Response to Report of the First Referee – EF11592/McCracken

(ref comments are emphasized)

We sincerely appreciate the time and care the referee took in reviewing our manuscript. We believe the changes we made to the manuscript in response to the referee's comments have led to a better overall manuscript. Our specific responses to each comment are as follows.

In this paper the authors propose a new method for exploratory causal analysis. They define two indicators, causal penchant and causal leaning, and apply them to several synthetic and real data examples.

Overall the manuscript is well written and clearly understandable. The analysis seems to be carried out carefully and the conclusions drawn are reasonable. I can recommend the manuscript for publication in PRE if the issues listed below are addressed.

Major issues:

This is already quite a comprehensive proposal sufficient for a paper on its own, although there remains a lot of work to be done. Given the already substantial length oft the current manuscript this might indeed better be left for future studies.

However, a more specific outlook would still be very helpful. The two most important issues still missing are:

- Explicit comparisons with the other four approaches mentioned, if possible on the same examples. Most convincing would of course be examples where the other methods fail but this one works fine. Currently this is only reasoned (with supporting citations).

We agree that explicit comparisons with other time series causality tools is a necessary part of exploring the usefulness of these techniques. Such comparisons, however, can be quite involved and lengthy. We are currently drafting another manuscript that precisely addresses these comparisons. Our hope was that the lagged cross-correlation comparison shown in Section VB and the brief citations to other techniques (to which the referee has alluded in their comments) was sufficient to satiate, at least temporarily, such questions in the readers' mind until the next manuscript is published. To further address the concern in this manuscript, we have expanded the discussion in Section VE about previous authors' work applying Granger causality and convergent cross-mapping techniques to that dynamical system.

The referenced additions to the manuscript are as follows,

Added to Section VE: Sugihara et al. [20] also discuss how a naive [1] application of Granger causality to the system described in Eqn. 24 may lead to conclusions that do not agree with intuition, while CCM does. The causal inference suggested by the leaning calculations of this subsection implies both CCM and leanings may be useful time series causality tools in situations where Granger causality is not. It has also been shown that CCM may fail to agree with intuition in example systems for which it has already been shown that leaning calculation do, e.g., Section VC [14].

- How would an estimate of significance look like? This very important issue is only alluded to in the Discussion.

We believe statistical tests, such has traditional t-tests, may or may not be useful for causal inference with time series causality tools. Such tests will depend on the sampling procedures

[1] Sugihara et al. [20] do not explore any of the numerous non-linear extenstions of Granger causality. The theoretical foundations of Granger causality are independent of its practical implementations, and failures of Granger causality may be failures of a specific implementation, e.g., using linear forecast model with non-linear data [8].

used to create the set of leanings, and it may not be possible to develop a single, reliable test for all situations where the leaning may be a useful causal inference tool. We have used the phrase "exploratory" causal inference throughout the manuscript to emphasize the lack of formal statistical testing (in analogy with Tukey's "exploratory data analysis"). Developing useful test statistics may rely on mathematical properties of the leaning that have not yet been fully explored and will, most likely, require another full length manuscript to explore. We also plan to discuss this issue in more detail in the future publication mentioned above. We have added text to the end of Section VI to address these concerns.

The referenced additions to the manuscript are as follows,

Added to Section VI: This example is the first for which a set of leanings has been used for causal inference, which may imply statistical testing should be used. The sample mean was used for causal inference and happened to agree with intuition for this example, but would the same conclusion be drawn using a formal hypothesis test? How should the null hypothesis and test statistic be constructed? Such questions can be subtle (see, e.g., [10]). The sampling procedure used to produce Figure 9(a) produces 2,796 defined leanings, 95% of which are below 4.2×10^{-3} and 5% of which are below -4.0×10^{-3}. A 90% confidence that the leaning falls in the interval $[-4.0 \times 10^{-3}, 4.2 \times 10^{-3}]$, however, is not a strong indication that the data supports the intuitive causal structure. The mean of the set is $\mu = 3.5 \times 10^{-3}$, and the variance is $\sigma^2 = 7.0 \times 10^{-5}$. If it is assumed that the leaning in this example is distributed as $\mathcal{N}(\mu, \sigma^2)$, then a 95% confidence interval may be $[\mu - 2\sigma, \mu + 2\sigma] = [-4.9 \times 10^{-3}, 5.6 \times 10^{-3}]$, which, again, does not strongly support the intuitive causal inference for this example. Approximately 40% of the leanings in this example are negative, which may imply that there is only a 60% confidence that this data supports the intuitive causal inference, given the tolerance domains and cause-effect assignments.

Suppose a null hypothesis is defined as $\langle \lambda_1^z \rangle = 0$. The standard error is $\text{SE} = \sigma/\sqrt{n} = 5.0 \times 10^{-4}$, from which the t-test statistic follows [3] as $t = \mu/\text{SE} = 7.08$. A two-tailed t-test (i.e., calculating the p-value with an alternative hypothesis of $\langle \lambda_1^z \rangle \neq 0$) returns a p-value of approximately zero [2], which implies the null hypothesis should be rejected in favor of the alternative at any significance level. A right-tailed t-test (i.e., the alternative hypothesis is $\langle \lambda_1^z \rangle > 0$) also returns a p-value of approximately zero. A left-tailed t-test (i.e., the alternative hypothesis is $\langle \lambda_1^z \rangle < 0$) returns a p-value of approximately one, which implies the null hypothesis cannot be rejected in favor of the alternative at any significance level. These hypothesis tests seem to imply the population mean of the sampled leanings calculated in this example is likely not zero (which implies the time series pair has some causal structure) and is likely greater than zero (which implies the causal inference made with the leaning agrees with intuition). These conclusions, however, depend on whether or not the t-test is applicable to this example. For example, the assumption that the sample mean of the leanings can be assumed to follow a normal distribution based on the central limit theorem [3] may rely on the sampled time series from which the leaning were calculated being independent and identically distributed, which may not be true. The assumptions used in these statistical tests are intended to be illustrative. Such assumptions should be explored in depth to formally develop a statistical test for causal inference using leanings. The sampling procedure used in this example may not be applicable to other data sets for which the leaning may still be a useful causal inference tool. Thus, it may not be possible that a single statistical test will be appropriate for all sets of leaning calculations.

A bootstrapping [4] procedure can be set up with the sample of leaning calculations, whereby 10^6 means are calculated from new sets (of the same size as the original set) of leanings that have been sampled (with replacement) from the original set. This procedure yields no negative means; the null hypothesis that the mean leaning value is actually negative (i.e., $\langle \lambda_1^z \rangle < 0$) can be rejected with a p-value less than 10^{-6}. The 90% confidence interval for the mean of the 10^6

[2]This calculation, and all t-test calculations discussed in the section were perform with the MATLAB function *ttest*.

bootstrapped means is $[3.48 \times 10^{-3}, 3.57 \times 10^{-3}]$, which, again, implies the mean leaning for this example is positive. A more rigorous causal inference of this data set using leanings will be explored in future work.

In the end the results of these future analyses will decide about the final usefulness of the proposed method. But even like this I feel that the present paper adds a valuable contribution to the field of causal analysis.

Minor issues:

Many Sections end with the motivation for the next Section. I think it could be helpful if these paragraphs would be moved to the beginning of these next Sections (this way each Section is self-contained).

Agreed. We have made the suggested change throughout the manuscript.

Sections IV E and F are kind of obvious. Is it really necessary to introduce the asymmetry and the weighting in Sections of their own? (but here I wouldnt insist)

We agree that the weighted mean observed leaning discussion fits naturally in the mean observed leaning discussion. We also agree that the cause-effect independence discussion does not flow naturally with the rest of the text, but we feel as though the discussion cannot be excluded and does not easily fit into any other section. So, we have left it as its own section.

Section IV. A. Maybe refer to Eq. 5 again

The reference has been added to the beginning of Section IV(A).

Page 3: assumed to be the same length (is this not a constraint?)

This is a constraint of the algorithm used to estimate the probabilities used in the leaning calculation. It is not, however, a constraint on the penchant and leaning calculations themselves. Those calculations only require the conditional probabilities for the cause-effect pairs. We have estimated those probabilities in this manuscript through a simple binning procedure. However, one could imagine an algorithm that estimates the probabilities from the data without requiring the time series lengths to be equal, e.g., by assuming each point in the time series is generated by a known distribution with unknown parameters that must be estimated from the data. We have removed the referenced parenthetical from the manuscript to avoid confusion.

Page 5 (and many other times before and after): t is supposed to be integer, so why not use t = 0,1,,9 which would imply this strongly. In example V.B when t is a multiple of f you use already a similar notation. The current notation is not very precise.

Agreed. The change has been made throughout the manuscript.

Page 9: will be labelled I_2^{peak} Where?

This is a typo left over from an earlier draft that had all the peak values written out explicitly. The phrase has been removed in the manuscript.

Page 11, Fig. 6: Maybe write the expected signs explicitly. Could help the reader. Subplot captions should repeat case number as well.

The requested text has been added to Figure 6.

Page 13, Fig. Caption 9: Why is the B_Z treatment necessary? Please provide background.

This treatment of B_z follows from early observations of the solar wind magnetic field in conjuction with measrements of the magnetic field on the surface of Earth. It was found that when B_z in the solar wind was southward for an extended period of time, the magnetic field measured on the surface of Earth became disturbed. When the magnetic field in the solar wind was the same magnitude but northward for the same period of time, the magnetic field measured on the surface of Earth showed very little response. A very crude model of the connection between the solar wind magnetic field and the magnetic field measured on the surface of Earth is that the magnetosphere acts as a low pass filter of the rectified B_z in the solar wind. With such a model, one can predict about 20% of the variance of the magnetic field measured on Earth's surface [http://www-ssc.igpp.ucla.edu/personnel/russell/papers/rectify/]. The physical explanation for why only southward B_z affects the magnetosphere was given by J.W. Dungey, Phys. Rev. Lett. 6, 47, (1961). A citation has been added to the manuscript to help provide background for the reader.

Always mention the system (equation number) in the Figure and Table Caption. Sometimes this is done already but sometimes not (e.g. Table I). Captions should be more self-contained.

All of the figure captions have been edited to include equation references.

Many even more minor issues:
Page 3: and defining Incomplete sentence.

Fixed.

Page 5: of of −− > of

Fixed.

Page 5: following sections below –¿ the following sections

Fixed.

Page 6: empirical and synthetic –¿ synthetic and empirical

Fixed.

Page 6, Fig. 1: Axis labels slightly overlap with tick labels. Shift axis label a bit to the right. This would also help in some other Figures (e.g. Fig. 4).

There appear to be a few issues with the plot fonts and label spacing. Our plan is to address these issues while the referees are reviewing our responses to the other comments.

Page 7: Introduce the abbreviation CCM (Convergent cross-mapping).

The first use of the term "cross convergent mapping" has been changed to "cross convergent mapping (CCM)".

Page 8, Fig. 3: Maybe move legend to the right of the plot (vertically stacked).

See response to the other tick label comment above.

Page 8: I(t), in equation and in the text below

Fixed.

Page 11: Why not case 1,2,3?

We agree that numbering the cases from zero is a little awkward. We have changed the case numbering throughout.

Page 11: unable identify –¿ unable to identify

Fixed.

Page 12: the normalized the time series –¿ the normalized time series

Fixed.

Page 12: The time series pair is shown in Fig. 7. Not really necessary anymore.

We agree the sentence is redundant. It has been removed.

Page 13: The starting points for each sampled time series are sampled −− > The starting points for each time series are sampled (should still be clear)

We have removed the redundant "sampled".

Page 14: may lead to causal –¿ may lead to spurious causal (?)

We agree that the additional "spurious" is helpful to the reader and have made the change.

Page 14: as it as −− > as it has

Fixed.

Page 15, Ref. 22: will used −− > will use

Fixed.

Response to Report of the Second Referee – EF11592/McCracken

(ref comments are emphasized)

We appreciate the time the referee took in reviewing our manuscript and writing their report. It seems as though some of the deliberate decisions we made during the drafting of our manuscript (e.g., using only simple, straightforward examples) may not have gone far enough to show the efficacy of our results to the referee, so we hope the subsequent changes we have made to the manuscript make our ideas clear. Our specific responses to each comment are as follows. We also note that we considered many of the additions suggested by the reviewer but decided that they made the manuscript too long, diffuse, or dense.

The paper introduces a simple causality measure and some numerical and real examples documenting for its ability to identify the expected causal effect. The simple form of the measure based on probabilities easily estimated by relative frequencies of scalar and two dimensional variables is appealing. However, the reasoning of the proposed measure is not clear and the examples are very simple, and the same holds for the real world examples.

Causality, and in particular Granger causality in terms of time series, is a timely topic that has gained much attention in the recent years. Given the large number of proposed Granger causality measures, a new measure should have good theoretical and practical grounds and beat other simpler measures at least in some particular cases of practical interest. This presentation is short in these aspects, as explained below.

1. The presentation of the measure is in terms of arbitrary cause and effect quantities, called C and E, and it is not clear how these relate to observed variables and time series. Even when this is shown in Sec.IV the assignment of C and E to e.g. x(t-1) and y(t) seems arbitrary.

The initial derivation of the penchant in Section II and the leaning in Section III is intentionally left in terms of the vague quantities of "cause" C and "effect" E. This decision comes directly from the probabilistic definition of causality as $P(E|C) > P(E|\bar{C})$ [21] (or, e.g., [9]), which was meant to convey a philosophical idea about the relationship between an object capable of generally being called a cause, i.e., C, and the associated object capable of generally being called an effect, i.e., E. Granger actually introduced his causality measure in a similar fashion using the same probabilistic causality definition [8]. The expression for the penchant derived in Section II has been used elsewhere in the literature (see footnote 23 of our manuscript and, e.g., the *causal significance* of Kleinberg et al. [13]). Our use of the expression differs mainly in our introduction of the leaning and our interpretation of how to make the terms C and E "operational" (as Granger puts it [8]). Our introduction of one possible interpretation of C and E in the first paragraph of Section IV is indeed arbitrary. We hoped this point was implied by the language we used to introduce the first cause-effect assignment (i.e., "it follows that a natural assignment may be" rather than, e.g., "the correct casue-effect assignment is"), as well as our discussion of this matter at the end of Section III and the introduction of other possible cause-effect assignments in Section IVF and VF. We believe the definition of the penchant and leaning in terms of arbitrary cause-effect assignments is a strength of the tool, allowing its application to wide variety of problems (including, possibly, those outside of physics). This motivation is similar to Granger's original arbitrary formulation of causality [8], which has allowed his ideas to be extended beyond the linear models he originally used to make the definition 'operational" (see, e.g., [1]). We have added text to end of Section III to help better explain this idea.

The referenced additions to the manuscript are as follows,

Added to Section III: The leaning is a function of four probabilities, $P(C)$, $P(E)$, $P(C|E)$, and $P(E|C)$. The usefulness of the leaning for causal inference will depend on an effective method for estimating these probabilities from times series and a more specific definition of

the cause-effect assignment within the time series pair. An operational definition of C and E will need to be drawn directly from the time series data if the leaning is to be useful for causal inference. Such assignments, however, may be difficult to develop and may be considered arbitrary without some underlying theoretical support. For example, if the cause is x_{t-1} and the effect is y_t, then it may be considered unreasonable to provide a causal interpretation of the leaning without theoretical support that $\mathbf{X}$ may be expected to drive $\mathbf{Y}$ on the time scale of $\Delta t = 1$. This issue is, however, precisely one of the reasons for divorcing the causal inference proposed in this work (i.e., exploratory causal inference) from traditional ideas of causality, as was explained in the second paragraph of the introduction. Statistical tools are associational, and cannot be given formal causal interpretation without the use of assumptions and outside theories (see [9] for an in-depth discussion of these ideas). In practice, many different potential cause-effect assignments may be used to calculate different leanings, which may then be compared as part of the causal analysis of the data.

In this article, the probabilities required for the leaning calculation will be estimated from the data straightforwardly through a counting/binning procedure, and the cause-effect assignment may be varied but will always use a simple lag structure to avoid unnecessarily complex computations.

2. The causal penchant is an ambiguous measure of causality. It can be negative and still E can depend on C, i.e. $P(E|C)$ may be smaller than $P(E|\bar{C})$ but still C may be a cause for E. This is particularly true if E can have a number of causes, so that C is only a fraction, but a significant one, of the set of causal factors to E. Even for bivariate time series, in which case $\bar{C}$ contains only the past of the response variable Y (and E is the present response), the past response may have more effect on the present response than the past of the driving variable X (which is named C), but still a causal effect from X to Y exists.

We agree that the penchant is ambiguous. We discuss this issue in the last two paragraphs of Section II and the second paragraph of Section III. We introduce the leaning as the tool for causal inference because of the ambiguity associated with the penchant definition (although many of the philosophical issues with the penchant still exist with the leaning). In this comment the referee has touched on one of the many arguments against interpreting expressions such as the penchant, leaning, or any time series causality tool as "true" causality (see, e.g., [15] and [9] for in-depth discussions of the philosophical shortcomings of these types of expressions). As we state at the end of Section II, "we emphasize, however, that we use terms such as cause, effect, causal inference, and related terms to specifically refer to the penchant and leaning quantities.". We have added text to the end of Section II to make these ideas more clear.

The referenced additions to the manuscript are as follows,

Added to Section II: In this article, we seek to determine if the penchant is a useful quantity for the identification of causality relationships between time series. Our goal is identify the usefulness of the penchant (and leaning introduced below) for *exploratory causal inference*, i.e., inference intended to determine if (and what) causal structure may be present in a time series pair but not to *prove* or *confirm* such structure. There are scenarios in which any time series causality tool such as the penchant may incorrectly assign causal structure or may incorrectly not assign causal structure [9]. Furthermore, proof of causal relationships is often considered impossible with data alone [15, 9, 10]. The goal of this work is to draw as much information as possible from the given data to, e.g., guide the design of future experiments.

3. The causal leaning $l(EC) = r(EC) - r(CE)$ is not well defined as the probabilities $p(E|C)$ and $p(C|E)$ (and the same for the supplemental conditional arguments) are not comparable. For time series, this means that the inherent dynamics of X and Y may be different, so the effect of the past of the response may be different at each case ($X- > Y$ and $Y- > X$). All but

one examples in the paper do not include inherent dynamics, which is standard in time series problems. In particular, only the example in Sec V.C involves inherent dynamics, and the results there indicate exactly this problem (see Fig.5b-d) as the authors acknowledge (and do not explain) in the text (paragraph at the end of column 1 of page 10). For the effect of inherent dynamics see, e.g. Papana A, Kugiumtzis D, Larsson PG (2011) "Reducing the Bias of Causality Measures", Physical Review E, 83, 036207; Palus M (2014) "Cross-Scale Interactions and Information Transfer", Entropy 16(10), 5263-5289.

The lack of synthetic data examples with coupled dynamics (beyond that of Section VE) was intentional. Such systems are often difficult to interpret causally (as discussed in the third paragraph of Section VE) and, we believe, would make interpreting the efficacy of the leaning for causal inference unnecessarily confusing. We tried to deliberately use systems for which the causal structure was strongly intuitive (e.g., in Section VA, intuitively $\mathbf{X}$ clearly drives $\mathbf{Y}$). Without such intuitively simple examples, the "correctness" of the causal inference derived from the leaning may be suspect. We have shown previously that relying only on coupled systems to test a time series causality tool may lead to such tools failing for the intuitive cases [14]. We also point out that studying coupled systems such as Henon maps and the coupled logistic map presented in Section VE is considered standard in nonlinear time series analysis (see, e.g., [12]), but many of the simpler examples, e.g., Section VB, presented in the manuscript may be found in standard time series analysis textbooks such as [2]. A more in-depth presentation of the example shown in Section VE will be shown in a future manuscript, but we believe the addition of any other examples to the current manuscript would take away from the core ideas.

Mathematically, $P(E|C)$ and $P(C|E)$ are both well defined probabilities and thus comparable. We agree that only one of the examples includes "inherent dynamics" as only one of the synthetic data examples includes coupling between $\mathbf{X}$ and $\mathbf{Y}$ (that example is Section VE, not Section VC, which may be a typo as Section VE does include the referenced paragraph on page 10 and Fig. 5). Both of the referee's references use only examples involving coupling (Rossler systems in the Palus paper and Mackey-Glass, Henon, and generic coupled nonlinear systems in the Papana et al. paper). The point that $\mathbf{X}$ may drive $\mathbf{Y}$ and vice versa is part of our motivation for introducing the leaning and is discussed in the first two paragraphs of Section III. The leaning is meant to quantify which times series may be seen as the stronger driver. The leaning is not intended to quantify "how much" $\mathbf{X}$ drives $\mathbf{Y}$ versus "how much" $\mathbf{Y}$ drives $\mathbf{X}$.

4. The definition of causal leaning l(EC)=r(EC)-r(CE) is confusing. r(CE) is not defined in the same way as r(EC) by interchanging C to E and vice versa. Rather C is always the past of the driving variable and E the presence of the response. Moreover, the past of the driving variable is represented by a single lag, which renders the probability estimation possible using binning, but reduces the power of the measure (it evaluates the effect of a single lagged variable and not collectively as other measures do).

We believe this referee comment is referring to the fact that if the leaning is defined, e.g., with the cause-effect assignment $\{C, E\} = \{x_{t-1}, y_t\}$, then the two penchants in the leaning are defined as $\rho_{EC} = \rho_{y_t, x_{t-1}}$ (i.e., with $\{C, E\} = \{x_{t-1}, y_t\}$) and $\rho_{CE} = \rho_{y_{t-1}, x_t}$ (i.e., with $\{C, E\} = \{y_{t-1}, x_t\}$). This is the example first presented in Section IV. The referee seems to be implying that the leaning under this cause-effect assignment should be defined with $\rho_{EC} = \rho_{y_t, x_{t-1}}$ (i.e., with $\{C, E\} = \{x_{t-1}, y_t\}$) and $\rho_{CE} = \rho_{y_t, x_{t-1}}$ (i.e., with $\{C, E\} = \{y_t, x_{t-1}\}$). While such a definition of the leaning may be mathematically symmetric, it violates the underlying assumption that a cause must precede the effect (as discussed in Section IVB of the manuscript). The cause-effect assignment is, rather, symmetric in the structure of the assumed cause and effect. For example, the example cause-effect assignment of $\{C, E\} = \{x_{t-1}, y_t\}$ is identifying the structure of a single time step lag as the "cause" and the current time step as the "effect".

The leaning applies this structure to both $\mathbf{X}$ and $\mathbf{Y}$. We believe this is the only way in which the cause-assignment can be considered symmetric because any direct interchange of the mathematical definition of the cause and effect will lead to effects preceding causes in the data, which violates the initial assumption of such a thing not being possible and makes causal interpretations difficult. Text has been added to the beginning of Section IV to discuss these ideas. Regarding the past being represented by a single lag, we refer to our discussion of the referee comment 2 and emphasize that the cause-effect assignments used in the manuscript were intentionally simple, as our main goal was to determine the efficacy of the leaning as a causal inference tool. The text added to the end of Section III discusses this issue (see the response to referee comment 1).

The referenced additions to the manuscript are as follows,

Added to Section IV: The cause-effect assignment is an assignment of a given structure or feature of the data in one time series as the "cause" and another structure or feature of the data in the other time series as the "effect". For example, in the l-standard cause-effect assignment, the cause is the lag l time step in one time series and the effect is the current time step in the other. The leaning compares the symmetric application of these cause-effect definitions to the time series pair. So, for the above example of $\{C, E\} = \{x_{t-l}, y_t\}$, the first penchant will be calculated using $\{C, E\} = \{x_{t-l}, y_t\}$ and the second will be calculated using $\{C, E\} = \{y_{t-l}, x_t\}$. The second penchant is not the direct interchange of $C \Leftrightarrow E$ from the first penchant because such an interchange would violate the assumption that a cause must precede an effect. For example, if the first penchant in the leaning calculation is calculated using $\{C, E\} = \{x_{t-l}, y_t\}$, then the second penchant is not calculated using $\{C, E\} = \{y_t, x_{t-l}\}$ because the definition of the effect, x_{t-l}, precedes the definition of the cause, y_t.

5. The so-called tolerance domain is actually the bin width. This is not clear in the presentation. Moreover, the literature about the selection of bin width is ignored and the authors instead present three methods to estimate it with relation to a particular numerical example (page 7).

The leaning calculations require various probabilities to be estimated from the data (as discussed at the end of Section III). We use counting/binning to make these estimations in our examples, and we agree with the referee that the tolerance domains may be thought of as bin widths. We also agree that it was an oversight not to mention the existing literature on bin width selection. However, we decided to use "tolerance domain" rather than "bin width" because we believe that the language better represents the problem as it specifically relates to the leaning calculations. The tolerance domains are the bin widths of what the analyst is willing to consider a "cause" and "effect" in the data; i.e., the tolerance domains can be interpreted as the set of values that an analyst is willing to consider equivalent causes or effects. For example, if the lag l time step of $\mathbf{X}$, i.e., x_{t-l}, is the assumed cause of some assumed effect (e.g., the current time step of $\mathbf{Y}$, y_t), then an x-tolerance domain of $[x_{t-1} - a, x_{t-1} + b]$ may be thought of as an analyst's willingness to consider all values that fall within that domain equivalently as the assumed cause of that assumed effect. There is no requirement that such domains be symmetric or equal for both time series. As such, the problem of setting tolerance domains is different from typical bin width selection problems, which are usually concerned with finding the optimal bin width to represent an empirical distribution of a single data set (see, e.g., [23]). We have added text to the end of Section IVF to discuss this issue.

The referenced additions to the manuscript are as follows,

Added to Section IVF: As mentioned in Section III, the probabilities required for the leaning calculations are estimated in this example (and all the following ones) through straightforward counting of the data. As such, the tolerance domains may be thought of as bin widths for the probability estimations. Work has been done on optimal and data-driven bin width selection procedures (see, e.g., [23]), usually in the context of finding histograms. Tolerance

domains, however, may be thought of in terms of the causal inference for which the leaning is intended. The tolerance domain for the "cause" (or "effect") is the domain in which an analyst considers the data may still reasonably be identified as a "cause" (or "effect"). It is not required to be symmetric, and the tolerance domain for one time series is not required to be equal to the tolerance domain for the other (which is seen in the example above and most of the examples that follow).

6. The selected examples are simple and non-standard for time series analysis. As mentioned above inherent dynamics are missing and all examples are low-dimensional, involving only one time lag for the causal effect. These are far from realistic.

As discussed in our response to referee comment 3, the examples presented in this manuscript are intentionally simple. The main goal was to show that the leaning leads to causal inferences that agree with intuition, which requires examples for which the causal intuition is straightforward. We have discussed the lack of examples with "inherent dynamics" in our response to referee comment 3. We agree with the referee that testing the leaning on realistic systems is important. We believe we addressed such concerns in Section VC (in which the synthetic data was generated from the well-known dynamics of an RL circuit), Section VE (which has been in previous time series causality studies in part because it is a good approximation to population dynamics [20]), and in Section VI (which used empirical data sets for which a "causal truth" was known from outside theory). While the argument that the examples are "non-standard" may apply to some of the example systems we used, we do not believe the "standard-ness" of an example system has any impact of the efficacy of the leaning as a causal inference tool, so we did not select our examples based on such criteria. We agree, however, that the application of penchants and leanings to "standard" nonlinear and chaotic dynamics, such as Henon maps, typically found in nonlinear time series analysis textbooks (e.g., [12]) would be both interesting and an important test of the limitations of causal inference with such tools. We believe such work is best left for future manuscripts.

7. The comparison to other Granger causality measures is very limited. A fair comparison would include some representative measures, e.g. a simple linear measure, such as the lagged cross-correlation and a nonlinear one, such as the lagged cross-mutual information, and possibly other linear and nonlinear measures (say VAR based Granger causality index and transfer entropy). The comparison should take place over all simulated examples, and not at selected cases as done here.

We agree that the comparison to Granger causality is limited. We believe explicit comparisons with other time series causality tools (not only Granger causality) is a necessary part of exploring the practicality of the leaning. Such comparisons, however, can be quite involved and lengthy, and we believe adding such material to the current manuscript might confuse the main goal of the work (which was the introduction and efficacy discussion of the penchant and leaning calculations). We are currently drafting another manuscript that precisely addresses the comparison of the leaning to several other time series causality tools (across many of the examples listed in this manuscript). Our hope was that the lagged cross-correlation comparison shown in Section VB and the brief citations to other techniques covered the topic sufficiently until the next manuscript is published. To address the concern in this manuscript, we have expanded the discussion in Section VE about previous authors' work applying Granger causality and convergent cross-mapping techniques to that dynamical system.

The referenced additions to the manuscript are as follows,

Added to Section VE: Sugihara et al. [20] also discuss how a naive [3] application of

[3]Sugihara et al. [20] do not explore any of the numerous non-linear extenstions of Granger causality. The

Granger causality to the system described in Eqn. 24 may lead to conclusions that do not agree with intuition, while CCM does. The causal inference suggested by the leaning calculations of this subsection implies both CCM and leanings may be useful time series causality tools in situations where Granger causality is not. It has also been shown that CCM may fail to agree with intuition in example systems for which it has already been shown that leaning calculation do, e.g., Section VC [14].

8. The two real examples are not presented in full. The temperature and snow fall example is shown only with respect to the lag parameter (l) for selected tolerance levels. The same holds for the OMNI time series. Moreover, the time series for temperature and snow fall are very large compared to the lengths used in the simulated examples. On the other hand, for the OMNI set the results on the whole time series are not given, but instead the large time series is segmented to many small time series giving a set of leaning values that spread around zero. However, the slightly positive mean is used to document the correct answer.

We agree that the empirical data examples of Section VI are not presented in full. We discussed this issue in paragraphs 4, 6, and 13 of that section, where, e.g., we state that the analysis presented in the manuscript was not intended to be thorough. Instead, this section was intended to show that the leaning could produce results that agree with the causal "truths" using empirical data sets. Text has been added to the beginning of Section VI to clarify this point.

The referenced additions to the manuscript are as follows,

Added to Section VI: The examples shown in this section are intended to demonstrate that causal inference using leaning calculations can agree with the causal "truth" in empirical data sets. The analysis shown here is not expected to illustrate how the leaning may be used for exploratory causal analysis of empirical data for which there is no causal "truth". Such analysis is expected to be more complicated than that which is shown here (e.g., involving multiple tolerance domain calculations and the comparison of different cause-effect assignments).

9. The latter point above makes clear that a significance test for the leaning measure is needed. The authors leave this out for future work as they comment in the discussion section. This is peculiar and in my opinion it shows that the authors have limited overview of the literature on Granger causality. Simply, a new measure cannot stand without a significance test.

We believe statistical tests, such has traditional t-tests, may or may not be useful for causal inference with time series causality tools. Such tests will depend on the sampling procedures used to create the set of leanings, and it may not be possible to develop a single, reliable test for all situations where the leaning may be a useful causal inference tool. We also plan to discuss this issue in more detail in the future publication mentioned above. We have added text to the end of Section VI to address these concerns. We would also like to briefly disagree with the statement that a new time series causality tool "cannot stand without a significance test." Granger's early work on causality was concerned with developing a "testable" definition of causality but did not, as far as we can tell, involve the construction of formal hypothesis tests with respect to his causality tool (see, e.g., [6]). Granger published a paper in 1963 that discussed the use of hypothesis testing for causality analysis, but the test statistic (the development of which he credits to Whittle) differs from those used in modern Granger causality hypothesis testing [7]. Hypothesis testing was, however, a standard part of Granger causality studies by

theoretical foundations of Granger causality are independent of its practical implementations, and failures of Granger causality may be failures of a specific implementation, e.g., using linear forecast model with non-linear data [8].

the late 1970s (see, e.g., [16]). The same time period also saw many authors argue against the use of such testing in causality studies (see, e.g., [19, 11]), and we agree, at least in part, with many of the issues those authors raised (e.g., concern about an analyst's over-confidence in the results of such tests). Many (most?) modern time series causality tools are introduced (and frequently used) without any kind of hypothesis testing or discussion of statistical significance, including transfer entropy [18], convergent cross mapping [20], and lagged cross-correlation (see, e.g., [17, 5]). Granger causality actually seems to be unique in its amenability to formal statistical tests. It, and its derivatives, are the only major time series causality tools that use forecast models for causal inference. Such techniques, including linear regression and VAR modeling, readily lend themselves to existing test statistics such as those used in Wald tests (see, e.g., [22]). We believe developing test statistics for the penchant and leanings should be done rigorously with special care taken to ensure there are no underlying statistical assumptions (such as obeying the central limit theorem) that are violated by either the leaning calculation itself or the sampling procedure used to create the sample of leanings upon which the test is being conducted. We do not believe such care can be taken in the current manuscript without increasing the length significantly. The text added to the end of Section VI discusses these issues. We also posit that formal statistical testing should not be a part of the "exploratory causal analysis" that is being discussed in this work. This idea will be explored further in the future manuscript discussed above.

The referenced additions to the manuscript are as follows,

Added to Section VI: This example is the first for which a set of leanings has been used for causal inference, which may imply statistical testing should be used. The sample mean was used for causal inference and happened to agree with intuition for this example, but would the same conclusion be drawn using a formal hypothesis test? How should the null hypothesis and test statistic be constructed? Such questions can be subtle (see, e.g., [10]). The sampling procedure used to produce Figure 9(a) produces 2,796 defined leanings, 95% of which are below 4.2×10^{-3} and 5% of which are below -4.0×10^{-3}. A 90% confidence that the leaning falls in the interval $[-4.0 \times 10^{-3}, 4.2 \times 10^{-3}]$, however, is not a strong indication that the data supports the intuitive causal structure. The mean of the set is $\mu = 3.5 \times 10^{-3}$, and the variance is $\sigma^2 = 7.0 \times 10^{-5}$. If it is assumed that the leaning in this example is distributed as $\mathcal{N}(\mu, \sigma^2)$, then a 95% confidence interval may be $[\mu - 2\sigma, \mu + 2\sigma] = [-4.9 \times 10^{-3}, 5.6 \times 10^{-3}]$, which, again, does not strongly support the intuitive causal inference for this example. Approximately 40% of the leanings in this example are negative, which may imply that there is only a 60% confidence that this data supports the intuitive causal inference, given the tolerance domains and cause-effect assignments.

Suppose a null hypothesis is defined as $\langle \lambda_1^z \rangle = 0$. The standard error is $\text{SE} = \sigma/\sqrt{n} = 5.0 \times 10^{-4}$, from which the t-test statistic follows [3] as $t = \mu/\text{SE} = 7.08$. A two-tailed t-test (i.e., calculating the p-value with an alternative hypothesis of $\langle \lambda_1^z \rangle \neq 0$) returns a p-value of approximately zero [4], which implies the null hypothesis should be rejected in favor of the alternative at any significance level. A right-tailed t-test (i.e., the alternative hypothesis is $\langle \lambda_1^z \rangle > 0$) also returns a p-value of approximately zero. A left-tailed t-test (i.e., the alternative hypothesis is $\langle \lambda_1^z \rangle < 0$) returns a p-value of approximately one, which implies the null hypothesis cannot be rejected in favor of the alternative at any significance level. These hypothesis tests seem to imply the population mean of the sampled leanings calculated in this example is likely not zero (which implies the time series pair has some causal structure) and is likely greater than zero (which implies the causal inference made with the leaning agrees with intuition). These conclusions, however, depend on whether or not the t-test is applicable to this example. For example, the assumption that the sample mean of the leanings can be assumed to follow a normal distribution based on the central limit theorem [3] may rely on the sampled time series

[4] This calculation, and all t-test calculations discussed in the section were perform with the MATLAB function *ttest*.

from which the leaning were calculated being independent and identically distributed, which may not be true. The assumptions used in these statistical tests are intended to be illustrative. Such assumptions should be explored in depth to formally develop a statistical test for causal inference using leanings. The sampling procedure used in this example may not be applicable to other data sets for which the leaning may still be a useful causal inference tool. Thus, it may not be possible that a single statistical test will be appropriate for all sets of leaning calculations.

A bootstrapping [4] procedure can be set up with the sample of leaning calculations, whereby 10^6 means are calculated from new sets (of the same size as the original set) of leanings that have been sampled (with replacement) from the original set. This procedure yields no negative means; the null hypothesis that the mean leaning value is actually negative (i.e., $\langle \lambda_1^z \rangle < 0$) can be rejected with a p-value less than 10^{-6}. The 90% confidence interval for the mean of the 10^6 bootstrapped means is $[3.48 \times 10^{-3}, 3.57 \times 10^{-3}]$, which, again, implies the mean leaning for this example is positive. A more rigorous causal inference of this data set using leanings will be explored in future work.

References

[1] Nicola Ancona, Daniele Marinazzo, and Sebastiano Stramaglia. Radial basis function approach to nonlinear granger causality of time series. *Phys. Rev. E*, 70:056221, Nov 2004.

[2] G.E.P. Box, G.M. Jenkins, and G.C. Reinsel. *Time Series Analysis: Forecasting and Control*. Wiley Series in Probability and Statistics. Wiley, 2013.

[3] G. Casella and R.L. Berger. *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning, 2002.

[4] B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1994.

[5] Mahmoud El-Gohary and James McNames. Establishing causality with whitened cross-correlation analysis. *Biomedical Engineering, IEEE Transactions on*, 54(12):2214–2222, 2007.

[6] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):pp. 424–438, 1969.

[7] C.W.J. Granger. Economic processes involving feedback. *Information and Control*, 6(1):28 – 48, 1963.

[8] C.W.J. Granger. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2(0):329 – 352, 1980.

[9] P. Illari and F. Russo. *Causality: Philosophical Theory Meets Scientific Practice*. Oxford University Press, 2014.

[10] G.W. Imbens and D.B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

[11] Rodney L. Jacobs, Edward E. Leamer, and Michael P. Ward. Difficulties with testing for causation*. *Economic Inquiry*, 17(3):401–413, 1979.

[12] H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*. Cambridge nonlinear science series. Cambridge University Press, 2004.

[13] S. Kleinberg. *Causality, Probability, and Time*. Causality, Probability, and Time. Cambridge University Press, 2012.

[14] James M. McCracken and Robert S. Weigel. Convergent cross-mapping and pairwise asymmetric inference. *Phys. Rev. E*, 90:062903, Dec 2014.

[15] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[16] David A. Pierce and Larry D. Haugh. Causality in temporal systems. *Journal of Econometrics*, 5(3):265 – 293, 1977.

[17] David Rogosa. A critique of cross-lagged correlation. *Psychological Bulletin*, 88(2):245, 1980.

[18] Thomas Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85:461–464, Jul 2000.

[19] G. William Schwert. Tests of causality. *Carnegie-Rochester Conference Series on Public Policy*, 10:55 – 96, 1979.

[20] George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. Detecting causality in complex ecosystems. *Science*, 338(6106):496–500, 2012.

[21] Patrick Suppes. *A Probablistic Theory of Causality*. North Holland Publishing Company, 1970.

[22] Hiro Y. Toda and Peter C. B. Phillips. Vector autoregression and causality: a theoretical overview and simulation study. *Econometric Reviews*, 13(2):259–285, 1994.

[23] M. P. Wand. Data-based choice of histogram bin width. *The American Statistician*, 51(1):59–64, 1997.