This manuscript proposes a new method for classification of flood generation mechanisms using machine learning that provides the information on the importance of different indicators on the generation of the particular flood event. The method has a potential to overcome the subjective choice of classification thresholds of the previously developed methods. It was tested across European catchments with particular focus on how flood mechanisms were changing in the past decades.

The proposed method has certainly some clear advantages compared to the previous methods and provides a new perspective on classification of flood events. It a very well-written manuscript and I have enjoyed reading it very much. I am certain that this is a substantial contribution to the current knowledge on flood generation processes and their changes. However, although the proposed method has the potential to avoid subjective thresholds, I do not think that the authors have succeeded in overcoming this issue completely. A more prominent attention has to be paid to this issue in the manuscript. I also have some minor suggestions that might help to improve the manuscript and clarify its novelty. Please see my detailed comments below.

Kind regards,

Larisa Tarasova

Response:

We thank the reviewer for the thoughtful comments and suggestions which we believe will greatly improve the manuscript. In the revision, the limitation regarding subjectivity has been emphasized appropriately. The replies to the comments follow.

General comments

Abstract: I think the abstract puts too much focus on the changes in flood mechanisms in Europe that is not the most novel findings and instead fails to elaborate on the machine learning approach used here and its advantages compared to previously existing methods. The method implemented in this study is the real novelty, while there are already several studies on changes in flood generation processes in Europe. Therefore, I suggest the authors to consider to put more stress on the methodological aspects in abstract to show how this study stands out.

Response:

Thank you for pointing this out. The explainable machine learning methodology was originally proposed in our previous paper (https://doi.org/10.1029/2021WR030185), and the present manuscript aimed to apply the developed framework to tackle practical problems (i.e., the changes in flood mechanisms in Europe). Therefore, we focused more on the scientific aspect instead of the methodology itself. However, your suggestions are appreciated, and we will make appropriate modifications to highlight more on the methodology by adding "*Recent years have witnessed the increasing prevalence of machine learning in hydrological modeling and its predictive power has been demonstrated in numerous studies. Machine learning makes hydrological predictions by recognizing generalizable relationships between variables, which, if explained properly, may provide us with further scientific insights into hydrological processes*". Moreover, we will modify the last sentence in the abstract as "*Overall, the study provides a new perspective on understanding changes in weather and climate extreme events by using explainable*

machine learning and demonstrates the prospect of artificial intelligence-assisted scientific discovery in the future."

Selection of thresholds: The proposed methodology has a very strong advantage that it can avoid arbitrary decisions on how the indicators and their threshold are selected for the event classification. However, the authors did not avoid that issue as they have selected the periods for which the effect of recent and antecedent precipitation was accumulated to avoid additional computational effort. This pragmatic choice is understandable and is in line with the subjective choices previous classification studies were making, but it has to be properly stated in the manuscript and a sensitivity analysis on the effect of this choice on the results of the study will be very welcome. Please also see my detailed comments to the corresponding part of the manuscript.

Response:

Thank you for pointing this out. We agree that choosing a 7-day window to separate between antecedent and recent precipitation will introduce subjectivities and uncertainties, as we admitted in the original manuscript "The method has reduced the need for accurate catchment wetness estimates, yet such uncertainty is not eliminated completely, particularly since we chose a 7-day window to separate between antecedent and recent precipitation".

Your suggestion about analyzing the sensitivity of the selection of the separating window is appreciated. In the revision, we supplemented an analysis that uses a 5-day window to separate recent contributions and antecedent contributions, and we found that the main conclusion is not affected by this change. We added the analysis into the new section "3.6 Limitations and outlooks", as:

"*In the clustering procedure, we used a 7-day window to aggregate the daily IG scores into a low-dimensional contribution vector for the sake of efficiency in clustering lengthy time series, which could induce evitable uncertainties and subjectivity. However, additional tests indicate that our findings will not be compromised if we separate contributions of a variable in recent days and an earlier antecedent period by using a 5-day window, which is also a common interval to consider flooding drivers (e.g., Rottler et al., 2021). Based on the 5-day window, the events identified with snowmelt, recent precipitation, or antecedent precipitation as the primary causes account for 15.0%, 47.9%, and 37.1% of all the 55,828 annual maximum peak discharges, which is only slightly different from using a 7-day window. As for the three mechanisms in individual catchments, decreasing the window length has the least impact on identifying snowmelt-driven floods, with the absolute changes in their proportions within 1% for 84.5% of catchments and within 5% for 98.7% of catchments. In comparison, the proportion changes for two other flooding types are more sensitive, with changes within 5% for 83.2% (82.7%) of catchments in terms of recent (antecedent) precipitation-driven flooding. However, this does not affect the conclusion regarding the respective trends in flooding mechanisms (see Fig. S4 in the supplementary material), indicating the robustness of the methodology. Despite this sensitivity analysis, we would like to emphasize that the selection of the separating window remains somewhat subjective, and further exploration is needed to avoid a possible bias due to arbitrary judgments in identifying flooding mechanisms.*"

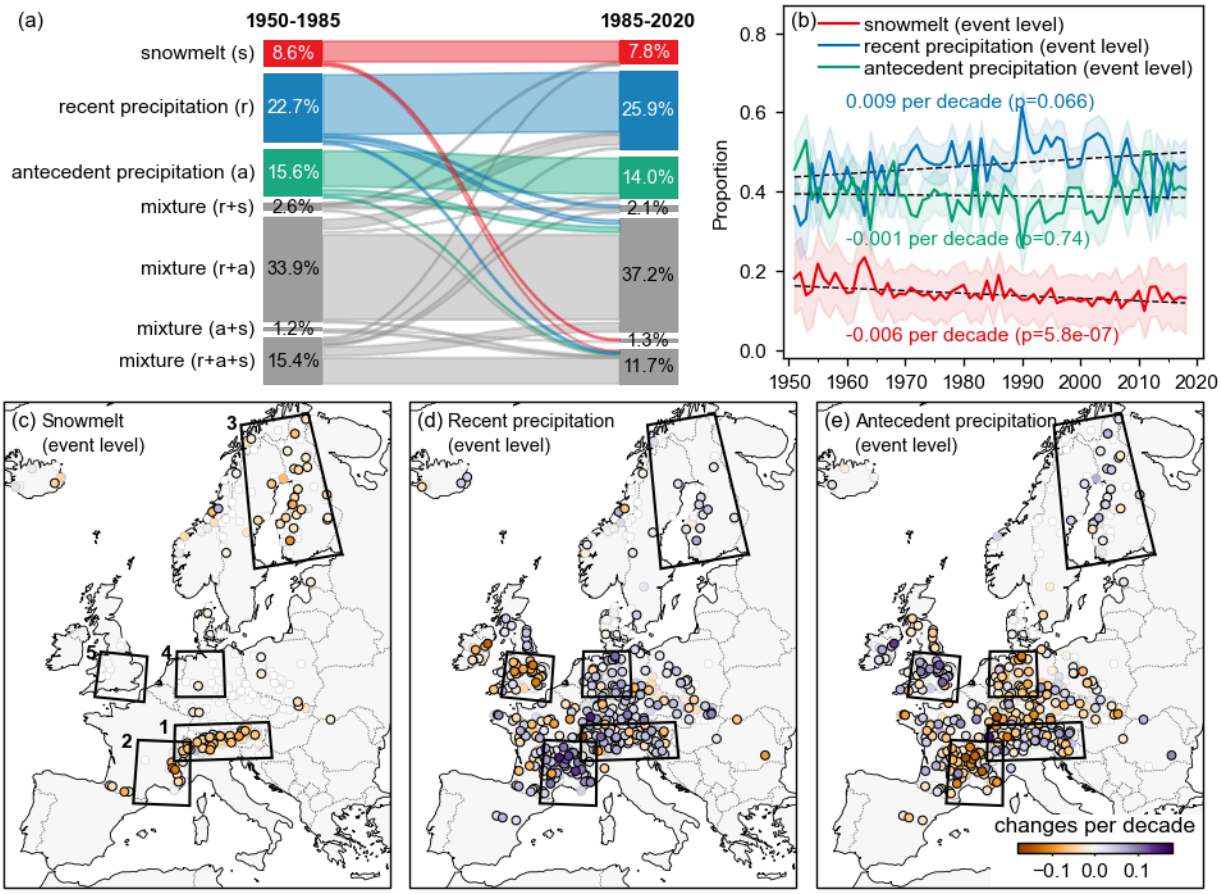

Figure S4: The same case as in Fig. 7 in the main text, but a 5-day window was used to separate contributions of a variable in recent days and an earlier antecedent period.

Detailed comments

Line 40-42 I suggest to also mention here the study of Kemter et al (2020) on changes of flood mechanisms in Europe and global analysis of Stein et al (2020)

Response:

Thank you for the suggestion. We will add "*Using a multicriteria approach, Kemter et al. (2020) identified the flooding mechanisms in Europe by classifying approximately 174,000 flood peaks and revealed their trends over the past 50 years. Likewise, Stein et al. (2020) analyzed flood events over 4,155 catchments worldwide and classified each event into one of 5 hydro-climatological flood generating processes*."

Line 48-49: Here I miss mentioning the study of Kemter et al (2020) that did exactly that.

Response:

Please refer to our response to the previous comment.

Line 88: Please indicate if the size of catchments was limited to avoid the effect of human influence or was there any other reason for this selection?

Response:

Thank you for the suggestion. We will add the reason: "*The catchment areas range between 8 km^2 and 10,000 km^2, with overly large catchments being excluded where the effect of spatial heterogeneity of flood drivers tends to be substantial*."

Line 100: Please indicate also the lower boundary of catchment sizes to clarify if the size study catchments comparable with the spatial resolution of the hydrometeorological datasets.

Response:

Thank you for pointing this out. In addition to indicating the lower boundary of catchment sizes (see the response to the previous comment), we will add a clarification "*Note that smaller catchments under 100 km^2 (approximately 0.1° × 0.1°) may encounter unexpected uncertainties due to the relatively coarser spatial resolution of the meteorological data. Nonetheless, those catchments with large uncertainties will not be considered for the subsequent attribution analysis if ML models cannot capture the relationship between inputs and outputs accurately*."

Line 103: Please elaborate how the catchment boundaries from two datasets were merged. Are they identical?

Response:

We will add the clarification "*The catchment boundaries were obtained from readily available GRDC (Lehner, 2012) and GSIM (Do et al., 2018) databases, with GRDC being prioritized when the boundary of a catchment was available in both databases.*"

Line 104-106: I miss here a more motivated choice for the day duration as an indicator for classification. It seems to me that essentially it is a combination of the location and day of the year information. Please provide more information on the nature of the preliminary test performed, particularly if other potential indicators were examined.

Response:

In the revision, after "Day length was included in the study since it was shown to improve model accuracy in a series of preliminary tests", we will add "*..., including the cases where only precipitation and temperature were used or day length was additionally incorporated. Catchments with obvious accuracy improvements are primarily located in northern Europe and the Alps*." The role of day length is as we explained in the original manuscript: "The role of day length implies that the magnitude of these peak discharges can be partially explained by the seasonality presented by day length, which peaks around the June solstice."

Figure 2: The interpretation arrow is not so clear, why does it return back to the input layer? At this point in the manuscript the meaning of the integrated gradients for the features is not yet explained and looks confusing in this Figure. Please add clarification in the caption. Consider indicating the target maximum annual flood in panel a as a point and not as a window. The panel c is rather confusing as there is only one event is being displayed in the panels a and b and the cluster plot is not set in any particular space (i.e., the axes are not indicated). Consider omitting this panel, I think that idea of clustering is understandable without this example only brings more confusion.

Response:

Thank you for the suggestions. We removed panel c in the figure. Replacing the window that shows the target peak sample with a point may be confusing because it contains both the observed (black) and predicted (orange) values. To clarify the used window, we will add "*The window in the time series of discharge highlights the target output (which is a point)...*" into the caption instead. We further simplified the arrows used and rewrote the caption. The figure and caption now read:

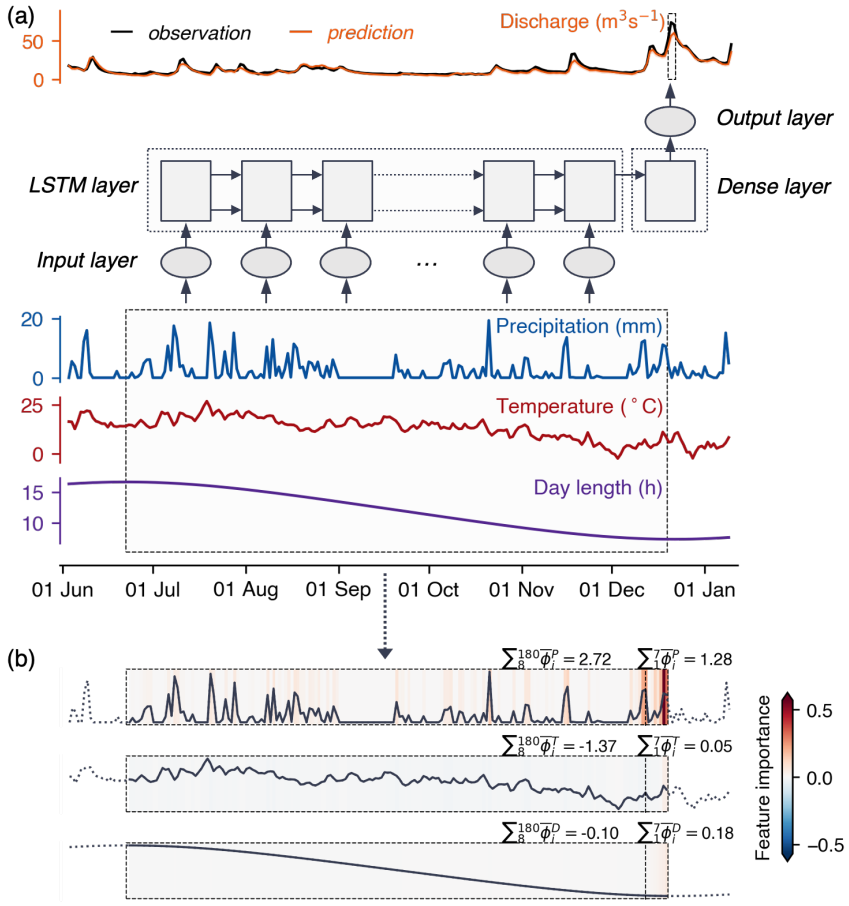

Figure 2: The workflow of using explainable ML methods for attributing flood peaks (annual maxima of river discharge) to their drivers. (a) Diagrammatic representation of the used LSTM models. The window in the time series of discharge highlights the target output (which is a point) and the window in the inputs indicates the input features used to predict the illustrated peak discharge sample. (b) The feature importance of the inputs for predicting the peak discharge shown in (a), which was obtained by using the ML interpretation technique (namely integrated gradient). The vertical dashed lines separate the feature

importance into a recent 7-day period and an earlier period to calculate the aggregated feature contributions (see main text).

Line 139-144: I agree that presenting LSTM model in detail is not necessary here, but I think I more detailed explanation on how the structure of LSTM suited for capturing short-term and long-term interaction will be very helpful here, as it can provide the readers with the insights on why particularly this method is more applicable than classifications that are based on subjectively selected thresholds.

Response:

Thank you for the good suggestion. We will add a more detailed explanation as: "*The effectiveness of the LSTM is partially due to the comparability of its formulation to the hydrological behavior of a catchment. Specifically, the backbone of the LSTM network is formed by recurrent cells that can store previous information from input sequences, which is conceptually similar to the way meteorological information (e.g., precipitation) is stored in the form of soil moisture or snow depth (Lees et al., 2022). The physically realistic mapping from inputs to outputs facilitates gaining hydrologically meaningful insights from subsequent model interpretations.*"

Line 145: It is not clear. Does it mean for each catchment? Please clarify.

Response:

The original sentence will be revised as "*To improve the robustness of model evaluation and analysis, we fitted 10 independent LSTM models for each of the 1,077 catchments.*"

Line 153: What is the sample in this case? Maximum annual floods? Please clarify.

Response:

The original sentence will be revised as "*...which allows for obtaining the time-wise feature importance of the three input variables for each sample of the output (i.e., daily discharges).*"

Line 185-187: I do understand authors' arguments on why they had to make this decision and restrict quantification of the effect to 7 and 180 days only. Although, I find it somewhat disappointing. The authors have stated earlier in the manuscript the main advantage of the proposed ML-based method is that one can avoid selecting subjective indicators and their thresholds. In my opinion selecting here 7 and 180 days is nothing else but exactly that kind of subjective threshold that partially impairs the main advantage of the method. If clustering indeed is very time consuming (which is actually surprising to me as in my experience k-mean clustering is not the most time consuming procedure and computational power is hardly a limitation with cluster resources available) at least a sensitivity analysis has to be performed to analyze how the selection of these thresholds affects the results.

Response:

Thank you for pointing this out. Please refer to our response to the general comment.

For the efficiency of clustering time series, K-mean clustering with Euclidean distance metric is indeed not a time-consuming procedure, but we have to point out that time series clustering tends to be much more complex as it usually needs Dynamic Time Warping (DTW) as the distance metric. The DTW has a quadratic complexity with respect to the length of sequences (in our study it is 180) compared with the linear complexity of Euclidean distance. Furthermore, in our preliminary tests, we have tried clustering the whole time series as did in our previous study that first attempted such methodology in the US catchments (https://doi.org/10.1029/2021WR030185). The preliminary results that we obtained are similar to those reported in the present study (three clusters). However, since we anticipate the methodology to be used in a larger-scale analysis in future studies, in this study, we improve the efficiency of the original framework proposed in our previous study by introducing a separating window.

Line 188-189: I cannot agree with this statement. The duration will be strongly affected by catchment size and mechanism. The build-up period of snowmelt floods in larger catchments can take up to several months. I also do not think that the provided reference is up to date. Please revise.

Response:

Thank you for pointing this out. We were meant to describe the hydrological response time to precipitation and snowmelt events, instead of the build-up periods. We will rewrite our original statements to justify the selection of 7 days, as "*The separating interval has to cover the period of precipitation and snowmelt leading to each peak discharge, which depends highly on the local characteristics. Following a check of the relationship between catchment area and mean event response time, Stein et al. (2020) suggested a synoptic window of 7 days should be sufficient to guarantee the response time for large catchments. As a result, this study used a 7-day period, similar to the practice in most studies that examined flooding causes (e.g., Blöschl et al., 2017; Berghuijs et al., 2019). However, using a shorter period (e.g., 5 days) will not affect subsequent conclusions about dominant flooding mechanisms and their trends (see discussion in Section 3.6).*"

Line 189-190: It is consistent with previous studies, but they also did not examine if these thresholds are appropriate. Please revise.

Response:

Please refer to our response to the previous comment.

Line 192: It is not clear what is the role of multiple-peak discharges here and how they were considered. Please clarify.

Response:

Sorry for the confusion. Rather than multiple-peak discharges, we referred to multiple peak-discharges here. To avoid ambiguity, we will remove "multiple" in the revision.

Line 197-203: Please clarify if clustering procedure was performed for all catchments simultaneously or if they were considered individually. If it was performed simultaneously for all catchments, does it mean that if a catchment has very local and specific mechanisms they likely not to be detected by the procedure?

Response:

Yes, the clustering procedure was performed for all peak discharges pooled from all catchments simultaneously. To clarify it, the original sentence will be revised as: "*To obtain an overall picture from the individual aggregated feature contributions, we used the K-means method to cluster the results for all annual maximum peak discharges pooled from the 1,009 catchments*." In the section presenting the cluster results, we further added "*Note that the clustering results reflect only major patterns widespread in data, with certain local and specific mechanisms unlikely to be detected*."

Line 206, 277, 402 and elsewhere: Are these maximum annual peak discharges? If yes, please indicate it clearly here and elsewhere.

Response:

Thank you for pointing this out. We will specify the peak discharges as "*annual maximum peak discharges*" throughout the manuscript.

Figure 3: Does this figure display NSE only for annual maxima or for the complete streamflow time series? Please clarify.

Response:

Thank you for pointing this out. In the figure caption, we will add "*The NSE values were calculated using all samples in respective testing datasets.*"

Line 294: I think it is a rather a stretch to call streamflow generation that occurs due to excess of soil storage capacity and heavy precipitation as we cannot guarantee that heavy precipitation generates overland flow. In case it is first contribute to increase of soil moisture storage the physical process of streamflow generation will be the same for both drivers. Please revise.

Response:

Thank you for pointing this out. We will remove this statement in the revision.

Line 303 and elsewhere: I think "mixed mechanisms" is not an accurate term here as it refers to the occurrence of different mechanisms in the same catchment, but not necessarily simultaneously. Consider using "mixture of mechamisms" instead.

Response:

Thank you for the suggestion. We will replace "mixed mechanisms" with "*mixture of mechanisms*" throughout the revised manuscript.

Figure 6: Please add an explanation for the mixtures in the caption. Please also clarify how the classes for two processes are formed, do the corresponding two processes have to generate more than 70% of annual maxima?

Response:

We defined a catchment dominated by a mixture of two processes as the case where the difference between the proportions of the two processes is less than 70% (say one is 35% and another is 65%), in order to distinguish it from single mechanisms. We will add an explanation in the figure caption, as: "*Mixture means the associated catchments are dominated by two or more flooding mechanisms. For example, mixture (r+s) indicates either recent precipitation (r) or snowmelt (s) is the primary cause of the annual maximum discharges for the associated catchments, and the difference between the two proportions is less than 70%.*"

Line 338-340: I think it will be helpful to relate here to the findings of Stein et al (2021) (doi: 10.1029/2020WR028300) on the controls of catchment characteristics on the dominance of different flood mechanisms

Response:

Thank you for the suggestion. We will add "*In addition to elevation, slope, and snow fraction, the study by Stein et al. (2021) on catchments in the United States demonstrated that other catchment characteristics (e.g., aridity, precipitation seasonality, and mean precipitation) also significantly influence flood generating processes. An in-depth investigation of how geographic and climatic characteristics affect flood mechanisms in European catchments is expected in future studies.*"

Line 350: Consider using term "pre-defined" criteria instead of "manual" as it is not so clear.

Response:

Replaced as suggested.

Table 1: Consider also adding catchment sizes to the comparison as I expect that there is a difference between these studies also in that regard.

Response:

Will be added.

Line 379: I think the study of Kemter et al 2020 also should be mentioned here.

Response:

We will add the reference as suggested.

Line 381-383: This note would be more helpful earlier before the comparison of the results. Consider moving this part up.

Response:

We will move it up.

Line 385-389: I think it might be worth mentioning here the work of Tarasova et al 2020 (doi: 10.1029/2019WR026951) that investigates how using different data sources for the same indicator affects event classification

Response:

Thank you for providing the useful reference. We will add "*A work worth mentioning is Tarasova et al. (2020), which conducted a rigorous uncertainty analysis of input data for a runoff event classification framework, emphasizing the importance of developing novel indicators to reduce these uncertainties*."

Line 404-407, 427-431: These parts would be more suitable in the dedicated Method section

Response:

Thank you for the suggestion. We moved these parts to the new subsection "*2.4 Trend analysis of flooding mechanisms*" in Method.

Figure 7: Please indicate how many catchments are the basis for Sankey plot in the caption. Please also clarify the origins of the p value in the caption. The information provided on methodological aspect of trends in this caption is not sufficient. Please add a corresponding section in the Methods. Panel b: I am wondering if the results of trend analysis are not so clear due to regional differences in the direction of trends. Looking at the results of Kemter et al (2020) it seems that there are disparate trends for different regions that can be obscured when mixed together. Perhaps something worth mentioning in the corresponding text.

Response:

Thank you for the suggestions.

In the caption, we will add "*The proportions were calculated based on 846 catchments, where at least 15 years of records were available in each period*." and "..., *with their significance being assessed by the modified Mann-Kendall test.*" The details will be provided in the new section "2.4 Trend analysis of flooding mechanisms".

For panel b, before introducing the results of trends in individual catchments, we will add "*Note that Fig. 7b only presents the overall trends in flooding mechanisms at the continental scale, while disparate trends may exist in different regions that could cancel each other out*."

Line 455-459: It would be helpful if this information is provided in the dedicated Method section.

Response:

Thank you for the suggestion. We will move these parts to the new subsection "*2.4 Trend analysis of flooding mechanisms*" in Method.

Figure 8: It is not clear why the lines of the plot do not correspond to the whole extent of time axis. Please clarify or correct. In region 1 and region 2 it seems that there is certain periodicity in the data, it would be helpful if the authors would add a short discussion on suitability of monotonic trends analysis in such cases. Please also consider adding geographical indications for regions instead of numbers. This will make this figure easier to interpret. Please also add the number of catchments in each of the considered regions.

Response:

For the time axis, because the 20-year moving window is used, the range reduces to from 1950-2020 to 1960-2010. In the caption, we will add, "*The proportions were calculated by a 20-year moving window, while precipitation and temperature were smoothed by using a 20-year moving average window, with their values at central positions in time windows*."

For the possible periodicity in data, we will add "*Note that here we merely examined the monotonic trends within data over the 70 years, while the trends may vary piecewise (e.g., the changes in maximum weekly precipitation in the Alps and southeast France), the impact of which on flooding mechanisms deserves further research.*"

For the geographical indications and number of catchments, we will add them to the figures as suggested.

Line 492, 499. Caption of Figure 9: It is not clear which length is meant here. Please clarify.

Response:

In the caption of Figure 9 (now Figure B1 in Appendix B), we explained the length as "*The mean resultant length is a measure in circular statistics between 0 and 1 that reflects the spread of a circular variable, with 0 representing the spread of flood dates evenly distributed over the year and 1 representing the spread concentrated at one day.*"

Line 486-504: This part is not very well connected to the previous narration and provides yet another new results for which methods were not clearly elaborated in the Method section. Consider omitting it or revise.

Response:

Thank you for pointing this out. We will move the description of Figure 9 into the Appendix and simplify the part in lines 486-504, and it now reads: "*A change in flooding mechanisms may affect the seasonality and magnitude of flooding, which might ultimately impair the current flood risk management measures. For example, in catchments previously dominated by snowmelt, increasing floods from extreme precipitation and soil moisture excess may lead to shifted flood mean dates and less concentrated seasonal patterns (as exemplified in Fig. B1 in Appendix B). By simulating daily discharge for a*

reference period (1961–1990) and a future period (2071–2099), Vormoor et al. (2015) predicted that floods in some Nordic catchments could even shift from spring to autumn as rain replaced snowmelt as the dominant flood-inducing process. These results suggest that, in a warmer climate, flood risk predictions in snowmelt-affected catchments should consider the interconnection between changes in flooding drivers and seasonality."

Line 539-543: I would recall here how "recent" and "antecedent" precipitation were defined in this study, because despite what this part claims the definition of these two indicators were set arbitrary by selecting corresponding number of days during which the effect was evaluated.

Response:

Thank you for the suggestion. In the revision, we will update the relevant sentence to read as follows: "*With the ML-captured feature importance of precipitation, temperature, and day length for predicting annual maximum discharges, we aggregated driver contributions in the recent 7 days and an earlier period (back to 180 days) and then applied cluster analysis to group them based on similar patterns.*"

Line 549: The term "perspective of catchment average" is not very clear here without the context. I think it would be clearer to just indicate that these methods did not perform an event-based classification and instead identified one single dominant driver per catchment.

Response:

Thank you for the suggestion, we will modify "...some of which were obtained taking a perspective on catchment averages" as "*...some of which did not perform event-based classifications but rather identified the overall mechanisms within individual catchments*."

Conclusion section: A statement about the dependence of the results on the performance of the ML model for the proposed classification method would be very welcome in this section. Moreover, same as for abstract more focus on the newly developed ML-based classification method instead of changed in the mechanisms will be welcome here to highlight the novelty of this study.

Response:

Thank you for the suggestion. To clearly highlight the novelty, in the first paragraph, we will replace the relevant sentences with "*To investigate whether flooding mechanisms changed in European catchments, this study introduced a novel explainable ML method to identify flooding mechanisms. Compared with conventional classification approaches, where the results are highly dependent on appropriate flood process definitions and sensitive to the selected indicators and threshold parameters, the combination of explainable ML and cluster analysis is able to avoid such predefinitions and reduces subjectivities in identification processes. With the ML-captured feature importance of precipitation, temperature, and day length for predicting annual maximum discharges, we aggregated driver contributions in the recent 7 days and an earlier period (back to 180 days) and then applied cluster analysis to group them based on similar patterns*."

Moreover, at the end of the conclusion section, we will replace the original outlook with an outlook for improving the methodology, and it now reads "*Overall, this study highlights the usability of explainable ML in helping uncover complex and possibly non-linear changes in weather and climate extreme events in the warming Earth system. With more large-sample hydrometeorological datasets becoming readily accessible, one next step is to extend the research to a larger scale for a better understanding of variations in flooding mechanisms globally. Still, many challenges remain for future work, forming exciting research opportunities. For example, the clustering procedure can be improved by adopting algorithms to aggregate daily feature importance adaptively, which would allow the predefined separating window to be avoided while maintaining high efficiency. Moreover, regional LSTM models with static catchment attributes incorporated can be employed to capture the spatial variations in flooding mechanisms and quantify the influence of catchments' geographical and climatic conditions on flooding processes. In addition to the integrated gradient method used in this study, other interpretation techniques might be explored further to uncover potentially valuable information when more input variables are included*."

Line 563-565: I think the authors have to be more cautious here with this statement, because there might be strong regional differences (i.e., there are disparate patterns in precipitation changes in Europe). Moreover, the term extreme precipitation is much more often related to very short precipitation (i.e., less than 1 day), while 7-day long precipitation can substantially affect the storage of the catchment and lead to soil moisture excess floods and hence the resultant magnitude of the flood will depend much more on the initial storage conditions compared to the floods that are generated by short and extreme precipitation. Finally, the authors have examined here maximum annual 7-day precipitation which does not guarantee that this is the same 7-day precipitation sum that have caused a maximum annual flood in the corresponding year.

Response:

Thank you for pointing this out. In the revision, we will remove the relevant statement and replace it with an outlook for improving the methodology. Please refer to our response to the previous comment.

Editorial comments

Line 264: regions with winter snowpack accumulation

Response:

Modified as suggested.

Line 276: catchments associated

Response:

Modified as suggested.