Causation and information flow with respect to relative entropy

X. San Liang

Citation: Chaos **28**, 075311 (2018); doi: 10.1063/1.5010253
View online: https://doi.org/10.1063/1.5010253
View Table of Contents: http://aip.scitation.org/toc/cha/28/7
Published by the American Institute of Physics

Causation and information flow with respect to relative entropy

X. San Liang[a)]
Nanjing Institute of Meteorology, Nanjing 210044, China

(Received 24 October 2017; accepted 7 June 2018; published online 24 July 2018)

Recently, a rigorous formalism has been established for information flow and causality within dynamical systems with respect to Shannon entropy. In this study, we re-establish the formalism with respect to relative entropy, or Kullback-Leiber divergence, a well-accepted measure of predictability because of its appealing properties such as invariance upon nonlinear transformation and consistency with the second law of thermodynamics. Different from previous studies (which yield consistent results only for 2D systems), the resulting information flow, say T, is precisely the same as that with respect to Shannon entropy for systems of arbitrary dimensionality, except for a minus sign (reflecting the opposite notion of predictability vs. uncertainty). As before, T possesses a property called principle of nil causality, a fact that classical formalisms fail to verify in many situation. Besides, it proves to be invariant upon nonlinear transformation, indicating that the so-obtained information flow should be an intrinsic physical property. This formalism has been validated with the stochastic gradient system, a nonlinear system that admits an analytical equilibrium solution of the Boltzmann type. *Published by AIP Publishing.* https://doi.org/10.1063/1.5010253

To identify the causal relation between two dynamical events is a fundamental problem in science and forms a direct objective in many research fields. It is also an important problem in philosophy, as it provides "guides to higher understanding."[1] During the past years, it has been realized that causality actually can be rigorously derived in terms of information flow from first principles, rather than axiomatically proposed as an ansatz.[2] The formalism with respect to Shannon entropy, or absolute entropy as it is called, results in a concise formula for causality measure, which, in the linear limit, unambiguously asserts that causation implies correlation, but correlation does not imply causation. It has been validated with many touchstone causal inference problems (e.g., Ref. 3). It has also been applied to real world systems with remarkable success, among which are the reversing causal direction between CO_2 and global warming,[4] and the long forgotten story about "Seven Dwarfs" competing with IBM the "Giant" for computer market.[36] Considering that the classical Granger causality is originally formulated as predictability improvement, and that relative entropy possesses some appealing properties (such as consistency with the second law of thermodynamics) and hence has been proposed as a natural measure of predictability,[5] in this study, the above approach is extended to re-establishing the formalism with respect to relative entropy. Different from previous studies along this line, the resulting formula is precisely the same as that with respect to absolute entropy, except for a minus sign, reflecting the opposite notion of predictability vs. uncertainty. Besides, it proves to be invariant upon nonlinear transformation, indicating that the so-obtained causality measure represents an intrinsic physical property.

I. INTRODUCTION

Historically causality analysis is formulated (arguably) by Granger[6] as a statistical hypothesis testing problem. The resulting metric and its varieties have been referred to as Granger causality. On the other hand, a physical notion, namely, information flow or information transfer, has been under development in parallel for more than three decades (e.g., Refs. 2 and 7–12) and has caught wide attention from different disciplines.[44] Now, it is generally agreed that information flow provides a natural way of measuring causality; indeed, it is this very logical association that makes the former gain such wide interest. Like Granger causality, there are also different measures for information flow, the most popular one being the axiomatically proposed transfer entropy.[9] Interestingly, transfer entropy and Granger causality turn out to be equivalent for Gaussian variables (up to a factor of 2).[13]

For a formalism of causality to be faithful, the following observational fact must be verified: If the evolution of an event, say, X_1, is independent of another one, X_2, then the causality from X_2 to X_1 is zero. This is actually the only quantitatively stated fact about causality; Liang[2] refers to it as *principle of nil causality*. All the formalisms proposed so far, including the classical Granger causality,[14,15] transfer entropy,[16–18] symbolic transfer entropy,[19] and the new ones such as the momentary information transfer,[11] causation entropy,[12] convergent cross mapping,[20] predictability improvement,[24,25] etc., have been tested with this principle in applications. Recently, a rigorous formalism has been developed for information flow within the framework of dynamical systems. Rather than axiomatically proposed as an ansatz, it is derived from first principles.[2] Most of all, the principle of nil causality appears in this formalism naturally as a proven theorem. This line of work begins some 12 years ago with two-dimensional (2D) deterministic systems.[10] The basic idea

[a)]Electronic address: sanliang@courant.nyu.edu

can be best illustrated with a nonlinear system

$$\frac{dx_1}{dt} = F_1(x_1, x_2, t), \tag{1}$$

$$\frac{dx_2}{dt} = F_2(x_1, x_2, t), \tag{2}$$

with randomness limited to its initial condition. Here, the convention in physics is followed which does not distinguish random and deterministic variables; in statistics, they are usually differentiated with upper-case and lower-case symbols. Suppose we are about to find the information flow from x_2 to x_1. We essentially need to check how the marginal entropy of x_1, written H_1, evolves. Here by entropy or Shannon entropy we mean differential Shannon entropy; throughout this paper, we will keep this convention unless confusion may arise. This evolution could be driven by two different mechanisms: the internal mechanism due to x_1 its own and the external influence from x_2. The latter is the very information flow from x_2. If we write the former as $\frac{dH_1^*}{dt}$ and the latter as $T_{2\to1}$, then $\frac{dH_1}{dt} = \frac{dH_1^*}{dt} + T_{2\to1}$. Now the problem is converted into finding dH_1^*/dt, since, given a deterministic system, there is a Liouville equation for the probability density function (pdf), and hence dH_1/dt can be obtained. In Ref. 10, dH_1^*/dt is acquired through an intuitive argument based on a concise law established therein:

$$\frac{dH}{dt} = E(\nabla \cdot \mathbf{F}), \tag{3}$$

where H is the joint entropy of (x_1, x_2), $\mathbf{F} = (F_1, F_2)$, ∇ is the gradient operator with respect to (x_1, x_2), and E signifies mathematical expectation. With this, Liang and Kleeman[10] argue that

$$\frac{dH_1^*}{dt} = E\left(\frac{\partial F_1}{\partial x_1}\right).$$

Subtraction of this from dH_1/dt then follows the rate of information flow from x_2 to x_1:

$$T_{2\to1} = -E\left(\frac{1}{\rho_1}\frac{\partial F_1\rho_1}{\partial x_1}\right), \tag{4}$$

where ρ_1 is the marginal density of x_1. We remark that this setting is rather generic; the only assumption is the differentiability of the vector field (F_1, F_2).

Equation (4) is later on rigorously proved[21,22] and has been remarkably successful. Particularly, it possesses a property which is later on realized to be the very principle of nil causality. (That is to say, within this framework, the principle of nil causality is no longer an issue.) The same approach has been extended to formulating the information flow between two subspaces.[23] This formalism, however, is only for 2D deterministic systems. For systems with more than two components, the above intuitive argument does not work anymore. Even for a 2D system with stochasticity, this does not work, either, because no such a simple law as (3) exists. The generalization to multidimensional stochastic systems has just been fulfilled; see Ref. 2.

The Liang-Kleeman formalism is with respect to Shannon entropy, or absolute entropy as it is called. In information theory, there is another quantity, namely, relative entropy or Kullback-Leibler divergence, which has been shown advantageous over Shannon entropy in that it possesses some appealing properties such as invariance upon nonlinear transformation, and in the context of Markov chain, consistency with the second law of thermodynamics (e.g., Refs. 5, 26, and 27). It therefore has been proposed as a better measure of predictability[5] as compared to Shannon entropy or absolute entropy.[43] We hence examine in this study how the problem may be re-formulated with respect to relative entropy. The difficulty is, even with a 2D deterministic system (1)–(2), there is no such nice law as (3) for Kullback-Leibler divergence. Previously, this was examined in Ref. 28 within the early version of this framework,[21,22] which results in a consistent information flow for 2D systems, but a different one for systems with dimensionality greater than 2 (though similar in form). As we have fulfilled a rigorous formalism in Ref. 2 for systems with both stochasticity and multi-dimensionality, one naturally wonders how this may result within the updated framework. This makes the major objective of this study. The resulting flow/transfer will be referred to as *information flow with respect to relative entropy*, or simply *information flow* when no confusion arises. In the following, we first take a brief stroll through the recent rigorous formalism (Sec. II), then do the derivation (Sec. III). An application is demonstrated in Sec. IV with a nonlinear stochastic system which possesses an analytical equilibrium solution. This study is summarized in Sec. V.

II. A STROLL THROUGH THE RECENT DEVELOPMENT OF THE RIGOROUS FORMALISM WITH RESPECT TO SHANNON ENTROPY

As mentioned above, causality forms the key to information flow, and information flow is the logical measure of causality. Since information flow is a real physical notion (not just something in statistics), (arguably) it should be formulated on a rigorous footing,[2,3,29] rather than be axiomatically proposed as an ansatz (such as the existing metrics). The following is a brief presentation of the main results of such a rigorous formalism.

Consider a dynamical system

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(t;\mathbf{x}) + \mathbf{B}(t;\mathbf{x})\dot{\mathbf{w}}, \tag{5}$$

where $\mathbf{x}$ and $\mathbf{F}$ are n-dimensional vector, $\mathbf{B}$ is an $n \times m$ matrix, and $\mathbf{w}$ is an m-vector of standard Wiener process ($\dot{\mathbf{w}}$ is a vector of white noise). Note that $\mathbf{B}$ can be a function of both $\mathbf{x}$ and time t. Throughout this study, $\mathbf{F}$ and $\mathbf{B}$ are assumed to be differentiable in x and t. We have the following theorem:[2]

Theorem II.1. *For the dynamical system (5), the rate of information flowing from X_2 to X_1 is*

$$T_{2\to1} = -\int_{\mathbb{R}^n} \rho_{2|1}\frac{\partial(F_1\rho_{\not 2})}{\partial x_1}d\boldsymbol{x} + \frac{1}{2}\int_{\mathbb{R}^n} \rho_{2|1}\frac{\partial^2(g_{11}\rho_{\not 2})}{\partial x_1^2}d\boldsymbol{x}, \tag{6}$$

where E stands for mathematical expectation, and $\rho_1 = \rho_1(x_1)$ is the marginal probability density function (pdf) of X_1, $\rho_{2|1}$ the conditional pdf of X_2 on X_1, $\rho_{\not 2} = \int_{\mathbb{R}} \rho dx_2$, and $g_{11} = \sum_{j=1}^m b_{1j}b_{1j}$.

Ideally if $T_{2\to1}=0$, then X_2 is not causal to X_1; otherwise, it is causal (for either positive or negative information flow).

Corollary II.1. *When $n=2$,*

$$T_{2\to1}=-E\left[\frac{1}{\rho_1}\frac{\partial(F_1\rho_1)}{\partial x_1}\right]+\frac{1}{2}E\left[\frac{1}{\rho_1}\frac{\partial^2 g_{11}\rho_1}{\partial x_1^2}\right]. \quad (7)$$

In the absence of noise, this is precisely Eq. (4), the result based on the heuristic argument in Ref. 10.

The information flow (6) [and hence (7)] possesses a nice property:

Theorem II.2. (Principle of nil causality) *If in the system (5) both F_1 and g_{11} are independent of X_2, then $T_{2\to1}=0$.*

Note this is the very principle of nil causality, an observational fact that defies the classical formalism in many situations. But, remarkably, here it appears as a proven theorem! For its proof, refer to Ref. 2.

In the case when only two time series are given, the causality between them can be estimated using maximum likelihood estimation.[3]

Theorem II.3. *Given two time series X_1 and X_2, under the assumption of a linear model with additive noise, the maximum likelihood estimator (mle) of the rate of information flowing from X_2 to X_1 is*

$$\hat{T}_{2\to1}=\frac{C_{11}C_{12}C_{2,d1}-C_{12}^2C_{1,d1}}{C_{11}^2C_{22}-C_{11}C_{12}^2}, \quad (8)$$

where C_{ij} is the sample covariance between X_1 and X_2, and $C_{i,dj}$ is the sample covariance between X_i and a series derived from X_j using the Euler forward differencing scheme: $\dot{X}_{j,n}=(X_{j,n+k}-X_{j,n})/(k\Delta t)$, with $k\geq1$ some integer.

This result is important in that it bridges the gap between theory and real applications. Considering that in history there is a long-standing debate over correlation versus causation, the above may also be written in terms of linear correlation coefficients. A direct corollary is that, in the linear sense,[3]

Causation implies correlation, but correlation does not imply causation.

Causality can be normalized so as to reveal the relative importance of a causal relation; see Ref. 36 for details. Also, statistical significance test can be performed for Eq. (8), which is referred to Ref. 3.

The above formalism has been validated with touchstone problems in causal inference. For example, the anticipatory system problem described in Ref. 30 turns out to be straightforward with the concise formula (8), though it is highly nonlinear. More validations have been evidenced in different applications with benchmark systems such as baker transformation, Hénon map, Kaplan-Yorke map,[31] Rössler system,[32] truncated Burgers-Hopf system,[33] etc.; see Ref. 2 for these examples.

The formalism has also been put to application with success to many real world problems. These include the El Niño-Indian Ocean Dipole relation study,[3] tropical cyclone genesis prediction,[34] near-wall turbulence study,[35] financial time series analysis,[36] regional and global climate change,[4,37,38] to name but a few. Among them stands out the causality study between CO_2 and global warming.[4] It is found that, during the past century, indeed CO_2 emission drives the recent global warming; the causal relation is one-way, i.e., from CO_2 to global mean atmosphere temperature. Moreover, the one-way causality is not homogeneously distributed, with much enhanced warming over drylands, just as reported in different previous studies.[39] However, on a paleoclimate time scale (1000 years or above), the causality is completely reversed: it is global warming that causes CO_2 concentration to rise!

Another application,[36] among many interesting ones, is regarding the relation between the two corporations IBM and GE, using the time series of US stocks downloaded from $^{YAHOO!}_{\text{finance}}$. It is found that their causal relation generally varies with time. On the whole, the causality seems to be insignificant, but if we do a running time analysis, there appears a strong, almost one-way causality from IBM to GE (i.e., $|T_{IBM\to GE}|\gg|T_{GE\to IBM}|$) in the 1970s, starting from 1971. This abrupt one-way causality rise reveals to us an old story about "Seven Dwarfs and a Giant" which has almost been forgotten: In the 1950s–1960s, GE was believed to be the biggest computer user outside the U.S. Federal Government; to avoid relying on IBM the computer "Giant", it together with six other companies ("Seven Dwarfs") began to build mainframes. But in 1970, GE sold its computer division. So starting from 1971, it had to rely on IBM again. That is why there is such a jump in $T_{IBM\to GE}$ from 1970 to 1971. While the story has almost gone to oblivion, this finding, which is solely based on a simple causality analysis of two time series with Eq. (8), is really remarkable.

III. INFORMATION FLOW WITH RESPECT TO KULLBACK-LEIBLER DIVERGENCE

A. Strategy for the derivation

Kullback-Leibler divergence,[40] also known as relative entropy, is defined as, for two joint probability density functions ρ and q of $\mathbf{x}=(x_1,x_2,\ldots,x_n)$,

$$D=D(\rho||q)=\int\rho\log\frac{\rho}{q}=E_\rho\log\frac{\rho}{q}, \quad (9)$$

where the integral is over the whole sample space; in this study, it is $\mathbb{R}^n$. It can be used to measure the distance between ρ and q. Let q be the initial or equilibrium pdf, Kleeman[5] has established that it is a natural measure of predictability for the dynamical system with which ρ evolves. Besides, different from Shannon entropy, D is invariant upon nonlinear transformation and is in accordance with the second law of thermodynamics. Likewise, the marginal relative entropy D_1 for $\rho_1(x_1)$ and $q_1(x_1)$ is

$$D_1=D_1(\rho_1||q_1)=E_{\rho_1}\log\frac{\rho_1}{q_1}. \quad (10)$$

Now consider the stochastic nonlinear dynamical system (5). Without loss of generality, it suffices to examine the information flow from x_2 to x_1; for any other pair, say (x_i,x_j), we may make a transformation by simply moving the two components to the first two slots. Here what we need to study is the evolution of the predictability of x_1, i.e., dD_1/dt. Note how it differs from that with respect to H_1. The involvement of q

complicates the problem in that, even for a two-dimensional deterministic system, there is no evolutionary law of D as (3) and hence the heuristic argument used by Liang and Kleeman in Ref. 10 does not work anymore (as mentioned in the introduction). Following what we have done before in fixing the problem (e.g., Ref. 2), dD_1/dt can be exclusively decomposed into two parts, one being the evolution of D_1 with the effect of x_2 excluded, written $\frac{dD_{1\nsubseteq}}{dt}$, another being the influence from x_2, i.e., the information flowing from x_2. Symbolically this is

$$\frac{dD_1}{dt} = \frac{dD_{1\nsubseteq}}{dt} + T^D_{2\to 1}, \tag{11}$$

where $T^D_{2\to 1}$ is the rate of information flow from x_2 to x_1 with respect to predictability. In the following, we will simply denote it as $T_{2\to 1}$ unless otherwise indicated.

The problem is hence changed to finding $\frac{dD_1}{dt}$ and $\frac{dD_{1\nsubseteq}}{dt}$. For any dynamical system, correspondingly there is a Fokker-Planck equation governing the pdf of the state. From this equation in principle, $\frac{dD_1}{dt}$ can be derived. The key is the derivation of $\frac{dD_{1\nsubseteq}}{dt}$. We will see this soon in Subsection III C.

B. Time evolutions of D and D_1

The following are some laws on the evolution of Kullback-Leibler divergence. For simplicity, we will suppress the subscript ρ in E_ρ; all expectations are henceforth understood with respect to ρ unless otherwise noted.

Theorem III.1. *Consider the dynamical system (5). Let ρ, q, ρ_1, and q_1 be the probability density functions as introduced in the preceding subsection, and suppose that they are twice differentiable. Then*

$$\frac{dD}{dt} = -E(\nabla\cdot\boldsymbol{F}) - E[\boldsymbol{F}\cdot\nabla\log q] - \frac{1}{2}E[\boldsymbol{G}:\nabla\nabla\log q], \tag{12}$$

$$\frac{dD_1}{dt} = E\left[F_1\frac{\partial(\log\rho_1/q_1)}{\partial x_1}\right] + \frac{1}{2}E\left[g_{11}\frac{\partial^2(\log\rho_1/q_1)}{\partial x_1^2}\right], \tag{13}$$

where $\mathbf{G}=\mathbf{B}\mathbf{B}^T$, with $g_{ij}=\sum_{k=1}^m b_{ik}b_{jk}$ being the entries.

Remark: Note when $g_{ij}=0$, i.e., in the absence of stochasticity, we have obtained these results in Ref. 28. Also note that if q is another pdf co-varying with ρ, then $dD/dt=0$, as established in Ref. 42. But such a D is not the predictability measure introduced in Ref. 5.

Proof. We here derive them from the Fokker-Planck equation:

$$\frac{\partial\rho}{\partial t} = -\nabla\cdot(\mathbf{F}\rho) + \frac{1}{2}\nabla\nabla:(\mathbf{G}\rho), \tag{14}$$

where the double dot is defined such that, for two dyadics $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A}:\mathbf{B}=\sum_{i,j}a_{ij}b_{ji}$. So

$$\begin{aligned}\frac{dD}{dt} &= -\frac{dH}{dt} - \int\frac{\partial\rho}{\partial t}\log q d\mathbf{x}\\ &= -E(\nabla\cdot\mathbf{F}) + \int_{\mathbb{R}^n}\left[\nabla\cdot(\mathbf{F}\rho) - \frac{1}{2}\nabla\nabla:(\mathbf{G}\rho)\right]\log q\, d\mathbf{x},\end{aligned}$$

where the concise formula[10]

$$\frac{dH}{dt} = E(\nabla\cdot\mathbf{F}) \tag{15}$$

has been used. Integrate by parts, and (12) follows.

To derive (13), integrate the Fokker-Planck equation with respect to $x_2,\ldots,x_n$ to get

$$\frac{\partial\rho_1}{\partial t} = -\int\frac{\partial(F_1\rho)}{\partial x_1}dx_2\cdots dx_n + \frac{1}{2}\int\frac{\partial^2 g_{11}\rho}{\partial x_1^2}dx_2\cdots dx_n.$$

Following the same procedure as above, we arrive at (13). □

C. Time change of D_1 with x_2 frozen as a parameter

To arrive at $\frac{dD_{1\nsubseteq}}{dt}$, the above approach is not through since now the system has been modified with x_2 frozen at time t and hence there is no such a Fokker-Planck equation in the usual sense for the corresponding pdf $\rho_{1\nsubseteq}$. We have to go back to the basics such as Frobenius-Perron operator (cf. Appendix). The following proposition gives the result:

Proposition III.1. *For the system in Theorem III.1, the time change of D_1 with x_2 frozen as a parameter is*

$$\begin{aligned}\frac{dD_{1\nsubseteq}}{dt} &= E\left[F_1\frac{\partial\log\rho_1/q_1}{\partial x_1}\right] + E\left[g_{11}\frac{\partial^2\log\rho_1/q_1}{\partial x_1^2}\right]\\ &\quad - E\left[\frac{1}{\rho_1}\int_{\mathbb{R}^{n-2}}\frac{\partial F_1\rho_{\nsubseteq}}{\partial x_1}dx_3\cdots dx_n\right]\\ &\quad + \frac{1}{2}E\left[\frac{1}{\rho_1}\int_{\mathbb{R}^{n-2}}\frac{\partial^2 g_{11}\rho_{\nsubseteq}}{\partial x_1^2}dx_3\cdots dx_n\right],\end{aligned} \tag{16}$$

where $\rho_{\nsubseteq}=\int_{\mathbb{R}}\rho(x_1,x_2,\ldots,x_n)dx_2$.

The proof is given in the Appendix.

D. Information flow with respect to relative entropy

Subtracting (16) from (13), we arrive at the information flow from x_2 to x_1 with respect to predictability. This is summarized in the following theorem.

Theorem III.2. *For the system in Theorem III.1, the information flow from x_2 to x_1 is*

$$\begin{aligned}T_{2\to 1} &= \frac{dD_1}{dt} - \frac{dD_{1\nsubseteq}}{dt}\\ &= E\left[\frac{1}{\rho_1}\int_{\mathbb{R}^{n-2}}\frac{\partial F_1\rho_{\nsubseteq}}{\partial x_1}dx_3\cdots dx_n\right]\\ &\quad - \frac{1}{2}E\left[\frac{1}{\rho_1}\int_{\mathbb{R}^{n-2}}\frac{\partial^2 g_{11}\rho_{\nsubseteq}}{\partial x_1^2}dx_3\cdots dx_n\right]\end{aligned} \tag{17}$$

$$\begin{aligned}&= \int_{\mathbb{R}^n}\rho_{2|1}(x_2\mid x_1)\frac{\partial F_1\rho_{\nsubseteq}}{\partial x_1}d\boldsymbol{x}\\ &\quad - \frac{1}{2}\int_{\mathbb{R}^n}\rho_{2|1}(x_2\mid x_1)\frac{\partial^2 g_{11}\rho_{\nsubseteq}}{\partial x_1^2}d\boldsymbol{x},\end{aligned} \tag{18}$$

where $g_{11}=\sum_{k=1}^m b_{1k}b_{1k}$, and $\rho_{\nsubseteq}=\int_{\mathbb{R}}\rho dx_2$.

Interestingly, this is precisely the same as that with Shannon entropy, i.e., Eq. (6), except for a minus sign. This is quite different from the previous studies along this line with the early version of the formalism (e.g., Ref. 28), where only for 2D systems the two are consistent. The nice properties of

the Shannon entropy-based formalism, such as principle of nil causality, disappearing effect with additive noise, etc., are then inherited here.

Theorem III.3. *The information flow in (17) or (18) is invariant upon coordinate transformation of* $(x_3, x_4, \ldots, x_n)$.

Remark 1: Here we cannot make transformation for x_1 and x_2 otherwise we would talk about information flow between different events.

Remark 2: This asserts from an aspect that the so-obtained T is an intrinsic physical property.

Proof. We first do the proof with a deterministic system. Let $\Phi : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{x} \mapsto \boldsymbol{\xi}$, such that

$$\begin{aligned}
\xi_1 &= x_1,\\
\xi_2 &= x_2,\\
\xi_3 &= \Phi_3(x_3, \ldots, x_n)\\
&\vdots \qquad\qquad \vdots\\
\xi_n &= \Phi_n(x_3, \ldots, x_n).
\end{aligned}$$

That is to say, here actually (x_1, x_2) are not involved in the transformation. The original system $d\mathbf{x}/dt = \mathbf{F}(\mathbf{x})$ is changed to

$$\begin{aligned}
\frac{d\xi_1}{dt} &= F_1[\Phi^{-1}(\boldsymbol{\xi})]\\
\frac{d\xi_2}{dt} &= F_2[\Phi^{-1}(\boldsymbol{\xi})]\\
\frac{d[\Phi^{-1}(\boldsymbol{\xi})]_3}{dt} &= F_3[\Phi^{-1}(\boldsymbol{\xi})]\\
\cdots &\qquad\qquad \cdots
\end{aligned}$$

where the components #3 − n are rather complex but we will not need them. Let the Jacobian of Φ be J, then J is actually equal to the Jacobian of $(\Phi_3, \ldots \Phi_n)$, written $J_{\mathfrak{y}\mathfrak{x}}$, which is independent of (x_1, x_2). By the Frobenius-Perron operator result, the density of $\boldsymbol{\xi}$ is $\rho_\xi(\boldsymbol{\xi}) = \rho(\mathbf{x})|J^{-1}| = \rho(\mathbf{x})|J^{-1}_{\mathfrak{y}\mathfrak{x}}|$. By (18), the information flow from ξ_2 to ξ_1, written $\tilde{T}_{2\to1}$, is

$$\begin{aligned}
\tilde{T}_{2\to1} &= \int_{\Phi(\mathbb{R}^n)} \frac{\rho(\xi_1,\xi_2)}{\rho_1(\xi_1)} \cdot \frac{\partial\{F_1[\Phi^{-1}(\boldsymbol{\xi})]\int \rho_\xi(\boldsymbol{\xi})d\xi_2\}}{\partial\xi_1} d\boldsymbol{\xi}\\
&= \int_{\Phi(\mathbb{R}^n)} \frac{\rho(\xi_1,\xi_2)}{\rho_1(\xi_1)}\\
&\quad \cdot \frac{\partial\{F_1[\Phi^{-1}(\boldsymbol{\xi})]\int \rho[\Phi^{-1}(\boldsymbol{\xi})]|J^{-1}_{\mathfrak{y}\mathfrak{x}}|d\xi_2\}}{\partial\xi_1} d\boldsymbol{\xi}\\
&= \int_{\Phi(\mathbb{R}^n)} \frac{\rho(\xi_1,\xi_2)}{\rho_1(\xi_1)} \cdot \frac{\partial\{F_1[\Phi^{-1}(\boldsymbol{\xi})]\int \rho[\Phi^{-1}(\boldsymbol{\xi})]d\xi_2\}}{\partial\xi_1}\\
&\quad \cdot |J^{-1}| d\boldsymbol{\xi}\\
&= \int_{\mathbb{R}^n} \frac{\rho(x_1,x_2)}{\rho_1(x_1)} \cdot \frac{\partial[F_1(\mathbf{x})\int \rho(\mathbf{x})dx_2]}{\partial x_1} d\mathbf{x}\\
&= T_{2\to1}.
\end{aligned}$$

Likewise, the stochastic case can be proved. So the information flow is invariant upon coordinate transformation. □

IV. AN APPLICATION TO THE STOCHASTIC GRADIENT SYSTEM: THE DEPENDENCE OF RELATIVE CAUSAL EFFECT ON STOCHASTICITY

As a preliminary application, we now examine how stochasticity may affect the information flow and causality within a dynamical system. The system we will be examining belongs to the class of stochastic gradient systems, which have vector fields in a gradient form. We choose such a class of systems because the corresponding equilibrium probability density functions are of the Boltzmann type and can be explicitly obtained. For convenience, consider a simple stochastic perturbation $\mathbf{B} = b\mathbf{I}$ with $\mathbf{I}$ being the identity matrix and b a tunable constant. The governing equation is thence

$$\begin{aligned}
\frac{d\mathbf{x}}{dt} &= \mathbf{F}(\mathbf{x}, t) + b\dot{\mathbf{w}},\\
\mathbf{F} &= -\nabla V,
\end{aligned}$$

with $V = V(\mathbf{x})$ some potential function. Correspondingly, the Fokker-Planck equation (14) is

$$\frac{\partial\rho}{\partial t} - \nabla\cdot(\rho\nabla V) = \nabla\cdot\left(\frac{1}{2}b^2\nabla\rho\right).$$

In the equilibrium state, $\partial/\partial t = 0$. It is easy to obtain

$$\rho = \frac{1}{Z}e^{-2V/b^2}, \tag{19}$$

where Z is the normalizer (or partition function as is called in statistical physics). Here, we consider a 3D case, with a potential function

$$V = \frac{1}{2}(x_1^2x_2^2 + x_2^2x_3^2 + x_1^2 + x_2^2 + x_3^2). \tag{20}$$

This is the case which has been briefly examined in Ref. 2; we hence avail us of that result (except for a minus sign). Shown in Fig. 1 are some of the density distributions with different stochastic perturbation amplitudes b.

The vector field $\mathbf{F}$ resulting from the potential function is

$$F_1 = -x_1x_2^2 - x_1, \tag{21}$$

$$F_2 = -x_2x_3^2 - x_2x_1^2 - x_2, \tag{22}$$

$$F_3 = -x_3x_2^2 - x_3. \tag{23}$$

By the symmetry and the principle of nil causality, it is easy to see that

$$\begin{aligned}
T_{3\to2} &= T_{1\to2},\\
T_{2\to1} &= T_{2\to3},\\
T_{3\to1} &= T_{1\to3} = 0.
\end{aligned}$$

And these have been verified in the computation in Ref. 2. We hence only need to look at the three information flows: $T_{2\to1}$, $T_{1\to2}$, and $T_{3\to1}$ (=0). By (18), the rate of information flow

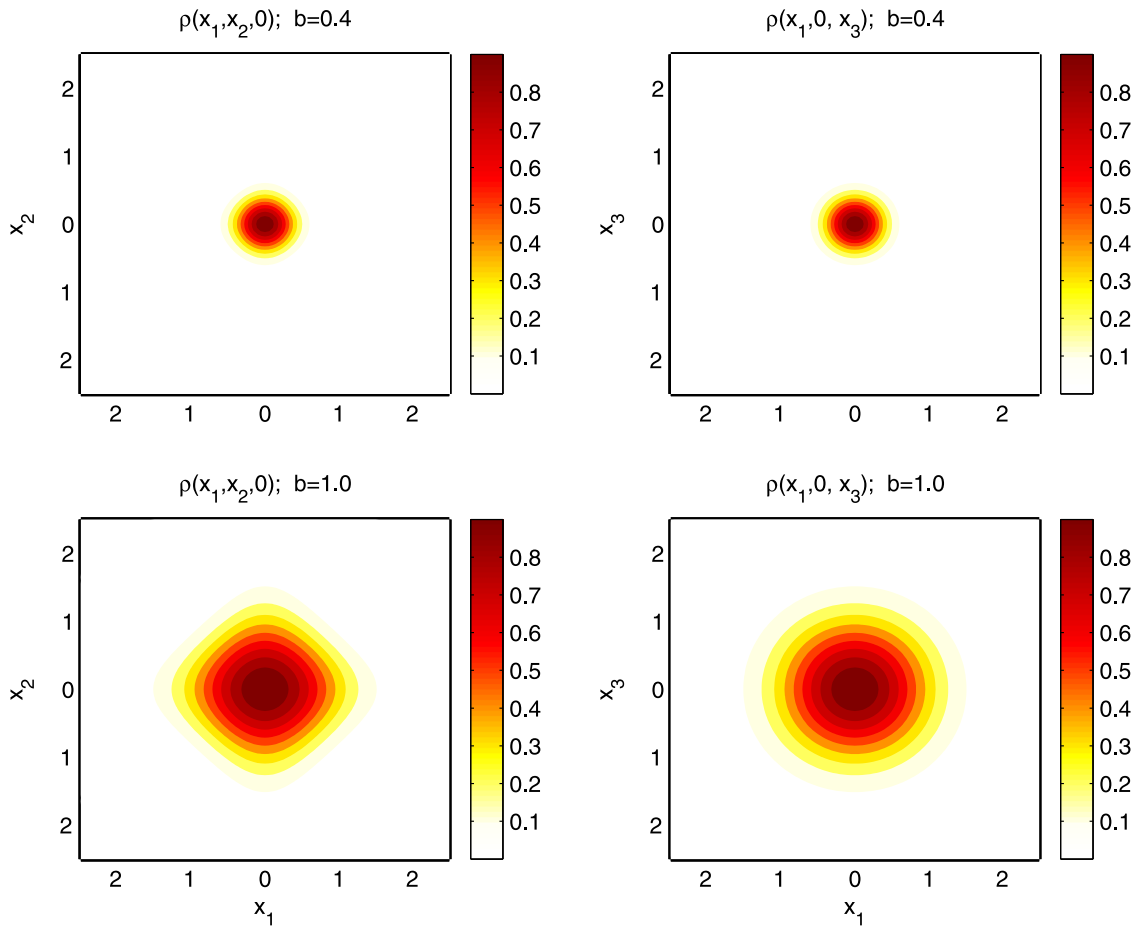

FIG. 1. Density distribution with the potential function (20) and stochastic perturbation amplitude $b = 0.4$ (top panel) and $b = 1.0$ (bottom panel). Here, the numbers are not normalized.

from x_j to x_i is

$$T_{j\to i} = \int_{\mathbb{R}^3} \rho_{j|i}(x_j \mid x_i) \frac{\partial F_i \rho_{\hat{\jmath}}}{\partial x_i}\, d\mathbf{x} - \frac{1}{2}\int_{\mathbb{R}^n} \rho_{j|i}(x_j \mid x_i) \frac{\partial^2 (b^2 \rho_{\hat{\jmath}})}{\partial x_i^2} d\mathbf{x}.$$

For this 3D case, $\rho_{\hat{\imath}} = \rho_{jk}(x_j, x_k)$, with i, j, and k running on the set $(1, 2, 3)$ in a cyclic way. Since b is a constant, it is easy to show that the second term on the right hand side is gone, as proved in Ref. 29. However, this does not mean that stochastic perturbation has no contribution; in fact, it plays a hidden role by affecting the deterministic variables in the first term, and that is what we want to examine here.

A remark on the computation. Theoretically, the sample space is $\mathbb{R}^3$ but in computation we can only deal with a finite domain. Let this domain be $[-\delta, \delta] \times [-\delta, \delta] \times [-\delta, \delta]$. Then to choose an appropriate δ and the spacing size Δx for discretizing the domain may be important in making the computation efficient. As shown in Fig. 1, when b is small, to adequately resolve the effective support in the sample space, Δx should be small. When b is large, δ should be chosen largely to cover the whole domain. We hence consider only a small range for b to vary: [0.3, 0.8]. In Ref. 2, a domain $[-5, 5] \times [-5, 5] \times [-5, 5]$ and a spacing size $\Delta x = 0.05$ have been chosen. We redo the computation using $\delta = 2.5$ and $\Delta = 0.05$ and the results look similar. Note here our results should differ from those in Ref. 2 by a sign. To avoid confusion, we just show the absolute values. As displayed in Fig. 2 (left), $T_{3\to 1}$ (and $T_{1\to 3}$) is identically zero, just as expected. Both $|T_{2\to 1}|$ and $|T_{1\to 2}|$ increase with b; they have

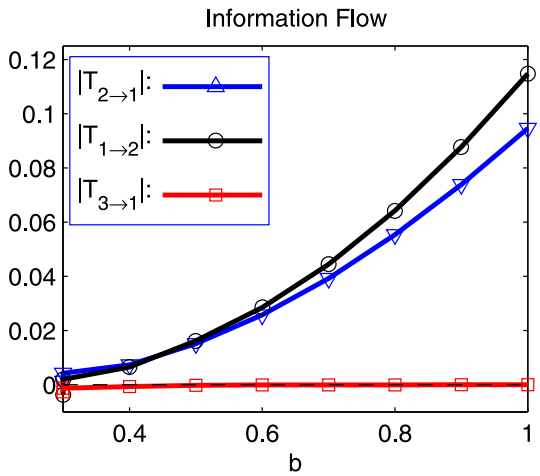

FIG. 2. Rate of information flow (T) within a gradient system with the potential function (20). Shown here are the absolute values. (Units: nats per unit time.)

been validated in Ref. 2 in the context of thermodynamics (b^2 functions like temperature).

An interesting observation is that, as b increases, the relative role of x_1 and x_2 changes. At $b = 0.4$, $|T_{2\to1}| = 0.736 \times 10^{-2}$ is slightly larger than $|T_{1\to2}| = 0.651 \times 10^{-2}$; when $b < 0.4$ $|T_{2\to1}|$ is much larger. That is to say, as b is small, x_2 is more important to x_1 than x_1 is to x_2. This is understandable. In (21)–(23), F_2 differs from F_1 by a term $-x_2x_3^2$; otherwise, x_1 and x_2 are symmetric, and hence, $|T_{2\to1}|$ should be the same as $|T_{1\to2}|$ in the absence of stochasticity. The addition of $-x_2x_3^2$ weakens the role of x_1 to x_2 hence reduces $|T_{1\to2}|$ (relative to $|T_{2\to1}|$).

This scenario changes when stochasticity becomes significant. As b increases, $|T_{1\to2}|$ exceeds $|T_{2\to1}|$: x_1 becomes more important to x_2 that x_2 is to x_1. This is interesting as usually one would expect that noise would take a share but would not change the information flow structure. A detailed explanation of how this may happen in general is beyond the scope of this study. For this particular example, by the entropy evolution laws,[42] the positivity of x_3^2 functions to reduce the uncertainty of the system, and it then may weaken the noise effect as well.

V. CONCLUDING REMARKS

Recently, a rigorous formalism has been established for information flow and causality within dynamical systems with respect to absolute or Shannon entropy. In this study, we re-establish the formalism with respect to Kullback-Leibler divergence or relative entropy, which is a well-accepted measure of predictability, and is advantageous over absolute entropy in that it possess such nice properties as consistency with the second law of thermodynamics.

For easy reference, here we rewrite the major result (Theorem III.2) as follows. Consider an n-dimensional nonlinear stochastic system

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(t;\mathbf{x}) + \mathbf{B}(t;\mathbf{x})\dot{\mathbf{w}},$$

where $\mathbf{F}$ is a differentiable n-vector, $\mathbf{w}$ is an m-vector of standard Wiener process ($\dot{\mathbf{w}}$ is a vector of white noise), and $\mathbf{B}$ is a differentiable $m \times m$ matrix. Let ρ be the probability density function of $\mathbf{x}$ and write $g_{ij} = \sum_{k=1}^{m} b_{ik}b_{jk}$, $\rho_{\mathfrak{X}} = \int_{\mathbb{R}} \rho(\mathbf{x})dx_2$. Then the rate of information flowing from x_2 to x_1 with respect to relative entropy is

$$T_{2\to1} = E\left[\frac{1}{\rho_1}\int_{\mathbb{R}^{n-2}} \frac{\partial(F_1\rho_{\mathfrak{X}})}{\partial x_1} dx_3\cdots dx_n\right] - \frac{1}{2}E\left[\frac{1}{\rho_1}\int_{\mathbb{R}^{n-2}} \frac{\partial^2\left(g_{11}\rho_{\mathfrak{X}}\right)}{\partial x_1^2} dx_3\cdots dx_n\right].$$

Different from the previous study along this line,[28] this formula is precisely the same as that with respect to Shannon entropy [cf. (6); also see Ref. 2], except for a minus sign. The nice properties of the Shannon entropy-based formalism, such as principle of nil causality, disappearing effect with additive noise, etc., are inherited accordingly. Besides, we have also established that the so-obtained $T_{2\to1}$ is invariant upon nonlinear transformation of $(x_3,\ldots,x_n)$.

We have validated the formalism with a 3D stochastic gradient system which, though nonlinear, has an analytical solution of the Boltzmann type for its equilibrium state. As expected, the above $T_{j\to i}$ can give in a precise sense the underlying causality. With the same system, it is found that the causal influence of one component relative to another can undergo a radical change with the introduction of white noise.

All in all, information flow as a real physical notion can be rigorously derived from first principles, rather than axiomatically proposed. We want to mention that the assumption involved so far in the formulation is rather weak; only differentiability for $\mathbf{F}$ and $\mathbf{B}$ is assumed. We expect that this generic setting will allow it to find broad applications in different fields.

ACKNOWLEDGMENTS

This study was partially supported by the Jiangsu Provincial Government through the 2015 Jiangsu Program for Innovation Research and Entrepreneurship Groups and through the Jiangsu Chair Professorship, and by the State Oceanic Administration through the National Program on Global Change and Air-Sea Interaction (GASI-IPOVAI-06).

APPENDIX: PROOF OF PROPOSITION III.1

To begin, we first need to introduce the concept of Frobenius-Perron operator, or F-P operator for short. By an F-P operator, we mean a mapping[41] $\mathscr{P} : L^1(\mathbb{R}^n) \to L^1(\mathbb{R}^n)$ corresponding to $\Phi : \mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{x} \mapsto \mathbf{y}$, which takes $\rho(\mathbf{x})$ to $\rho(\mathbf{y})$, such that, for any $\omega \subset \mathbb{R}^n$,

$$\int_{\omega} \mathscr{P}\rho(\mathbf{x})d\mathbf{x} = \int_{\Phi^{-1}(\omega)} \rho(\mathbf{x})d\mathbf{x}. \tag{A1}$$

We also need the following result:

$$E\psi(\mathbf{y}) = E\psi[\Phi(\mathbf{x})], \tag{A2}$$

for any differentiable function $\psi : \Omega \to \Omega$, with Ω the sample space (R^n here). Note that the expectation operator E on the right hand side applies to a function of $\mathbf{x}$; it is thence with respect to $\rho(\mathbf{x})$. On the left hand side E is with respect to $\rho(\mathbf{y}) = \mathscr{P}\rho$, where $\mathscr{P}$ is the F-P operator associated with the mapping Φ. See Ref. 2 (pp. 3–4) for a proof.

The proposition is about the derivation of $\frac{dD_{1\mathfrak{X}}}{dt}$, i.e., the Kullback-Leibler divergence of x_1 with x_2 frozen as a parameter instantaneously at time t. To prove, consider Eq. (5) on a small interval $[t, t+\Delta t]$. Euler-Bernstein differencing,[41]

$$\mathbf{x}(t+\Delta t) = \mathbf{x}(t) + \mathbf{F}(t;\mathbf{x})\Delta t + \mathbf{B}(t;\mathbf{x}\Delta\mathbf{w}). \tag{A3}$$

To avoid confusion, write $\mathbf{x}(t+\Delta t)$ as $\mathbf{y}$, and reserve $\mathbf{x}$ for $\mathbf{x}(t)$. We need to find

$$(\mathscr{P}_{\mathfrak{X}}\rho)_1 = \int_{\mathbb{R}^{n-2}} \mathscr{P}_{\mathfrak{X}}\rho dx_3dx_4\cdots dx_n,$$

and its logarithm. By the result of Liang (2016) (see Ref. 2, p.18, l. 1 from bottom),

$$\begin{aligned}\log(\mathscr{P}_{\natural}\rho)_1(y_1) &= \log\rho_{1\natural}(y_1)\\ &\quad - \frac{\Delta t}{\rho_{1\natural}}\int_{\mathbb{R}^{n-2}}\frac{\partial F_1\rho_{\natural}}{\partial y_1}dy_3\cdots dy_n\\ &\quad + \frac{\Delta t}{2\rho_{1\natural}}\int_{\mathbb{R}^{n-2}}\frac{\partial^2 g_{11}\rho_{\natural}}{\partial y_1^2}dy_3\cdots dy_n + o(\Delta t).\end{aligned}$$

So

$$\begin{aligned}D_{1\natural}(t+\Delta t) &= E\log(\mathscr{P}_{\natural}\rho)_1(y_1) - E\log q_1(y_1)\\ &= E\log\rho_{1\natural}(y_1)\\ &\quad - E\left[\frac{1}{\rho_{1\natural}}\int_{\mathbb{R}^{n-2}}\frac{\partial F_1\rho_{\natural}}{\partial y_1}dy_3\cdots dy_n\right]\Delta t\\ &\quad + \frac{1}{2}E\left[\frac{1}{\rho_{1\natural}}\int_{\mathbb{R}^{n-2}}\frac{\partial^2 g_{11}\rho_{\natural}}{\partial y_1^2}dy_3\cdots dy_n\right]\Delta t\\ &\quad - E\log q_1(y_1) + o(\Delta t).\end{aligned}$$

The expectation on the right hand side is with respect to ρ at t. Recall that $\rho_{1\natural} = \rho_1$ at time t, and in the second and third terms $\mathbf{y}$ can be replaced by $\mathbf{x}$ with error going to higher order terms, i.e.,

$$\begin{aligned}&\frac{1}{\rho_{1\natural}(y_1)}\int_{\mathbb{R}^{n-2}}\frac{\partial F_1(\mathbf{y})\rho_{\natural}(\mathbf{y}_{\natural})}{\partial y_1}dy_3\cdots dy_n\\ &\quad = \frac{1}{\rho_1(x_1)}\int_{\mathbb{R}^{n-2}}\frac{\partial F_1(\mathbf{x})\rho_{\natural}(\mathbf{x}_{\natural})}{\partial x_1}dx_3\cdots dx_n + o(\Delta t),\\ &\frac{1}{2}\frac{1}{\rho_{1\natural(y_1)}}\int_{\mathbb{R}^{n-2}}\frac{\partial^2 g_{11}(\mathbf{y})\rho_{\natural}(\mathbf{y}_{\natural})}{\partial y_1^2}dy_3\cdots dy_n\\ &\quad = \frac{1}{2}\frac{1}{\rho_1(x_1)}\int_{\mathbb{R}^{n-2}}\frac{\partial^2 g_{11}(\mathbf{x})\rho_{\natural}(\mathbf{x}_{\natural})}{\partial x_1^2}dx_3\cdots dx_n + o(\Delta t).\end{aligned}$$

Besides,

$$\begin{aligned}\log q_1(y_1) &= \log q_1(x_1 + F_1\Delta t + \mathbf{B}_1\Delta\mathbf{w})\\ &= \log q_1(x_1) + \frac{\partial\log q_1}{\partial x_1}(F_1\Delta t + \mathbf{B}_1\Delta\mathbf{w})\\ &\quad + \frac{1}{2}\frac{\partial^2\log q_1}{\partial x_1^2}\mathbf{B}_1\Delta\mathbf{w}\Delta\mathbf{w}^T\mathbf{B}_1^T + o(\Delta t)\end{aligned}$$

and notice that, by definition of the Wiener process,

$$\begin{aligned}E\Delta\mathbf{w} &= 0,\\ E\Delta\mathbf{w}\Delta\mathbf{w}^T &= \Delta t\mathbf{I}_{m\times m}.\end{aligned}$$

So

$$\begin{aligned}D_{1\natural}(t+\Delta t) &= E\left[\log\rho_1(x_1) + \frac{\partial\log\rho_1}{\partial x_1}(F_1\Delta t + \mathbf{B}_1\Delta\mathbf{w})\right.\\ &\quad \left. + \frac{1}{2}\frac{\partial^2\log\rho_1}{\partial x_1^2}\mathbf{B}_1\Delta\mathbf{w}\Delta\mathbf{w}^T\mathbf{B}_1^T\right]\\ &\quad - E\left[\frac{1}{\rho_1}\int_{\mathbb{R}^{n-2}}\frac{\partial F_1\rho_{\natural}}{\partial x_1}dx_3\cdots dx_n\right]\Delta t\\ &\quad + \frac{1}{2}E\left[\frac{1}{\rho_1}\int_{\mathbb{R}^{n-2}}\frac{\partial^2 g_{11}\rho_{\natural}}{\partial x_1^2}dx_3\cdots dx_n\right]\Delta t\\ &\quad - E\log q_1(x_1) - E\left[F_1\frac{\partial\log q_1}{\partial x_1}\right]\Delta t\\ &\quad - E\left[g_{11}\frac{\partial^2\log q_1}{\partial x_1^2}\right]\Delta t + o(\Delta t)\\ &= D_1(t) + E\left[F_1\frac{\partial\log\rho_1/q_1}{\partial x_1}\right]\Delta t\\ &\quad + \frac{1}{2}E\left[g_{11}\frac{\partial^2\log\rho_1/q_1}{\partial x_1^2}\right]\Delta t\\ &\quad - E\left[\frac{1}{\rho_1}\int_{\mathbb{R}^{n-2}}\frac{\partial F_1\rho_{\natural}}{\partial x_1}dx_3\cdots dx_n\right]\Delta t\\ &\quad + \frac{1}{2}E\left[\frac{1}{\rho_1}\int_{\mathbb{R}^{n-2}}\frac{\partial^2 g_{11}\rho_{\natural}}{\partial x_1^2}dx_3\cdots dx_n\right]\Delta t\\ &\quad + o(\Delta t).\end{aligned}$$

Take the limit to get

$$\begin{aligned}\frac{dD_{1\natural}}{dt} &= \lim_{\Delta t\to 0}\frac{D_{1\natural}(t+\Delta t) - D_1(t)}{\Delta t}\\ &= E\left[F_1\frac{\partial\log\rho_1/q_1}{\partial x_1}\right] + E\left[g_{11}\frac{\partial^2\log\rho_1/q_1}{\partial x_1^2}\right]\\ &\quad - E\left[\frac{1}{\rho_1}\int_{R^{n-2}}\frac{\partial F_1\rho_{\natural}}{\partial x_1}dx_3\cdots dx_n\right]\\ &\quad + \frac{1}{2}E\left[\frac{1}{\rho_1}\int_{R^{n-2}}\frac{\partial^2 g_{11}\rho_{\natural}}{\partial x_1^2}dx_3\cdots dx_n\right].\end{aligned}$$

□

[1] A. P. Dempster, "Causality and statistics," J. Stat. Plan. Infer. **25**, 261–278 (1990).
[2] X. S. Liang, "Information flow and causality as rigorous notions ab initio," Phys. Rev. E **94**, 052201 (2016).
[3] X. S. Liang, "Unraveling the cause-effect relation between time series," Phys. Rev. E **90**, 052150 (2014).
[4] A. Stips, D. Macias, C. Coughlan, E. Garcia-Gorriz, and X. S. Liang, "On the causal structure between CO_2 and global temperature," Nat. Sci. Rep. **6**, 21691 (2016).
[5] R. Kleeman, "Measuring dynamical prediction utility using relative entropy," J. Atmos. Sci. **59**, 2057–2072 (2002).
[6] C. Granger, "Investigating causal relations by econometric models and cross-spectral methods," Econometrica **37**, 424 (1969).
[7] K. Kaneko, "Lyapunov analysis and information flow in coupled map lattices," Physica D **23**, 436–447 (1986)..
[8] J. A. Vastano and H. L. Swinney, "Information transport in spatiotemporal systems," Phys. Rev. Lett. **60**, 1773 (1988).
[9] T. Schreiber, "Measuring information transfer," Phys. Rev. Lett. **85**(2), 461–464 (2000).
[10] X. S. Liang and R. Kleeman, "Information transfer between dynamical system components," Phys. Rev. Lett. **95**(24), 244101 (2005).
[11] B. Pompe and J. Runge, "Momentary information transfer as a coupling measure of time series," Phys. Rev. E **83**, 051122 (2011).

[12]J. Sun and E. Bolt, "Causation entropy identifies indirect influences, dominance of neighbors, and anticipatory couplings," Physica D **267**, 49–57 (2014).
[13]L. Barnett, A. B. Barrett, and A. K. Seth, "Granger causality and transfer entropy are equivalent for Gaussian variables," Phys. Rev. Lett. **103**(23), 238701 (2009).
[14]C. W. J. Granger, "Testing for causality: A personal viewpoint," J. Econ. Dyn. Control **2**, 329–352 (1980).
[15]H. Nalatore, M. Ding, and G. Rangarajan, "Mitigating the effects of measurement noise on Granger causality," Phys. Rev. E **75**, 031123 (2007).
[16]D. A. Smirnov, "Spurious causalities with transfer entropy," Phys. Rev. E **87**, 042917 (2013).
[17]J. T. Lizier and M. Prokopenko, "Differentiating information transfer and causal effect," Eur. Phys. J. B **73**(4), 605–615 (2010).
[18]J. Runge, J. Heitzig, N. Marwan, and J. Kurths, "Quantifying causal coupling strength: A lag-specific measure for multivariate time series related to transfer entropy," Phys. Rev. E **86**, 061121 (2012).
[19]M. Staniek and K. Lehnertz, "Symbolic transfer entropy," Phys. Rev. Lett. **100**, 158101 (2008).
[20]G. Sugihara *et al.* "Detecting causality in complex ecosystems," Science **338**, 496–500 (2012).
[21]X. S. Liang and R. Kleeman, "A rigorous formalism of information transfer between dynamical system components. I. Discrete mapping," Physica D **231**, 1–9 (2007).
[22]X. S. Liang and R. Kleeman, "A rigorous formalism of information transfer between dynamical system components. I. Continuous flow," Physica D **227**, 173–182 (2007).
[23]A. J. Majda and J. Harlim, "Information flow between subspaces of complex dynamical systems," Proc. Natl. Acad. Sci. **104**(23), 9558–9563 (2007).
[24]U. Feldmann and J. Bhattacharya, "Predictability improvement as an asymmetrical measure of interdependence in bivariate time series," Int. J. Bifurcat. Chaos Appl. Sci. Eng. **14**(2), 505–514 (2004).
[25]A. Krakovská and F. Hanzely, "Testing for causality in reconstructed state spaces by an optimized mixed prediction method," Phys. Rev. E **94**, 052203 (2016).
[26]R. Kleeman and A. J. Majda, "Predictability in a model of geostrophic turbulence," J. Atmos. Sci. **62**, 2864–2879 (2005).
[27]R. Kleeman, "Limits, variability and general behavior of statistical predictability of the mid-latitude atmosphere," J. Atmos. Sci. **65**, 263–275 (2008).
[28]X. S. Liang, "Local predictability and information flow in complex dynamical systems," Physica D **248**, 1–15 (2013).
[29]X. S. Liang, "Information flow within stochastic dynamical systems," Phys. Rev. E **78**, 031113 (2008).
[30]D. W. Hahs and S. D. Pethel, "Distinguishing anticipation from causality: Anticipatory bias in the estimation of information flow," Phys. Rev. Lett. **107**, 128701 (2011).
[31]J. L. Kaplan and J. A. Yorke, *Functional Differential Equations and Approximations of Fixed Points*, Lecture Notes in Mathematics Vol. 730 (Springer-Verlag, 1979).
[32]O. E. Rössler, "An equation for continuous chaos," Phys. Lett. **57A**(5), 397–398 (1976).
[33]A. J. Majda and I. Timofeyev, "Remarkable statistical behavior for truncated Burgers-Hopf dynamics," Proc. Natl. Acad. Sci. U.S.A. **97**(23), 12413–12417 (2000).
[34]C. Bai, R. Zhang, S. Bao, X. S. Liang, and W. Guo, "Forecasting the tropical cyclone genesis over the Northwest Pacific through identifying the causal factors in the cyclone-climate interactions," J. Atmos. Ocean Tech. **35**, 247–259 (2018).
[35]X. S. Liang and A. Lozano-Durán, "A preliminary study of the causal structure in fully developed near-wall turbulence," in *Proceedings of the Summer Program 2016* (Center for Turbulence Research, 2016), pp. 233–242.
[36]X. S. Liang, "Normalizing the causality between time series," Phys. Rev. E **92**, 022126 (2015).
[37]B. H. Vaid and X. S. Liang, "The changing relationship between the convection over the western Tibetan Plateau and the sea surface temperature in the northern Bay of Bengal," Tellus A: Dyn. Meteorol. Oceanogr. **70**(1), 1440869 (2018).
[38]I. Hoyos, J. Ca nón-Barriga, T. Arenas-Suárez, F. Dominguez, and B. A. Rodríguez, "Variability of regional atmospheric moisture over Northern South America," Clim. Dyn. (2018).
[39]J. Huang, Y. Li, C. Fu, F. Chen, Q. Fu, A. Dai, M. Shinoda, Z. Ma, W. Guo, Z. Li, L. Zhang, Y. Liu, H. Yu, Y. He, Y. Xie, X. Guan, M. Ji, L. Lin, S. Wang, H. Yan, and G. Wang, "Dryland climate change: Recent progress and challenges," Rev. Geophys. **55**, 719–778 (2017).
[40]T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, Inc., 1991).
[41]A. Lasota and M. C. Mackey, *Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics* (Springer, New York, 1994).
[42]X. S. Liang, "Entropy evolution and uncertainty estimation with dynamical system," Entropy **16**, 3605–3634 (2014).
[43]T. Schneider, S. M. Griffies, "A conceptual framework for predictability studies," J. Clim. **12**, 3133–3155 (1999).
[44]J. M. Amigó, R. Monetti, N. Tort-Colet, and M. V. Sanchez-Vives, "Infragranular layers lead information flow during slow oscillations according to information directionality indicators," J. Comput. Neurosci. **39**, 53–62 (2015).