Reviewer 1:

The manuscript 'Empirical assessment of normalized information flow for quantifying causal contributions' by Cheng and Redfern considers how causal sensitivity could be measured through information flow in the context of climate science. The main potential contribution lies in the empirical definition of measuring causal influences between variables (i.e. causal sensitivity) as a product of a constant maximal causal sensitivity and a modified normalized information flow.

This version of the manuscript is much improved over a previous one that I reviewed, especially at the beginning. However, with the Results section when first test cases are evaluated the clarity is again lost (for me, at least). I also still find the mathematical notation confusing, possibly because I am not familiar with a few papers frequently cited in the manuscript. Variables seem to come and go, become multidimensional and scalar as you wish, and I don't see how different locations and time lags are fed into the overall picture. A good example is also Figure 1 where a thorough mathematical notation would allow the reader to immediately comprehend the meaning. I suggest introducing a consistent notation in a subsection that differentiates between all these points, which is then used throughout the manuscript and its figures.

***Response**: We thank the reviewer for going through this manuscript and appreciate the reviewer's comment. Please refer to below responses regarding the comments on notation.*

Specific comments:

- l. 68: allow for a comparison? ***Response**: Thanks. Yes.*
- l. 69: 'for simplicity' probably suffices/is simple enough.

***Response**: Thanks. agree. However, we have re-written some parts of the manuscript and this sentence has been deleted.*

- l. 79-81: I am slightly confused wrt the notation here. If X and Y are single variables (as implied by the text), how can then there be a maximum that is different from the same term itself? The text appears to imply that this is a statement over different locations and times, however, should this not lead to some sort of vector or matrix notation? L. 83/84 also seem to imply that there are at least multiple X. Bold notation needed?

Response**: yes, the "maximal" causal sensitivity refers to the "maximum" of causal sensitivity throughout tested durations and locations (how it is determined in the designed tests is illustrated in the Supplementary Information: Data Processing). We note the suggestion about the vector or matrix notation. In this manuscript, the most important "direction" is the "causal direction", which is expressed by the arrow sign "→". The second most important direction is whether the two variables are positively or negatively correlated, for which we add a subscript a1 to nIF as nIF_{a1}. For the spatial or temporal direction between cause- and effect- variables, it may seem logical to express them using matrix notation (bold capital* **X), but it may also cause additional confusion, while vector notation will likely create further confusion. <u>Firstly, for all our 1D tests, spatial direction is not even considered at all</u>. But we estimate running nIFs (e.g. one estimate of nIF for each 100 units of time, running throughout 1000 time-units). In this case, we will still have a "maximal causal sensitivity" throughout the entire period but the use of matrix notation is not appropriate. Secondly, even if spatial direction is involved, it simply distinguishes a causal influence from a cause-variable at a certain location to an effect variable somewhere. For example, there is a teleconnection (remotely connected) causality* *from the sea surface temperatures (SSTs) in the ENSO region to the temperatures and precipitation far away, but it does not seem necessary to use capital bold notation for gridded SST. The key is not the dimensionality of the data, but the "common maximal causal sensitivity" in the designed function regardless of the location. A different notation between 1D and 3D examples would, we feel, tend to distract the reader away from our key emphasis. Lastly, the time series is not a vector. A vector relationship does not apply to these remotely connected variables with complicated mechanisms. In brief, this paper considers the nIF between only two time series, although we can change any of the two time series by other time series, as long as the same causal function applies.*

We feel that the use of matrix notation is largely unnecessary and even distracting. As far as the notation is concerned, we chose to keep it consistent, just using X and Y, instead of switching between X and **X***, Y and* **Y***. However, to prevent a possible confusion (that is when the reader tries to conceptually distinguish a 1D variable from a 3D matrix and might expect different notation), we chose to briefly clarify this point at the beginning of section 2.2.*

- l. 110-115: here it is implied that somehow local or non-local does not play a role, so I am starting to wonder here how this relates to the notation above. I also don't understand at this point the link between the interchangeability over causes at different times and locations/identification of particular causes and how that links to natural methane emissions and global mean temperature. Is the global mean not exactly the opposite of identifying locations where specific processes lead to methane emissions? Maybe only rephrasing is needed? Could you clarify?

Response*: yes, the nIF between two time series is not affected by whether it is local or non-local causal influence. We could estimate the spatial distribution of causal contributions to a common effect variable, which allows us to compare (to a certain extent) between local and non-local influence. We have rewritten the part about "interchangeability". In brief, such interchangeability holds only when a "common maximal causal sensitivity" applies. The logic behind the revised discussion about the "common maximal causal sensitivity" and "normalizing causal function" should now be much clearer. We have also rephrased the methane example but moved that example to a specific section about "examples".*

- l. 137-140: this raises the question as to why one would not use lagged relationships of X and Y in the set of variables? Is the discrete nature of lags a problem?

Response: *We are not sure if you are referring to method like time-lagged correlation. This manuscript does not evaluate or compare with other causal methods. We assume that the user has no idea about the estimate of time-lag (or lead-time) prior to analyzing the causality between two time series. If you are referring to adjusting the time series after we have an estimated lead-time, this is possible, but we do not go into this in our paper. In addition, we are uncertain if it is practicable since it is quite possible to incorrectly estimate the lead-time. Nevertheless, for l. 137-140 of the earlier draft, concerning the influence of IF(X→Y) and IF(non-X→Y) on IF(Y→Y), we have elaborated further, as well as explaining the different normalizing factors in more detail.*

- Eqs (11)-(13) my first impression would be that the importance of shared causal influences will be problem-dependent. Could you clarify how each treatment would help/or not/ to generalize the concept, i.e. if problem-dependence in how information flows would affect the validity of the choice for Z?

Response: *The different choices of Z (eqns 11-13 in the earlier draft reviewed), have now been explained in much more detail in the revised manuscript. The explanation also covers what its* *individual term means, based on the extended hypothesis. The explanation in section 2.3 of the revised draft is also verified in the result section.*

- l. 185: word 'noise' missing here somewhere?

Response: *Yes, very weak self-dependency and independent noise contributions*

- l. 187: I don't understand why 21 steps of time lag equal 21% of each time analysed window? Can you explain the idea?

Response: *Revised. 21-steps of time-lag (i.e. 21% of each analyzed running time-window of 100 time-units)*

- l. 190: might be worth explaining that teleconnections stand for spatial interactions here? Again, how could the additional spatial dependency be better included in the mathematical notation employed here?

Response: *Noted, we have rephrased the relevant sentences, but we chose to keep the original notation for the reasons given above.*

- l. 198-199: similar – suddenly X1, X2, X3 and Y1, Y2, Y3 are introduced, which I assume should indicate a problem with three X-variables and three Y-variables? Why this choice? Where was this introduced?

Response: *These are still two-time series 1D tests, not a matrix, and not associated with shared causes of 3 X- or Y- variables. We have rewritten the part about our test using a 1:2:3 ratio, especially its role as the extended assessment criterion 2, and have improved the Figure presentation.*

- Figure 1: again clarity of notation, e.g. in (b) what is the meaning of multiple arrows between X and Y? Representing multiple variables? Time lags? Spatial points? I don't understand why there are no teleconnections here, but there are in the other subfigures? Somehow this has to do with the crossing arrows, but I doubt that many will understand why this is a way to symbolize teleconnections (or how they are imagined here). For me, the notation throughout the manuscript is confusing and still reduces the clarity too much. There are multiple processes (X and Y) which are related spatially and temporally (with potential lags)? However, how do these differ in the notation, how are they made obvious? Maybe write a subsection where you formally introduce the notation you are using and be consistent afterwards.

Response: *We have improved the figure captions as well as the related main text to clearly describe the multiple arrows, the presence/absence of teleconnections, and crossing arrows. Regarding the comments on notation, please refer back to our response above.*

- Figure 2 - I am lost here: what do the different colours stand for? How is this a test? I need a clear instruction as to how to read this plot. Why are the results good? Why do they confirm the hypothesis? Why does the second column sometimes look like a flat line? Which lines should be the same? The previous section already became less clear, but latest here I have literally no idea what is going on anymore. The reader has to work really hard to keep track. This needs to be improved.

Response: *We have revised the Figures and their captions to make them more self-explanatory. We also explain better how the assessment criteria are verified through our test-results.*

- Figure 3 same.

Response: *Figure is revised.*

- This could partly be helped, of course, by more clearly explaining what is going on in the Results sections here.

Response: *Thank you for the suggestion. We hope the revised manuscript has improved the clarity.*

- Figure 5 same.

Response: *Figure is revised.*