. Remarkably, the quality of the approximation increases with the separation between the time scales.

The idea is to replace $\mathbf{x}_F$ by

$$\tilde{\mathbf{x}}(t) = \mathbf{x}(t) - \int_{-\infty}^{\infty} ds \, \mathcal{G}(t-s)\mathbf{x}(s) , \tag{13}$$

with

$$\mathcal{G}(t) = \frac{e^{-t^2/2T_w^2}}{\sqrt{2\pi}\, T_w} , \tag{14}$$

i.e. subtracting from the full dynamics a suitably defined running average. Here $T_w$ is the characteristic time-window of the Gaussian filter. The idea, not new[41], is that the filtered signal mimics the behaviour of the slowly varying components, so that $\tilde{\mathbf{x}}(t)$ can be seen as a "surrogate" of $\mathbf{x}_F(t)$; unlike $\mathbf{x}_F(t)$, however, $\tilde{\mathbf{x}}(t)$ can be easily computed from empirical data.

One of the advantages of using a Gaussian filter[42] relies on the possibility to show analytically that

$$\langle \tilde{\mathbf{x}}(t)\tilde{\mathbf{x}}^T(t') \rangle \simeq \langle \mathbf{x}_F(t)\mathbf{x}_F^T(t') \rangle + O(\max\{\tau_0/T_w, T_w^2/\tau_1^2\}) . \tag{15}$$

The details of the proof, which involves easy but tedious computations, are reported in the Supplemental Material, Sec. "Evaluation of the error". Here, the main point of the computation is the possibility to always find a $T_w$ such that both $\tau_0/T_w$ and $T_w^2/\tau_1^2$ are small, provided that the time-scale separation between $\tau_0$ and $\tau_1$ is large enough. The optimal order of magnitude for the width of the window is given by

$$T_w \sim (\tau_0 \tau_1^2)^{1/3} , \tag{16}$$

which ensures

$$\frac{\tau_0}{T_w} \simeq \frac{T_w^2}{\tau_1^2} \ll 1 . \tag{17}$$

Thus, the correlation functions in Eq. (11) can be written in terms of the quantities (13), which can be straightforwardly obtained form the time series. In other words, one has
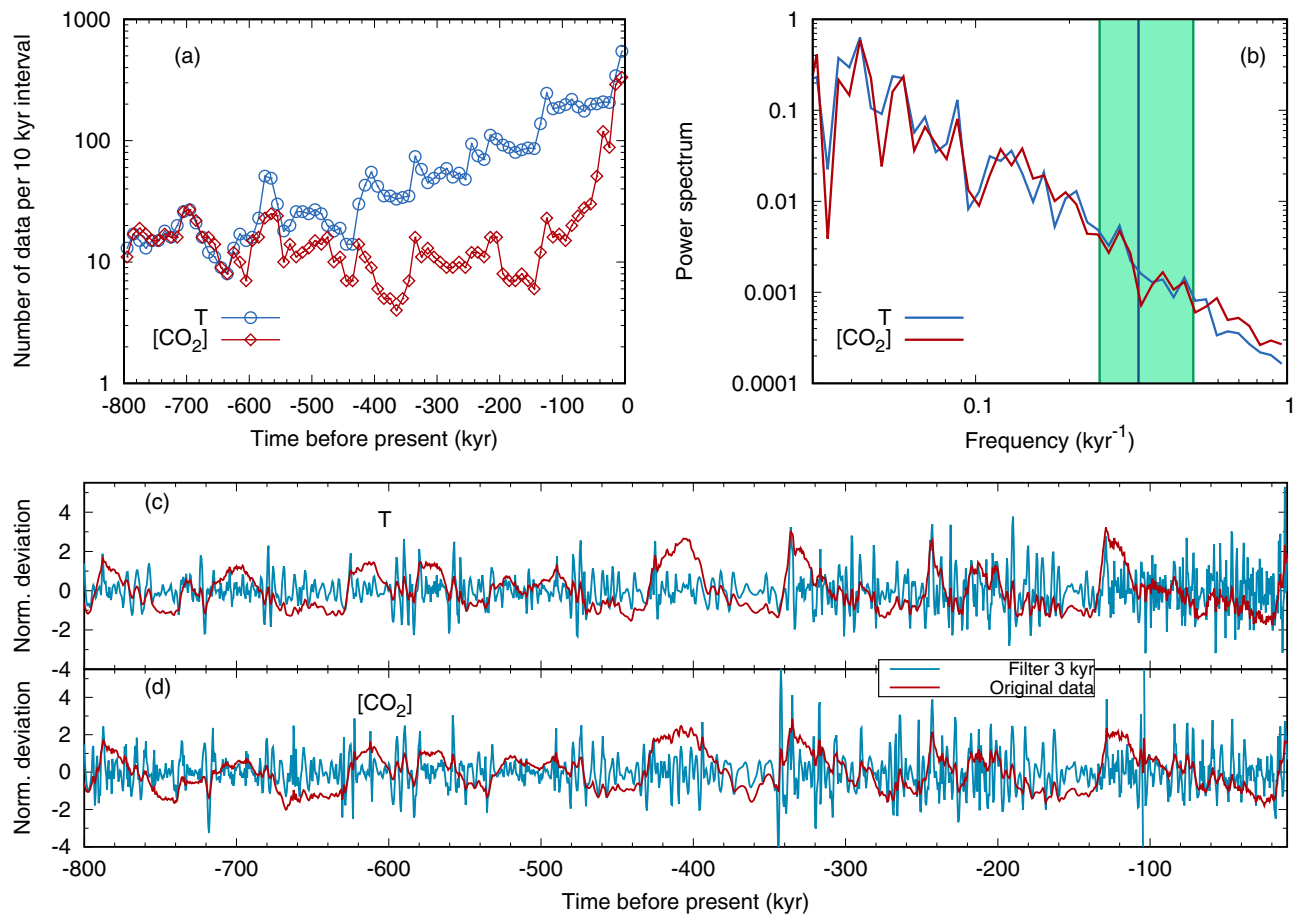
**Figure 3.** (**a**) Temporal resolution of the temperature (azure circles) and $CO_2$ concentration (red diamonds) along the record. In the analysis, we have degraded the temporal resolution of temperature to match that of $CO_2$. (**b**) Power spectrum of the time series for $T$ (azure) and $[CO_2]$ (red). The vertical green band indicates the frequencies corresponding to the three values of $T_w$ employed in the "Results" section. (**c**), (**d**) High-pass-filtered versions of the T and $[CO_2]$ signals, for the whole 800-kyr period, obtained by using a threshold $T_w = 3$ kyr, compared to the original datasets. Data are shown as normalized deviations from average.

$$\mathbb{R}(t) \simeq \tilde{\mathbb{C}}(t)\tilde{\mathbb{C}}^{-1}(0) \tag{18}$$

with $\tilde{\mathbb{C}}(t) = \langle \tilde{\mathbf{x}}(t)\tilde{\mathbf{x}}^T(0) \rangle$.

Section "Numerical examples" of the Supplemental Material illustrates how the above proposed combination of filtering and generalized FDR works in practice. We also show, numerically, that the method is robust with respect to the presence of (small) nonlinear terms in the fast dynamics.

**Application to paleoclimate.** To apply the proposed analysis to paleoclimate dynamics we assume, as a working hypothesis, that the dynamics of temperature and $[CO_2]$ can be approximated by a model of the form (5). Here $\tau_1$ is the typical time-scale of the Milankovitch series (approximately 40 kyr), while $\tau_0$ is a characteristic time of the fast dynamics of $T$ and $[CO_2]$, which we expect to be of the order of the kyr. The time-scale separation should then allow the application of our analysis, at least at a qualitative level. This scenario is supported by consistency checks which will be described in the remaining part of this Section.

*Data pre-processing.* Our study of generalized FDR on paleoclimate time series has required a pre-processing to make the data ready for the analysis. First, we operated a "data alignment", since the time series needed to be synchronized to make the generalized FDR applicable, while the original data were obviously not. In particular, the data of temperature and $[CO_2]$ were recorded on different set of times $a = \{t_1, t_2, \ldots, t_n\}$, $b = \{t'_1, t'_2, \ldots, t'_k\}$. We used a spline interpolation method to align the two datasets, in such a way that the resulting series were characterised by time intervals of 0.5 kyrs between consecutive entries (close to the original average time interval).

A relevant point concerns the fact that temperature and $CO_2$ have different temporal resolution, with $CO_2$ showing more sparse data than temperature, in most parts of the record. This is reported in Fig. 3a, which displays the temporal resolution of the $T$ and $[CO_2]$ records. In any case, the temporal resolution rarely becomes lower than 1 kyr, thus affecting only the shortest lag considered (0.5 kyr). The different abundance of data between the two signals may introduce spurious statistical effects when interpolating: points generated from a set with lower density are more correlated, and this may affect the subsequent analysis, at the shortest time scales. To
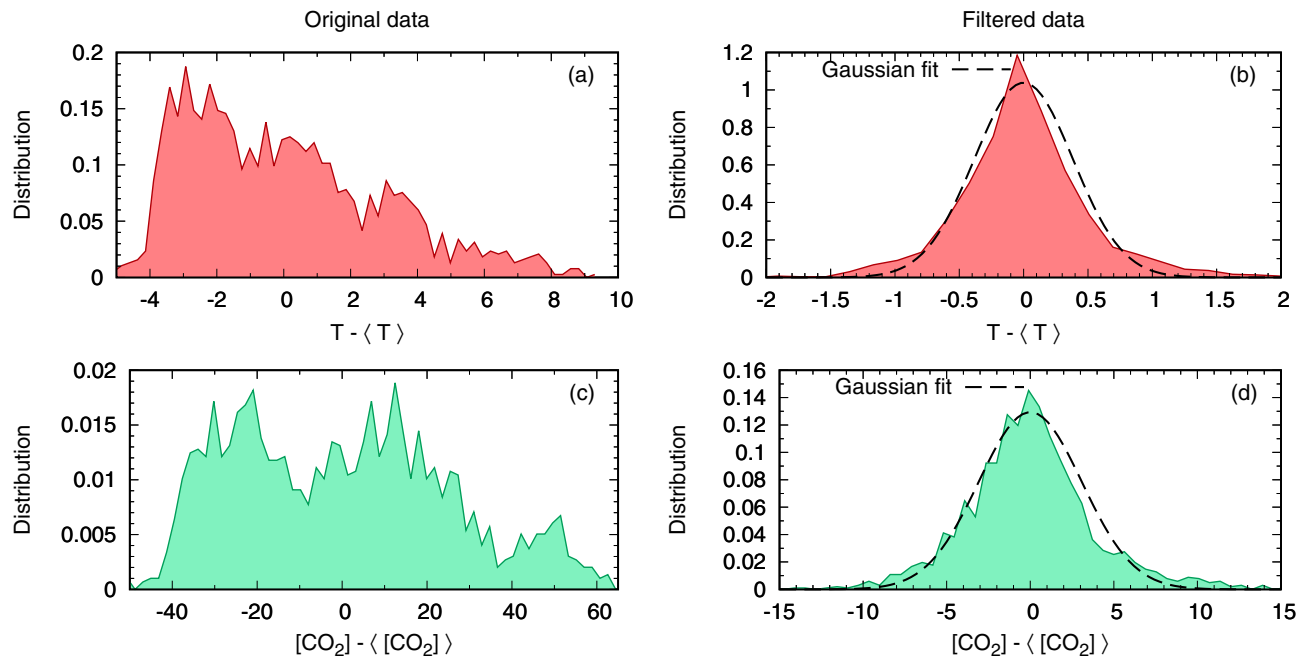
**Figure 4.** Histograms of the centered variables $T - \langle T \rangle$ and $[CO_2] - \langle [CO_2] \rangle$ before (Panels **a, c**) and after the filtering (Panels **b, d**). The filter makes the distribution of the signal Gaussian-like (compare with the Gaussian fits, dashed line in the right column); at the same time it reduces the excursion of the signal, since it removes the large oscillations due to the slow external forcing. Binning: for each plot, 60 bins of equal size are considered, ranging from the lowest to the highest recorded value.

avoid this kind of effects we degraded the temporal resolution of the records in such a way that they were always locally comparable. In particular, we divided the total observation interval (800 kyr) into 80 equal segments. For each of these 10 kyr intervals, we compared the number of available data for $T$ and $[CO_2]$, and we deleted from the larger sample a number of entries equal to the difference. The data to be deleted were taken at regular intervals in the sequence.

The analysis of the data with undegraded temporal resolution confirms the findings reported here, and it is discussed in Sec. 5 of the Supplemental Material, see also Fig. S5. The analysis of the two halves of the signal, shown in Figs. S6 and S7 of the Supplementary Material , was performed on the same temperature signal with reduced temporal resolution used here.

After filtering, we rescaled the values of temperature and $[CO_2]$ in order to be standardized: zero average and unitary variance. This is necessary to get rid of the degree of freedom due to the arbitrary choice of measure units, and allows to make comparisons between response functions relative to different physical quantities (see Ref.[30] for a discussion on this point).

*Width of the time window.* The power spectra of $T$ and $[CO_2]$, Fig. 3b, show a common regime, smoothly decreasing with almost power-law dependence at time scales shorter than about 10 kyr. As such, there is no spectral gap at a precise frequency. We applied the high-pass filter at a time scale that is much shorter than the scale of the astronomical forcing. In the "Results" sections we take as a reference value $T_w = 3$ kyr, and we repeated the analysis also for $T_w = 2$ kyr and $T_w = 4$ kyr. The effect of this filtering procedure on the data series can be estimated by looking at Fig. 3c,d, where the signals before and after applying the filter are plotted for the whole dataset. In general, the millennial-scale oscillations tend to be more pronounced during the glacial periods.

*Effect of the filter.* The action of the Gaussian filter (13) is clearly visible in Fig. 4, reporting the histograms of $T - \langle T \rangle$ and $[CO_2] - \langle [CO_2] \rangle$, before and after filtering. What can be deduced by the comparison between the original and the final distributions is that the filtering has a twofold effect: it makes the distributions Gaussian-like and, at the same time, it reduces the excursion of the signal.

The former fact can be seen as an hint (although not a proof) that the filtered variables have an almost-linear dynamics, which is our working hypothesis. The latter indicates that most of the variability is on longer time scales, where the effects of the slow forcing and/or of stronger nonlinear climatic responses (both being removed when the signal is filtered), are non-negligible.

Finally, in Fig. 5 we show the cross correlations of the two signals, before and after the filtering procedure. From these plots it is clear that the effect of the filter consists in removing from the analysis the large contributions coming from correlations on longer time scales. It should also be remarked that the cross correlations alone, even after the filter, are not very informative about the causal relations between the signals: in order to give insightful information on the causal structure of the system they must be properly combined with the self correlations, as prescribed for instance by the generalized FDR formalism.
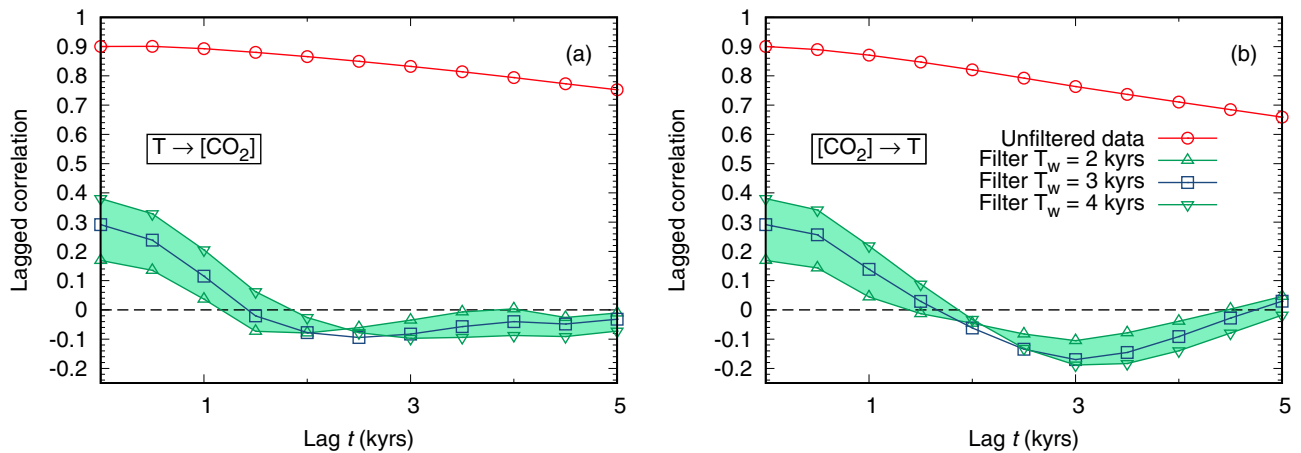
**Figure 5.** Lagged cross correlations for the T and $[CO_2]$ signals, before and after applying the Gaussian filter discussed in the text.

**Transfer entropy: a complementary approach.** Transfer entropy was introduced by Schreiber[27] as an indicator of the information which a given time-dependent signal $x_1(t)$ provides about a second variable $x_2(t)$. The basic idea is to measure how much information is lost about the distribution of $x_2(t)$ when the knowledge of $x_1(t)$ is ignored.

For a two-variable Markovian system, the TE with lag $t$ is defined as

$$TE_{1\to2}(t) = H[x_2|x_2](t) - H[x_2|x_1, x_2](t) \tag{19}$$

where

$$H[y|x](t) = -\int dx\,dy\,P(x, 0; y, t)\ln P(x, 0; y, t) + \int dx\,P(x)\ln P(x) \tag{20}$$

is the conditional Shannon entropy. Here $P(x)$ represents the marginal $x$ probability density functions, while $P(x, 0; y, t)$ is the joint distribution of $x$ at time 0 and $y$ at time $t$ assuming stationarity. If the dynamics is linear the above relations can be simplified, as shown in Ref.[43] (see Sec. 4 of the Supplemental Material for details): in particular, it can be shown that TE is a (complicated) function of two-points correlations functions.

From the point of view of the application to paleoclimatic series, the study of TE presents therefore the same difficulties encountered in the case of generalized FDR: the system under study is not Markovian, and the limited amount of data does not enable to determine a reliable functional form for the probability density functions appearing in Eq. (20). It may thus be expected that the above-discussed filtering procedure, by isolating the fast component of the correlation functions, allows to get rid of the spurious long-time-scale correlations, as in the case of the generalized FDR. This expectation seems to be confirmed by Fig. 1.

It is worth mentioning that an alternative rigorous attempt to assess causation was due to Granger[44], who suggested that the link $x_1 \to x_2$ holds if the knowledge of the past history of $x_1$ enhances the ability to predict future values of $x_2$. Remarkably, Granger causality and TE have been shown to be equivalent in linear auto-regressive systems[45].

With respect to the detection of causal links in a given system, TE and Granger's approach can be regarded as complementary to responses. While the former focuses on our ability to *predict* future values of the considered process, the latter aims at defining the interaction mechanisms internal to the system[30].

## Data availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

## References

1. Pierrehumbert, R. *Principles of Planetary Climate* (Cambridge University Press, 2010).
2. Ghil, M. & Lucarini, V. The physics of climate variability and climate change. *Rev. Mod. Phys.* **92**, 035002. https://doi.org/10.1103/RevModPhys.92.035002 (2020).
3. Benzi, R., Parisi, G., Sutera, A. & Vulpiani, A. Stochastic resonance in climatic change. *Tellus* **34**, 10–16 (1982).
4. Ghil, M. Cryothermodynamics: The chaotic dynamics of paleoclimate. *Physica D* **77**, 130–159 (1994).
5. Jin, F., Neelin, J. & Ghil, M. El Nino on the devil's staircase: Annual subharmonic steps to chaos. *Science* **264**, 70–72. https://doi.org/10.1126/science.264.5155.70 (1994).
6. Tziperman, E., Stone, L., Cane, M. & Jarosh, H. El Nino chaos: Overlapping of resonances between the seasonal cycle and the pacific ocean-atmosphere oscillator. *Science* **264**, 72–74. https://doi.org/10.1126/science.264.5155.72 (1994).
7. Boers, N., Ghil, M. & Rousseau, D.-D. Ocean circulation, ice shelf, and sea ice interactions explain Dansgaard–Oeschger cycles. *Proc. Natl. Acad. Sci. USA* **115**, E11005–E11014 (2018).

8. Lucarini, V. & Bodai, T. Transitions across melancholia states in a climate model: Reconciling the deterministic and stochastic points of view. *Phys. Rev. Lett.* **122**, 158701. https://doi.org/10.1103/PhysRevLett.122.158701 (2019).
9. Runge, J. *et al.* Inferring causation from time series in earth system sciences. *Nat. Commun.* **10**, 2553 (2019).
10. Runge, J., Petoukhov, V. & Kurths, J. Quantifying the strength and delay of climatic interactions: The ambiguities of cross correlation and a novel measure based on graphical models. *J. Clim.* **27**, 720–739 (2014).
11. Moon, W. & Wettlaufer, J. S. A unified nonlinear stochastic time series analysis for climate science. *Sci. Rep.* **7**, 1–9 (2017).
12. Hays, J., Imbrie, J. & Shackleton, N. Variations in the Earth's orbit. Pacemaker of the ice ages. *Science* **194**, 1121–1132 (1976).
13. Imbrie, J. *et al.* On the structure and origin of major glaciation cycles. 2. The 100,000-year cycle. *Paleoclimatol. Paleoceanogr.* **8**, 699–735 (1993).
14. Lisiecki, L. & Raymo, M. A Pliocene–Pleistocene stack of 57 globally distributed benthic $\delta^{18}$o records. *Paleoceanogr. Paleoclimatol.* **20**, PA1003 (2005).
15. Raymo, M., Lisiecki, L. & Nisancioglu, K. Plio-Pleistocene ice volume, Antarctic climate, and the global $\delta^{18}$o record. *Science* **313**, 492–495 (2006).
16. Milankovitch, M. Canon of insolation and the iceage problem. Koniglich Serbische Akademice Beograd Special Publication **132** (1941).
17. Petit, J. *et al.* Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* **399**, 429–436 (1999).
18. EPICA. Eight glacial cycles from an Antarctic ice core. Nature **429**, 623–628 (2004).
19. Luthi, D. *et al.* High-resolution carbon dioxide concentration record 650,000–800,000 years before present. *Nature* **453**, 378–382 (2008).
20. Bereiter, B. *et al.* Revision of the EPICA Dome C $CO_2$ record from 800 to 600 kyr before present. *Geophys. Res. Lett.* **42**, 542–549 (2015).
21. Siegenthaler, U. *et al.* Supporting evidence from the EPICA Dronning Maud Land ice core for atmospheric $CO_2$ changes during the past millennium. *Tellus B* **57**, 51–57. https://doi.org/10.3402/tellusb.v57i1.16774 (2005).
22. Parrenin, F. *et al.* Synchronous change of atmospheric $CO_2$ and Antarctic temperature during the last deglacial warming. *Science* **339**, 1060–1063 (2013).
23. van Nes, E. H. *et al.* Causal feedbacks in climate change. *Nat. Clim. Change* **5**, 445–448. https://doi.org/10.1038/nclimate2568 (2015).
24. Bauska, T. K., Marcott, S. A. & Brook, E. J. Abrupt changes in the global carbon cycle during the last glacial period. *Nat. Geosci.* **14**, 91–96. https://doi.org/10.1038/s41561-020-00680-2 (2021).
25. Dansgaard, W. *et al.* Evidence for general instability of past climate from a 250-kyr ice-core record. *Nature* **364**, 218–220 (1993).
26. Stips, A., Macias, D., Coughlan, C., Garcia-Gorriz, E. & San Liang, X. On the causal structure between $CO_2$ and global temperature. *Sci. Rep.* **6**, 1–9 (2016).
27. Schreiber, T. Measuring information transfer. *Phys. Rev. Lett.* **85**, 461–464. https://doi.org/10.1103/PhysRevLett.85.461 (2000).
28. Pearl, J. *Causality* (Cambridge University Press, 2009).
29. Aurell, E. & Del Ferraro, G. Causal analysis, correlation-response, and dynamic cavity. *J. Phys. Conf. Ser.* **699**, 012002 (2016).
30. Baldovin, M., Cecconi, F. & Vulpiani, A. Understanding causation via correlations and linear response theory. *Phys. Rev. Res.* **2**, 043436 (2020).
31. Falcioni, M., Isola, S. & Vulpiani, A. Correlation functions and relaxation properties in chaotic dynamics and statistical mechanics. *Phys. Lett. A* **144**, 341–346 (1990).
32. Baldovin, M., Caprini, L. & Vulpiani, A. Handy fluctuation–dissipation relation to approach generic noisy systems and chaotic dynamics. *Phys. Rev. E* **104**, L032101. https://doi.org/10.1103/PhysRevE.104.L032101 (2021).
33. Kawamura, K. *et al.* State dependence of climatic instability over the past 720,000 years from Antarctic ice cores and climate modeling. *Sci. Adv.* **3**, e1600446 (2017).
34. Caillon, N. *et al.* Timing of atmospheric $CO_2$ and Antarctic temperature changes across Termination III. *Science* **299**, 1728–1731 (2003).
35. Pedro, J., Rasmussen, S. & van Ommen, T. Tightened constraints on the time-lag between Antarctic temperature and $CO_2$ during the last deglaciation. *Clim. Past* **8**, 1213–1221. https://doi.org/10.1038/ngeo1985 (2012).
36. Landais, A. *et al.* Two-phase change in $CO_2$, Antarctic temperature and global climate during termination II. *Nat. Geosci.* **6**, 1062–1065 (2013).
37. Gottschalk, J. *et al.* Mechanisms of millennial-scale atmospheric $CO_2$ change in numerical model simulations. *Quat. Sci. Rev.* **220**, 30–74 (2019).
38. Shin, J., Ahn, J., Chowdhry Beeman, J., Lee, H.-G. & Brook, E. Millennial variations of atmospheric $CO_2$ during the early holocene (11.7–7.4 ka). *Clim. Past Discuss.* https://doi.org/10.5194/cp-2021-113 *(2007)*.
39. Loulergue, L. *et al.* New constraints on the gas age-ice age difference along the EPICA ice cores, 0–50 kyr. *Clim. Past* **3**, 527–540 (2007).
40. Berger, A. & Loutre, M. Insolation values for the climate of the last 10 million years. *Quat. Sci. Rev.* **10**, 297–317 (1991).
41. Baldovin, M., Puglisi, A. & Vulpiani, A. Langevin equations from experimental data: The case of rotational diffusion in granular media. *PLoS One* **14**, e0212135 (2019).
42. Blinchikoff, H. J. & Zverev, A. I. *Filtering in the Time and Frequency Domains* (Krieger Publishing Co. Inc., 1986).
43. Sun, J., Taylor, D. & Bollt, E. M. Causal network inference by optimal causation entropy. *SIAM J. Appl. Dyn. Syst.* **14**, 73–106 (2015).
44. Granger, C. W. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, 424–438 (1969).
45. Barnett, L., Barrett, A. B. & Seth, A. K. Granger causality and transfer entropy are equivalent for Gaussian variables. *Phys. Rev. Lett.* **103**, 238701. https://doi.org/10.1103/PhysRevLett.103.238701 (2009).

## Acknowledgements

## Author contributions

All authors planned the research and wrote the manuscript. M.B. made the data analysis and F.C. supervised the work.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-18406-2.

**Correspondence** and requests for materials should be addressed to M.B.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.