*Throughout this response, the editor's and reviewers' comments are presented in blue, and our responses are presented in black. The line numbers in red correspond to those in the clean version of the revised manuscript.*

## To Editor:

The two reviewers both provided detailed reviews for your manuscript and agree that your work is very interesting and the manuscript well written. I am looking forward to reading a revised version of your manuscript that addresses the points raised by the two reviewers.

**Response:**

Thank you very much for handling our manuscript. We are grateful for the constructive comments from the two reviewers, which are invaluable to us in improving the quality of the manuscript. We have carefully addressed all the comments received.

## To Reviewer 1:

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.

**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 have made appropriate modifications to highlight more on the methodology by adding "*Recently, numerous studies have demonstrated the skill of machine learning (ML) for predictions in hydrology, e.g., for predicting river discharge based on its relationship with meteorological drivers. The relationship, if explained properly, may provide us with new insights into hydrological processes*" (**lines 11-14**). Moreover, we modified the last sentence in the abstract as "*Overall, the study offers a new perspective on understanding changes in weather and climate extreme events by using explainable ML and demonstrates the prospect of future scientific discoveries supported by artificial intelligence*" (**lines 23-25**).

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 explicitly indicated 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 "**3.7 Limitations and outlooks**", as:

"*In the clustering procedure, we chose to use 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 inevitable uncertainties and subjectivity. Despite this, additional tests indicate that our findings are similar when using a 5-day window, which is also a common interval to consider flooding drivers (e.g., Rottler et al., 2021). Specifically, 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*" (**lines 582-595**).

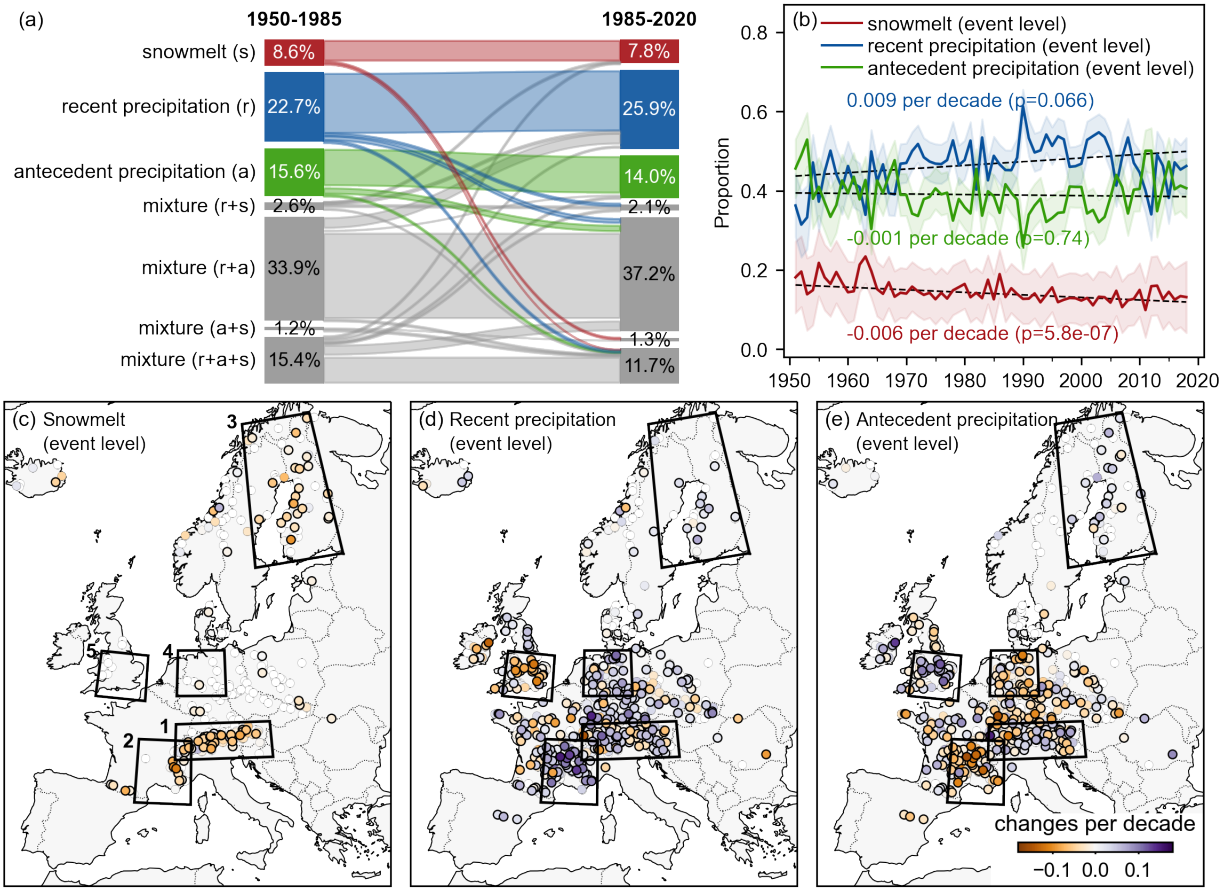


*Figure S4: The same case as in Fig. 7 in the main text, but a 5-day window was used to separate contributions of a variable in recent days and an earlier antecedent period.*

## 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. In **lines 56-59,** We added "*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 them into five 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 added the reason in **lines 95-97**: "*The catchment areas range between 8 km$^2$ and 10,000 km$^2$ — very large catchments, where the effect of spatial heterogeneity of flood drivers tends to be substantial, were not considered*".

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 added a clarification "*Note that for smaller catchments under 100 km$^2$ (approximately 0.1° × 0.1°), uncertainties may exist 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 effectively*" (**lines 110-113**).

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

**Response:**

We added 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*" (**lines 108-110**).

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 added "*..., including the cases where only precipitation and temperature were used and day length was additionally incorporated. Catchments where day length largely improves accuracy are mainly located in northern Europe*" in **lines 114-116**. 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 added "*The window in the time series of discharge highlights the target output (which is a point)...*" into the caption instead (**lines 137-138**). We further simplified the arrows used and rewrote the caption. The figure and caption now read:

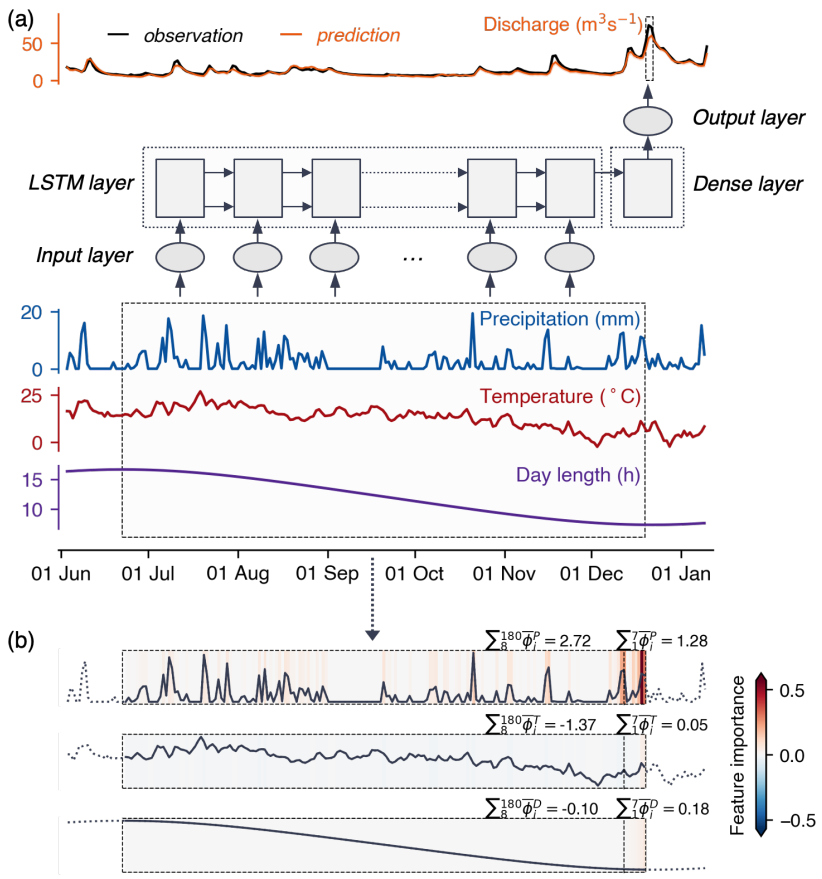


*Figure 2: The workflow of using explainable ML methods for attributing flood peaks (annual maxima of river discharge) to their drivers. (a) Diagrammatic representation of the used LSTM models. The window in the time series of discharge highlights the target output (which is a point) and the window in 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 in the windows 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 added 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 composed of 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 snowpack (Lees et al., 2022). The physically*

*realistic mapping from inputs to outputs facilitates gaining hydrologically meaningful insights from subsequent model interpretations*" (**lines 146-151**).

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

**Response:**

The original sentence has been revised as "*To improve the robustness of model evaluation and analysis, we fitted 10 independent LSTM models for each of the 1,077 catchments*" (**lines 162-163**).

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

**Response:**

The original sentence has been 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)*" (**lines 175-176**).

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. First, please see above 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 implemented such methodology for catchments in the US (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 the present study we improve the efficiency of the original framework proposed in our previous study by introducing a separating window.

In the Conclusion section, we further added a remark as "*the clustering procedure can be improved by developing algorithms to aggregate daily feature importance adaptively, thereby avoiding the predefined separation window while maintaining high efficiency*" (**lines 633-635**).

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 meant to describe the hydrological response time to precipitation and snowmelt events, instead of the build-up periods. We revised our original statements to justify the selection of 7 days, as "*The separating window size should cover the period of precipitation and snowmelt events leading to each peak discharge, which depends highly on the local characteristics. After examining 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) does not affect the conclusions about dominant flooding mechanisms and their trends (see discussion in Section 3.7)*" (**lines 214-220**).

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:**

We apologize for the confusion. Rather than multiple-peak discharges, we referred to multiple peak-discharges here (note the position of the hyphen). To avoid ambiguity, we removed "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 has been 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 all considered catchments*" (**lines 223-224**). In the section presenting the cluster results, we further added "*It should be noted that the clustering results here only reveal major patterns widespread in data, with certain local and specific mechanisms unlikely to be detected*" (**lines 294-295**).

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 specified 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 added "*The NSE values were calculated using all samples in respective testing datasets*" (**line 270**).

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 have removed 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 including figures.

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 added 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%*" (**lines 360-362**).

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 added "*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 can be expected in future studies*" (**lines 391-395**).

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

**Response:**

Replaced as suggested (**line 405**).

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:**

Thank you for the good suggestion. We added the catchment sizes for comparison in **Table 1** and in **lines 429-431**, we further added "*In addition to methodological differences, the inconsistent catchment samples are also responsible for the divergent attribution results in different studies. As shown in Table 1, the catchments examined in this study are generally smaller, which tend to be more susceptible to rainfall with high intensity*".

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

**Response:**

We added the reference as suggested (**line 446**).

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

**Response:**

Thank you for the suggestion. We have moved the notes next to other explanations for the divergences in results (**lines 431-435**), which we thought might be more appropriate.

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 added "*Tarasova et al. (2020) 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*" (**lines 435-437**).

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 added "*The proportions were calculated based on 846 catchments, where at least 15 years of records were available in each period*" (**lines 474-475**) and "..., *with their significance being assessed by the modified Mann-Kendall test*" (**lines 477-478**). The details have been provided in the new section "***2.4 Trend analysis of flooding mechanisms***", as "*Specifically, at the continental scale, we estimated the overall trends of various flooding mechanisms based on their respective proportions within all the annual maximum peak discharges per year. At the catchment scale, to capture the variations of flooding mechanisms over different periods, we calculated the proportion series using a 20-year moving window in each catchment. The 20-year time frame was used to ensure an adequate sample size for reliably estimating the intra-period proportions and also to guarantee enough periods to observe decadal variability (Pagano and Garen, 2005). Only proportions that were calculated with at least 10 years of peak discharge data in each window were used to estimate the trend slope*" (**lines 238-244**).

For panel b, before introducing the results of trends in individual catchments, we added "*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*" (**lines 481-482**).

Line 455-459: It would be helpful if this information is provided 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 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. For clarification, in the caption, we added "*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*" (**lines 518-520**).

For the possible periodicity in data, we added "*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*" (**lines 532-535**).

For the geographical indications and number of catchments, we added 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*" (**lines 654-656**).

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 moved the description of Figure 9 into **Appendix B** and simplify the part original 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*" (**lines 537-544**).

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 updated 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*" (**lines 603-606**).

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 modified "...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*" (**lines 616-617**).

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 replaced the relevant sentences with "*To investigate whether flooding mechanisms have 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 usually 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*" (**lines 599-606**).

Moreover, at the end of the conclusion section, we replaced 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, providing potential research opportunities. For example, the clustering procedure can be improved by developing algorithms to aggregate daily feature importance adaptively, thereby avoiding the predefined separation window while maintaining high efficiency. Moreover, regional LSTM models that incorporate static catchment attributes 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*" (**lines 629-638**).

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:**

We appreciate you pointing this out and we agree with you. In the revision, we have removed the relevant statement and replaced 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 (**lines 314-315**).

Line 276: catchments associated

**Response:**

Modified as suggested (**line 327**).

**To Reviewer 2:**

This paper presents a large-sample study to detect different flooding mechanisms across Europe. I have to admit that initially, I was skeptical about this study. But while reading the manuscript for this review, I became quite excited about the presented work.

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.

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.

In general, I think this is a very interesting study that fits into the scope of HESS and I only have a few general comments. Note, I already spoke to the first author during the EGU GA but for the sake of transparency, I will add all points here again. Please also note that, due to the overlap of this research with our own research in the past, in many of the studies that I reference I'm either the first or a co-author. I do not mention these studies here, because I want them to be cited in the manuscript, but I think they help to explain my reasoning.

**Response:**

We thank the reviewer for the positive comments and insightful suggestions of our work, which will not only help us to improve the present manuscript but also provide us with good ideas for future works. The replies to the comments follow.

### 1. Training setup

You train LSTM models individually for each basin, instead of one model on the combined data of all basins. Again, I'm very biased on this topic but I think there are multiple studies that show that the recommended way for training LSTMs is the latter (on all basins at once, using meteorological timeseries features and static attributes). The regional modeling setup was introduced in 2019 (see Kratzert et al. 2019) and further discussed in Nearing et al. (2021) (see Fig 2). One study that follows the regional training scheme is even cited in the manuscript (Lees et al. 2021).

The question is, is this important in the context of this study? This is a good question that I asked myself quite a lot over the last few days. On one side, I think it is important to follow best practices when working with any model. The benefit of the LSTM is that it can learn a very general understanding of the underlying processes if it is trained on a variety of basins. Nobody would probably train a conceptual model in a regional calibration scheme, if she/he is only interested in a particular basin. On the other hand, the authors are not interested in getting the best-possible streamflow performance, but to learn about flooding-mechanisms from the model.

From my experience, I would assume that changing the training setup would not change the results of this study (i.e. the clusters of different flooding mechanisms found here). What might change is the number of

basins that are considered in their study (because of the NSE threshold). However, even (or especially?) if the results of this study do not change, I would suggest training an LSTM on the combined data of all basins and to re-run the analysis, to reflect the best-practices of the chosen model. During the EGU, I offered Shijie Jiang help with setting up such a run as I have the code + resources available. I would be happy to help and don't want/expect any co-authorship/acknowledgements for that.

**Response:**

Thanks for the insightful and enlightening suggestions that are well worth considering.

First of all, we agree with you that training a regional model (i.e., training one single model for all catchments, including both meteorological time series and static catchment attributes as inputs) is a good practice, with the benefit that the ML model can learn relationships from a large sample of hydrological variability and enables to use the learned relationships for better predictions in individual catchments. 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.

The key idea of the study is to identify flood generation mechanisms based on distinguishable patterns of meteorological variables' contributions. 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. Moreover, the feature importance of meteorological variables may also be obscured by the importance of catchment characteristics, for the sake of explaining (possibly larger) spatial variation. This can result in changed contribution patterns of meteorological variables to individual peak discharges.

To demonstrate the above point, we have conducted a small-scale test on around half of the studied catchments (530 catchments) using the regional LSTM model, in which static attributes were directly concatenated with meteorological variables at every time step. The static variables contain 12 catchment attributes available in the GSIM dataset, such as the area, slope, elevation, snow fraction, climate type, etc. Here we tested a random subset of catchments instead of all the 1,077 catchments mainly because of the limitation of our memory resources (the median sample size of individual catchments is 20,455, and each sample has a dimension of [180, 15]). As a result, we found that the feature importance pattern of meteorological variables had changed significantly because the feature importance had been re-assigned to some catchment attributes to reflect the difference in flooding schemes across catchments. For example, the model outputs show a large reliance on elevation in the Alps area. However, elevation can affect a variety of aspects of flow behavior because mountainous catchments are typically smaller, receive greater precipitation, and have a higher snow fraction, resulting in a possible confounding factor to different flooding mechanisms. Introducing confounding factors will make it challenging to distinguish flooding mechanisms based on drivers, which is however out of the scope of the present study.

As a clarification that we used individual models instead of the regional model, we added a short discussion in the new section "**3.7 Limitations and outlooks**": "*In this study, we trained LSTM models individually for each catchment, while 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). In the latter case, both meteorological time series and static catchment attributes are used as inputs to distinguish response behaviors across time and space, with the benefit that the ML model can learn more general relationships from a larger sample of*

*hydrological variability (Kratzert et al., 2019b). However, introducing catchment attributes may prevent the identification of flood generation mechanisms based on distinguishable patterns of meteorological variables' contributions, which is the main objective of this study. 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. Therefore, here we employed local models. Nevertheless, we note that using regional models can provide insight into how flooding mechanisms vary spatially, particularly for how the spatial distribution is affected by the geographic and climatic characteristics of catchments, and it merits more exploration in future studies*" (**lines 561-571**).

In the Conclusion section, we added an outlook as "*Moreover, regional LSTM models that incorporate static catchment attributes can be employed to capture the spatial variations in flooding mechanisms and quantify the influence of catchments' geographical and climatic conditions on flooding processes*" (**lines 634-636**).

2. Input variables

In this study, you only use precipitation, temperature and day-length (as a proxy for solar radiation) as model inputs. I agree that these are the main-drivers for flooding mechanisms but I think the exciting thing about data-driven models is that if there is anything else, these models are pretty good at finding it. That is, if provided with more input features, the models would find other flooding mechanisms, if they are deducible from the data. When I discussed this point with Shijie Jiang at the EGU, he mentioned that he has/had a hard time to interpret the contribution signal for different features, and I agree, it is much simpler and more intuitive to reason about the meaning of high feature importance of e.g. temperature in the recent days. However, how exciting would it be to find that the model finds a flooding mechanism that is not straightforwardly explainable with the patterns we already know? If you have easy access to more input features, I think it could be interesting to run the models with more input features. If not, I think it is ok not to do it, but then it could be worth adding this to the discussion.

**Response:**

Thank you for the constructive suggestions. 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. However, considering the purpose of the study, we simplified the inputs to include only precipitation, temperature, and day length as input variables in the final manuscript, since their results corresponded to three well-known flood mechanisms at the catchment level. This allows us to directly compare our findings with those of other studies conducted with classical methods and perform the subsequent trend analyses. The performance did not drop much by not including more variables.

To highlight the possibility that the methods can provide insight into previously seldom known patterns, in the new section "**3.7 Limitations and outlooks**", we add "*Moreover, to strike the balance between model interpretability and accuracy, we only selected daily precipitation, temperature, and day length as meteorological inputs. The combination of inputs results in uncovering three well-known flooding mechanisms, which allows us to make a direct comparison with findings from other studies that used*

*classical methods. However, with more input variables incorporated into the model, the methodology may be able to recognize more distinct patterns in terms of input contributions. In principle, this could allow identifying flooding mechanisms that are often overlooked and may not be easily explainable with well-known processes. However, it would likely be much more challenging to make sense of the flooding mechanisms from the likely much more complicated patterns in input feature importance, which we leave for future studies*" (**lines 573-580**).

3. Data split

In L 138f, you mention that some hyper parameters were determined considering the model performance and efficiency. To me it is unclear which data was used for this hyperparameter tuning. Usually, the validation split (note, here I am referring to a 3-fold split with a train, validation and test set) is used for these kinds of hyperparameter tunings. However, from the explanation in L146, I can only estimate that this was done with a 2-fold split, thus on the test data?

**Response:**

Sorry for the unclear statements. In the original manuscript, we omitted to report that we actually performed a 3-fold split (training:validation:testing = 49%:21%:30%) on the dataset when training models. We modified the original sentences as: "*Each independent model was trained and tested based on samples that were randomly split in a 7-to-3 proportion. During the training process, a portion of the training data (70%) was repeatedly used to update the model parameters every epoch until no further decrease in the loss function was observed on the remaining 30% (also known as validation data). The trained models were independently evaluated on the testing datasets*" (**lines 163-166**).

4. Data split pt. 2

In L146, you say that the timeseries data was randomly split in a "7-to-3 proportion". I am not 100% sure but I think randomly splitting timeseries data is not optimal, especially when considering the overlap of different samples because of their input window size. I think much more common is a k-fold cross validation in time (none random).

**Response:**

Thank you for pointing this out. We performed a random split instead of the more common temporal k-fold cross-validation due to two reasons. 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. Secondly, using a random split rather than temporal k-fold cross-validation is based on the purpose of our study, which facilitates our subsequent trend analysis for flooding mechanisms, as we emphasized in the revised manuscript, "*Note that here we adopted a random sampling strategy instead of the time-series splitting strategy with fixed time intervals in order to enable capturing the overall hydrometeorological variability observed across various periods*" (**lines 166-168**). In **lines 168-170**, we further added a remark as "*It should be emphasized that while the random sampling strategy is appropriate with respect to the purpose of this study, it might not be the best practice if the models were developed for prediction tasks, particularly if they were to be applied to new datasets*".

5. Data split pt. 3

I am curious, if you train/eval the models in a 7-to-3 split random fashion, how could you a) guarantee an equal number of model predictions per timestep and b) guarantee that every time step was e.g. evaluated at least once and c) guarantee that the flood peak was in the validation and not the training period? Because from L 169, it seems like you used all 10 models for all peaks, but some peaks are certainly in the training period, right?

**Response:**

Thank you for highlighting this unclear point in the original manuscript. The average feature importance score for one peak was computed by averaging the scores from all 10 models, regardless of whether the peak appeared in the training or testing dataset. 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. The generalizable relationship should hold not only for the testing dataset but also for the training dataset. Because of this, the interpretation results from the training dataset were not excluded, similar to other geoscientific studies that aimed to gain insights through machine learning interpretations (e.g., Barnes et al., 2020; Toms et al., 2020).

To better clarify this point, we will add Figure S1-S3 in the supplementary material in the revision, which gives the respective feature importance scores from the 10 models used to derive the examples of Figure 2b and Figure 4 in the main text. Some of the scores are derived when the flood peaks are in the training dataset (the solid lines), while others are in the testing dataset (the dashed lines). It can be seen that these feature importance scores present consistent patterns. We will add the following description: "*Note that the averaged IG scores for an individual peak were computed by averaging the scores obtained from all the independent 10 models, regardless of whether the peak was part of the training data or the testing data in the models. Overall, the IG scores extracted from the 10 models for each target peak discharge generally follow a similar pattern, though with inevitable differences due to randomness and uncertainties in training processes (see Figs. S1–S3 in the Supplementary Material for examples)*" (**lines 198-202**).

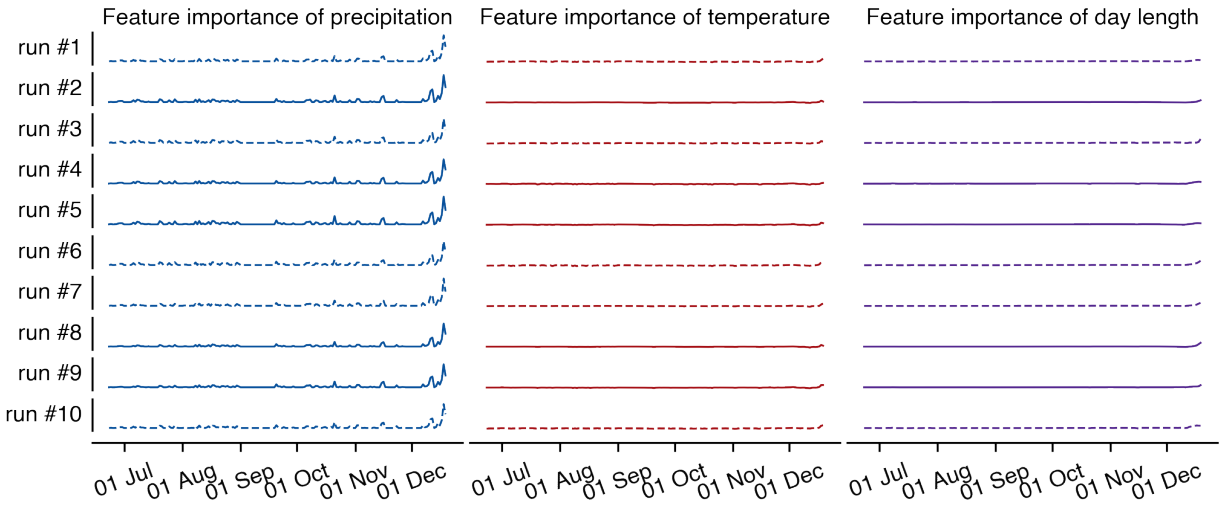


*Figure S1: The integrated gradient (IG) scores of precipitation, temperature, and day length extracted from 10 independent models for predicting the peak discharge that is illustrated in Fig. 2a in the main text. The 10 models were trained and tested with different randomly split datasets, where the target discharges in*

*runs #1, #3, #6, #7, and #10 were in the testing dataset (indicated by dashed lines), and the target discharges in the remaining runs were in the training dataset (solid lines). All the y-axes use the same scale for better comparisons. The average of the 10 sequences across different models generates the heatmaps shown in Fig. 2b in the main text.*

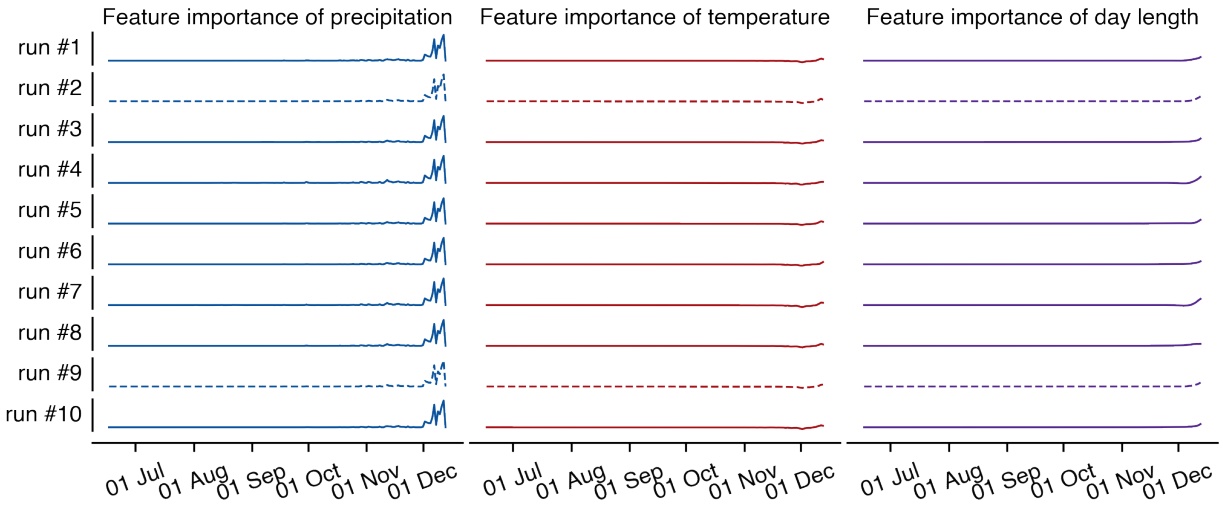


*Figure S2: The same case as in Fig. S1 to generate heatmaps illustrated in Fig. 4a in the main text.*

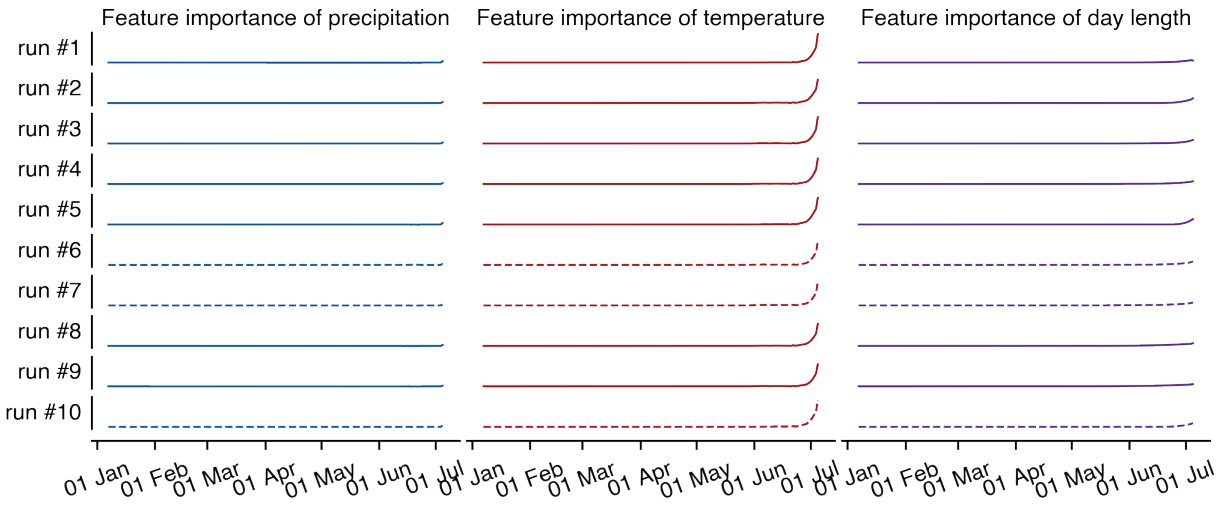


*Figure S3: The same case as in Fig. S1 to generate heatmaps illustrated in Fig. 4b in the main text.*

**References:**

Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020). Physically interpretable neural networks for the geosciences: Applications to Earth system variability. Journal of Advances in Modeling Earth Systems, 12, e2019MS002002. https://doi.org/10.1029/2019MS002002

Barnes, E. A., Toms, B., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., & Anderson, D. (2020). Indicator patterns of forced change learned by an artificial neural network. Journal of Advances in Modeling Earth Systems, 12, e2020MS002195. https://doi.org/10.1029/2020MS002195