Reviewer 2:

The paper is sort of a continuation of the methodological work by Liang about information flow. Is starts with the concept of causal sensitivity, based on partial derivatives of an effect variable Y as dependent on a causal variable X, time, and other variables, and also potentially containing noise.

Response: *First of all, we would like to thank the reviewer for going through the manuscript in detail and for the useful questions and comments raised. We have revised the manuscript in response to these. Please refer to the comments below.*

Full comprehension of the framework is not warranted without referring to Liang's papers, and the reader is left alone why the ratio of a partial derivative of Y to the total time derivative of X should bear the word "causal". Also, notation-wise, the total derivative of Y with respect to time would be a sum of several partial derivatives of Y wrt all variables (X1, X2, ...) times the total derivative of them:
dY/dt = partial Y / partial X1 * dX1/dt + partial Y / partial X2 + ... + partial Y / partial t
so it one wants to single out the part which is due to X1 (say), eq. (1) in line 81 would simply be
nCS = | partial Y / partial X1 | - or otherwise, the reviewer doesn't understand the notation \partial Y(X)) / partial t
but the main problem is why this would be called "causal" sensitivity?

Response: *There seems to have two levels of confusion here. Firstly, the "n" in nCS refers to normalization. The question why it is called "causal" sensitivity does not involve normalization. To prevent such confusion and for better clarity, we have split the equation of nCS into two in the revised manuscript, one for defining CS, and another for normalization. Secondly, the word "causal" is used, simply because X serves as the cause variable and Y serves as the effect variable, and they are associated through certain causal functions. In the revised manuscript, we use an example of a simplified linear causal function $Y = mX + n$, as well as another example of methane-climate feedback sensitivity, to illustrate the "causal sensitivity".*

Regarding the notation, yes, our defined CS appears to be the same as the $\partial Y/\partial X$ in the reviewer's given expression. However, if we are not mistaken, the use of a partial derivative usually implies cause variables (e.g. the X1, X2 ... in reviewer's given expression) to be independent of each other. This is not the case here. In our tests, we even need to consider single / bidirectional causality. If it is bidirectional they are not independent. We also need to consider the time gap between back and forth of feedback directions. The use of $\partial Y/\partial X$ notation does not seem to be appropriate, as we do not obviously see how "time" and "causality" is involved in this notation. We therefore choose to keep the notation CS. To improve the clarity, we have added a phrase to describe the instantaneous CS: "the change in Y due to the change of per unit X at time t". In equation 1 of the revised manuscript, we also choose to keep the /dt and /∂t, to show its most fundamental definition. Note that we also express the change of cause-variable as full derivative rather than partial derivative.

Also, in applications in the Earth Sciences where time series of observations are available, how do you calculate the partial derivatives from them if you are ignorant of the underlying processes?

Response: *As answered above, in the application, we use the known full derivative of cause-variable, and the (modified) nIF between two time-series X and Y, to estimate the causal contributions and explore potential underlying processes.*

The connection to IF is presented only indirectly - by stating the hypothesis that nCS is approximately equal to |nIF|. If there is no alternative way to calculate nIF, how would you be able to test this hypothesis? Where is the independent definition of nIF ?

Response: *The IF and nIF can be estimated between two time-series, therefore we can test whether nIF can represent nCS through comparing the designed and estimated causal contribution. For better clarity, we have split the definition of nIF and the hypothesis into two equations (equations 4 and 5 in the revised manuscript), and how |IF| and |nIF| are estimated are in equations 6-14 in the revised manuscript.*

Later, one special case is considered - linear models (very unlikely to work for complex systems) where IF has a representation through covariance values. In addition, it seems that the authors believe in a decomposition of the normalization factor into a simple sum of IF, a noise component and a "self-dependence" term. Why should it be that simple? And how would you be able to discern the three terms, given only the time series of X and Y? The surprising answer is in eq. (8) to (10) which show that the self-dependence of Y is dependent on X, and the noise term (contribution of other variables) is also dependent on X. How is that possible? What are the assumptions about the phase space structure, stationarity etc. which go into that?
The reviewer also notes that md3 = md2 whenever the sign in the absolute bracket of eq. (13) is positive, and =2 md1 - md2 in the opposite case, so in which sense is md3 anything new once you have md1 and md2?

Response: *We noted the difficulty in understanding the concept of normalization here, and the potential confusion between “self-dependency term of Y” in our designed equations and the “self-representing” nIF*$_{(Y\rightarrow Y)}$*, as well as another confusion between “IF*$_{(non\text{-}X\rightarrow Y)}$*” and the “external noise contribution in the designed function* $\partial Y(n)/\partial t$*. They are conceptually different. In the revised manuscript, we have explained these terms in more detail.*

Note that the simple sum of IF for the normalization factor and equations 8-10 (9-11 of the revised manuscript) are obtained from Liang’s earlier work where these are derived. On the issue of “simple sum of the three terms for the normalizing factor”, it should be clear that our work does not agree to Liang’s proposal, as the best normalizing factor in our test is md$_3$*Z.*

Regarding why eq. (8) to (10) (9-11 in the revised manuscript) show that the self-dependence of Y is dependent on X, and the noise term is also dependent on X, we have added a brief explanation: “Note that since these two terms must remain comparable to the IF$_{(X\rightarrow Y)}$*, they could be seen as “relative information flows” from the effect variable itself and from the noise as compared to IF*$_{(X\rightarrow Y)}$*. It is hence* *not surprising for them to be a function of cause variable X (e.g. via C_{XX}, or C_{YX})." Other than checking Liang's paper, we also subsequently try to explain the $IF_{(non\text{-}X\rightarrow Y)}$, $IF_{(Y\rightarrow Y)}$, the various modifications of normalizing factors, and the $|md_1Z - md_2Z|$ terms, by extending our hypothesis of $nCS_{(X\rightarrow Y)} \approx nIF_{(X\rightarrow Y)}$ to these terms. We hope that it is clear to the reader that all these terms can be understood for their relevance in normalizing the causal function. That is why these terms are all comparable to $nIF_{(X\rightarrow Y)}$.*

Regarding the comment that such "linear models are unlikely to work for complex systems", we would like to clarify that the "linearity" in the "linear model" here refers to the proportional relationship between dX/dt and ∂Y(X)/∂t, but such proportionality (CS) can be "non-linearly changing", such as via combination of trigonometric terms in our designed examples, or via a combination of various physical processes acting on the effect-variable in the Earth system. You may refer to our paper on methane-climate feedback for how this is applied for a real world climate problem. In other words, as long as the causal relation between two variables can be reasonably and approximately described by proportionality, even if the proportion keeps changing, this method will be useful.

The "empirical tests" chapter lists no less than 8 artificially generated time series ("designed mock-up data sets") without, however, providing any details. The curious reader might want to reproduce the numerical results shown in the Figures, but there is no clue how to do that. What exactly did you choose as (say) "1D example with fluctuating self-dependency noise-contribution and a sinlge causal direction"? One has to refer to the supplement (not referenced to in lines 181-190 of the manuscript) to find answers to these questions - however, also this is difficult since mathematical notation is wrong (example: what does the sum "_2 ^nt 1/nt" mean? The summation index (n) can't be the upper limit of the sum itself, and if the user has to choose nt first, i.e. nt is a constant for the sum, the latter is juts (nt -1)/nt, which hardly makes sense? Fundamentally, if you have the partial derivative of X1(Y1) explicitly given as a time-varying function, you simply can't require that the partial derivative of Y1(X1) would be exactly zero, contradicting the inverse function theorem. What is going on here?

Response: *Thank you for the comment. We have added a sentence referring to Supplementary Table for the actual designed functions. Regarding the comment on the notation "$\sum_2^{nt}\frac{1}{nt}$", for nt=2, it means (1/2), for nt=3, it means (1/2+1/3), until nt=1000, it means (1/2+1/3+…..+1/1000). The nt refers to the number of time-steps. At the top of this supplementary Table, we have stated "One-Dimensional: nt = 2 to 1000; (or split into 2-200, 201-1000) initial conditions (i.e. nt = 1) are zeros". We are not sure what causes the confusion here. Anyway, we would like to highlight that this sub-term is not important. One could remove it and the value of designed function will still oscillate, and the general findings in this manuscript will not be affected. For some examples, the partial derivative of X1(Y1), i.e. $\frac{\partial X_1(Y_1)}{\partial t}$, is set as zero (note that we have swapped and X and Y in this revised manuscript, so we are referring to the partial derivative of Y1(X1) in reviewer's question). This simply means that X1 is independent of Y1. In these examples, we are testing 1-directional causality from X1 to Y1, i.e. $\frac{\partial Y_1(X_1)}{\partial t}$. We are not sure whether your comment about the "inverse function" suggests that Y=f(X), then X=f^{-1}(Y) so that x and y cannot be fully independent. But bear in mind that we are defining their partial derivative directly instead of defining X or Y, and then we numerically integrate these partial derivatives and sum them to obtain X* *and Y. We do not see what stop us from designing a zero $\frac{\partial Y_1(X_1)}{\partial t}$. Note that this also explains why we do not express $CS_{(X \rightarrow Y)}$ as ∂Y/∂X, to prevent such confusion related to the inverse function.*

In the previous chapter, dependence on other variables was considered as "noise", and there was the self-dependence term. But now, in l. 177, self-dependent terms are suddenly also noise, adding to the confusion.

Response: *The reviewer may have been confused by the different contexts between designed functions and normalizing IF. In the revised manuscript we have addressed this issue. In short, in the context of designed causal function, the noise includes the self-dependency terms. But, in the context of nIF, the "self-dependency" or $IF_{(Y \rightarrow Y)}$ could be better understood as a "self-representing information flow", which is conceptually different from the "self-dependency" of our designed causal function.*

The reviewer was fully lost when there was talk about the "1:2:3" ratio for 1D examples where X1, X2, X3, Y1, Y2 and Y3 are occurring. How is that a "1D" example?

Response: *We have restructured the discussion. The 1:2:3 ratio was first mentioned together with the "extended criterion" for assessing the hypothesis. It is 1D, with no matrix involved. The main comparison is between designed and estimated $\frac{\partial Y(X_1)}{\partial t}$, or between designed and estimated $\frac{\partial X(Y_1)}{\partial t}$. We have changed our Figures for 1D tests to focus on that. For the point on the 1:2:3 ratio between the designed $\frac{\partial Y_1(X_1)}{\partial t}$: $\frac{\partial Y_2(X_1)}{\partial t}$: $\frac{\partial Y_3(X_1)}{\partial t}$, it is shown only in Fig. 2k, l, and Supplementary Figures.*

The figures in the results section are not illuminating to the reviewer. It seems like one has to recognize a 21-units time lag from Fig. 3; however, even if one happens to know which panels have to be compared here, there are 1000 time steps shown, so the 21-units lag would only make a difference of 2.1%, you would need a magnifier to see anything here (the text in l. 187 talks about a 21% effect, which seem to indicate that there were somehow windows of length 100 analyzed each time, but this is mentioned nowhere else and is not visible in the Figures).

Response: *The length of 100 time steps in each analyzed window was mentioned in the Supplementary Information. In the revised manuscript we have briefly mentioned that in the main text (when we use bullet point to describe test 6 in section 2.6). Regarding the difficulty to identify 21-units time lag from Fig. 3, we have put the designed and estimated trends in the same sub-figure, so that the comparison is easier. Note that we have emphasized that the "the tendency of capturing the time-gap between cause- and effect- variable" supports our hypothesis, but we have also pointed out that is just a tendency. Although it may have real application in estimating such lead-time, it could easily go wrong, for example under the influence of noise contribution.*

Eq. (20) seems to be a differential equation for Xadj, but not even the units fit here (unless both X and t are dimensionless, which wouldn't be case in any applications). Admittedly, the reviewer didn't even get the "25-75% split" mentioned in l. 249.
The only way to come thtough the material presented is by going through the (uncommented!) Matlab scripts provided as fileshare by the authors. Do you really expext this from a reviewer, let alone a "normal" reader?
Response: *We appreciate the reviewer's effort looking into the Matlab scripts and pointing out our shortcomings. This was our mistake and we have corrected it in the revised manuscript. By the way, as the "preliminary removal of noise by subtracting a reference contribution" is not the key message in this paper, we have put these details into the Supplementary Information and added Supplementary Figure S1 to explain how we do it.*

As the reviewer doesn't see an easy fix to render the paper comprehensible., there are no detailed comments to the text (there would be many!). Still, there could be some interesting ideas, not the least since causal inference approaches (like Granger causality, CCM, PCMCI etc.) are quite fashionable in recent years in the Earth Sciences, and the concept of "higher order dependency" might be interesting. However, in its current form, the concepts are not communicated in a way that would render the paper acceptable for publication.

Response: *We have gone through a major revision and believe that all the reviewers' comments have been addressed. We look forward to further comments on the revised manuscript, particularly to the four questions we raised in the general response. Thank you.*