Reviewer #2 Comment 1: (hereafter referred to as R2C1, R2C2…) *The study applies a panel research design to estimate the causal effect of three hypothesized human-related drivers (urban extent, cropland extent and reservoir regulation) of annual flood peaks in China. While the methodological contributions of the study are (in my view) limited compared to recent other studies using panel regressions in a similar context (e.g., Blum 2020 and Davenport 2020 cited in the study), the study is nonetheless valuable in that it provides important insights on how these process operate in conjunction, using a very large dataset in China. The study is in my view appropriate for publication in HESS, provided the author address the following major concerns that I have*

A: Thank you for your constructive comments. We have carefully considered your suggestions and addressed your concerns. We have made the following major revisions in the method and data.

First, we added 3-day and 30-day total precipitation before flood peaks for each catchment in the regression to account for individual time-varying confounders. The reason for such a revision is that the delineated climate regions cannot fully control climatic confounders since the climatic drivers of floods have sub-regional spatial variability. Therefore, the regression equation has been revised as:

$$log(Q_{i,t}) = \alpha_i + g_1\left(Urban_{i,t}\right) + g_2(Crop_{i,t}) + g_3(RI_{i,t}) + \pi_{r,t}D_rD_t + D_r\left(\varphi_r P_{i,t}^{(3)} + \lambda_r P_{i,t}^{(30)}\right) + \varepsilon_{i,t}$$

where $P_{i,t}^{(3)}$ is the 3-day total precipitation before the flood peak in year t of catchment i, which accounts for the rainfall that causes the flood; $P_{i,t}^{(30)}$ is the 30-day total precipitation before the flood peak in year t of catchment i, which accounts for the soil moisture and snowmelt that cause the flood. The coefficients of $P_{i,t}^{(3)}$ and $P_{i,t}^{(30)}$, namely φ_r and λ_r, are assumed to be constant within a climatic region r. The original region term $\pi_{r,t}D_rD_t$ accounts for omitted time-varying regional confounders other than $P_{i,t}^{(3)}$ and $P_{i,t}^{(30)}$.

Second, we selected 757 non-nested catchments to fit the regression model so that the residuals of the model were not highly correlated. This revision avoids uncorrected inference about the regression coefficients due to the underestimation of their standard deviations using correlated flood samples.

Note that the results do not substantially change after the methodology revision above. While we believe the revised method and data are more convincing and solid.

R2C2: *1. Potentially misleading map figures. To be clear, the panel approach does *not* allow to estimate heterogeneous treatment effects. It allows to estimate one average effect of (say) urban expansion on flow peaks (i.e. one single value of beta, if g() is linear) across the whole sample. It does *not* allow to say that urban expansion has a larger effect on flood peaks in some regions than in others. Yet the maps in figures 6 and 7 (and their discussion throughout the paper) appear to suggest exactly that, which I find misleading. The spatial variability in the "effect" of crop/urban on floods represented in these maps only emerges because changes crop and urban cover are themselves varying across regions. Figure 6 is nothing more than a map of urban cover change, scaled by a constant factor (the estimated beta) representing the linear effect it has on flood peaks. This point is important to clarify throughout the text, at the very least by specifying the estimated value of beta and theta in the captions of Figures 6 and 7 (see minor comments for other suggestions).*

A: Thank you very much for your critical comment. As we know, the panel regression in this paper is unable to estimate heterogeneous effects across different catchments. The "average effect" of a factor on floods you mentioned is shown in Fig. 4 and Fig. 5, where the homogeneous sensitivities of floods to human factors across all catchments are presented. Fig. 6 and Fig. 7 show the accumulated changes in floods due to the changes in human factors in a long period for each catchment rather than the heterogeneous effects. To avoid misunderstanding, in the captions of Fig. 6 and Fig. 7 in the revised manuscript, we will emphasize the changes in floods as the "accumulated changes" and clarify that the changes are calculated by Eq. (7), i.e., $\Delta Q(\%) = \exp(g(X_{i,t_2}) - g(X_{i,t_1})) - 1)$.

R2C3: *2. Fixed Effects. I am wondering why you use "regions" as space fixed effects, and not the individual basins themselves. For Blum et al., this approach made sense because they interact the treatment (X) with covariates (e.g., soil permeability, etc) but I don't really see the point of doing that here. I am concerned that it might introduce a bias associated with varying confounding factors within the regions (e.g., basin altitude can vary within regions and affect both the treatment crop or urban cover and flood magnitude). Adding a specification with basin-level fixed effect (i.e. setting k==number of basins) as robustness check might help alleviate my concern.*

A: Thank you for your important comment. We use "regions" to account for omitted time-varying confounders such as vegetation changes. If we set *k*=number of basins, i.e., we use a basin-level dummy variable interacted with a year dummy to represent time-varying confounders, the number of regression coefficients will be larger than the number of flood peak observations, which makes the estimation of coefficients infeasible. Therefore, we use a region-level dummy variable interacted with a year dummy (i.e., $\pi_{r,t}D_rD_t$) to control omitted time-varying effects. As for basin-level confounders, the time-varying effect has been controlled by the event precipitation ($P_{i,t}^{(3)}$ and $P_{i,t}^{(30)}$, see R2C1), and the time-invariant effect has been controlled by the individual-specific intercept (α_i). Basin altitude, the sub-regional confounder you mentioned, has been included in α_i.

R2C4: *3. Heterogeneous treatment effect: I am wondering if your results are affected by heterogenous treatment effects in the sense that most basins of the sample likely have little impervious surface cover. (By the way, please add a table with descriptive statistics for the reader to assess that). If the deviates (even slightly) from the three arbitrary functional forms that you impute to g(), this may potentially bias your average estimates. A way to control for this (perhaps) would be to do a robustness check by running the analysis to a subset of highly (lowly) impervious basin to see how sensitive the effect is.*

A: Thank you very much for your valuable comment. We agree with you that the results depend on the selection of basins. However, we believe the difference in the results brought by basin selection is more related to sampling uncertainties rather than heterogeneous treatment effects. If we manually select a subset of basins with low impervious areas to fit the model, we may get a model with a low signal-to-noise level since the changes in floods caused by changing impervious areas are far small than the model errors. Therefore, in the revision, we calculated the confidence intervals of the regression coefficients by bootstrapping, which resampled all pooling flood samples to fit the model 1000 times. Using bootstrapping, the confidence intervals of the regression coefficients accounted for the sampling uncertainties related to basin selection and year selection. We will add the bootstrapping part in the revised manuscript.

In addition, we will add a summary table of catchment characteristics (Table R1) in the revised manuscript.

Table R1. Summary of catchment characteristics for 757 catchments in 1992-2017. The summaries of *RI* and Δ*RI* are calculated based on 207 catchments with at least one large and medium dam.

Variables	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Area **(km^2)**	29	499	1096	3341	2763	142372
Urban **(%)**	0	0.06	0.30	1.52	1.10	65.07
Δ***Urban*** **(%)**	0	0.05	0.23	1.14	0.85	24.66
Crop **(%)**	0	10.63	24.71	32.75	48.99	99.58
Δ***Crop*** **(%)**	-21.58	-0.81	-0.02	0.38	0.87	32.04
RI	0.01	0.09	0.21	0.51	0.61	7.45
Δ***RI***	0	0	0	0.17	0.07	7.44

R2C5: *4. Nestedness: Finally, the ordinary least square estimator that (I assume) you are using only provides an unbiased estimate of standard errors if residuals (epsilon) are independant. In your case, I am concerned that many of your observations might be nested (i.e. taken along different reaches of a same river), which might introduce a correlation in the epsilon. For instance a time- and space- specific shock on flow peaks observed in a headwater catchment will likely affect flow peaks observed at several gauges along that river. The fact that errors congregate around specific basins in Figure 8 is actually a strong indication of that effect! This effect might lead you to underestimate the standard errors on your regression coefficient and find a significant effect where there is none. A way to address that would be to use the topology of your river network to specify the structure of your variance-covariance matrix (see, e.g., Muller and Thompson 2015) which you can then incorporate in your estimation via Generalized Least Square or Restricted Maximum Likelihood. Alternatively, you could do a robustness check where you run your OLS estimation on multiple subsets of your full sample, for which you made sure that all observations are from different catchments âA˘T hopefully the results will be similar.*

A: Thank you very much for your comment. We indeed used an ordinary least square estimator to fit the model. So we agree with you that nested catchments cause dependence between model residuals, and thus produce wrong inference about the regression coefficients. In order to select non-nested catchments and include as many catchments with dams as possible, in the revision, we selected the most upstream catchments with large or medium dams (if possible) among overlapping catchments. We got 757 catchments from this selection, among which 207 catchments had as least

one dam. Although the results did not change substantially using the new subset of catchments, the regression model in the revised manuscript will be based on these 757 catchments.

Minor Comments.

R2C6: *The first sentence of the abstract is awkward ("because the knowledge and observations toward the effects are limited"). Please reformulate.*

A: Thank you very much for your comment. We will revise the sentence as "Quantifying the effects of human activities on floods is challenging because of limited knowledge and observations" in the revised manuscript.

R2C7: *L79: middle -> medium ?*

A: Thank you very much for your suggestion. We will change all "large and middle dams" to "large and medium dams" in the revised manuscript.

R2C8: *L94: It took me a while to realize that you *defined* your regions such that climate is homogeneous within them (as oppose to assuming that climate is homogeneous within a bunch of predetermined regions). Maybe clarify that here?*

A: Thank you very much for your comment. We will introduce how we define regions here in the revised manuscript.

R2C9: *Eqn 6: My understanding is that ΔQ varies of space but not time: if so, how to you "average over" the time index in the middle expression. Also, this would be an ideal place to clarify that ΔQ varies in space only because ΔX varies in space. Your estimation of g() is constant in space and time.*

A: Thank you very much for your comment. The ΔQ here is the effect is the sensitivity of Q to X rather than the accumulated change of floods along time. As we said in Line 133, this sensitivity is "the percentage change in Q given a fixed change in X". When we exhibited this sensitivity, we kept X and ΔX to be constant across all catchments (See Fig. 4 and Fig. 5), so the ΔQ here was constant not only in time but also in space. To avoid misunderstanding, in the revised manuscript, we will change Equation 6 to:

$$\Delta Q(\%) = \Delta Q/Q = e^{g(X+\Delta X)-g(X)} - 1$$

R2C10: *Fig 2: I agree with the other reviewer that p-values are an odd criteria for model selection. Either justify it, or use goodness of fit metric.*

A: Thank you very much for your comment. We agree with you that p-values are not appropriate for choosing models. In the revision, we used AIC to select effect terms (i.e, function $g_1(\cdot)$, $g_2(\cdot)$, and $g_3(\cdot)$) and found no change in the terms compared with the original manuscript. We will use AIC in the revised manuscript.

R2C11: *L155-160 and Fig 8. I find it a good idea to analyze the spatial distribution of model deviations (i.e. locations where variations in Q are not explained by the modeled drivers), but I find the approach chosen to identify these locations odd/arbitrary and challenging to understand. Wouldn't it be more straightforward to simply map the temporal variance of the residuals (i.e. Var_i(Eps_it))?*

A: Thank you very much for your suggestion. However, our purpose here is to see how floods change in catchments that are free from the impacts of urbanization and dam constructions. Therefore, we selected catchments with low *Urban* and *RI* to derive flood trends. Analyzing the spatial distribution of model deviations (e.g., to calculate $\mathrm{Var}(\varepsilon_{i,t})$) only tells us the places where some drivers are missing in the regression. Model deviations do not tell us the directions and magnitudes of changes in floods. We will clarify our purpose of Fig. 8 in the revised manuscript.

R2C12: *L294 "Coefficient of Variation" can be understood as the ratio between the standard deviation and the mean. I don't think that's what you mean here, so please reformulate.*

A: Thank you very much for your comment. We believe the "Coefficient of Variation" is correct here. As we said in Line 291-293, "the method derives a common percentage change in all flood peaks given changing human factors, which means no changes in coefficients of variation." For example, suppose we have random variable *Z*, a percentage change in *Z* (e.g., a 10% decrease) makes *Z* to be 0.9*Z*. In this case, the coefficient of variation is Std.(0.9*Z*)/Mean(0.9*Z*)= Std.(*Z*)/Mean(*Z*). Therefore, the causal effect in our study does not allow the changes in the coefficient of variation of floods.

R2C13: *L294. You provide a good illustrative example of the models inability to capture heterogeneous treatment in time, but here would also be a good opportunity to give an example of a heterogeneous treatment in space (i.e. a scenario where cropland might persistently have a stronger effect on flow peaks in some locations than in other). That would contribute alleviating my first major concern, above.*

A: Thank you very much for your comment. Omitting spatially heterogeneous effects has been mentioned as the first limitation of the method in Line 288 as "no interaction terms between human factors and regional or individual characteristics". To make it clear, we will change this sentence to "no interaction terms between human factors and regional or individual characteristics that produce significant spatially heterogeneous effects". We will also give an example of spatially heterogeneous effects, e.g., the effect of increasing urban areas on floods may be larger in regions with high soil permeability.

R2C14: *SI. Please add a descriptive statistics table with key stats on all the considered variable across your sample.*

A: We will add the table in the revised manuscript. Please also see R2C4.

Marc Muller

References Blum, A. Et al (2020) Causal Effect of Impervious Cover on Annual Flood Magnitude in the United States, GRL

Davenport et al. 2020 Flood Size Increases Nonlinearly Across the Western United States in Response to Lower Snow-Precipitation Ratios, WRR

Muller, M.F. and Thompson, S.E. (2015) "TopREML: a topological restricted maximum likelihood approach to regionalize trended runoff signatures in stream networks", HESS