Dear Professor Knepley,

Thank you for taking the time to review our manuscript. Please find our answers to your comments below.

Original comment: *This paper was intended to "propose a novel methodology combining deep learning (DL) and principles of causality research". However, I do not believe it does so. It reiterates a standard theorem from causal models describing a causally sufficient set for some node X of a probabilistic graphical model. Then the authors claim to choose carefully such a set. If it were possible to do so apriori, there would be no confounding and no need for the causality formalism. After choosing this set, the interpolation of the joint probability distribution with a neural network follows standard practice. Since there is no real use of the mathematical formalism of causality, this cannot justify publication. Moreover, since "An extensive discussion of our results on soil moisture-precipitation coupling in terms of physical processes (e.g. Seneviratne et al., 2010; Santanello et al., 2018) and a comparison with results from other studies (e.g. Seneviratne et al., 2010; Taylor et al., 2012; Guillod et al., 2015; Tuttle and Salvucci, 2016; Imamovic et al., 2017) are postponed to a second paper", no new physical results are presented. Thus I recommend that the paper be rejected, and the authors submit a paper with the new physical insights included.*

Answer: To the best of our knowledge, we are the first to combine the approach of using interpretable DL to gain new scientific insights with the theorem on causally sufficient sets. The interpretable DL approach has been applied in several recent geoscientific studies and has led to new scientific insights into the Earth system (see references in lines 38 and 39 of the submitted manuscript). However, so far, in the application the difference between causality and correlation has been neglected. To overcome this important limitation, we propose to combine the approach with the theorem on causally sufficient sets. There are multiple reasons, why we believe that the methodological focus of the manuscript is justified, and why we delegate the comprehensive discussion of results on soil moisture-precipitation coupling to a second paper. First, the considered theorem on causally sufficient sets has hardly received any attention in the geosciences (see lines 47-50 of the submitted manuscript), which warrants the focus on the general methodology, which is applicable to numerous Earth system processes. Second, as an extension of the approach of using interpretable DL to gain new scientific insights, the proposed methodology requires some care, i.e. suitable choices of loss functions and DL models as discussed in Sections 2.2.1 and 3.2, and the choice of DL model gradients as the interpretation method (rather than for example the common layerwise relevance propagation method (Bach 2015)), as detailed in Section 2.2.2..

We disagree with the comments that "if it were possible to do so [choose a causally sufficient set] apriori, there would be no confounding and no need for the causality formalism" and "there is no real use of the mathematical formalism of causality". In many Earth system applications, a causal graph can be constructed based on physical insights (e.g. in the described example of soil moisture-precipitation coupling; see also Massmann 2021). Although this graph may not always be exhaustive, this formalization of system dynamics has two particular uses in the context of the proposed methodology. First, the causal graph formally represents the assumptions underlying the respective application of the proposed methodology. Second, in the methodology, it is used to choose a causally sufficient set and prevent confounding (according to the considered theorem on causally sufficient sets). Note that Section 4, "Further analyses to assess the correctness of obtained results", of the manuscript also addresses the possibility of an incorrect underlying causal graph. The discussion of these aspects will be expanded in the revised manuscript.

Original comment: *In the paper itself, some claims could be better supported by evidence. The authors claim that simulations are always more expensive than their deep learning scheme, but no data is provided. Simulations at what resolution? Is the cost of DNN training included? More nuance here would be helpful.*

Answer: In the manuscript, we claim that “statistical approaches usually have much lower computational costs [than approaches based on numerical simulations]” (lines 31 to 32), which we believe to be true in the general context of Earth system applications and Earth system simulations. In the submitted manuscript, we analyze the causal effects of soil moisture changes at each of 120×80 target pixels on subsequent precipitation in the target region. To estimate the average causal effects, we average the causal effects over all time steps in two test years, constituting 2208 time steps. Performing an analogous study based on numerical simulations would require $120 \cdot 80 \cdot 2208 = 21196800$ 4-hourly simulations with the ECMWF Earth system model used to produce the considered ERA5 data (each simulation would be initialized with the state of the reference simulation at one of the 2208 considered time steps, the only difference being that soil moisture would be slightly increased or decreased at one of the 120×80 target pixels). This corresponds to simulating approximately 10000 years with the ECMWF Earth system model and is computationally infeasible. We will clarify this in the revised manuscript.

Original comment: *Derivatives calculated from the DNN solution are used to quantify sensitivities and errors, but how accurate are these estimates?*

Answer: The error in the approximation of the function from Eq. 13 in the submitted manuscript as well as in its derivatives is difficult to quantify explicitly. We address this in Section 4, “Further analyses to assess the correctness of obtained results”, of the submitted manuscript and state in lines 504 to 506 that “While these analyses cannot guarantee the correctness of obtained results, and developing further analyses is desirable, we believe that the proposed analyses provide a solid indication of the correctness of obtained results.” In an updated version of the manuscript, we will further clarify the sources of errors in the proposed methodology (namely errors in the approximation of the function from Eq. 13 as well as in its derivatives, and errors due to an incorrect underlying causal graph) and the difficulty to quantify these errors.

Original comment: *On page 17, the authors state that “In our example, the null hypothesis was rejected at a confidence level of 99 %”, however it is later stated that only two samples were taken. This seems misleading at best. Clarification of what is meant by the 99 % confidence level in this case would be very helpful.*

Answer: For this example, we detail the computation of confidence levels on lines 400 to 405 of the submitted manuscript. In total, 20 samples are produced by testing multiple instances of the DL model on the original and the modified test set, respectively. In lines 406 to 408, we also note that “for the validity of this test, it may be harmful that there are only two test years in our case and thus only one possible permutation of years apart from the original one.” Moreover, we describe a variation of the test that resolves this issue (but only allows for weaker conclusions) in lines 408-410.

We hope that we could resolve the concerns mentioned by you. We think that the presented research is of interest to many geo- as well as other scientists and deserves publication.

Sincerely,

Tobias Tesch

References

Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W (2015) On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLoS ONE 10(7): e0130140. https://doi.org/10.1371/journal.pone.0130140

Massmann A, Gentine P, Runge J (2021) Causal inference for process understanding in Earth sciences. ArXiv. https://arxiv.org/abs/2105.00912