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

Response to Editor

Editor: Thank you very much for your detailed responses to the reviewers' comments. Because of the disagreement of one of the former reviewers with the suitability of the (local) modeling strategy chosen, I looked for a third person to review your manuscript. This new reviewer agrees with the previous two reviewers and myself that the study is of interest to the HESS readership. They also reflect on the suitability of the local modeling approach chosen and think that the local approach is suitable to answer the research question at hand. Therefore, we will conclude this discussion and 'accept' the local modelling strategy as the strategy of your choice. The previous reviewer also criticized the data splitting. The new reviewer does not comment on this issue but I personally think that this point has sufficiently been addressed by testing the alternative splitting approach and presenting its results in the Supplementary Materials. The comments that remain to be addressed are the two minor comments by the new reviewer. Thanks for addressing these and I am looking forward to reading the next (and hopefully final) version of your manuscript.

Response: We are grateful to the editor for providing us with the opportunity to refute/clarify some arguments, which have significantly improved the manuscript and inspired new research questions. We also appreciate the three reviewers for their positive comments and constructive suggestions. In the revision, we made appropriate modifications and clarifications for the new comments.

Response to Reviewer 2

Reviewer 2: Dear authors, I am thankful for your responses and I want to say that I am sorry if I appear to be the “mean reviewer 2” in this revision process.

Let me start with an TLDR;

● I think we can stop arguing here and conclude that we need to agree to disagree. I think I could go on forever but it probably adds little to this review process. So to save us all time, I will stop arguing after this review and I only reply here to clarify a few of the statements/misunderstandings.

● That being said, I think the first point, regarding data splitting (TLDR; use the results of the test splits in the paper, not the averaged results of training periods) should be taken seriously.

● One last thing: I mentioned this to Shijie at the EGU in personal communication and I don’t want this review process to negatively impact what I said: I think this is an exciting study and I am happy to see people using LSTMs for these kinds of studies. I think it would have been easy

to make this paper even more exciting but I understand that I need to wait for another publication from your group to get answers to my questions.

Response: Once again, we wish to express our gratitude to Reviewer 2. We thought it was an honest and thought-provoking debate/arguing, which helped the formulation of new research questions for us. The reviewer's expertise was greatly appreciated, his thoughtful insights in the modeling would benefit not only this study but also our following ones.

Reviewer 2: Data splitting – I am very happy to hear that running your analysis only on the test splits did barely change the results. What I don't understand is, why you did not change the manuscript so that these results are used in the paper. I don't think that I need to search for literature here to say that in any modeling study, you should use independent test data (which you have!) to analyze/interpret your model. We already know that this is not affecting anything in your results, so it should be a no-brainer to change this and standard practices in statistical analysis with models.

Response: We are grateful to the reviewer for helping to make the analysis more rigorous and reasonable. In the last revision, we have updated the analysis results based on the new splits with the reviewer's comments.

Reviewer 2: Input selection – I think it is somewhat funny that on this point you argue with NSE performance (that it did not or did negatively affect the NSE, when adding more inputs), while in the other points (e.g. multi-basin model) you say that performance is not really important. Anyway, there is not much to add from my side, I think it would have been a great addition to the paper to include and analyze more input features, especially because it seemed like you already did the experiments. You yourself said in the first rebuttal "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", which to me sounded exciting and I was curious to hear more about clusters that are different from previous literature and about interpretations/hypotheses of these clusters. You decide to keep it simple, your choice. If I understand you correctly you want to come back to this in a future publication and I am looking forward to diving into the results then.

Response: As we responded in the last revision, we did not deny the value of adding more input variables for possibly new insights from interpretations/hypotheses of these clusters. What we were concerned about is that multicollinearity in meteorological drivers at daily scales is likely causing instability in model interpretation, making the clustering less robust. In spite of the challenges, we will attempt to provide a definite answer to the exciting problem in a future publication.

Reviewer 2: Model setup – Let us just say that we disagree. I certainly do believe that, no-matter what you want to do, if you use a tool, you should use that tool correctly. Following your argument (which tries to say the other extreme is possible) I could make such a study with uncalibrated hydrology models and then say that my results have any meaning?!

From a very basic point of view: I hope we can agree that a model that is better in modeling the task at hand (meteorological forcings in, discharge out) should have a better understanding of the underlying system/processes, right? Now if I want to make a study that investigates the underlying process understanding of a model, then why would I willingly pick a model that has a worse understanding of the underlying processes? If the model would not have a worse process understanding, it would not be generally worse in predicting discharge, or would it? Since your study is about model interpretation of process understanding, I do not see how this is decoupled from performance. But I see how this example will be used again to say "he only cares about performance", which is wrong. But I'm not so naive to think that good process understanding is decoupled from good performance.

Your line of argument is the ease of interpretation, I got it. But what I am saying is that this is a bad argument. And again, this is not "because the regional model has a higher NSE", but because, without any doubt, this model has a better process understanding. And your study is about analyzing such a process (flood generation). Funnily enough, the one thing where single-basin models have their biggest bias (compared to regional models) is flood peak prediction, which I also mentioned in my previous review (see point on saturation). One could argue that having a (negatively affecting bias) in flood predictions is also affecting the flood generating processes in the model

Response: We agree with the arguments that regional modeling is better in some cases, most of which we have already responded to in the previous revisions. However, we think the reviewer seems to have still ignored or misunderstood our concerns about regional modeling. We were not susceptive to the ability of regional modeling can have a better underlying process understanding, the challenge we thought of is how to disentangle the roles of meteorological drivers and catchment attributes given the possible confounding and multicollinearity resulting from static catchment attributes. In this study, we did not intend to address these challenges since they were beyond its scope, but we will have an exploration of in our next studies, as we stated in previously revised manuscript, "*in light of the benefit of regional modeling that can provide insights into how flooding mechanisms vary spatially by geographic and climatic characteristics of catchments, how to deal with these challenges in the interpretation merits more exploration in future studies*". We thank the reviewer for bringing it up.

Response to Reviewer 3

Reviewer 3: This study applied the explainable machine learning method to examine river flooding mechanisms between meteorological forcings and streamflow response. They worked on large sample European catchments, associated the patterns of three identified mechanisms with hydrological processes and further investigated the temporal trends. The ms is well-written, and I am fascinated by the topic and related analyses from hydrological perspective.

Response: We appreciate the positive comments from the Anonymous Referee.

Reviewer 3: Reading through the previous rounds of reviews and responses, I find the main question discussed is the impacts of using locally trained LSTMs (one LSTM for each catchment) versus regional LSTMs (one LSTM for all catchments with attributes used) on this study. In other words, is it appropriate to use locally trained models here to identify flooding mechanisms? I have been thinking about this interesting question for some time. Training regional LSTM with attributes is certainly a more appropriate way to get better performance, which has been shown in previous studies as the reviewer mentioned. Let's put these two types of models in the context of this study. For local LSTM, the streamflow responses only rely on meteorological forcings which totally determine the gradient contributions. Given that the local model has a performance gap to the regional LSTM model, we can safely infer that the used attributes play an important role in regional modeling and fairly contribute to the gradient dynamics. This implies that the gradient contributions of meteorological variables would behave differently in the local and regional models. As expected, the authors mentioned that different gradient patterns emerge when testing regional modeling.

It's interesting to think why there exist different gradient patterns between two modeling forms and which one is more reliable. Back to this study, the concerning question is whether the analyses done in one specific form are still valid given these differences. In my view this should be evaluated by the consistency between the identified mechanisms with our established domain knowledge. Reading through the authors' hydrological interpretations of identified mechanisms from ML, my feeling is that the results and related analyses are convincing to safely draw the conclusions. However, I to some extent agree with the point that one reviewer mentioned, due to not reporting these differences, the authors kind of limit themselves in the range of prior knowledge. Emerging differences may give us chances to better understand these models and learn new knowledge. Therefore, it would be nice to report some identified differences caused by training forms in the paper to inform readers and educe a more thorough examination in the future study. These are my thoughts and hopefully they can help to understand this question.

Response: Thanks to the reviewer for providing thoughtful comments and helpful suggestions. We agree that catchment attributes would play an important role in regional modeling and affect

the gradient dynamics in some way. Also, we agree that it would be interesting to examine the impact of modeling strategy (local vs. regional) on the interpretation of the models and ultimately on the understanding of flooding mechanisms. In spite of this, we prefer not to elaborate on differences in the manuscript that have only been identified preliminarily, in order to avoid jumping to a hasty conclusion and losing the original focus of the current study. As we indicated in the previous revisions, we have already planned to have a systematic examination and comparison in our next studies to give a detailed and insightful answer to this open question.

We appreciate the reviewers' suggestions about informing readers and stimulating a deeper examination. In **lines 588-592**, we supplemented the following statement, "*An immediate question to address is whether adopting different modeling strategies will result in different interpretations regarding the gradient contributions of meteorological forcings, which ultimately leads to alternative understandings of flooding mechanisms. The emerging differences may provide us with an opportunity to gain new insights into flooding mechanisms from these models.*"

Reviewer 3: Another point that concerns me is, in line 155, the authors state forcings over the past 180 days are used to predict the following day. The correct setup should be using forcings of 180 days to predict the last day's streamflow (180th, here larger number means the most recent time), not the next day (181st). The same-day precipitation is very important to predict the streamflow and should not be missed.

Response: Many thanks for the good suggestion. We agree with the reviewer that same-day precipitation is very important to predict streamflow. Our previous practice followed the common practice of using LSTM in rainfall-runoff prediction, which aims to build a predictive relationship between past meteorological data and the discharge the following day. Another consideration was the possible mismatch between the time resolution of precipitation and discharge, one of which is daily and the other may be only quasi-daily (depending on when the discharges were recorded every day in a gauge). Therefore, the prediction models only took into account the lagged meteorological forcings up till the day before each daily discharge.

However, taking precipitation on the day of the flood peak into account can be more appropriate for the purpose of flood classification, which would also be in line with common practice in similar studies. Therefore, we updated our results in the manuscript by using the new input series (which still are minor changes in reported numbers, and all previous conclusions hold). Moreover, we added the following statement for clarification in the revision, "*Note that we included predictors on the same day as the output in the model, since precipitation on that day could also affect the discharge, especially in small catchments with quick catchment response times. However, the conclusions do not change even if using LSTM models to predict discharge on the next day (i.e., the prediction models consider the lagged meteorological forcings up till the day before each daily discharge)*" (**lines 157-160**).

Reviewer 3: I am also a little confused at the separation of training and testing period. Line 258 mentioned "NSE value computed in the testing period". What's the testing period in time dimension? I didn't find a specific time span mentioned in the main text.

Response: As a result of the 10-fold cross-validation, the testing period is 1/10 the sample size of each catchment, while the exact length differs due to variable sample sizes of catchments. For clarification, we added "*The predictive performance of each model was evaluated independently based on testing data, i.e., 1/10 of the data for each catchment, which ranged from 2 to 7 years due to the 20-70 years of sample size available in studied catchments*" in the Methodology (**lines 170-172**).