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 superseded to varying degrees 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 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 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 will add a short discussion in the new section "*3.6 Limitations and outlooks*": "*In the 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). For such a case, both meteorological time series and static catchment attributes will be 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). In spite of this, the main objective of the current study is to identify flood generation mechanisms based on distinguishable patterns of meteorological variables' contributions, while introducing catchment*

attributes may obscure the role of meteorological variables in explaining temporal variations in flooding. Moreover, interpreting flooding mechanisms with regional LSTM models may become more challenging than with local LSTM models that use only meteorological variables, since some catchment attributes would confound the interpretation. Therefore, considering the main objective here we use local models. However, we anticipate using regional models can provide insights into how flooding mechanisms vary spatially, particularly for how the spatial distribution is affected by the geographic and climatic characteristics of catchments, and it deserves more exploration in future studies."

In conclusion, we will add "*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*."

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 "*3.6 Limitations and outlooks*", we will add "*Moreover, to strike the balance between model interpretability and accuracy, we only selected daily precipitation, temperature, and day length as the 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. This could identify flooding mechanisms that are often overlooked and may not be easily explainable with well-known processes. Despite the potential, it would be much*

more challenging to make sense of the flooding mechanisms from the likely much more complicated patterns in input feature importance, which we therefore leave for future studies."

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 will modify 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, 70% of the training data 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 by the testing datasets.*"

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 stated in the original manuscript, "*random sampling strategy enables capturing the overall hydrometeorological variability observed across various periods*".

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 Fig. S1–S3 in the supplementary material for examples).*"

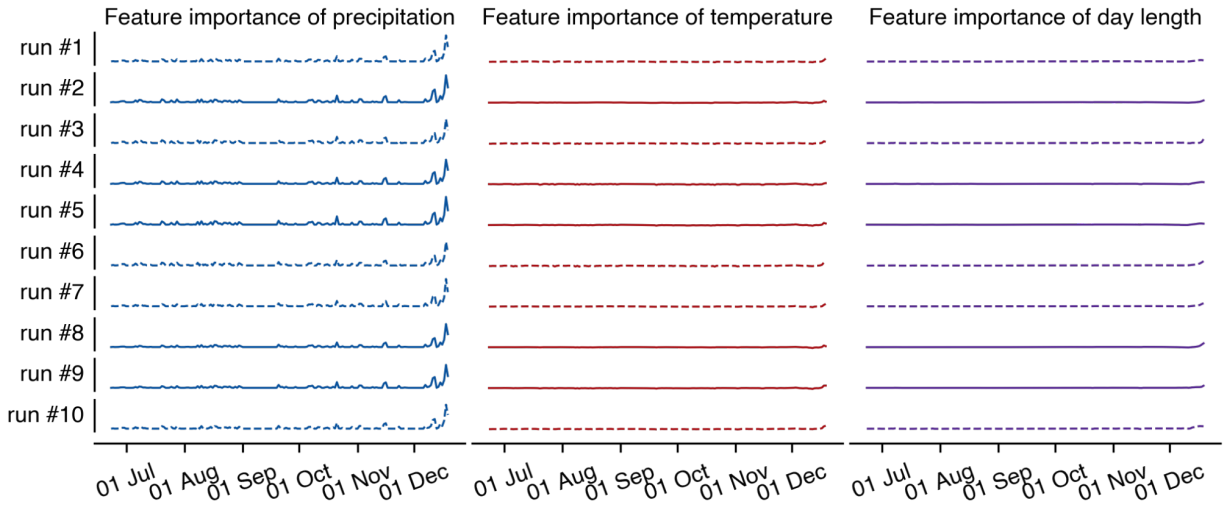

Figure S1: The integrated gradient (IG) scores of precipitation, temperature, and day length extracted from 10 independent models for predicting the peak discharge that is illustrated in Fig. 2a in the main text. The 10 models were trained and tested with different randomly split datasets, where the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

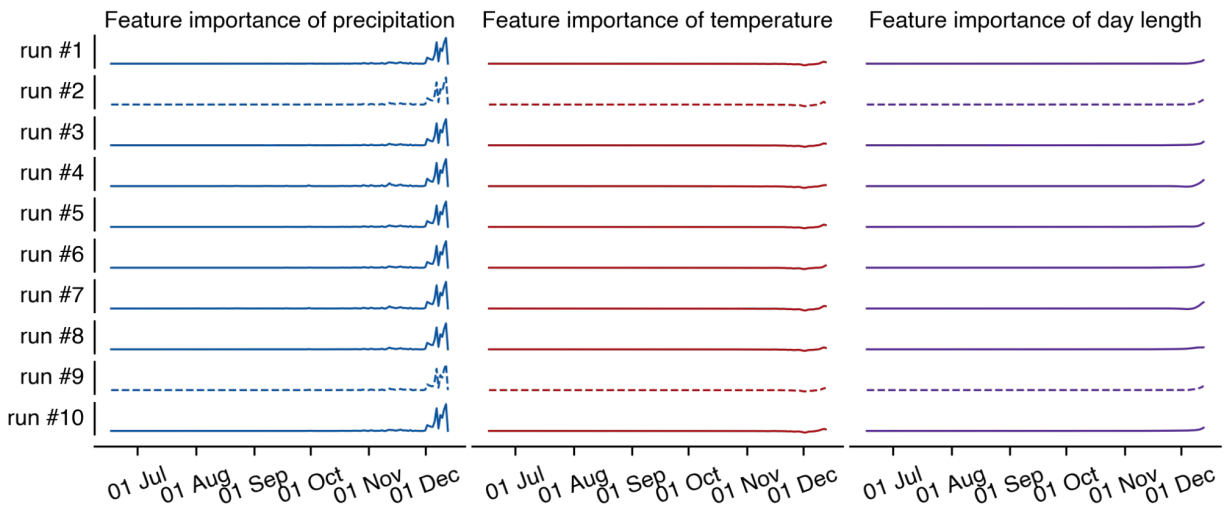

Figure S2: The same case as in Fig. S1 to generate heatmaps illustrated in Fig. 4a in the main text.

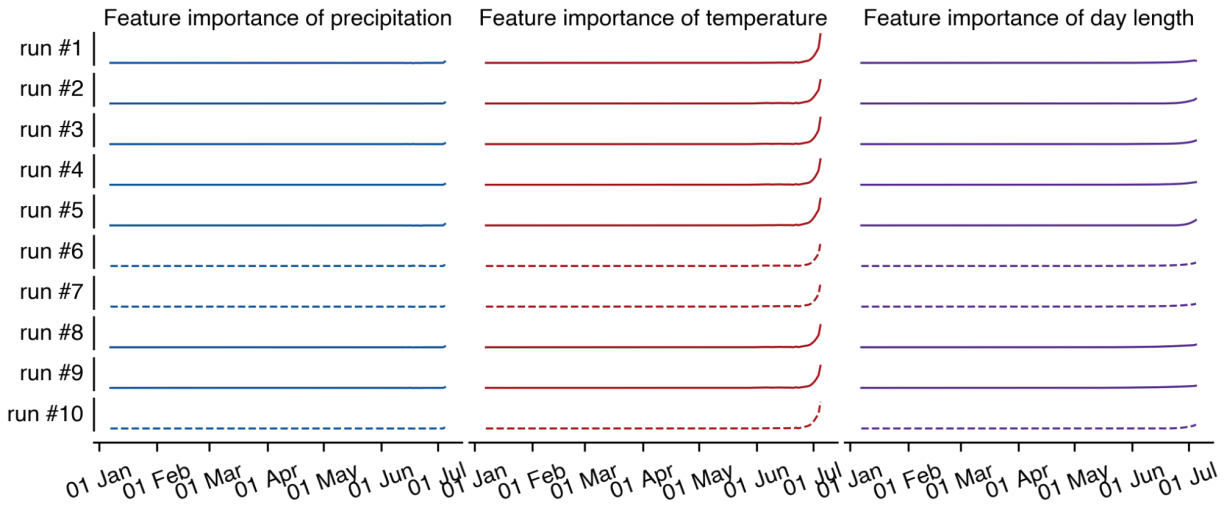

Figure S3: The same case as in Fig. S1 to generate heatmaps illustrated in Fig. 4b in the main text.

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