Revisiting causality using stochastics: 1. Theory

Demetris Koutsoyiannis[1], Christian Onof[2], Antonis Christofides[1] and Zbigniew W. Kundzewicz[3]

[1] Department of Water Resources and Environmental Engineering, School of Civil Engineering, National Technical University of Athens (dk@itia.ntua.gr)

[2] Department of Civil and Environmental Engineering, Faculty of Engineering, Imperial College London

[3] Meteorology Lab, Department of Construction and Geoengineering, Faculty of Environmental Engineering and Mechanical Engineering, Poznan University of Life Sciences, Poznań, Poland

Abstract Causality is a central concept in science, in philosophy and in life. However, reviewing various approaches to it over the entire knowledge tree, from philosophy to science and to scientific and technological application, we locate several problems, which prevent these approaches from defining sufficient conditions for the existence of causal links. We thus choose to determine necessary conditions that are operationally useful in identifying or falsifying causality claims. Our proposed approach is based on stochastics, in which events are replaced by processes. Starting from the idea of stochastic causal systems, we extend it to the more general concept of hen-or-egg causality, which includes as special cases the classic causal, and the potentially causal and anticausal systems. Theoretical considerations allow the development of an effective algorithm, applicable to large-scale open systems, which are neither controllable nor repeatable. The derivation and details of the algorithm are described in this paper, while in a companion paper we illustrate and showcase the proposed framework with a number of case studies, some of which are controlled synthetic examples and others real-world ones arising from interesting scientific problems.

Keywords causality, causal systems, stochastics, impulse response function, system identification

> *ὡς ἐγὼ οὐ νῦν πρῶτον ἀλλὰ καὶ ἀεὶ τοιοῦτος οἷος τῶν ἐμῶν μηδενὶ ἄλλῳ πείθεσθαι ἢ τῷ λόγῳ ὃς ἄν μοι λογιζομένῳ βέλτιστος φαίνηται.*
>
> *I am not only now but always a man who follows nothing but the reasoning which on consideration seems to me best.*
>
> Plato, Crito, 46b-47d, quoting Socrates

Postprint (2022-06-11) with updates on the analytical solution (p. 22, equations (30)-(37)) and a correction of a typo in equation (27)

1 Introduction

Causality (or causation[1]) is a central concept in science, in philosophy and in life, yet its meaning is not clear. Difficulties in disambiguating it extend over the entire knowledge tree, from philosophy and to science to scientific and technological applications. When it comes to science, an operational framework is well established, which is based on the notion of controlled and repeatable experiments. However, this framework is not applicable to large-scale open systems, which are neither controllable nor repeatable. The difficulties in identifying causality in such systems are amplified. Even though several algorithms have been proposed to identify causality in such systems, based on probabilistic considerations and statistical processing of data, they are mostly problematic.

Apparently, as giants in philosophy and science have not yet resolved these problems, one should not expect our humble set of two companion papers to do that. On the other hand, existing knowledge gaps offer us grounds for trying to make some headway in attempting to locate and elucidate those gaps (Section 2 of this paper) and propose a different identification framework applicable to open systems (Section 3 of this paper). We illustrate and showcase the proposed framework by means of a number of case studies, some of which are controlled synthetic examples and others real-world ones characterizing interesting scientific problems. To avoid a long paper, we present the case studies in the companion paper (Koutsoyiannis et al. 2022a).

2 Theoretical background

2.1 Philosophical background

There is a mystery in the concept of cause. While at first glance it seems clear what we mean with the word, whenever we consider it more closely, we find ourselves unable even to define it. Aristotle (384–322 BC) seems to have noticed this, for he wrote, among other things:

> *that which when present is the cause of something, when absent we sometimes consider to be the cause of the contrary; for example, we consider the absence of the captain to be the cause of the ship's capsizing, whereas his presence was the cause of the ship's rescue"* (*ὃ γὰρ παρὸν αἴτιον τοῦδε, τοῦτο καὶ ἀπὸν αἰτιώμεθα ἐνίοτε τοῦ*

[1] The terms "causation" and "causality" most of the times, also in this text, are typically used interchangeably as synonymous, meaning the existence of a cause-effect relationship. We note though that the two terms sometimes have been used with slightly differing meanings. The former sometimes (e.g. Sion 2010) denotes a deterministic cause-effect relationship (deterministic causality). The latter may also mean the principle that everything has a cause (e.g. Bunge 1979, p.3).

> *ἐναντίου, οἷον τὴν ἀπουσίαν τοῦ κυβερνήτου τῆς τοῦ πλοίου ἀνατροπῆς, οὗ ἦν ἡ παρουσία αἰτία τῆς σωτηρίας* [Bekker number 195a12-14]).

In some languages, the original meaning of the word appears to be "the one who is to blame". According to various dictionaries, the Greek word "αἴτιος", which comes from the verb "αἴνυμαι", "to grab", must originally have meant "he who has a part". Likewise, the Latin word "causa" is also the origin of "accuse". The German "Ursache", on the other hand, literally means "the original thing". This suggests two important aspects of what we understand as cause: (i) insofar as identifying a cause is identifying what is responsible for something, it provides an explanation; (ii) insofar as a cause is the origin/ground leading to some occurrence, it is connected with the idea of some process (physical or psychological). However, this common understanding of the notion of cause hides certain philosophical problems.

The Scottish philosopher David Hume (1711-1776) was the first to raise doubts about the existence of causes. His work has been so influential that almost all modern philosophical and scientific studies referring to causation start the thread from him. He pointed out that causal connections are not visible or otherwise directly perceivable by our senses. When a flame causes heat, we perceive the flame, we perceive the heat, but the causal connection between the flame and the heat we do not perceive; we deduce it. From the time of birth, we observe certain regularities in the world. For Hume, when we say that the flame is the cause and heat is the effect, we only express an observed regularity—a physical law. Here is a related passage from Hume (1748, Section VI, part I—with modernized spelling and punctuation):

> *Suppose a person, though endowed with the strongest faculties of reason and reflection, to be brought on a sudden into this world; he would indeed immediately observe a continual succession of objects, and one event following another; but he would not be able to discover anything further* […] *Suppose, again, that he has acquired more experience, and has lived so long in the world as to have observed familiar objects or events to be constantly conjoined together. What is the consequence of this experience? He immediately infers the existence of one object from the appearance of the other. Yet he has not, by all his experience, acquired any idea or knowledge of the secret power by which the one object produces the other; nor is it by any process of reasoning he is engaged to draw this inference.*

Hume's conclusion is that the concept of a cause is merely a way we use to describe regularities.

Hume's attack on the notion of cause served as a wake-up call for the German philosopher Immannuel Kant (1724–1804), who agreed with Hume's diagnosis but not his conclusion. Rather, he argued that Hume did not go far enough (Kant 1787/1998,

B788-9; pp. 653-4; see Gardner 1999, p.11). This was one of the triggers for Kant's so-called "*critical turn*" which introduced a revolutionary approach to how we understand what it is to be an object. While Kant did not doubt that knowledge of objects starts with the information we receive through our senses, the question of how this sensory input leads to the experience/knowledge/representation of an object is one that philosophy had never really provided a satisfactory answer to. Rather than assume, as his predecessors had done, that our knowledge of objects depends in some way (that has never been explained—e.g. how does raw sensory information lead to the representation of an object?) upon objects that are already given to us independently of our cognition, he proposed that our cognition is partly responsible for constituting such objects. So, features like space, time and causality are contributed by the subject as ways of structuring the raw sensory input. The concept of cause and effect in particular, and Kant's claim that "*All alterations occur in accordance with the law of the connection of cause and effect*" (Kant 1787/1998, B232, p. 304) play a fundamental role in the temporal structure of objective experience. For our purposes, what is important in Kant's understanding of causality is that (a) it is understood in terms of rule-governedness (i.e. that which is *regular* has a *rule* governing its behaviour), and (b) the temporal/causal order is irreversible (ibid. B237, pp. 306-7). These two characteristics will be used below.

The general question of how to define and identify causes remains however and this is the locus of the contemporary debate around the concept of cause. Many philosophers and mathematicians have attempted to answer it by developing theories of causation. These are based on interventions (A causes B if by deliberately creating A, B follows), counterfactuals (A causes B if B would not have occurred had A been absent), necessary and sufficient conditions (A causes B if A is necessary and sufficient for B to occur), or probability (A causes B if the presence of A increases the probability of B; see section 2.2). Combinations of these approaches have also been proposed. However, no completely satisfactory characterization has been formulated. Naturally, this philosophical perplexity is also reflected in the mathematical representation of causality, which defines the scope of our study. In particular, this paper and its companion are a contribution to the probabilistic approach to causality.

2.2 Probabilistic theories of causality

While the above considerations show that it is difficult to define what causality is, we have seen grounds for presupposing it is intimately connected with temporal asymmetry and irreversibility. Thus, Mehlberg (1983) explained that "*no causal process (i.e., such that of two consecutive phases, one is always the cause of the other) can be reversible*" and also presented the causal theory of time, according to which "*two events are simultaneous* by

definition *if there can be no causal action between them*". The issue of time directionality of causality was also discussed by Kline (1980).

Further, insofar as a causal link is governed by some rule, it is natural to turn to a mathematical representation of causation. In a deterministic framework, the definition of what constitutes a causal link can be formalised into a "logic of causation" (e.g. Sion, 2010). Application of the deterministic framework in describing a natural system is rather easy if the system is simple and the mechanisms acting on the system are well understood (e.g. gravity is a cause behind Earth's rotation around the Sun, as well as for chains of sequential events, such as the popular example of a falling row of dominoes, depicted on the cover page of the cited book by Sion). The problem of detecting and establishing causality becomes challenging when the mechanisms are complex and not well understood, and when inference by deduction together with empirical causal laws is not possible. In this case we have to resort to induction based on observations, use probabilistic logic and model the system by stochastics.

Among the first who connected causality with probability and statistics were Hopf (1934), and Birkhoff and Lewis Jr. (1935). The latter authors used the term "causal system", for which they noted:

> *In the practical calculation with actual causal systems, only a limited degree of accuracy is sought, since the laws of the system are at most an idealization of the actual laws, and the isolation of the system from other systems, which is always postulated, can never be more than imperfectly realized.*

Later, Wold (1954, 1960), as well as Strotz and Wold (1960) also studied causality in the framework of econometrics and made again, within this framework, a connection with probability and statistics.

Probabilistic definitions of causality are based on time asymmetry on the one hand and conditional probability on the other hand. Thus, Suppes (1970) defined it as follows

> *An event* $B_{t'}$ [occurring at time t'] *is a* prima facie *cause of the event* A_t [occurring at time t] *if and only if*
>
> (i) $t' < t$,
> (ii) $P(B_{t'}) > 0$,
> (iii) $P(A_t|B_{t'}) > P(A_t)$.

The notion of a "*prima facie* cause" is discussed below. In plain language, the cause must precede the effect and the conditional probability of the effect under the condition of the cause must exceed the unconditional probability. This definition does not exclude the possibility of more than one cause.

Suppes's third criterion conveys the idea that the presence of the cause raises the probability of occurrence of the effect. This idea is arguably better expressed as an inequality between conditional probabilities (Skyrms 1980):

(iii′) $P(A_t|B_{t'}) > P\left(A_t|\overline{B}_{t'}\right)$

where $\overline{B}_{t'}$ is the absence (non-occurrence) of event $B_{t'}$. However, using the obvious relationship $P(A_t) = P(A_tB_{t'}) + P(A_t\overline{B}_{t'})$, it can easily be shown that the two versions are equivalent. Cox (1992) points out that such a condition still allows for "spurious causality". The latter could only be eliminated by adding a condition such as:

(iv) there is no event $C_{t''}$ at time $t'' < t' < t$ such that $P(A_t|B_{t'}C_{t''}) = P\left(A_t\middle|\overline{B}_{t'}C_{t''}\right)$.

A version of this condition was also defined by Salmon (1998) within his statistical-relevance theory of explanation, as the key to distinguishing between statistical and causal relevance which he defines as:

(iv′) there is no event $C_{t''}$ at time $t'' < t' < t$ which "screens off" $B_{t'}$ from A_t such that $P(A_t|B_{t'}C_{t''}) = P(A_t|C_{t''})$.

Salmon's example of statistical relevance which is not causal and therefore defines a spurious correlation is if $A_t, B_{t'}, C_{t''}$ refer respectively to the occurrence of a storm, a barometer drop and an air pressure drop. However, conditions such as (iv) or (iv′) are pretty much impossible to verify satisfactorily in practice. This places limits upon the practical use of these characterisations of causation.

Another popular definition, given by Granger (1980), is the following: "Y_n *is said to cause* X_{n+1}, *if* $P(X_{n+1} \in A|\Omega_n) \neq P(X_{n+1} \in A|\Omega_n - Y_n)$ *for some* A." In this, Granger assumes discrete time which he denotes as n, while he denotes Ω_n "*the knowledge in the universe available at that time*" and Y_n the information composed of "*the values taken by a variable* Y_t *up to time* n, *where* $Y_n \in \Omega_n$" (with the last notation best rendered as $Y_n \subseteq \Omega_n$). Further, he provided three axioms, the first of which is equivalent to (i) above and the third highlights the constancy in causality direction throughout time.

In his earlier publication, which has been much more influential[2], Granger (1969) gave a different version of his definition in an attempt to be statistically testable. With his notation of the later definition stated above, the condition upon which the earlier definition is based reads $\text{var}[X_{n+1}|\Omega_n] < \text{var}[X_{n+1}|\Omega_n - Y_n]$. That is, the probability of an event here becomes variance and the inequality sign "≠" here becomes "<". Granger (1969) clarified his mathematical expression thus: "Y_t *is causing* X_t *if we are better able to predict* X_t *using all available information than if the information apart from* Y_t *had been used.*" Furthermore, Granger (1969) defined the feedback in this way (after replacing the

[2] 27 000 Google Scholar citations vs. 2300 of Granger (1980).

notation with that of Granger (1980)): "*If* $\text{var}[X_{n+1}|\Omega_n] < \text{var}[X_{n+1}|\Omega_n - Y_n]$, $\text{var}[Y_{n+1}|\Omega_n] < \text{var}[Y_{n+1}|\Omega_n - X_n]$ *we say that feedback is occurring* […], *i.e., feedback is said to occur when* Y_t *is causing* X_t *and also* X_t *is causing* Y_t".

Granger (1969) also proposed what has later been known as the "Granger causality test". This is based on the improvement in the prediction of a process $\underline{y}_\tau$ by considering the influence of a "causing" process $\underline{x}_\tau$. Notice that, at this point on, we do not follow Granger's original notational conventions; rather we make it clear that $\underline{x}_\tau$ and $\underline{y}_\tau$ are stochastic processes and we underline stochastic (random) variables and stochastic processes to distinguish them from common variables (representing single real numbers) and deterministic functions, respectively. The prediction equation is the Granger regression model:

$$\underline{y}_\tau = \sum_{j=1}^{\eta} a_j \underline{y}_{\tau-j} + \sum_{j=1}^{\eta} b_j \underline{x}_{\tau-j} + \underline{\varepsilon}_\tau \tag{1}$$

where a_j and b_j are the regression coefficients and $\underline{\varepsilon}_\tau$ is an error term. Notice that the equation (1) does not include the term $\underline{x}_\tau$ that is synchronous with $\underline{y}_\tau$ and thus it excludes what Granger calls "instantaneous causality". We note though that in reality this does not indicate "instantaneous causality" but treatment of discrete time as if it were continuous (we discuss this point in sections 3.3 – 4.) The test is based on the null hypothesis (H_0) that the process $\underline{x}_\tau$ is not actually causing $\underline{y}_\tau$, formally expressed as:

$$H_0: b_1 = b_2 = \cdots = b_\eta = 0. \tag{2}$$

Algorithmic details of the test are given in Gujarati and Porter (2009), among others. The test is quite popular and several software platforms include free applications to execute it[3]. The rejection of the null hypothesis is commonly interpreted in the literature with the statement "$\underline{x}_\tau$ Granger-causes $\underline{y}_\tau$".

It is clear that Granger's statement "Y_t *is causing* X_t *if we are better able to predict* X_t…" in reality identifies improvement of predictability with causality, or in other words, statistical association with causation. If this statement is taken together with his regression equation (1), in which the involved parameters are calculated through correlation coefficients, it eventually identifies correlation with causation. But the mantras "*association is not causation*" and "*correlation is not causation*" express a widely held opinion which we believe is correct.[4] Granger (1980) was clearly aware of this:

[3] For example, the function GRANGER_TEST is available for Excel by C. Zaiontz (Real Statistics Using Excel, http://www.realstatistics.com/; Real Statistics Examples Workbooks. http://www.real-statistics.com/free-download/real-statistics-examples-workbook/; accessed on 1 September 2020).

[4] Google counts 348 000 appearances of the former and 533 000 of the latter.

> *when discussing the interpretation of a correlation coefficient or a regression, most textbooks warn that an observed relationship does not allow one to say anything about causation between the variables.*

Replacing correlation with probability, as did Granger (1980) and before him Suppes (1970), does not change the essence in the problem. Perhaps this is the reason why Suppes used the term "*prima facie* cause" in his definition given above (the adjective prima facie, originating from Latin, means based on the first impression; accepted as correct until proved otherwise). Suppes attributed the expression to J. Hintikka but he did not explicitly explain it. Furthermore, he discussed *spurious causes* and eventually defined the *genuine cause* as a "prima facie *cause that is not spurious*"; he also discussed the very existence of genuine causes. The term "*prima facie* cause" was also used by Granger. In particular, Granger and Newbold (1986) noted that a cause satisfying a causality test "*still remains* prima facie *because it is always possible that, if a different information set were used, then* [it] *would fail the new test*". This is in line with the inductive, rather than deductive, character of statistical tests, insofar as the conclusion is never the confirmation of a hypothesis but only its non-rejection.

Despite these caveats, the term "Granger causality" is very popular, particularly in the expression "Granger causality test" (e.g., Gujarati and Porter, 2009). This terminology has misled many to understand the test as identifying causality and resolving the "correlation is not causation" problem. In fact, all it detects is correlation, not genuine causality.

Cohen (2014) clearly saw the problem when he suggested replacing the term "Granger causality" with "Granger prediction" after correctly pointing out that:

> *Results from Granger causality analyses neither establish nor require causality. Granger causality results do not reveal causal interactions, although they can provide evidence in support of a hypothesis about causal interactions.*

The ambition to identify genuine causes with statistical tools and thereby overcome the concern that "correlation is not causation" has motivated others to find statistics other than the correlation coefficients to characterize causality. For example, Liang (2016) used the so-called information flow (or information transfer) between two processes, while in later works this method has been called "*Liang causality*" (Stips et al., 2016). He asserted that "*causation implies correlation, but not vice versa*" (Liang, 2016) and "*causality actually can be rigorously derived in terms of information flow from first principles*" (Liang, 2018), . On the other hand, Koutsoyiannis and Kundzewicz (2020) asserted that:

> [The] *vanity* [of this approach] *to determine genuine causality is easy to infer: It suffices to consider the case where the two processes for which causality is studied are jointly Gaussian. It is well known that in any multivariate Gaussian process, the*

covariance matrix (or the correlation matrix along with the variances) fully determines all properties of the multivariate distribution of any order. For example, the mutual information in a bivariate Gaussian process is (Papoulis, 1991)

$$H[\underline{y}|x] = \ln\left(\sigma_y\sqrt{2\pi e(1-r^2)}\right) \quad (3)$$

where σ_y and r denote standard deviation, and correlation, respectively. Thus, using any quantity related to entropy (or information) is virtually identical to using correlation. Furthermore, in Gaussian processes, whatever statistic is used in describing causality is readily reduced to correlation. This is evident even in Liang (2016), where, e.g., in his Equation (102), the information flow turns out to be the correlation coefficient multiplied by a constant.

In a similar vein, Verbitsky et al. (2019) used a technique of distances of multivariate vectors to reconstruct the system dynamics. To do so, they assumed that "*each time series is a variable produced by its hypothetical low dimensional system of dynamical equations*". But if indeed the system dynamics were of low dimensionality, it would be preferable to model the system by deduction, rather than induction based upon doubtful statistical techniques. As pointed out by Koutsoyiannis and Kundzewicz (2020) (also referring to Koutsoyiannis, 2006),

such assumptions and techniques are good for simple toy models but, when real world systems are examined, low dimensionality appears as a statistical artifact because the reconstruction actually needs an incredibly high number of observations to work, which are hardly available. The fact that the sums of multivariate vectors of distances is a statistical estimator with huge uncertainty is often missed in studies of this type, which treat data as deterministic quantities, thereby obtaining unreliable results. We do not believe that the Earth system and Earth processes […] *are of low dimensionality.*

A more satisfactory framework was proposed by Hannart et al. (2016), based on the works by Pearl (2009) and Pearl et al. (2016). In it they used the so-called *causal graph* reflecting the assumed dependencies among the studied variables along with the notion of *exogeneity* (perhaps borrowed from Wold (1960), and Strotz and Wold (1960)). To define the latter, they stated that "*a sufficient condition for X to be exogenous wrt any variable is to be a top node of a causal graph.*" But importantly, this assumes that we already have a causal graph, i.e., a way of identifying causes.

Further, central to the framework of Hannart et al. (2016) is the notion of *intervention* of an experimenter (perhaps again borrowed from Strotz and Wold (1960)). But clearly, while experimentation is feasible in laboratory experiments, it is infeasible in natural (e.g. geophysical) processes. To bypass this fundamental obstacle, Hannart et al. resorted to the "so-called *in silico experimentation*". While this is indeed an impressive name, it simply means experimentation with a mathematical model that represents the

process. It is trivial to note that models, however sophisticated, are not identical to the real world. Hence, objectively the technique examines a hypothetical "causality" that is incorporated in the model rather than natural causality. Arguably, calculating probabilities by model simulations is inferior to inspecting the model's equations or code. The latter method would be more appropriate to reveal what "causality" is incorporated into the model through its construction.

Hannart et al. (2016), studied the probability of occurrence of an event *Y*, conditional upon the two-valued (binary) variable X_f, which indicates whether or not a forcing *f* is present, for which they stated:

> *The probability* $p_1 = P(Y = 1|X_f = 1)$ *of the event occurring in the real world, with f present, is referred to as factual, while* $p_0 = P(Y = 1|X_f = 0)$ *is referred to as counterfactual.* […] *The so-called fraction of attributable risk (FAR) is then defined as*
>
> $$\mathrm{FAR} = 1 - \frac{p_0}{p_1} \tag{4}$$
>
> *The FAR is interpreted as the fraction of the likelihood of an event that is attributable to the external forcing*.

They showed that, under some conditions, FAR is a probability, which they denoted PN and called *probability of necessary causality*. They stressed that it "*is important to distinguish between necessary and sufficient causality*" and they associated PN, "*with the first facet of causality, that of necessity*". They claimed to have "*introduced its second facet, that of sufficiency, which is associated with the symmetric quantity* $1 - (1 - p_1)/(1 - p_0)$"; they denoted it as PS and called it *probability of sufficient causality*.

However, the framework has several drawbacks and can fail, as illustrated by the following counter-example by Koutsoyiannis and Kundzewicz (2020): When the atmospheric temperature is high people wear light clothes and also sweat much more than when it is cold. Thus, the weight of clothes improves the prediction of the sweat quantity. Koutsoyiannis and Kundzewicz (2020) used the two-valued stochastic variables $\underline{x}, \underline{y}, \underline{z}$ to model the states of temperature, clothes weight and sweat, respectively, and assumed a hypothetical "artificial intelligence entity" (AIE) which decides on causality based upon the probability rules of Hannart et al. (2016). After assigning plausible values to the conditional probabilities of high sweat for the four conditions of cold/hot and heavy/light clothes, and following detailed numerical calculations of PN and PS, they obtained the absurd result that the AIE will decide that there is all necessary and sufficient evidence that light clothes cause high sweat. Hannart et al. (2016) might protest that the absurd result occurs because of improper assignment of the exogenous variable. But how could the AIE know that? How do we know the chain of causation a priori in order to create the causal graph?

The above critical summary of some recent and earlier studies on causality strengthens what was stated by Koutsoyiannis and Kundzewicz (2020), i.e., that "*identifying genuine causality is not a problem of choosing the best algorithm to establish a statistical relationship (including its directionality) between two variables*" and, ultimately, that "*the big philosophical problem of causality cannot be resolved by technical tricks*".

Therefore, here we focus on simpler problems, such as falsifying an assumed genuine causality and adding statistical evidence, in an inductive context, for potential causality and its direction.

3 Proposed framework

3.1 From seeking a definition to defining necessary conditions

Coming back to the probabilistic definitions of causality summarized in section 2.2, we may remark that they do not have the clarity and unambiguousness required in science. The only clear element, at least in a classical physical framework, is the time precedence of the cause from the effect. The conditional probability element of the definition or the related axioms do not help clarify real causality, if we assume that such a thing really exists. If the inequality $P(A_t|B_{t'}) > P(A_t)$ entails (*prima facie*) causality then the opposite one, $P(A_t|B_{t'}) < P(A_t)$ also does, because it can be written as $P\left(A_t\middle|\overline{B}_{t'}\right) > P(A_t)$. Indeed, using standard probability calculus and noting that $P(A_t|B_{t'}) < P(A_t)$ implies $P(A_tB_{t'}) < P(A_t)P(B_{t'})$ we find:

$$P\left(A_t\middle|\overline{B}_{t'}\right) = \frac{P\left(A_t\overline{B}_{t'}\right)}{P\left(\overline{B}_{t'}\right)} = \frac{P(A_t) - P(A_tB_{t'})}{1 - P(B_{t'})} > \frac{P(A_t) - P(A_t)P(B_{t'})}{1 - P(B_{t'})} = P(A_t) \qquad (5)$$

Therefore, the only case where (*prima facie*) causality is excluded is stochastic independence, in which $P(A_t|B_{t'}) = P(A_t)$ or, equivalently, $P(A_tB_{t'}) = P(A_t)P(B_{t'})$. It is thus understandable why Granger (1980) generalized the inequality order ">" in Suppes's (1970) definition, replacing it with "≠".

A similar argument can be applied by reversing the time inequality $t' < t$ to $t' > t$, and stating that in the latter case, provided that the events A_t and $B_{t'}$ are not independent, A_t (*prima facie*) causes $B_{t'}$.

Thus, in effect the existing definitions assert that *any* two events that are neither synchronous nor independent establish a causal relationship, with the direction of causality determined by the time order. This is too general to have any usefulness. Also, it is rather unnecessary, as it does not add anything important to the well-defined notion of (in)dependence.

There are additional problems with the usefulness of the above definitions, related to the estimation of probabilities from real-world data. One may assume that the notions of experimentation and its repeatability tacitly lie behind these definitions. And indeed,

these are possible in experimental physics in which laboratory experiments are used. A laboratory case represents a closed system where one can exclude influences from all kinds of external factors, which may even be very distant (cf. quantum entanglement, for instance). This, however, cannot be the case in open systems. Thus, in geophysics (a particular case of an open system) there is no repeatability because of the influence of these ever-changing external factors and the impossibility of controlling such large systems. The temporal evolution of a geophysical system is unique and unrepeatable, so that we cannot have observed samples. We can only have time series that cannot be regarded as a sample because consecutive measurements are never independent. Details about the differences between random samples and time series can be found in Koutsoyiannis (2021).

For these reasons, here we abandon the use of the notion of events and we reformulate the notion of causation on the basis of stochastic processes, which are families of (infinitely many) stochastic variables indexed by time. A series of observations from a natural process is termed a time series and is regarded as a single (and unique in geophysical processes) realization of a stochastic process.

Furthermore, given the philosophical problems that, as we have seen, characterise attempts to give a a definition of causality, we limit our scope of our investigation to providing necessary (and not sufficient) conditions of causality. We stress that necessary conditions are particularly useful in falsifying hypotheses of causality, rather than confirming it. An obvious necessary condition which we retain from all existing definitions of causality is the time precedence of the cause with respect to the effect. Other conditions are studied below. As the necessary conditions can hardly confirm causality, we use the term *potential causality* (cf. the Aristotelian notion of δύναμις—Latin: potentia; English: potency or potentiality).

3.2 Basic concepts and definitions

Let $\underline{x}(t)$ and $\underline{y}(t)$ denote two stochastic processes in continuous time t. We recall from stochastics (e.g. Papoulis, 1991, pp. 405, 508) that the two processes form a *causal system*, with $\underline{x}(t)$ being the cause and $\underline{y}(t)$ the effect, if they are related by:

$$\underline{y}(t) = \int_0^\infty g(h)\underline{x}(t-h)\mathrm{d}h \tag{6}$$

Here the deterministic function $g(h)$ is termed *impulse response function* (IRF), with h being a time lag. In the case of a causal system (sometimes also called nonanticipative system), $g(h) = 0$ for any $h < 0$. Noticeably, Papoulis did not provide a definition of causality per se, but used the concept of a causal system, defined through equation (6). The property characterizing a causal system is precisely defined by the zero values of IRF

for negative lags (Papoulis, 1991, p. 405). Notice that all values of $\underline{x}(t)$ that contribute to $\underline{y}(t)$ through the integral of the right-hand side of equation (6) correspond to a time period earlier than *t*, i.e. to the past. Also notice that, in theory, the entire past matters and hence the infinity in the upper limit of the integral. In practice the function $g(h)$, if determined from observations, has to be assumed zero beyond a certain value, i.e. the upper limit of the integral becomes finite. This, for example, has been the case in the application of the idea in hydrology, namely in the notion of the unit hydrograph (an implementation of the IRF in precipitation-runoff), even though its pioneers (Nash, 1959, Dooge, 1959) also used the full (infinite) range. Finally, notice the linearity of the relationship, which is discussed further below.

The theory of causal systems has been based upon a pioneering work by Kolmogorov (1941) followed by works by Wold (1948) and Wiener (1949). Notably, Wold (1938, 1948), influenced by Kolmogorov (see his interview by Hendry and Morgan, 1994), introduced the celebrated Wold decomposition, proving that any stochastic process can be decomposed into a regular process (i.e., a process linearly equivalent to a white noise process) and a predictable process (i.e., a process that can be expressed deterministically in terms of its past values). In none of these works did these pioneers use the term "causal system", nor did they explicitly speak about causality. However, each of them studied a form of the linear filter that was later to be called causal. The objective of these works was to enable stochastic prediction based on the past, a prediction which Kolmogorov and Wiener called "extrapolation". A little later, Bode and Shannon[5] (1950), drawing upon Kolmogorov's and Wiener's works, made the connection with causality, stating:

> *How is it possible to predict at all the future behavior of a function when all that is known is a perturbed version of its past history? This question is closely associated with the problems of causality and induction in philosophy and with the significance of physical laws.*

The connection with causality is also mentioned by Robbins (1959). In the 1960s, the term "causal system", earlier used with a different meaning by Birkhoff and Lewis Jr. (1935) as mentioned above, was connected with Kolmogorov's and Wiener's "extrapolation" filter (essentially our equation (6)), particularly in the literature of communication engineering (Drenick, 1963; Post, 1963; Sharnoff, 1964; Masani, 1966; Keats, 1967; Parzen, 1968; Clifton, 1968). But it was perhaps the book by Papoulis (1991, first edition – 1965), that disseminated the concept of a "causal system".

The relationship of equation (6) is an ideal that we can hardly meet, in a precise fashion, in a natural process. In fact, it can only be valid in a mathematical process that is

[5] It is relevant to note than two years earlier, Shannon (1948) had introduced the modern definition of entropy, while Wiener (1948) in his famous book *Cybernetics*, had used essentially the same definition (albeit with a negative sign) for information.

defined by equation (6). Therefore, if we keep equation (6) as a definition of a causal process, we will exclude causality in natural processes. Instead, here we call the system defined by equation (6) a *classic causal system* and we will relax the requirements for calling a system *causal*. What is meant by 'classic' is that the effect is (1) *fully explained* (2) by one *well-identified* cause. This condition is implicitly assumed in the absence of other additive terms in equation (6), either random $\underline{v}(t)$ (i.e., $\underline{y}(t) = \int_0^\infty g(h)\underline{x}(t-h)\mathrm{d}h + \underline{v}(t)$) or causal from a second cause $\underline{z}(t)$ (i.e., $\underline{y}(t) = \int_0^\infty g(h)\underline{x}(t-h)\mathrm{d}h + \int_0^\infty r(h)\underline{z}(t-h)\mathrm{d}h$).

We recall that, given *any* two stationary stochastic processes $\underline{x}(t)$ and $\underline{y}(t)$, we can write an equation relating them of the form (cf. Papoulis, 1991, equation (14.12)):

$$\underline{y}(t) = \int_{-\infty}^{\infty} g(h)\underline{x}(t-h)\mathrm{d}h + \underline{v}(t) \tag{7}$$

where, compared to equation (6), in the lower limit of the integral we have replaced 0 with $-\infty$ and also added a third stochastic process, $\underline{v}(t)$, assumed to be uncorrelated with $\underline{x}(t)$. The function $g(h)$ is no longer unique, but infinitely many such functions exist. The most interesting among them is the one that corresponds to the minimum variance of $\underline{v}(t)$, typically called the least-squares solution. Since this is a general property of any two processes, we can also write it in the reverse direction, i.e.,

$$\underline{x}(t) = \int_{-\infty}^{\infty} g_1(h)\underline{y}(t-h)\mathrm{d}h + \underline{v}_1(t) \tag{8}$$

where again the most interesting of the infinitely many solutions is the one yielding the minimum variance of $\underline{v}_1(t)$. Note that, $g_1(h)$ in equation (8) is different from $g(h)$ in (7)—there is no symmetry. Likewise, the process $\underline{v}_1(t)$, which is now uncorrelated to the process $\underline{y}(t)$, is different from $\underline{v}(t)$. Naturally, the selection of the optimally applicable equation between equations (7) and (8) depends upon which of them gives the minimum variance in relative terms, i.e. as a proportion of the variance of $\underline{y}(t)$ or $\underline{x}(t)$, respectively. In what follows we will exclusively use equation (7) which denotes a causality direction $x \to y$. When we examine the reverse direction, $y \to x$, instead of explicitly using equation (8), we interchange processes $(\underline{x}(t), \underline{y}(t))$ and again use equation (7). Equivalently, the final choice of the direction $x \to y$ or $y \to x$ depends upon which maximizes the *explained variance ratio*, defined as

$$e \coloneqq 1 - \frac{\mathrm{var}[\underline{v}]}{\mathrm{var}[\underline{y}]} \tag{9}$$

We will discuss some additional desiderata for the two IRFs below, which also define additional criteria for the selection of the best solution.

Further explanations on the motivation for the use of equation (7) as necessary condition for causation are provided in Supplementary Information (Section SI1.2), including a justification for its linear form. The linearity of the equation is kept from the original definition of a causal system by Papoulis (1991) (equation (6)). Certainly, linearity could be regarded by many as a limitation of our approach and possible future nonlinear extensions thereof could be considered. However, it is our opinion that linearity may suffice for most problems, for the following reasons:

- We use a stochastic approach, in which the meaning of linearity vs. nonlinearity is dramatically different from that in deterministic approaches, something not often recognized in literature. In stochastics, linearity is rather a powerful characteristic enabling the study of demanding problems, rather than a limitation. For example, stochastic dynamics need not be nonlinear to produce realistic trajectories and change. Conversely, in a deterministic system with linear dynamics, any perturbation of initial conditions dies off, as does the potential for change—and hence the importance of nonlinearity in deterministic approaches (Koutsoyiannis 2014a, 2021; Koutsoyiannis and Dimitriadis, 2021),
- In stochastics, linearity is not an (over)simplification of the dynamics but has some sound justification, as indicated by the already mentioned Wold decomposition, in which the stochastic component (the regular process) is linearly equivalent to a white noise process (i.e. a *linear* combination of white noise terms; Wold, 1938, 1948; Papoulis 1991).
- In addition, linearity in a stochastic description results from maximum entropy considerations (under plausible conditions; e.g. Papoulis, 1991) and hence it is related to the most powerful mathematical and physical principle of maximum entropy (Jaynes, 1991; Koutsoyiannis et al. 2008; Koutsoyiannis 2014b).
- In a stochastic approach, a deviation from linearity can be conveniently incorporated through an error term, which is already included in our proposed equation (7), in order to generalize Papoulis' (1991) original equation (6).
- The fact that linearity is not regarded as a severe limitation in causality assessment is indirectly reflected in the popularity of Granger's (1969) approach, which is also linear (equation (1)).
- In the companion paper (Koutsoyiannis et al., 2022 and its Supplementary Information), we show that the linear form of the framework effectively captures the important characteristics of causality, even in cases that the true dynamics is a priori known to be nonlinear.

We further note that our proposed bivariate approach to causality could allow for the possibility of "spurious" causality, where changes in both variables are affected by another cause, possibly with different time delays and response functions. This is not a

drawback insofar as our framework of detecting necessary, rather than sufficient, conditions. But further, the inclusion of an error term in it allows for such more remote causes to be represented in the framework. Additional clarifications on multiple causes are provided in Supplementary Information (Section SI1.2),

Following the above considerations, and assuming that a least-squares solution of equation (7) has been determined for the system $(\underline{x}(t), \underline{y}(t))$, we will call that system:

1. *potentially causal* if $g(h) = 0$ for any $h < 0$, while the explained variance is non negligible;
2. *potentially anticausal* if $g(h) = 0$ for any $h > 0$, while the explained variance is non negligible (this means that the system $(\underline{y}(t), \underline{x}(t))$ is potentially causal);
3. *potentially hen-or-egg* (*HOE*) *causal* if $g(h) \neq 0$ for some $h > 0$ and some $h < 0$, while the explained variance is non negligible;
4. *noncausal* if the explained variance is negligible.

These cases are graphically illustrated in Figure 1. We note that the term "negligible" can be quantified in statistical terms, e.g. by invoking statistical significance. However, here we will treat this in a practical manner leaving the related theoretical reflections for future research. In the hypothetical case that the explained variance reaches its upper limit, i.e., 1 (100%), it may be justified to replace the term "potentially" with "classic", in accordance with the definition given above for a classic causal system. Further, we note that in some texts the term "noncausal" is used for systems which here we call "potentially HOE causal". In other texts, the potentially HOE causal systems are treated as causal systems with feedback.

In this respect, in a HOE causal system, earlier realizations of $\underline{x}(t)$ affect the current realization of $\underline{y}(t)$, but also earlier realizations of $\underline{y}(t)$ affect the current realization of $\underline{x}(t)$. Thus, each one of the processes $\underline{x}(t)$ and $\underline{y}(t)$ is correlated to both the past and the future of the other one. This may seem paradoxical in terms of a conventional way of thinking about causality, but it is not more paradoxical than the expression "hen-or-egg", first used by Plutarch (*Moralia, Quaestiones convivales*, B, Question III). Clearly, Plutarch (and subsequent users of this expression) did not mean one particular hen and one particular egg; in this case the existence or not of a causal relationship would be easy to tell. Rather, he meant the sequences of all hens and all eggs, something similar with what the abstract term "process" used here represents. For further explanation of the term "hen-or-egg" see Koutsoyiannis and Kundzewicz (2020).

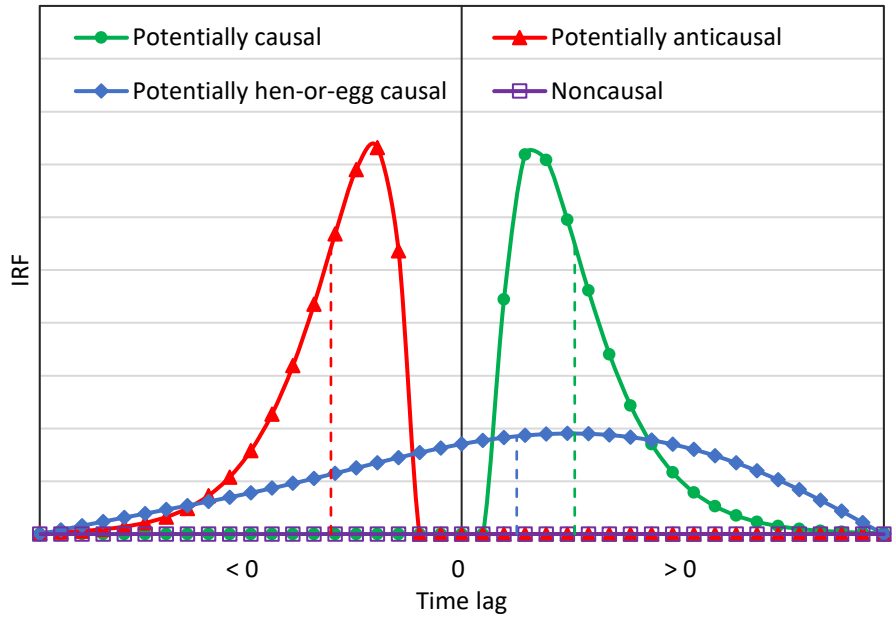

Figure 1 Explanatory sketch for the definition of the different potential causality types. For each graph the mean μ_h is also plotted with dashed line.

It is often stated that in causal systems the present of the process $\underline{y}(t)$ does not stochastically depend on the future of $\underline{x}(t)$ (i.e., $\underline{x}(t+h), h > 0$). This may be intuitive, but it is also clearly wrong: the intuition involves a confusion between causal and stochastic dependence. To see this, we define the autocovariance $c_{xx}(h)$ of process $\underline{x}(t)$,

$$c_{xx}(h) \coloneqq \mathrm{cov}\left[x(t+h), \underline{x}(t)\right] \tag{10}$$

where $\mathrm{cov}\left[\underline{a}, \underline{b}\right] \coloneqq \mathrm{E}\left[\left(\underline{a} - \mathrm{E}\left[\underline{a}\right]\right)\left(\underline{b} - \mathrm{E}\left[\underline{b}\right]\right)\right]$ denotes the covariance of any stochastic variables $\underline{a}, \underline{b}$ and $\mathrm{E}\left[\underline{a}\right]$ denotes the mean of $\underline{a}$. Likewise, we define the autocovariance $c_{yy}(h)$ of the process $\underline{y}(t)$. The cross-covariance of the two processes is

$$c_{yx}(h) \coloneqq \mathrm{cov}\left[\underline{y}(t+h), \underline{x}(t)\right] \tag{11}$$

It is easily seen that the autocovariance is an even function of the lag h, i.e., $c_{xx}(-h) = c_{xx}(h)$ and that this does not hold for the cross-covariance, $c_{yx}(h)$, which is generally an asymmetric function of h; here the symmetry appears if we change the order of the variables, i.e., $c_{xy}(h) = c_{yx}(-h)$. Furthermore, it is shown in the Supplementary Information (Section SI1.3; see also Papoulis, 1991) that the autocovariance and cross-covariance functions are related by

$$c_{yx}(h) = \int_{-\infty}^{\infty} g(a) c_{xx}(h-a) \mathrm{d}a \tag{12}$$

Now, in a potentially causal system the latter equation takes the form

$$c_{yx}(h) = \int_0^\infty g(a)c_{xx}(h-a)\mathrm{d}a \tag{13}$$

The covariance of $\underline{y}(t)$ with the future variable $\underline{x}(t+h)$, $h>0$ is $\mathrm{cov}\left[\underline{x}(t+h), \underline{y}(t)\right] = \mathrm{cov}\left[\underline{x}(t), \underline{y}(t-h)\right] = c_{yx}(-h)$ and is determined as

$$c_{yx}(-h) = \int_0^\infty g(a)c_{xx}(-h-a)\mathrm{d}a = \int_0^\infty g(a)c_{xx}(h+a)\mathrm{d}a \tag{14}$$

This is clearly nonzero, which proves that independence of the current value $\underline{y}(t)$ from the future of $\underline{x}(t)$ does not hold. There is only one trivial exception, i.e., when $c_{xx}(h+a) = 0$ for any a, which is the case only if the cause $\underline{x}(t)$ is white noise. This exception, along with the fact that the future $\underline{x}(t+h)$ does not functionally appear in equation (6), seems to have been the culprit for misleading our intuitions.

Clearly, in a potentially causal system the time order is explicitly reflected in the definition. In a potentially HOE causal system the time order needs to be clarified by defining the principal direction. This could be done in several ways, the most natural being the following:

1. The time lag $h = h_\mathrm{c}$ maximizing the (absolute value of) cross-covariance $c_{yx}(h)$, (equation (11)).
2. The mean (time average) of the function $g(h)$, defined as:

$$\mu_h \coloneqq \int_{-\infty}^\infty h\, g(h)\mathrm{d}h\,/H_0, \qquad H_0 \coloneqq \int_{-\infty}^\infty g(h)\mathrm{d}h \tag{15}$$

3. The median $h_{1/2}$ of the function $g(h)$, implicitly defined by:

$$\int_{-\infty}^{h_{1/2}} g(h)\mathrm{d}h = \frac{1}{2}H_0 \tag{16}$$

The index h_c is independent of the function $g(h)$, while the other two depend on it. The indices μ_h and $h_{1/2}$ are meaningful for $H_0 \neq 0$ and for $g(h) \geq 0$, respectively. It is reasonable to expect that, unless a system is noncausal (with $H_0 = 0$), all three variants, $h_\mathrm{c}, \mu_h, h_{1/2}$, will have the same sign, which determines a principal direction in the HOE causality. Thus, if the sign is nonnegative, the principal causality direction is $\underline{x}(t) \to \underline{y}(t)$. The principal direction is crucial for the understanding of the system studied. For example, in a system characterized as $\underline{x}(t) \to \underline{y}(t)$, if we speak about a positive feedback, we would mean that the effect of $\underline{x}(t)$ on $\underline{y}(t)$ is magnified, rather than vice-versa.

By taking expectations in equation (7), it is readily seen that,

$$\mu_y = H_0\mu_x + \mu_v \tag{17}$$

where

$$\mu_y \coloneqq \mathrm{E}\left[\underline{y}(t)\right], \qquad \mu_x \coloneqq \mathrm{E}\left[\underline{x}(t)\right], \qquad \mu_v \coloneqq \mathrm{E}\left[\underline{v}(t)\right] \tag{18}$$

Additional bulk characteristics of the IRF—more specifically, its temporal means and higher moments in relation to those of the processes $\underline{x}(t)$and $\underline{y}(t)$—are given in the Supplementary Information (section SI1.4).

3.3 Properties and desiderata for IRF

In contrast to Granger's analysis of causality (section 2.2), which treats the processes in discrete time by definition, here we treat them in continuous (i.e. natural) time, and we only convert them to discrete time for estimation purposes. If we think of the processes in natural time, we understand that a causality relationship is not an instantaneous one. In other words, if $\underline{x}(t')$ affects $\underline{y}(t)$, where $t' < t$, it is reasonable to assume that, for small h, $\underline{x}(t' \pm h)$ will also affect $\underline{y}(t)$. Therefore, the IRF, $g(h)$, is not a Dirac delta function, but one with some domain, $\mathbb{h} \subseteq \mathbb{R}$, of nonzero (and potentially infinite) measure, where $g(h) \neq 0$ for $h \in \mathbb{h}$. It is also reasonable to assume that $g(h)$ is a continuous function and has the same sign for all $h \in \mathbb{h}$. The latter can be justified as follows. If $\underline{x}(t')$ is positively correlated with $\underline{y}(t)$, then it is reasonable that $\underline{x}(t' \pm h)$ are also positively correlated with $\underline{y}(t)$. Without loss of generality, in what follows we will assume that $g(h) \geq 0$ for $h \in \mathbb{h}$ (if it were $g(h) \leq 0$, we would reflect $\underline{x}(t)$, i.e. replace it with $-\underline{x}(t)$, and hence $g(h)$ would also be reflected becoming nonnegative).

Here we clarify that the problem of identifying causality is different from that of recovering the full system dynamics. The former and not the latter, is the scope of our study. We note that, while there exist oscillatory nonlinear systems, in which the sign of $g(h)$ could alternate, we avoid subsuming them under the causality notion, particularly when causality is inferred from data in an inductive manner. This choice is consistent with Cox's (1992) conditions for causality, according to which the effect *"shows a monotone relation with 'dose'"* of the cause. Here we note that in our framework the "dose" is not regarded as an instantaneous event, but one with some time span (see details in Supplementary Information, section SI1.2).

The continuity desideratum can be quantified by defining a measure of roughness and demanding that it be restricted below a threshold. We may define such a roughness index by means of the squares of second derivative (cf. Koutsoyiannis, 2000). Specifically, we define a roughness index as

$$E \coloneqq \int_{-\infty}^{\infty} \left(g''(h)\right)^2 \mathrm{d}h \tag{19}$$

In summary the desiderata for the IRF are:

- an adequate time span $\mathbb{h}$;

- nonnegativity, $g(h) \geq 0$ for all $h \in \mathbb{h}$;
- smoothness, $E \leq E_0$, where E_0 is a real number;
- minimum variance of $\underline{v}(t)$.

3.4 Estimation of IRF

The literature offers several methods for estimating an IRF in terms of auto- and cross-correlations (Young, 2011, 2015) or their Fourier transforms, i.e., power spectra and cross-spectra (e.g. Papoulis, 1991). Here we seek a more direct method that can work with time series of observations per se, being easily understandable and reproducible by any reader using simple computational means, and can also host our desiderata for the IRF.

In dealing with observations, we first note that they are necessarily made in discrete time τ, representing the time period $(\tau - 1)D$ to τD (where τ is integer and D is the discretization time step). It is assumed that each measurement represents the time average of the process in this period (other cases are also discussed in the Supplementary Information, Section SI1.1). Thus,

$$\underline{x}_\tau \coloneqq \frac{1}{D} \int_{(\tau-1)D}^{\tau D} \underline{x}(t)\mathrm{d}t \tag{20}$$

and likewise for $\underline{y}_\tau$ and $\underline{v}_\tau$. The discrete-time version of equation (7) becomes

$$\underline{y}_\tau = \sum_{j=-\infty}^{\infty} g_j \underline{x}_{\tau-j} + \underline{v}_\tau \tag{21}$$

Specifically, using the definition of the discrete-time processes in equation (20) we show in the Supplementary Information (Section SI1.1) that equation (21) follows from (7) and that the discrete-time version of the IRF is related to the continuous-time one by

$$g_j = \frac{1}{D}\Big(G\big((j-1)D\big) - 2G(jD) + G\big((j+1)D\big)\Big) \tag{22}$$

where

$$G(b) \coloneqq \int_{-\infty}^{b} \int_{-\infty}^{a} g(h)\mathrm{d}h\,\mathrm{d}a \tag{23}$$

Second, the observation period L is finite and hence the series g_j sought should necessarily be assumed of finite length too. We thus formulate the estimation equation as

$$\hat{\underline{y}}_\tau = \sum_{j=-J}^{J} g_j \underline{x}_{\tau-j} + \mu_v \tag{24}$$

where $\hat{\underline{y}}_\tau$ is the estimate of $\underline{y}_\tau$ given the series of g_j and J is an integer chosen as $J \ll L$ and μ_v is determined from equation (17). This estimation results in an error:

$$\underline{v}_\tau \coloneqq \underline{y}_\tau - \underline{\hat{y}}_\tau = \underline{y}_\tau - \sum_{j=-J}^{J} g_j \underline{x}_{\tau-j} - \mu_v \tag{25}$$

Assuming that we have simultaneous observations of the $\underline{x}_\tau$ and $\underline{y}_\tau$ series, for a length L the estimator of the variance of $\underline{v}_\tau$ is

$$\underline{\hat{\gamma}}_v \coloneqq \frac{1}{L-2J} \sum_{\tau=J+1}^{L-J} \left(\underline{y}_\tau - \sum_{j=-J}^{J} g_j \underline{x}_{\tau-j} - \mu_v \right)^2 \tag{26}$$

and the explained variance ratio is

$$\underline{e} \coloneqq 1 - \frac{\underline{\hat{\gamma}}_v}{\underline{\hat{\gamma}}_y} \tag{27}$$

As already mentioned, the above least-squares-based determination of the ordinates g_j is not the only technique for the identification of the IRF; additional techniques can be found in Young (2011, 2015 and references therein). A well-known weakness of determining numerous ordinates is that it is an over-parameterized problem, which is typically addressed by assuming a parametric model (such as a Box-Jenkins model or an autoregressive moving average exogenous—ARMAX—model; Young, 2011, 2015). Here we prefer to use a nonparametric approach and we tackle the over-parameterization problem by imposing the roughness threshold, as discussed above. An additional parametric method, formulated in terms of parameterizing the IRF per se in continuous time is also discussed and compared to the proposed non-parametric method in the Supplementary Information of the companion paper (Koutsoyiannis et al., 2022; sections SI2.3 and SI2.4).

Now the roughness index of the IRF can be formulated by replacing the continuous second derivative with the discrete one as

$$E \coloneqq \sum_{j=-J+1}^{J-1} \left(g_{j-1} - 2g_j + g_{j+1} \right)^2 \tag{28}$$

and can also be expressed as a standardized index by

$$\varepsilon \coloneqq \frac{E}{8 \sum_{j=-J}^{J} g_j^2} \tag{29}$$

The constant 8 in the denominator has been introduced in order for the standardized roughness index ε of nonnegative g_j to have a maximum value of 1 (which becomes 2 without the nonnegativity constraint), while its least value is obviously zero. The zero value corresponds to the case where all g_j are identical. The value 1 corresponds to the case where J is large (theoretically, tends to infinity) while the IRF is saw-like with $g_{2j} = g_0(1 - (j/\lceil J/2 \rceil)^2)$ (even values of the index) and $g_{2j+1} = 0$ (odd values of the index).

It is computationally convenient to write the above equations in vector form, also replacing estimators with estimates and stochastic processes with time series of observations (cf. Koutsoyiannis, 2000). Thus, combining equations (24) and (17) we write $\underline{\hat{y}}_\tau - \mu_y = \sum_{j=-J}^{J} g_j(\underline{x}_{\tau-j} - \mu_x)$ and in vector form we will have

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{g} \tag{30}$$

where $\hat{\boldsymbol{y}} \coloneqq \left[\hat{y}_{J+1} - \mu_y, \dots, \hat{y}_{L-J} - \mu_y\right]^{\mathrm{T}}$ is the vector of estimates of $\boldsymbol{y} \coloneqq \left[y_{J+1} - \mu_y, \dots, y_{L-J} - \mu_y\right]^{\mathrm{T}}$, $\boldsymbol{g} \coloneqq \left[g_{-J}, \dots, g_0, \dots, g_J\right]^{\mathrm{T}}$ is a vector with the $2J+1$ elements of IRF, and $\boldsymbol{X}$ is a matrix with $L-2J$ rows, numbered $J+1$ to $L-J$, and $2J+1$ columns whose row numbered τ is $\underline{\boldsymbol{x}}_\tau \coloneqq \left[x_{\tau+J} - \mu_x, \dots, x_\tau - \mu_x, \dots, x_{\tau-J} - \mu_x\right]$.

Now, combining equations (25) and (17), the estimate of the variance of $\underline{v}_\tau$ is

$$\hat{\gamma}(\boldsymbol{g}) = \frac{1}{L-2J}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{g})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{g}) \tag{31}$$

while the roughness index is

$$E = \boldsymbol{g}^{\mathrm{T}}\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi}\boldsymbol{g} \tag{32}$$

where $\boldsymbol{\Psi}$ is a matrix with $2J-1$ rows and $2J+1$ columns whose ijth entry is

$$\psi_{ij} = \begin{cases} 2 & j = i+1 \\ -1 & |j-i-1| = 1 \\ 0 & \text{otherwise} \end{cases} \tag{33}$$

In this way, the estimation of IRF is formulated as the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & \hat{\gamma}(\boldsymbol{g}) = \frac{1}{L-2J}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{g})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{g}) \\ \text{subject to} \quad & \boldsymbol{g}^{\mathrm{T}}\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi}\boldsymbol{g} \le E_0 \\ & \boldsymbol{g} \ge \boldsymbol{0} \end{aligned} \tag{34}$$

If we ignore the last constraint (nonnegativity) and combine the second one (small roughness) with the objective function using a weight (multiplier) $\lambda/(L-2J)$, we obtain an unconstrained optimization problem, i.e.

$$\text{minimize } f(\boldsymbol{g}) \coloneqq (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{g})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{g}) + \lambda\boldsymbol{g}^{\mathrm{T}}\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi}\boldsymbol{g} \tag{35}$$

which has an analytical solution. Indeed, $f(\boldsymbol{g})$ has derivative

$$\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{g}} = 2\boldsymbol{g}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + 2\lambda\boldsymbol{g}^{\mathrm{T}}\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi} - 2\boldsymbol{y}^{\mathrm{T}}\boldsymbol{X} \tag{36}$$

Equating the derivative to $\boldsymbol{0}$ and solving for $\boldsymbol{g}$ we find

$$\boldsymbol{g} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{37}$$

One may notice in equation (37) that the term $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ represents autocovariance estimates of the process $\underline{\boldsymbol{x}}_\tau$ while the term $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ represents cross-covariance estimates of the

processes $\underline{\boldsymbol{x}}_\tau$ and $\underline{\boldsymbol{y}}_\tau$. By increasing the weight λ of the roughness term $\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi}$ we can decrease the roughness of $\boldsymbol{g}$ and thus with a proper choise of λ we can satisfy the roughness constraint.

If the solution in equation (37) also satisfies the nonnegativity constraint, then we have determined the IRF sought analytically. Otherwise, we have to abandon the analytical solution and solve the problem numerically, as no analytical solution is available for the nonnegativity constraint (Chen and Plemmons, 2009). However, all software platforms, including common spreadsheet software, provide solvers that can easily tackle the full optimization problem in equation (34) within seconds or minutes, depending on the time series length.

The discussion of the analytical solution helps us to understand that the autocovariance, and hence the autocorrelation, even though it does not explicitly appear in the numerical version of the optimization procedure, strongly influences the parameter estimation. It is relevant to mention that high autocorrelation results in increased estimation uncertainty and may even result in spurious causality claims. This is illustrated in the Supplementary Information of the companion paper (Koutsoyiannis, 2022, Section SI2.2) along with techniques to handle such situations and avoid false conclusions.

4 Discussion and conclusions

We have briefly examined various approaches to the notion of causality and the criteria for identifying an event as causally responsible for another. When dealing with physical quantities whose values are only known with some degree of certainty, a probabilistic approach to identifying causal links is required. However, attempts to define probabilistic necessary and sufficient conditions have all been found to have limitations. In particular, it is clear that no sufficient condition for concluding to a causal link has ever been identified.

This therefore suggests that the focus should be exclusively upon identifying necessary conditions for causation. Additionally, in view of the additional problem of validating probabilistic statements about a unique occurrence of a putative causal link, it is practically necessary to resort to considering stochastic processes rather than uniquely temporally located events.

Drawing upon Papoulis's proposal for a causal system consisting of two stochastic processes $\underline{x}(t)$ and $\underline{y}(t)$, which turns upon the existence of an impulse response function (IRF) connecting the two, we identified necessary conditions for the existence of a (linear) causal link either from $\underline{x}$ to $\underline{y}$ or $\underline{y}$ to $\underline{x}$, or of a hen-or-egg (HOE) situation in which causal influences appear to go in both directions. The distinction between these depends upon the nonexistence of nonzero weights of the linear relationship for negative (respectively positive) lags for the $x \to y$ causation (respectively the $y \to x$ causation) with HOE

otherwise. Additionally, the IRF must enable enough of the variance of the caused process to be explained for a causal link to be a possibility. Additional constraints of smoothness and nonnegativity of the IRF are included in the method.

An additional benefit of the proposed method, albeit not discussed above, is its direct applicability to the simulation of bivariate processes that exhibit time directionality and causality. Conventional stochastic models generate time symmetric processes. The problem of simulating a scalar process with time directionality has been tackled recently (Koutsoyiannis, 2019, 2020). The present framework provides direct methods to simulate time-directional vector processes with two variates, as well as hints for multivariate processes—a problem to be studied in future research.

The methodological framework proposed herein features substantial differences from existing methods, such as those discussed in section 2.2. A first difference is in its epistemological background which leads to a less ambitious objective, that of seeking necessary conditions of causality rather than sufficient ones. The usefulness of this objective lies in its ability to falsify an assumed causality and to add statistical evidence, in an inductive context, for potential causality and its direction.

A second difference is that our focus is upon maximizing not the predictability per se, but the lucidity in identifying the (potentially causal) relationship between two processes $\underline{x}_\tau$ and $\underline{y}_\tau$. This can be seen by comparing Granger's expression in equation (1) with our expression in equation (24). To estimate y_τ, the former includes terms y_i for times earlier than τ while the second does not. Such terms may increase predictability but say nothing about a potentially causal relationship between the two processes $\underline{x}_\tau$ and $\underline{y}_\tau$; rather, they may obscure that relationship, as autocorrelation is by definition symmetric in time.

Furthermore, by its construction, our framework can detect not only mono-directional (potentially) causal relationships, but also causality of HOE type. Notably, our method is formulated from the outset for the latter case, while the former case will be obtained as a result if in one of the two directions the IRF weights are zero. Further, like other methods (e.g. Granger's) our method allows testing in two directions, $x \to y$ and $y \to x$. The results in each direction are not anti-symmetrical in terms of the estimated IRFs and thus the method provides two different views of the (potentially) causal relationship, thus becoming more insightful.

A fourth difference of our method from many other methods lies in the recognition that natural time is continuous rather than discrete (nb., some methods, e.g. Liang, 2016, also use continuous time). The discrete-time relationships, which are necessary in estimation based on observations, are deduced from the continuous-time formulation, rather than taken as such from the outset. To understand the importance of this difference in foundation, consider a classic causal system in continuous time. If we considered the

system in discrete time from the outset, then the weight for lag zero would be zero to exclude synchrony (cf. Granger's expression in equation (1)). But considering continuous time, it becomes clear from equation (22) that $g_0 = \big(G(D)\big)/D \neq 0$ and this is not a violation of the axiom of time precedence.

The properties of an adequate time span of causality, instead of an instantaneous action, along with the nonnegativity and roughness (or smoothness) constraints are additional specific features of the method. Their importance derives from their enabling the true dynamics of the system $\left(\underline{x}_\tau, \underline{y}_\tau\right)$ to be revealed, in addition to just identifying the time lags—at least for systems consistent with the constraints, i.e. not those with oscillatory or excessively rough actual dynamics. This importance will become evident in the second part of this study, devoted to applications (Koutsoyiannis et al., 2022). On the negative side, the constraints certainly make the formal statistical testing more challenging. This would certainly be feasible with a Monte Carlo approach, but it is not within the scope of this paper.

Acknowledgments: The constructive comments of two anonymous reviewers helped us to substantially improve our work. We also thank the Board Member Graham Hughes for the processing of the paper and the favourable decision, as well as Keith Beven for his advice on a preliminary version of the manuscript.

Funding: This research received no external funding but was motivated by the scientific curiosity of the authors.

Conflicts of Interest: We declare no conflict of interest.

References

Birkhoff, G. D. & Lewis Jr., D. C. 1935 Stability in causal systems, *Philosophy of Science* **2**(3), 304-333, https://www.jstor.org/stable/184526 (accessed 21 August 2021).

Bode, H. W. & Shannon, C. E. 1950 A simplified derivation of linear least square smoothing and prediction theory. *Proc. of the IRE* **38**(4), 417-425.

Bunge, M. 1979 *Causality and Modern Science*. Third edition, Dover Publications.

Chen, D. & Plemmons, R. J. 2009. Nonnegativity constraints in numerical analysis. In *The Birth of Numerical Analysis* (ed. A. Bultheel & R. Cools), pp. 109-139, World Scientific, https://doi.org/10.1142/7075

Clifton, D. L. 1968. *The use of the transfer function and impulse response in acoustic scattering problems*. Defense Research Laboratory, The University of Texas at Austin, Austin, Texas, USA. https://apps.dtic.mil/sti/pdfs/ADA033260.pdf (accessed 21 August 2021).

Cohen, M. X. 2014 *Analyzing Neural Time Series Data: Theory and Practice.* Cambridge, MA, USA: MIT Press.

Cox, D.R., 1992 Causality: Some statistical aspects. *J. Roy. Stat. Soc. A* 155(2), 291-301.

Dooge, J. C. 1959 A general theory of the unit hydrograph. *J. Geoph. Res.* **64**(2), 241-256.

Drenick, R. F. 1963 Random processes in control and communications, *IEEE Trans. on Military Electronics*, MIL-**7**(4), 275-280 , doi: 10.1109/TME.1963.4323091.

Gardner, S. 1999 *Kant and the Critique of Pure Reason*. London: Routledge.

Granger, C. W. 1969 Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, **37**(3), 424-438.

Granger, C. W. 1980. Testing for causality: a personal viewpoint. *J. Econ. Dynamics and Control* **2**, 329-352.

Granger, C. & Newbold, P. 1986 *Forecasting Economic Time Series.* San Diego, CA, USA: Academic Press,.

Gujarati, D. N. & Porter, D. C. 2009 *Basic Econometrics.* 5th ed., Boston, MA, USA: McGraw Hill.

Hannart, A., Pearl, J., Otto, F. E. L., Naveau, P. & Ghil, M. 2016 Causal counterfactual theory for the attribution of weather and climate-related events. *Bull. Amer. Met. Soc.* **97**(1), 99-110.

Hendry, D. F. & Morgan, M. S. 1994 The ET interview. Professor H.O.A. Wold: 1907-1992. *Econometric Theory* **10**, 419-433, http://korora.econ.yale.edu/et/interview/wold.pdf (accessed 21 August 2021).

Hopf, E. 1934 On causality, statistics and probability, *J. of Mathematics and Physics,* 13, doi: 10.1002/sapm193413151.

Hume, D. 1748 *An enquiry concerning human understanding.*

Jaynes, E.T. 1957 Information theory and statistical mechanics. *Physical Review*, 106 (4), 620-630

Kant, I. 1787/1998 *Critique of Pure Reason.* (transl. P. Guyer & A. W. Wood), Cambridge: Cambridge University Press.

Keats, R. G. 1967 A note on the black box problem. *Appl. Prob.* **4**(1), 113-122, https://www.jstor.org/stable/3212303 (accessed 21 August 2021).

Kline, A. D. 1980 *Are there cases of simultaneous causation? PSA* Proc. Bienn. Meet. Philos. Sci. Assoc., 292–301, Philosophy of Science Association.

Kolmogorov, A. N. 1941 Interpolation und extrapolation, *Izv. Akad. Nauk SSSR Ser. Mat.* **5**, 3-14 (Republished in Kolmogorov, A. N. & Shiryayev, A. N. 1992 *Selected works of A. N. Kolmogorov*: Vol. 2, Probability theory and mathematical statistics. Kluwer Academic, pp. 272-280).

Koutsoyiannis, D. 2000 Broken line smoothing: A simple method for interpolating and smoothing data series. *Environ. Modelling and Software* **15**(2), 139–149.

Koutsoyiannis, D. 2006 On the quest for chaotic attractors in hydrological processes. *Hydrol. Sci. J.* **51**(6), 1065–1091, doi: 10.1623/hysj.51.6.1065.

Koutsoyiannis, D. 2014a Random musings on stochastics (Lorenz Lecture). *AGU 2014 Fall Meeting*, American Geophysical Union, San Francisco, USA , doi: 10.13140/RG.2.1.2852.8804.

Koutsoyiannis, D. 2014b Entropy: from thermodynamics to hydrology, *Entropy*, 16 (3), 1287–1314, doi:10.3390/e16031287.

Koutsoyiannis, D. 2019 Time's arrow in stochastic characterization and simulation of atmospheric and hydrological processes. *Hydrol. Sci. J.* **64**(9), 1013–1037, doi:10.1080/02626667.2019.1600700,.

Koutsoyiannis, D. 2020 Simple stochastic simulation of time irreversible and reversible processes. *Hydrol. Sci. J.* **65**(4), 536–551, doi:10.1080/02626667.2019.1705302.

Koutsoyiannis, D. 2021 *Stochastics of Hydroclimatic Extremes - A Cool Look at Risk.* Athens: Kallipos, ISBN: 978-618-85370-0-2, 333 pp; http://www.itia.ntua.gr/2000/.

Koutsoyiannis, D. & Dimitriadis, P. 2021 Towards generic simulation for demanding stochastic processes, Sci, 3, 34, doi: 10.3390/sci3030034.

Koutsoyiannis, D. & Kundzewicz, Z. W. 2020 Atmospheric temperature and CO_2: Hen-or-egg causality? *Sci* **2**(4), 83, doi:10.3390/sci2040083.

Koutsoyiannis, D., Onof, C. Christofides, A. & Kundzewicz, Z. W. 2022 Revisiting causality using stochastics: 2. Applications, *Proceedings of the Royal Society A*, this issue.

Koutsoyiannis, D., Yao, H. & Georgakakos, A. 2008 Medium-range flow prediction for the Nile: a comparison of stochastic and deterministic methods, *Hydrol. Sci. J.*, 53 (1), 142–164, doi:10.1623/hysj.53.1.142.

Liang, X. S. 2016 Information flow and causality as rigorous notions ab initio. *Phys. Rev. E* **94**, 052201.

Liang, X.S., 2018. Causation and information flow with respect to relative entropy. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7), 075311.

Masani, P. 1966 Wiener's contributions to generalized harmonic analysis, prediction theory and filter theory. *Bull. Amer. Math. Soc.* **72**(1), 73-125.

Mehlberg, H. M. 1983. *Time, Causality, and the Quantum Theory*. Dordrecht, Holland: Reidel,.

Nash, J.E. 1959 Systematic determination of unit hydrograph parameters. *J. Geoph. Res.* **64**(1), 111-115.

Papoulis, A. 1991 *Probability, Random Variables and Stochastic Processes*, 3rd ed.; New York, NY, USA: McGraw-Hill (1st edition 1965).

Parzen, E. 1968 *Multiple Time Series Modeling*. Technical Report 22, Department of Statistics, Stanford University, Stanford, CA, USA, https://apps.dtic.mil/sti/pdfs/AD0627023.pdf (accessed 21 August 2021).

Pearl, J. 2009 Causal inference in statistics: An overview. *Statistics Surveys* **3**, 96-146, doi: 10.1214/09-SS057.

Pearl, J., Glymour, M. & Jewell, N. P. 2016 *Causal Inference in Statistics: A Primer*. Chichester, UK: Wiley.

Post, E. J. 1963 Note on phase-amplitude relations. *Proc. of the IEEE* **51**(4), 627-627.

Robbins, H. 1959 An extension of Wiener filter theory to partly sampled systems. *IRE Transactions on Circuit Theory* **6**(4), 362-370, doi: 10.1109/TCT.1959.1086575.

Salmon, W. 1998 *Causality and Explanation*, New York: Oxford University Press.

Shannon, C. E. 1948 The mathematical theory of communication. *Bell System Technical Journal* **27**(3), 379-423.

Sharnoff, M. 1964 Validity conditions for the Kramers-Kronig relations. *Am. J. Phys.* **32**, 40-44, doi: 10.1119/1.1970070.

Sion, A. 2010 *The logic of causation: Definition, induction and deduction of deterministic causality*. Avi Sion (self-publication), http://thelogician.net/LOGIC-OF-CAUSATION/Cover-page.htm (accessed 3 July 2021).

Skyrms, B., 1980. *Causal Necessity: A Pragmatic Investigation of The Necessity of Laws*. New Haven and London: Yale University Press.

Stips, A., Macias, D., Coughlan, C., Garcia-Gorriz, E., Liang, X. S. 2016 On the causal structure between CO_2 and global temperature. *Sci. Rep.* **6**, 21691, doi:10.1038/srep21691.

Strotz, R. H, & Wold, H. O. A. 1960 Recursive vs. nonrecursive systems: an attempt at synthesis (Part I of a triptych on causal chain systems). *Econometrica* **28**(2), 417-427, https://www.jstor.org/stable/1907731 (accessed 21 August 2021).

Suppes, P. 1970 *A Probabilistic Theory of Causality.* Amsterdam, The Netherlands: North-Holland Publishing.

Verbitsky, M. Y., Mann, M. E., Steinman, B. A. & Volobuev, D. M. 2019 Detecting causality signal in instrumental measurements and climate model simulations: global warming case study. *Geosci. Model Dev.* **12**, 4053–4060, doi: 10.5194/gmd-12-4053-2019.

Wiener, N. 1948 *Cybernetics or Control and Communication in the Animal and the Machine.* Cambridge, MA, USA: MIT Press, 212 pp.

Wiener, N., 1949. *Extrapolation, Interpolation and Smoothing of Stationary Time Series, with Engineering Applications.* New York, NY, USA: Wiley.

Wold, H.O. 1938 A Study in the Analysis of Stationary Time-Series. Ph.D. Thesis, Almquist and Wicksell, Uppsala, Sweden.

Wold, H. O. A. 1948 On prediction in stationary time series. *Ann. Math. Statist.* **19**(4), 558 – 567, doi: 10.1214/aoms/1177730151.

Wold, H. O. A. 1954 Causality and econometrics. *Econometrica*, **22** (2), 162-177, https://www.jstor.org/stable/1907540 (accessed 21 August 2021).

Wold, H. O. A. 1960 A generalization of causal chain models (Part III of a triptych on causal chain systems). *Econometrica* **28**(2), 443-463, https://www.jstor.org/stable/1907733 (accessed 21 August 2021).

Young, P.C. 2011. *Recursive Estimation and Time Series Analysis*, Berlin, Heidelberg: Springer-Verlag.

Young, P.C., 2015. Refined instrumental variable estimation: Maximum likelihood optimization of a unified Box-Jenkins model. *Automatica*, **52**, 35–46.