The line numbers indicated here are consistent with those in the CLEAN (i.e., no changes tracked) version of the revised manuscript.

Response to Editor

Editor: Your manuscript has again been reviewed by the initial two reviewers. While reviewer 1 is very happy with the revisions, reviewer 2 points out that major revisions are still needed. Specifically, he points out that model performance and validity would profit from (1) a regional instead of a local model fit and (2) improvements in the cross-validation strategy. I would like to ask you to address these concerns in the revised version of your manuscript.

Response: It is our sincere gratitude to the two reviewers who helped improve our manuscript. It is unfortunate that we were unable to convince Reviewer 2 of the reasonability of local models in the first-round revision. We appreciate the critical review that has motivated us to improve the analysis, but we disagree with those comments that suggest the regional model is the only valid approach. Much of what the reviewer writes seems to imply that one should always use the model following the "best" practice shaped for prediction tasks (i.e., achieving the highest performance/predictive accuracy), independent of the research questions, such as "*no matter what you want to do with a model, you should always set up a model in the state-of-the-art for whatever model you use*". We strongly disagree with this presumption. A regional model and a local model can be considered different modeling choices in this application, and we deliberately chose a local model as it is more suitable for our research question. In particular, a local model allows a relatively straightforward interpretation of the flood-generating process at the catchment level. Using a regional model, however, would have completely changed the paper and diverted attention from the main objective of our study: identifying flooding mechanisms from the local and interpretable relationship between meteorological drivers and flood responses.

The reviewer argues a regional model would be a better choice because it captures relationships that can be generalized to ungauged basins and unprecedented events. This may be true, but we cannot agree with the implication that only such a model (i.e., using one mixed-effects framework to fit all available data) can be used to gain physical insights. In general, from a scientific perspective, even parsimonious models can help generate scientific insights. Otherwise, a lot of simplified models (intermediate complexity climate models, integrated assessment models, simple correlation analyses, etc.) would not exist. Importantly, in our case, we decided not to change the paper towards using a regional-modeling approach because employing such a model would require introducing catchment attributes in the modelling, which could confound the interpretation of local flooding mechanisms and therefore make physical insights uncertain or less concise. This was explained in the manuscript and in the previous rebuttal, but it seems the reviewer missed that point. We agree that a regional model may also generate interesting scientific insights related to a different research question than what we ask in our work, and we had planned to employ such a regional approach for future work to avoid overloading the paper.

The same argument as above holds for the selection of input features. Since our goal is interpretability, we restricted ourselves to a few input features whose effects can be relatively

easily interpreted and linked to fundamental physical processes. In fact, as we explicitly expressed in the first-round revision, we do not negate the value of including more input features to potentially discover unknown patterns, but a comprehensive investigation of these unknown patterns and mechanisms is outside the scope of the present study.

For the problem of cross-validation, we do admit that the first version may not have addressed the concern raised by the reviewer. In this round, we reran the models with a more rigorous data split approach and update the results throughout the manuscript. Despite minor changes in most reported numbers, all previous conclusions still hold. However, we appreciate the reviewer's comment, which has helped to make the conclusions more robust.

Overall, we respect and appreciate the comments from Reviewer 2 and try to reconcile the specific research question in the present study with the expectation of using a regional model with more variables, which is undoubtedly our long-term goal of using interpretive deep learning. In response to the suggestions, we tried to address the concerns without losing focus or without the paper becoming unnecessarily long. It is our hope that the revised version will be positively received.

Response to Reviewer 1

Reviewer 1: I have received a revised manuscript of Jiang et al. In the revised version the authors have addressed most of my main concerns. I think the manuscript has been considerably improved, I have only several editorial suggestions.

Response: We are grateful to Reviewer 1 for providing positive comments for the present paper and good suggestions. The responses to the comments follow.

Reviewer 1: Line 37: should it be Stein et al 2020 instead of Stein et al 2021?

Response: Thank you for pointing it out. Yes, it should be Stein et al 2020 and we have corrected it. (**line 37**)

Reviewer 1: Line 97: 20,455 time steps

Response: Added as suggested. (**line 97**)

Reviewer 1: Line 246: The selection of catchments is not clear here. Please clarify on which basis they were selected and attributed to different regions.

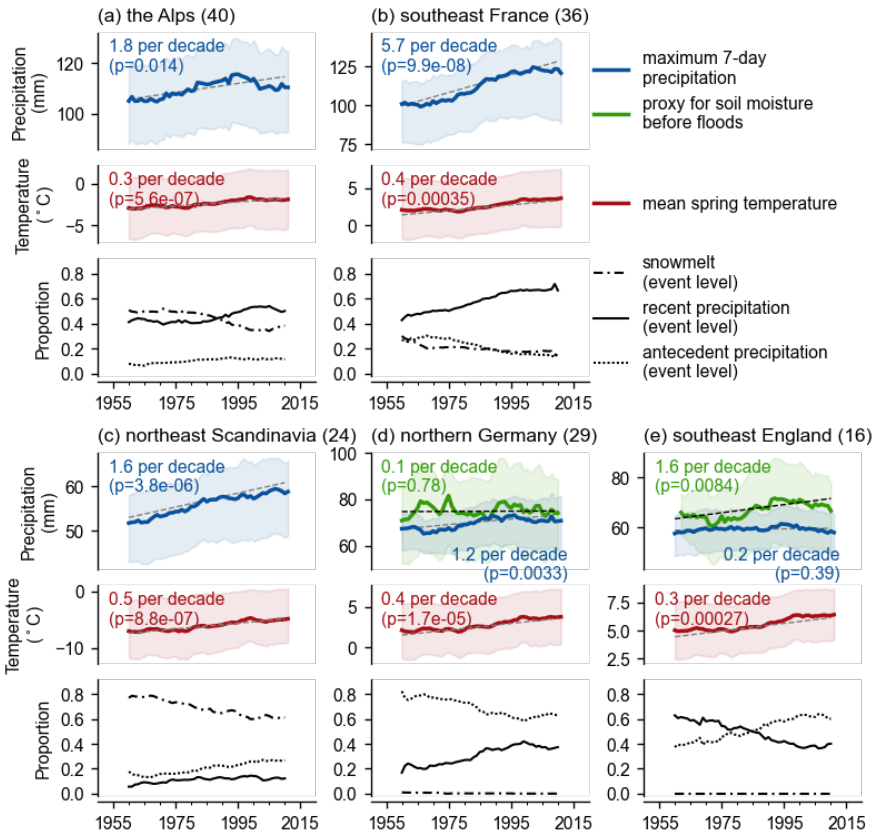
Response: We have rewritten the original sentence as "*Moreover, in order to analyze the possible causes of trends, we selected a number of regions where most catchments present consistent trends in certain mechanisms. We investigated those catchments exhibiting significant changes in flooding mechanisms and compared the temporal regional changes in flooding mechanisms with changes in potential flooding drivers.*" (**lines 246-248**)

Reviewer 1: Figure 7: Please clarify in the caption how confidence intervals were computed.

Response: In the caption, we added "*The shades denote the 95% confidence interval of the proportions, which was calculated as* $\hat{p} \pm 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ *($\hat{p}$ is the estimated proportion and n is the sample size).*" (**lines 477-478**)

Reviewer 1: Figure 8: I suggest to add 25th and 75th percentiles to indicate spatial variability within regions.

Response: Thank you for the suggestion. On the basis of Figure 8, we additionally added new **Figure S6** to show the 25th and 75th percentiles of changes in meteorological drivers. To refer to the figure in the main text, we added "*These figures are robust against spatial variability within regions (see Fig. S6 in the Supplementary Material).*" (**line 539**)



(new) ***Figure S6****. The temporal changes of the event-level mechanisms in relevant catchments within the five selected regions (see Fig. 7c in the main text), as well as the changes in average extreme precipitation, mean spring temperatures, and antecedent soil moisture conditions prior to flooding. The notions are the same in Fig. 8 in the main text, except for the shades that further illustrate the 25th and 75th percentiles of the yearly changes in respective meteorological drivers (smoothed by a 20-year moving average window as well).*

Response to Reviewer 2

Reviewer 2: In the first rounds of reviews, I commented on three topics:

1. Model setup (single basin LSTMs vs regional LSTMs).

2. Selection of input variables.

3. Data splits.

In most parts, the authors argued that my concerns do not require any change to their manuscript apart from adding a new section on "Limitations and Outlooks" that basically lists these concerns.

In one point (selection of input variables), I think it is a pity that the authors limit themselves to their selection and do not report potential novel insights (see detailed comment below). On another point (model setup), I do not agree with the argumentation of the authors, why they think that an inferior model calibration procedure is justified, but in the end it is their decision.

The last point (data splits), is probably the most critical. Reading that "the model was used for statistical purposes instead of a prediction task", which according to the authors justifies the use of a data split that is not independent of the training data is in my opinion a red flag for any kind of study that involves a model. Even if the results would not change (what I sincerely hope for the authors) it is a scientifically wrong thing to do. And to justify it by saying a per-gauge time split is more complicated is in my opinion a really bad argument.

I added some specific comments to the three points below but in general, I have nothing much to add to what I said in the first round. Ultimately, it is up to the editor to decide how to proceed.

Response: In the first place, we would like to express our appreciation for the reviewer's comments, which aimed to help improve our manuscript. We regret that the reviewer did not appreciate our response to his first-round comments and the paper's rating has been dropped from excellent/good/excellent to good/fair/good and from minor revision to major revision, even though the paper has not fundamentally changed.

During the first round of review, the reviewer made positive comments like "*To my knowledge, it is the first time that such an analysis (detecting flooding mechanisms and analyzing the change over time) is made using a) deep learning models (here LSTMs) and b) methods from the field of explainable AI (here integrated gradients). My views on LSTMs is no secret and I*

have often said that you can do more than "just fitting streamflow records" with these models, so naturally, I am quite excited to see someone coming up with such an idea." and "*Additionally, I think this paper is exceptionally well written and at least for me personally, everything seemed pretty clear and reasonable. For example, the authors make a couple of assumptions (like grouping the integrated gradient signal into two groups of a) the last 7 days and b) all other days before that), but their reasoning for all these assumptions is clearly articulated and to me, they make sense.*" Overall, the reviewer gave a "minor revision", which is generally assumed to involve minor amendments rather than substantial revisions. We can assure that we had carefully considered every comment raised by the reviewer in the first round and we did not mean to address the issues half-heartedly by placing the work in "*Limitations and Outlooks*". Instead, we fully recognize the potential of using regional LSTM with more variables to help a better process understanding, which is what we had planned for future work as well. We decided not to change the manuscript in that way because it would have changed the paper into a completely different one, making it unnecessarily long, and diverting attention from the main objective of the study: identifying flooding mechanisms from the local and interpretable relationship between meteorological drivers and flood responses using a novel approach based on machine learning.

For the points of input variable selection and model setup, we argue that the criticism that "*they think that an inferior model calibration procedure is justified*" is not fair. The comment implies that an incomplete model is inferior in spite of the fact that even incomplete models can generate scientific insights as well. Model selection should follow the principle of parsimony, otherwise known as Occam's razor, which "*seeks to find an optimal trade-off between the ability of the model to fit data and the model's required complexity to do so*" (Höge et al., 2018). Of course, one could always add more predictors (including the static catchment attributes to build the regional model) to increase model accuracy. However, in light of the tradeoff between accuracy and interpretability, is this approach really useful for extracting concise insights about the local relationship between meteorological drivers and flood responses as intended by our study? In particular, because of the collinearity between the variables that we already have and the variables to be added, the effect of input features might be less interpretable and uncertain (see our reply to the specific comments below).

For the point of data splitting, we acknowledge that our previous practice was less rigorous. We re-ran the models with a more rigorous data split approach and updated the results throughout the manuscript. None of the conclusions have been affected by this change.

Again, we are grateful for the reviewer's comments intended to improve our work. For the comments that are beyond the scope of the specific research question in the present study, we made further clarification in the revision. We have tried to address the concerns without changing the manuscript to a completely different study and without losing its original focus.

Reference:

Höge, M., Wöhling, T., & Nowak, W. (2018). A primer for model selection: The decisive role of model complexity. Water Resources Research, 54, 1688– 1715.

Reviewer 2: 1. Model Setup

The topic of discussion is, whether it is necessary/recommended or not for this study, to use a regional training setup with LSTMs rather than the per-basin calibrated LSTMs the authors chose to use. Here you say "However, as you mentioned as well in your comment, the choice of which strategy to apply (local model vs. regional model) should be in line with the research purpose."

No, this is not what I have mentioned. I deeply believe that no matter what you want to do with a model, you should always set up a model in the state-of-the-art for whatever model you use. Again, you would probably not calibrate a conceptual hydrology model in a regional fashion if you are interested in a single basin.

There is no reason to believe that a model that is consistently and significantly better than another model (or the same model but trained with a different setup) should not have a better understanding of the underlying physical processes (i.e. flood generation).

Response: In the last paragraph above, it seems that the reviewer believes that "better" models are always equivalent to models that are capable of making better predictions. However, we disagree with this assumption. In particular, from a scientific perspective, "better" models should be models that are more suitable for the research question at hand. Hence, different purposes, i.e., research questions, will lead to different best models, even for the same data set (Tredennick et al., 2021). Consequently, we argue that it is inappropriate to ask us to adopt best practices for prediction problems (for which a regional model would be more suitable) in our study that, instead, aims at physical interpretability. In the context of our application, a regional model and a local model can be considered different modeling choices, and we deliberately chose a local model as it is more suitable for our research question. In particular, a local model allows a relatively straightforward interpretation of the flood-generating process at the catchment level.

Reference:
Tredennick, A. T., Hooker, G., Ellner, S. P., and Adler, P. B.. 2021. A practical guide to selecting models for exploration, inference, and prediction in ecology. Ecology 102(6):e03336.

Reviewer 2: LSTMs trained to local basins suffer from two problems:

1. They witness saturation effects (which is easily visible when plotting discharge time series and e.g. comparing the max peak the model is able to output during the training period vs validation period).

2. They are unable to model any process that did not happen in that particular basin during the training period. The solution to both of these problems is training LSTMs on a multi-basin dataset, which afaik should be the one and only way how LSTMs should be applied for rainfall-runoff modeling. Looking at e.g. Frame et al. (2022), one can see how the regionally trained LSTM is able to predict unprecedented events in individual basins (e.g. being trained only on

data with q < HQ5 and then tasked to predict > HQ100), which a locally trained model will never be able to do.

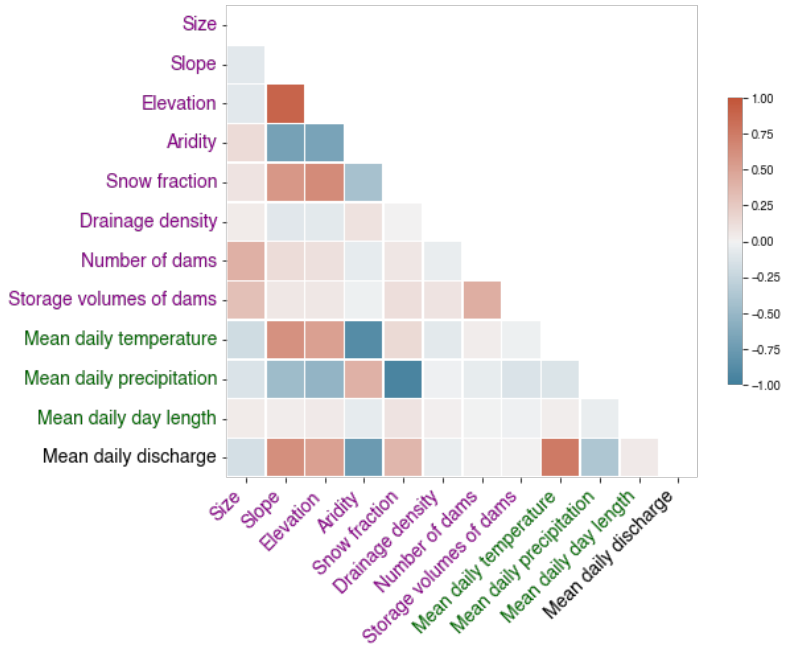
Response: The rebuttals against using local LSTM models may be true if concerns are about their ability in predicting unprecedented events, but we must emphasize that the argument has little relevance to the objective of the present study. We are not interested in processes that didn't happen in a given catchment as we want to identify catchment-level flood processes in the observational period. It is therefore a poor argument to suggest the invalidity of local modeling for our study, which focuses on interpretability instead of prediction. Assuming the argument holds, even linear approaches should be avoided in hydrology as they cannot accommodate and capture nonlinear processes in hydrological systems. Again, high predictive accuracy cannot be the only guiding principle for model selection.

Reviewer 2: "With the local models, since meteorological variables are the only inputs, we are able to focus on how they explain the temporal variation of discharges. In comparison, for a regional model that is supposed to capture both temporal and spatial variations in discharge peaks, it is challenging to distinguish how meteorological variables contribute to temporal variation in flooding within catchments from how they contribute to the spatial variation in flooding across catchments."

I think I disagree with everything you have said here. First, just because something is "more challenging", that doesn't mean in my opinion that the easier (but potentially wrong/biased) approach is justified. And in more detail, I also don't see the problem of looking at flood generating processes through integrated gradients, if static features are also model input arguments. Sure, the static arguments influence the model dynamics, but so does the fact that you have different model weights for every catchment. Just as an example, imagine you have a regionally calibrated model and two different basins A and B. Now we force the model with the same timeseries of meteorological features but using the different sets of static features, corresponding to basin A and basin B. Now imagine that in one basin, the model would produce a flood peak that, by looking at the integrated gradients of the meteorological features, is driven by the precipitation of the last few days, while in the other basin it produces a flood peak that is driven by temperature of the recent days and precipitation of some weeks ago. The difference here is certainly, because the static attributes are different. But does this have an influence on your analysis? I think not, and most likely this model has a much better understanding of the underlying processes. In your case, you would identify these two peaks as generated by different processes, which is most likely correct. So sure, static features have an influence on the model dynamics, but your analysis will still tell you which meteorological inputs are the most important for the peak flow. To be entirely honest, I fail to see what is the real challenge here.

Response: We fear the reviewer has misunderstood what we meant by the challenge and regret not clarifying it clearly enough in the first-round revision. What we were mostly concerned about is the issue due to the collinearity between the variables we already used and the variables to be added for regional modeling. For illustration, the new **Figure S7** shows the correlation heatmap between some common static catchment attributes (indicated by purple texts), daily

mean meteorological variables (indicated by green texts), and daily mean river discharges (indicated by black texts) for the 1,077 catchments. The catchment attributes include catchment area, slope, elevation, aridity, snow fraction, drainage density, number of dams, and storage volumes of dams. It is apparent that some catchment attributes and average meteorological drivers are highly correlated, such as slope vs. precipitation, aridity vs. precipitation, snow fraction vs. temperature, etc. The multicollinearity might not be problematic for prediction and forecasting tasks but can seriously impede interpretation, as the multicollinearity can affect the coefficients (weights) of independent variables in a way that limits the physical interpretation of the feature importance.



(new) ***Figure S7.*** *The Pearson correlation heatmap between some common static catchment attributes (the first eight attributes, written in purple), daily mean meteorological drivers (in green), and daily mean river discharges (in black) for the 1,077 catchments in the main text (see Fig. 1). The catchment size, slope, elevation, drainage density, number of dams, and storage volumes of dams were derived from the Global Streamflow Indices and Metadata Archive (GSIM, https://doi.org/10.1594/PANGAEA.887477), the aridity index and snowfall fraction were calculated from the catchment-averaged precipitation and temperature. The daily mean meteorological drivers include the daily mean value of catchment-averaged precipitation, temperature, and day length during 1950–2020. Daily mean river discharges were calculated by using the available discharge records during 1950–2020 and they have been represented in mm/d to exclude the effect by catchment size. Each grid represents the correlation between two variables across the 1,077 catchments, with the dark red or dark blue color denoting strong positive or negative correlations.*

We further disagree with the argument that including static catchment attributes will not impact the interpretation. Taking the example given by the reviewer, suppose basin A is located in a mountainous region with a steep slope, basin B is situated on a snowy plain, and peaks in basin A are overall higher than peaks in basin B. When one model predicts peak 1 in basin A and peak 2 in basin B, we would say that not only the variance in meteorological predictors, but

also the difference in catchment attributes, have explained the variance in the two peaks. In that case, it is hard to separate the effects of meteorological predictors from static catchment attributes especially when the number of catchments increases, which will come with a lot of spatial and temporal confounding. For instance, the weight of meteorological predictors will be confounded by catchment attributes due to their interactions. For the reasons outlined above, we believe local modeling is more suitable for the present study, which aims to make inferences about local catchments and avoids confounding and multicollinearity resulting from static catchment attributes.

Note here that we do not deny the value of regional modeling for making inferences. Instead, we do recognize that a regional model may also generate interesting scientific insights related to a different research question, and we had planned this for future work to avoid overloading the paper.

Reviewer 2: "...some studies have suggested that training a regional model for all catchments at once may be a better practice (e.g., Nearing et al., 2021)."

I come back to this point, because it sounds like this is just a "suggestion". Let me tell you from experience from the operational side (not only at Google but various companies/agencies) but also from literature: I am not aware of anyone that applies LSTMs that is not using regionally trained LSTMs. In my view, there is really no discussion happening around whether you should use regionally or single basin calibrated models. Again, it is not only about better performance but also why the model can achieve this higher performance (see list above). I don't want to step on someone's toes but from my point of view, the reason why people are sticking to single basin LSTMs is because it was hammered into their head for decades that single basin models are better at capturing local processes than regional models, which is true for conceptual models but it is not anymore for LSTM-based models.

Response: Again, we agree that regional modeling is better suited for prediction tasks, as was the practice in industry and a lot of the literature. In the present study, however, reliable inference on the effect of predictors is prioritized, which means we have to consider multicollinearity, confounders, etc. besides accuracy. These factors might not affect the predictive accuracy of regression methods, but they would impair interpretation. Basically, the fact we choose local modeling is a consequence of the objective of the study.

To better clarify the reasons why we chose local modeling instead of regional modeling, as well as the real challenge for future studies of using regional modeling for the investigation, we rewrote the discussion in the section "*Limitations and outlooks*":

"*In this study, we trained LSTM models in a local fashion (i.e., training the model individually for each catchment), rather than a regional fashion (training a single model across multiple catchments), since the main objective of the study is to identify distinguishable patterns of meteorological variables' contributions at local scales. From a prediction standpoint, particularly for unprecedented events and ungauged basins (Nearing et al., 2021; Frame et al., 2022), regional modeling may be a better choice because it is capable of learning more*

general relationships from a larger variety of hydrological data (Kratzert et al., 2019b). However, for the regional modeling, both meteorological time series and static catchment attributes are used as inputs to distinguish response behaviors across time and space. Adding such static attributes would introduce substantial multicollinearities among the considered variables (see Fig. S7 in the Supplementary Material for illustration). Multicollinearity might not be a problem for ML models when they are used for prediction, as long as the collinearity between variables remains stationary (Dormann et al., 2013). Nevertheless, for our study that aims to interpret the effects of predictors on responses, high multicollinearity in predictors indicates considerable information may be shared among the collinear sets. This would result in difficulties in separating the physical effects of these variables – this is also the case in traditional regression models (Hartono et al., 2020). Therefore, interpreting flooding mechanisms with regional LSTM models may become more challenging than with local LSTM models that use only meteorological time series, since some catchment attributes would confound the interpretation. In this study, we therefore employed simple local models, which avoids confounding and multicollinearity resulting from static catchment attributes. However, in light of the benefit of regional modeling that can provide insights into how flooding mechanisms vary spatially by geographic and climatic characteristics of catchments, how to deal with these challenges in the interpretation merits more exploration in future studies." (**lines 565-582**)

References:

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., and Gupta, H. V.: What Role Does Hydrological Science Play in the Age of Machine Learning?, Water Resources Research, 57, 2021.

Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., Qualls, L. M., Gupta, H. V., and Nearing, G. S.: Deep learning rainfall–runoff predictions of extreme events, Hydrol. Earth Syst. Sci., 26, 3377-3392, 2022.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrology and Earth System Sciences, 23, 5089–5110, 2019b.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., and Lautenbach, S.: Collinearity: a review of methods to deal with it and a simulation study evaluating their performance, Ecography, 36, 27-46, 2013.

Hartono, N. T. P., Thapa, J., Tiihonen, A., Oviedo, F., Batali, C., Yoo, J. J., Liu, Z., Li, R., Marrón, D. F., Bawendi, M. G., Buonassisi, T., and Sun, S.: How machine learning can help select capping layers to suppress perovskite degradation, Nat. Commun., 11, 4172, 2020.

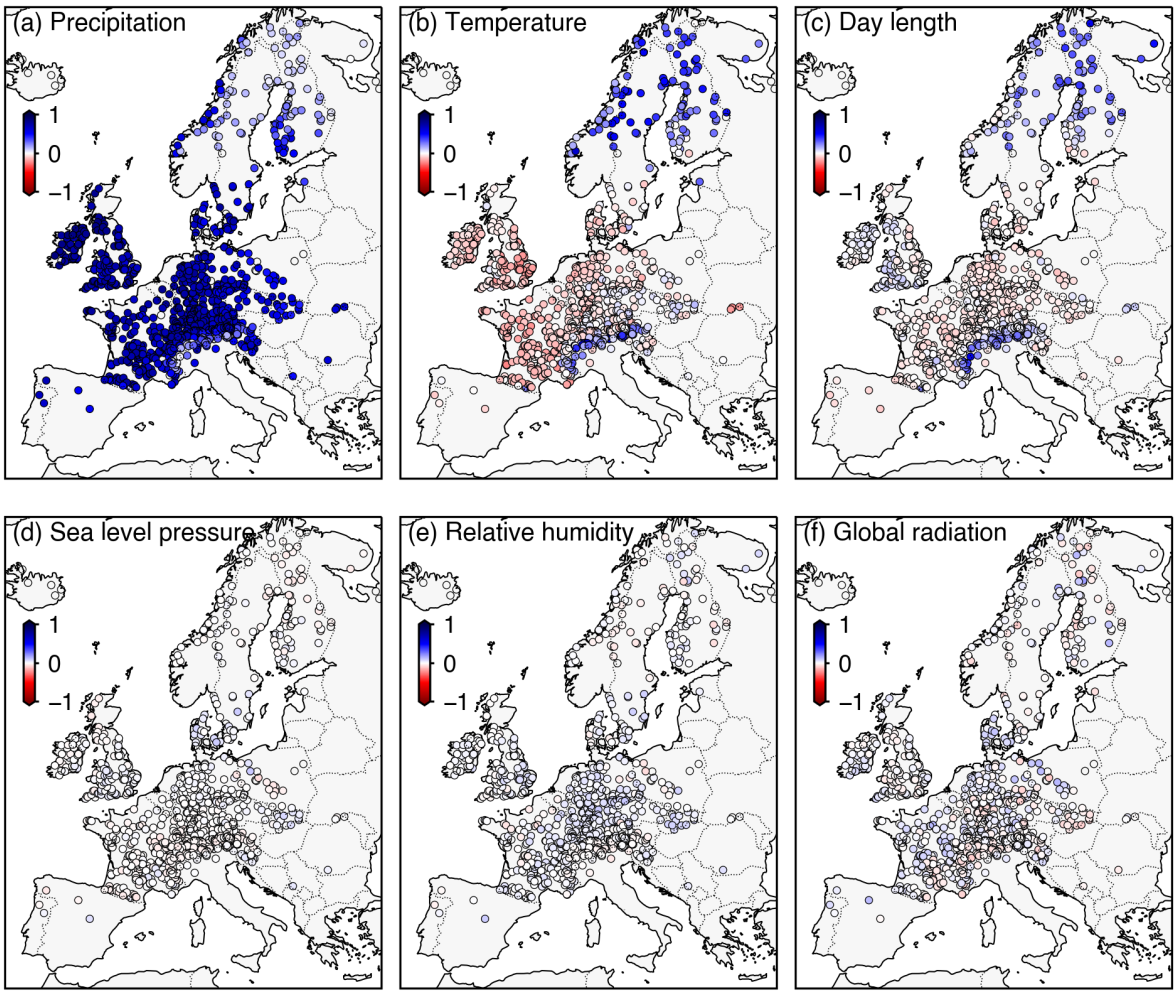
Reviewer 2: 2. Input feature selection

"We agree that including more inputs is beneficial to uncovering patterns related to flood mechanisms that are likely to be overlooked. During our preliminary tests, we had run models with daily averaged sea level pressure, relative humidity, and radiation as additional inputs, which did indeed lead to more clusters in terms of feature importance patterns"

Reading that you already did these experiments, which led to different results, but decided to not include them because they probably don't agree with the "known" patterns is in my opinion simply sad. I agree with the first reviewer on something he mentioned in a slightly different context, which is that you artificially limit yourself in this study. You have the potential to report new findings and even if you can't explain them, you could still present them and potentially start a new discussion in this area. Instead you limit yourself and the model to only check if your method identifies the same patterns as previous studies with different approaches.

"The performance did not drop much by not including more variables" I think this is not important here, but as you said yourself, the clusters of flood generating processes changed. So which one is "the truth"? I have nothing to add to this point and ultimately, it is the decision of the authors. I think it is just a missed opportunity and generally, I am not happy with the fact that all major reviewer comments are put into a new "Limitations and outlooks" section and that this suffices.

Response: The same argument as above holds for the selection of input features. Performance is not everything, and our goal is interpretability, which is not necessarily improved by more input features. For example, using the rigorous data splitting strategy in line with the third major concern, we have re-run the models by including daily averaged sea level pressure, relative humidity, and global radiation as inputs in addition to precipitation, temperature, and day length we used. In this round, adding the three kinds of variables still did not substantially improve the predictive accuracy. The mean NSE values for testing data in the 10-fold cross-validation changed from 0.641 to 0.644, while the median dropped from 0.716 to 0.702. We argue that the fact that performance did not change much is not unimportant here. Instead, it implies the newly added input features may not provide additional information to the model prediction. This can be validated by the new **Figure S8**, which shows the average of normalized aggregated contributions for the respective input variables in each catchment. The results indicate that compared to precipitation, temperature, and day length, the new input features do not show substantial contributions.



(new) ***Figure S8****. The average normalized aggregated contribution for respective input variables in each catchment if sea level pressure, relative humidity, and global radiation were added into the model in the main text in addition to precipitation, temperature, and day length. The sea level pressure, relative humidity, and global radiation were retrieved from E-OBS dataset (cited in the main text) and they were processed as catchment-averaged time series by the method described in the main text. Model settings and interpretations are also the same as in the main text, where only precipitation, temperature, and day length are used. Here the aggregated contribution indicates the contributions of a variable in all the180 days to the target peak discharges (i.e.,* $\Sigma_1^{180}\bar{\phi}_i$*, see the notions in Section 2.3 in the main text). The aggregated contribution for each variable has been normalized per peak discharge for comparability. The color reflects the average value of the normalized aggregated contribution for respective input variables in each catchment. Darker colors indicate that the specific variables have a greater impact (either positive or negative) on peak discharges. The figure implies that the sea level pressure, relative humidity, and global radiation are less important features.*

In spite of these interpretation results, we won't assert that the three variables have no effect on flooding, since the used interpretation technique (i.e., integrated gradients) does not measure how important a feature is in the real world, but simply how important a feature is to the model. It is perhaps the multicollinearity in the variables that makes some variables (e.g., radiation)

redundant to the model, but this might not be the case in reality. As a result, the interpretation results would need to be handled very carefully if no adequate physical knowledge exists to justify their inclusion. Therefore, instead of using more predictors that result in less interpretability, we restricted ourselves to few input features (temperature, precipitation, and day length) whose effect can be relatively easily interpreted and understood. We do admit it is a limitation of the present study and we have not negated the value of including more input features to potentially discover unknown patterns, as we already stated in the "*Limitations and outlooks*" section. However, our manuscript is not intended to cover everything about unknown patterns and implied mechanisms, for which a comprehensive investigation and understanding would fall outside the scope of the present study. For better clarification, we improved our statement in the "*Limitations and outlooks*", which now reads:

"*The multicollinearity also exists in meteorological drivers at daily scales, which requires careful handling of the interpretation results if adding more predictors. For example, radiation is usually an important driver of snowmelt that favors flooding (Merz and Blöschl, 2003), but the interpretation method might not assign it high importance when it is combined with day length as an additional predictor due to the high correlation between the two variables (see Fig. S8 in the Supplementary Material for an example). This is because the used interpretation technique does not measure how important a feature is in the real world, but how important it is to the model. Therefore, it is not necessarily better to add more input features to a model in terms of process understanding, which can be even misleading if the interpretation results are not justified by sufficient physical knowledge (Kroll and Song, 2013). In this study, instead of using more predictors that result in less interpretability, we restricted ourselves to few input features whose effect can be relatively easily interpreted and understood. Therefore, we only selected daily precipitation, temperature, and day length as meteorological inputs, the combination of which results in uncovering three well-known flooding mechanisms. The results are physically interpretable and comparable with findings from other studies that used classical methods. Incorporating more meteorological drivers into the model might, in theory, allow for the identification of additional flooding mechanisms that may be overlooked. However, multicollinearity and confounding can pose a challenge to interpretability, especially when the recognized patterns cannot be linked to fundamental physical processes. Therefore, we leave how to resolve the trade-off as an open question for future studies.*" (**lines 584-598**)

Reference:

Merz, R. and Blöschl, G.: A process typology of regional floods, Water Resources Research, 39, 2003.

Kroll, C. N. and Song, P.: Impact of multicollinearity on small sample hydrologic regression models, Water Resources Research, 49, 3756-3769, 2013.

Reviewer 2: 2. Data split I think this is probably the most worrisome point to me. I think the applied data splitting (random in time) introduces a severe data leakage. You are not testing your method on unseen data, not even in these cases where you predict on a "test sample". The

reason is that if you randomly select timesteps to be train/test data then three adjacent timesteps could e.g. be "train - test - train". Since each timestep is predicted from an input sequence of 180 days, e.g. the first train sample and the test sample only differ in a single time step of data. And without any doubt, the discharge data is highly auto-correlated. In Figure S1-S3, it is e.g. possible that for the models for which this time step appears to be in the test data (dashed lines) the previous and next timestep of that event is a training step. And in this case you can not at all argue that this is independent test data. And that these plots show that the signal is the same for all 10-folds could be just because of this effect, because there is not really any point in your input time series that wasn't seen during training of any of these models.

"Firstly, runoff data available in the GRDC dataset is not temporally complete in many catchments in Europe, with missing data sometimes occurring for several months or years irregularly. This complicates carrying out a unified temporal k-fold cross-validation across these catchments." This is related to what I wrote above: Just because something is "more complicated", this doesn't mean you shouldn't do it. In fact, it isn't that hard to loop over basins and do time series splits per basin. How difficult it is to include different data splits for each basin in your training pipeline depends on your code. It is a built-in function in the open source library NeuralHydrology (https://github.com/neuralhydrology/neuralhydrology/ Disclaimer: I am one of the developers) and would work out of the box if you would use this for training your models.

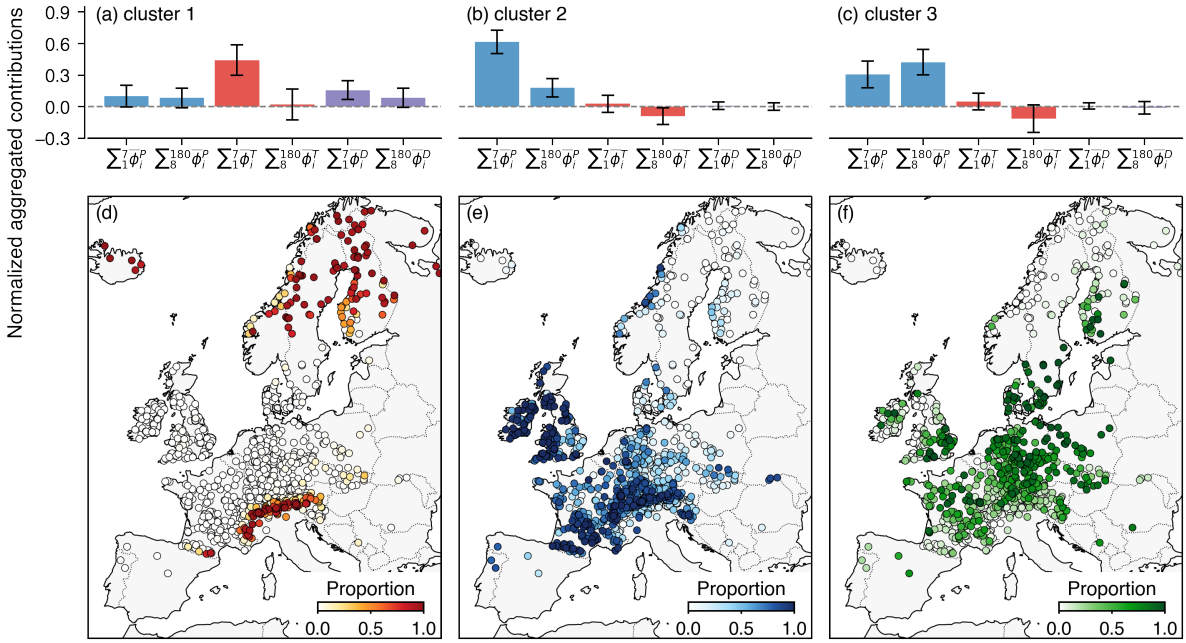
"We should emphasize that the model was used for statistical purposes instead of a prediction task, thus the split of the training dataset and the testing dataset is only to ensure the model has learned a generalizable relationship between variables." I don't understand your point here. For any kind of statistical analysis with any model (data driven or not) you want to make the analysis on an independent dataset. The point is, that your test dataset is not independent of the training dataset as there is data leakage.

"The generalizable relationship should hold not only for the testing dataset but also for the training dataset." This is not necessarily true. First, in the extreme case you could overfit on the training data, meaning your model remembers every sample and thus is not generalizing at all. Second, have you ever looked at e.g. the NSE of an LSTM during the training period and compared this to the test period (with a non-random splitting). You will see that the LSTM achieves a much higher NSE during the training period and I would be more than cautious to draw any conclusions from this on the models generalization capabilities.

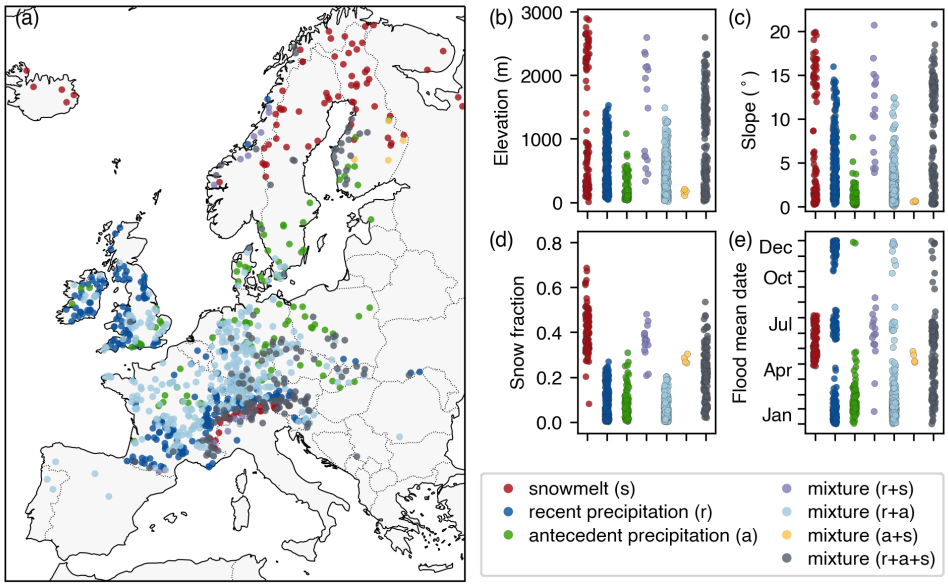
Response: On this point, we agree that the previous random splitting approach is less rigorous than the one suggested by the reviewer. Therefore, we re-ran our models with rigorous 10-fold cross-validation. In **lines 164-166**, the method description was modified as "*To improve the robustness of model evaluation and analysis, we fitted 10 independent LSTM models for each of the 1,077 catchments. Specifically, the data for each catchment was divided into 10 folds without shuffling the temporal sequence, and each fold was tested once with a model trained with the remaining 9 folds.*"

In the more rigorous experiments, the median NSE for the 1,077 catchments drops to 0.72 and the number of catchments to be analyzed is now 943 (using an average NSE above 0.5 in the testing periods as a criterion). However, the conclusions about the proportion and trends

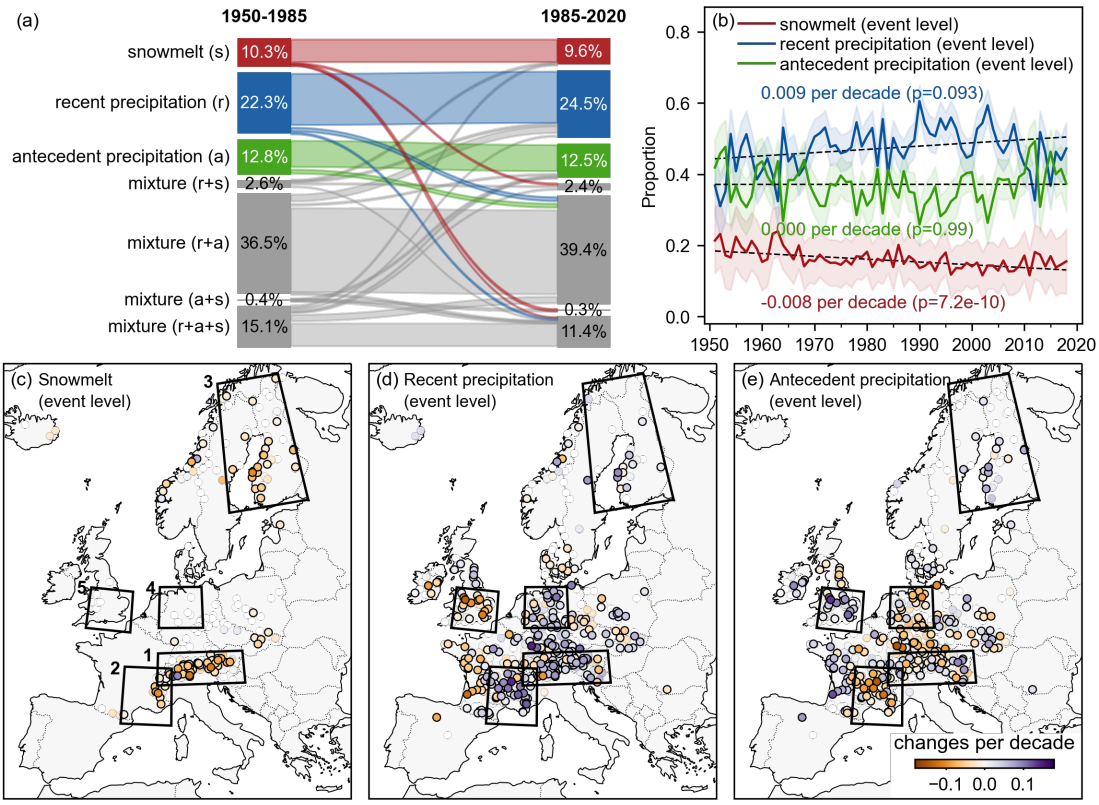
(including the overall trends and those in specific hotspots) are consistent with the original manuscript, except for the minor changes in numbers (see revised Figures 5-8 below). Besides, using the IG scores based on the peaks in testing datasets alone does not yield substantial impacts on our conclusion in subsequent analyses, either (see new **Figure S4-S5** below). In the revision, we have updated the manuscript throughout with the new results.



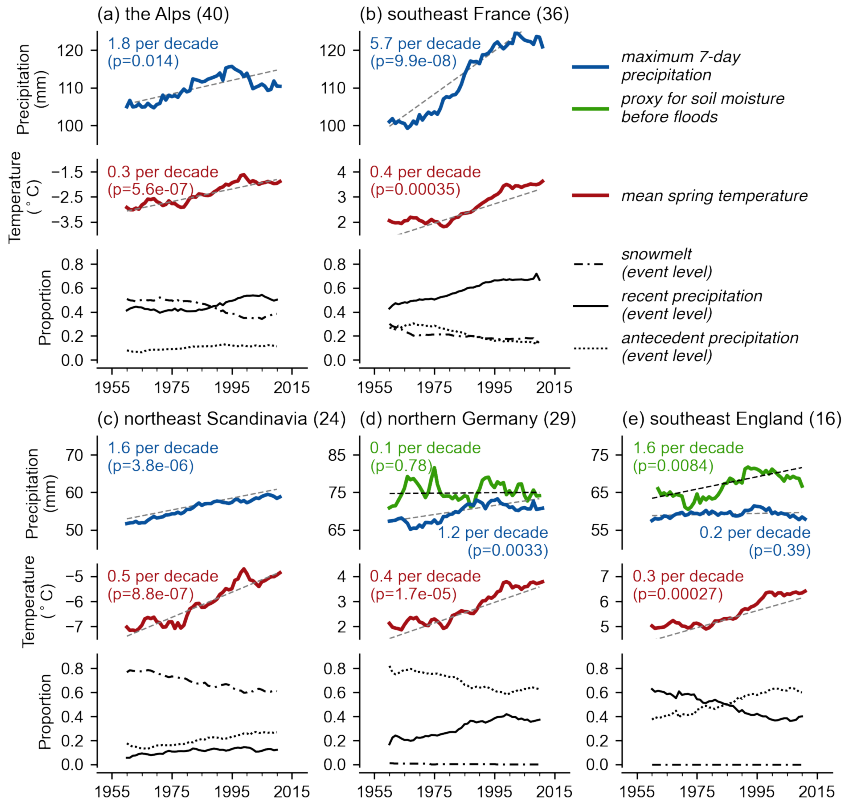
(revised) ***Figure 5****: The cluster centroids and variance for the three clusters and their respective proportions of all peak discharge events in each catchment….*



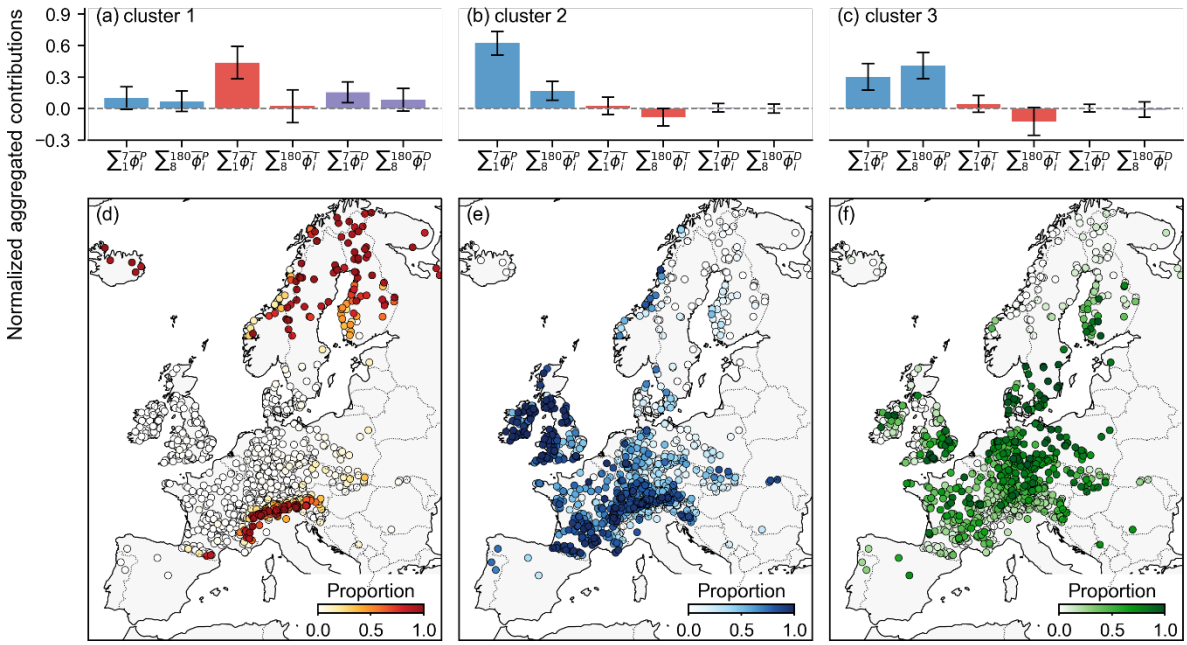
(revised) ***Figure 6****: The dominant flooding mechanisms and their relevance to catchment attributes and seasonality….*



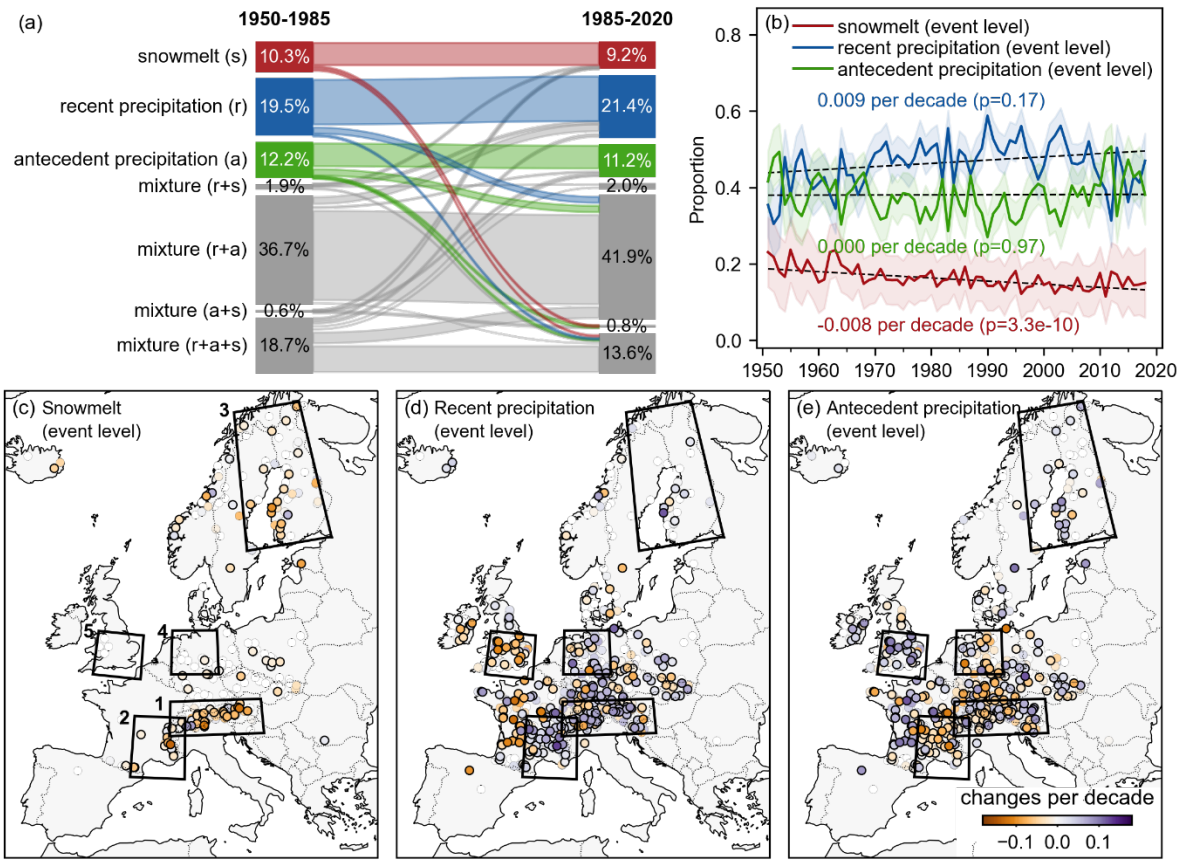
(revised) ***Figure 7****…*



(revised) ***Figure 8****: The temporal changes of the event-level mechanisms in relevant catchments within the five selected regions…*



(new) ***Figure S4****: The same case as in Fig. 5 in the main text, but we use the IG scores based on the peaks in testing datasets alone to perform cluster analysis. The events identified with snowmelt, recent precipitation, or antecedent precipitation as the primary causes account for 16.6%, 47.7%, and 35.7% of all the 52,247 annual maximum peak discharges, which is only slightly different from using the averaged IG scores from the 10-fold models for individual peaks.*



(new) ***Figure S5****: The same case as in Fig. 7 in the main text, but we use the IG scores based on the peaks in testing datasets alone to perform cluster analysis and trend analysis.*