Interpretable feature incorporation machine learning framework for flood magnitude estimation

Response to Editor and Reviewers

Editor decision: Reconsider after major revisions (further review by editor and referees):

[**Public justification (visible to the public if the article is accepted and published)**: I have assessed the overall contributions provided in the manuscript as well as the Reviewers' comments and the Authors' replies to these. All Reviewers provide a set of substantial criticisms, recommending a set of major revisions (one of them actually recommending rejection) to be implemented. I have seen the Authors have initiated a genuine and solid effort and I am willing to give them an opportunity to revise their work. It should be understood that, in case the Authors decide to undertake this route, there is no guarantee the manuscript will finally be accepted unless the totality of the Reviewers' concerns is satisfactorily and unambiguously addressed.]

We thank the Editor for the opportunity to revise the manuscript, and the Reviewers for their thorough and constructive feedback. Their detailed comments have helped us to significantly improve the manuscript's clarity, framing, and methodological rigour. We appreciate the recognition of the relevance of the topics addressed, as well as the thoughtful critiques regarding the use of weather patterns (WPs), the structure of our modelling framework, and the interpretation of results. In response, we have made very substantial revisions to the manuscript. These include a reframing of the paper to better reflect its original contributions, particularly regarding the limited marginal value of WPs in flood magnitude prediction, and a reframing **towards an interpretable feature incorporation framework**. We have improved the explanation of the **methodological design** and **evaluation strategy** and provide a more transparent and fair comparison between regional and national models. We have also clarified definitions, ensured consistent use of terminology, and addressed technical concerns raised in both the major and minor comments.

Wording changes are as follows:

- **From prediction to *estimation*:** We are predicting unseen future data outside of the training set, but our models are trained and tested within a historical flood magnitude envelope (1969-2021), not forward-forecasting in an operational way. The models are learning relationships between predictors/features and flood magnitude and estimating the magnitude of a flood event which has already been experienced (2011-2021). Therefore, we believe that flood 'estimation' more accurately reflects the nature of the task.
- **From model generation to *feature set*:** Originally, the expression 'generation 1-6' was used to describe successive model versions which each incorporate an additional set of features/predictors. However, readers may interpret generation as meaning a different model architecture. Furthermore, to align with the new title ("Interpretable feature incorporation machine learning framework for flood magnitude estimation"), we change this expression to 'feature set'. We also include an additional Feature Set 7 in the revised manuscript (in which redundant features are removed and remaining features are pruned).

Clarification of the focus on weather patterns (WPs): the manuscript's original research goal was to evaluate whether the UK Met Office WPs improve and contribute to the estimation of winter flood magnitudes in UK catchments, using a machine learning framework. This goal was important because of the use of the WPs in operational frameworks for coastal and fluvial flood risk across the UK, namely Coastal Decider (Neal et al., 2018) and Fluvial Decider (Richardson et al., 2020). To our knowledge, there had not been an assessment of the WPs for estimating high flood magnitudes > 99th percentile using UK observational streamflow data. Thus, even though we discovered that the Met Office WPs were not useful for flood magnitude estimation in the UK, this result is an important one to share, considering the WPs are used as tools in the UK.

Clarification of feature incorporation approach: we think it is important to explicitly show (through the feature incorporation approach) the influence of commonly used spatial identifiers, such as latitude and longitude, in the machine learning models. We showed that latitude and longitude as a baseline model provided some moderate skill, encapsulating inter-catchment variability and behaving as spatial proxies. When we then included static catchment characteristics (e.g., aridity index, runoff-ratio, baseflow index, area), the model skill did not increase significantly because latitude and longitude had already encapsulated the spatial and inter-catchment variability information. Finally, when we ablate (remove) latitude and longitude in the final (feature set 7) model, and apply VIF pruning, the overall model skill remains stable, demonstrating that predictive performance is driven by

physically meaningful variables rather than spatial proxies or redundant predictors. The ablation and pruning for feature set 7 does not substantially change model performance (e.g., R-squared) but does have implications for interpretability of important features.

Pruned model addition: to ensure interpretability and robustness, we have introduced a final pruned model (feature set 7) that removes non-physical predictors (latitude, longitude), the WPs and antecedent weather patterns (AWPs).

Model spatial scope: we also wanted to compare UK vs regional models to assess (a) predictive skill (whether a model with more diverse data, or one with more relevant local data performs better), and (b) consistency of feature importance results across different spatial scales.

Novelty: our research makes a novel contribution in the following ways:

- Comprehensive ML feature incorporation framework
- Assessment of catchment and feature importance insights from regional to national scale
- Exploration of the role of latitude and longitude and spatial/catchment identifiers
- Evidence based evaluation of the limited predictive value of the UK Met Office WPs, used in the Met Office DECIDER tool.

Uncertainty quantification: we have also incorporated an ensemble-based uncertainty quantification to assess model reliability.

Together, these revisions establish the study as a rigorous and transparent benchmark for integrating atmospheric and hydrological predictors within large-sample hydrology machine learning frameworks. They also directly address all reviewer concerns regarding model design, framing, and interpretation. Below, we provide detailed responses (in blue text) outlining substantial changes made in response to each of the reviewer's comments (plain text). We thank both reviewers for their important input.

Response to Reviewer 1

R1C1: In the manuscript, if weather patterns contribute to predicting winter flood magnitudes was discussed using machine learning. To my knowledge, flood is mainly caused by intensive rainfall and antecedent soil conditions. In this study, they also considered these two factors and add some other variables. Some main questions are below

Response: We thank the reviewer for highlighting this fundamental point regarding the drivers of flooding. We fully agree that flood generation is primarily governed by precipitation and antecedent wetness. These variables are central components of our study and are included in the feature set used for model training. Our specific focus on weather patterns (WPs) stems from the operational and research relevance of the 30-pattern classification developed by the UK Met Office (Neal et al., 2016). These WPs have been used in prior studies examining their links to precipitation extremes and droughts (e.g., Richardson et al., 2018; 2019; 2020) and are integrated into tools like the Met Office's Fluvial Decider for flood risk assessment (Richardson et al, 2020), and Coastal Decider (Neal et al., 2018). What sets our study apart is that we conduct the first large-scale, data-driven evaluation of the direct relationship between these WPs and observed fluvial flood magnitudes, using NRFA streamflow records. Our goal is to empirically assess whether synoptic-scale atmospheric circulation patterns, often more predictable at longer lead time than local-scale variables, contribute meaningfully to flood estimation frameworks when more immediate forcings are also included. Our results show that while WPs capture synoptic-scale atmospheric conditions, their marginal value in flood magnitude estimation is limited. This finding itself is important. It quantifies the redundancy of synoptic-scale indicators within data-driven hydrological models and helps refine which predictors offer true added value. We have clarified this in the introduction and the discussion of results. Please see the revised Introduction section.

R1C2: Table 2, I cannot understand the relationship between total event count, number of catchments and catchment average event count.

Response: Thank you for this comment, we have improved Table 1, which is now Appendix Table A3 (page 25) to make this clearer to the reader. The total event count is the total number of fluvial flood magnitude events across all catchments. The number of catchments is the number of unique catchment IDs with events. The

catchment average is the total events divided by the number of catchments. This is calculated for each regional sample.

R1C3: Line 154, 'pre-filtered to contain only extreme flood magnitude days', this will not ensure the flood event from beginning to the end.

Response: We appreciate the reviewer's comment and agree that filtering based solely on peak days does not capture the full duration or hydrograph shape of each flood event. However, our study is specifically designed to focus on flood magnitude. By this, we mean the maximum daily flow or peak, associated with each extreme event, rather than the full flood hydrograph or event duration. We have made it much clearer in the text that this is our purpose. We have revised the text in 'Identifying flood magnitudes and dataset creation' section 2.2, which should now be clearer for the reader. Please see line 133.

R1C4: Line 163, the categorize small, medium and large is not appropriate. Because in hydrology, there is a standard for definition of small, medium and large catchments.

Response: We thank the reviewer for this observation. While we acknowledge that hydrologists often use informal thresholds to classify catchment sizes, to the best of our knowledge there is no single, universally accepted standard for defining small, medium, and large catchments, specifically within the context of UK hydrology and the UKBN2 benchmark dataset. In this study, the terms "small," "medium," and "large" were used descriptively to indicate relative size classes, based on quantiles of the catchment area distribution within the UKBN2 dataset. This was done purely for interpretative convenience, rather than to imply strict hydrological classifications. We have now clarified this point in the revised text. We have stated that the classification is based on terciles of the UKBN2 area distribution and used solely for within-sample comparison. We address this in the 'Distribution of flood magnitudes' section 2.3.1. Please see this stated more clearly at line 168.

R1C5: Line 278, the WP associated with the most extreme precipitation, does not necessarily translate to the WP associated with extreme flood magnitude days across UK regions.' I cannot understand the intrinsic relations between WP, extreme precipitation and extreme flood magnitude days.

Response: We thank the reviewer for raising this important point. The relationship between WPs, extreme precipitation, and extreme flood magnitudes is not direct or linear. While previous research (e.g., Richardson et al., 2018) has shown that certain WPs are associated with heavy precipitation in the UK, extreme precipitation alone does not always result in extreme flooding. Flood magnitudes are influenced by several additional factors, including antecedent catchment conditions, catchment memory (timing and accumulation of prior rainfall), catchment specific properties (e.g., topography), and the temporal structure of rainfall, and how the rainfall interacts with unique catchment land-surface characteristics. Even when a WP is associated with high rainfall in aggregate (through conditional probability), regional variability in hydrological response can lead to different WPs dominating the flood magnitude signal in different parts of the country. For example, a WP that causes widespread rainfall may trigger flooding in steep, saturated western upland catchments, but not in drier eastern regions with higher infiltration capacity. Topography plays a key role as well. Mountainous areas can respond more rapidly and intensely than flatter regions to the same synoptic forcing. Therefore, finding that a WP most associated with extreme precipitation is not necessarily the same WP most associated with peak flood magnitudes highlights the critical modulating role of other hydrological processes. This supports our conclusion that WPs, while useful for understanding large-scale atmospheric conditions, offer limited explanatory power for flood magnitudes when catchment-specific variables are not considered. We have revised the text to clarify this distinction more explicitly throughout the manuscript.

R1C6: Line 363, CEE had the lowest baseline R2 (0.28), and only the final R2 of 0.37 in Generation 6 was statistically significant. Why the precision is so low? Are there any previous hydrological simulation in this region? Please compare this result with previous studies.

Response: We thank the reviewer for this important observation. The relatively low R^2 values for the Central and Eastern England (CEE) region reflect genuine modelling challenges in this area, driven by a combination of hydrological, data, and methodological factors. First, the CEE region is characterized by predominantly low-relief terrain, permeable soils, and mixed land use, leading to slower and more diffuse runoff responses to precipitation.

These characteristics make peak flood magnitudes less directly tied to single rainfall events, and thus more difficult to capture using event-scale predictors alone. We are not aware of any published regional ML flood estimation models specifically for CEE. Higher uncertainty in this region could be explained by the baseflow-dominated hydrology, low relief and highly permeable (often chalky) catchment characteristics. Considering this, our R^2 value of 0.37 in feature set 6 model and 0.33 in feature set 7 model, though modest, is consistent with low performance found in other studies, and represents a statistically significant improvement over the baseline model (latitude and longitude). We have added further discussion in the manuscript to contextualize the CEE performance within existing literature and acknowledge the challenges of modelling in low-gradient, infiltration-dominated catchments. We have addressed this in the revised text. We have also ensured to further contextualise and improve the discussion of the other regional models with the existing literature. For example, please see line 346.

R1C7: Line 370, 'The SE region's relatively lower sensitivity to antecedent precipitation and hydrometeorological inputs suggests that other factors, such as urbanization and engineered drainage systems, may dominate flood generation.' However, when you select the watersheds, they are not influenced by human activities.

Response: We appreciate this clarification and have revised the manuscript throughout, to remove any implication or mention that urbanization is a known factor in these specific sites.

R1C8: When using SHAP, you need to explain the definition of aridity, runoff ratio….

Response: We thank the reviewer for pointing this out. We agree that clear definitions of all predictor variables used in the SHAP analysis are essential for interpretability, especially for features like aridity index and runoff ratio that may vary slightly in definition depending on the context. We have made amendments to Table 1 (page 7) and Table 3 (now appendix Table A2) much clearer in the manuscript than they were originally. We have listed all the variables included in the SHAP analysis and changed the SHAP beeswarm and mean absolute feature importance plots to have clearer labels which match the variable names in Table 1 and Appendix Table A2, for consistency and clarity throughout the manuscript. Appendix Table A2 includes full variable names, units, definitions and data sources. All SHAP Figures have been reproduced for clarity and consistency (Figure 6, page 15, and Figure 7, page 16).

R1C9: Line 490, 'The SHAP summary plot further supports the limited contribution of the WPs. In traditional flood analysis, rainfall and soil moisture are the main contributors. We never consider WPs. In this study, WPs are focused, but still limited contribution. What is the innovation of this study?

We thank the reviewer for this important question, and we note that the SHAP summary plots no longer show the WPs as they were ablated from the final feature set 7 models. Yes, traditional flood analysis rightly emphasizes precipitation and antecedent soil moisture as the dominant drivers of flood events. Our study does not challenge this understanding, instead, it aims to quantify whether atmospheric circulation patterns (in our case, synoptic scale UK Met Office WPs) provide any additional explanatory or predictive value. The innovation of our study lies in the following key contributions. We provide a comprehensive ML feature incorporation framework, with assessment of feature importance insights from regional to national scale. We explore the role of latitude and longitude and spatial/catchment identifiers. We provide an evidence-based evaluation of the limited predictive value of the UK Met Office WPs, used in the Met Office DECIDER tool. We emphasise that this is the first empirical test of WPs for flood magnitudes at national scale. While WPs have been linked to rainfall, coastal flooding, and drought in past studies (e.g., Neal et al., 2016; Neal et al, 2018; Richardson et al., 2018), this is the first study to evaluate their direct relationship with flood magnitudes across a large sample of UK catchments using machine learning and observational streamflow data in near natural catchments, to isolate the noise of anthropogenic activities. We can quantify the redundancy of these synoptic-scale indicators by showing that WPs have limited marginal importance and are largely redundant in data-driven flood models. This is a useful insight, as these WPs are used operationally in UK forecasting tools such as Fluvial Decider and Coastal Decider.

In summary, while WPs were shown to contribute little additional predictive power in this specific context, our clear empirical evaluation of their role, and demonstration of their limited value, constitutes a meaningful and novel contribution to the field and to operational hydrology. We have revised the manuscript discussion to make this contribution more explicit, and they have been removed from the final models (feature set 7).

R1C10: Line 501, ‘Interestingly, precipitation on the day of the event consistently ranks higher than antecedent precipitation’, actually, this is a common sense.

Response: We thank the reviewer for this observation. We agree and have reworded this point.

We thank you for your time and efforts in improving this work. We have made substantial changes to address your useful comments!

Response to Reviewer 2

We sincerely thank you for your detailed and constructive review. Your comments were exceptionally thorough and helped us significantly improve the clarity, rigour, and alignment of this manuscript with its central findings. Your feedback prompted us to:

1. Reframe the paper and title, abstract, and introduction to better reflect the main contributions, as a systematic evaluation of the added value of large-scale weather patterns (WPs) relative to hydrometeorological and catchment predictors.
2. Ensure fairer model comparisons by recalculating UK–regional model performance on matched catchments, replacing Figures 5a/5b with fairer comparison versions for the new feature set 7 models.
3. Increase transparency and interpretability by clearly documenting the role of latitude/longitude as spatial encoders, providing final feature set 7 models and SHAP plots without those features, and adding ablation experiments to quantify their contribution (we also did this for the WPs).
4. Address potential concerns about model robustness by providing uncertainty quantification through ensemble uncertainty and the coefficient of variation.
5. Clarify limitations and future directions, emphasizing that while WPs add little predictive value in this framework, we believe the negative result is important to share with the flood forecasting community as the WPs are used in operational tools in the UK.

These changes have substantially strengthened the manuscript by improving transparency, reproducibility, and interpretability.

R2C1: The paper analyses two main relevant topics in Hydrology. The first one is the predictability of extreme flooding events. The second topic is the difference between national (UK) and regional models. The main finding from the first one is that WP is not relevant for prediction, mainly because other attributes and forcing already share the same information. In the second topic, the national model exhibits the overall best performance, but with considerable variability between some regional models, indicating that, in many cases, regional models capture the dynamics of the region more effectively. The results align with other research; however, some concerns arise from the framework and the presentation of the results.

Main comments

The title is not aligned with the framework and the results. The author used a progressive feature incorporation to explore the benefit of having them in the model. From this analysis, WP was not relevant or caused a deterioration of the performance in most of the regions. Therefore, the author should not use them; however, they insist on using them across the entire paper despite of the no-value. The same happens with other features in generations 2, 3, and 4. My suggestion is to reframe the title and the paper toward the characterization of the extreme events through different types of models, and keep just the relevant features, which will help to have a better interpretability of the results.

Response: We thank Reviewer 2 for the thoughtful evaluation of our manuscript. We greatly appreciate the recognition of the importance of our central themes, such as the predictability of extreme flooding events and the comparative performance of national (UK) versus regional models. We acknowledge the concern that the framing of the paper did not fully align with the results, particularly regarding the role of weather patterns (WPs) and other features with limited or negative marginal value. Our revisions have addressed this misalignment and improved the interpretability of the work. We have revised the title and abstract to emphasize that the study tests and ultimately questions the added value of large-scale climate indicators like WPs, as part of a feature incorporation framework. The revised title reflects the papers contribution better, regarding the model framework. We have clarified throughout the revised text that WPs and other weak predictors were retained as part of a systematic evaluation of their marginal contribution relative to hydrometeorological and catchment-scale predictors. Then, they were ablated/removed, and VIF pruning was applied in the final model (feature set 7) to enhance interpretability. We ensured a fairer comparison between UK-wide and regional models by using matched catchment sets and consistent evaluation metrics. We provided a version of the final model without the WPs, AWPs, and latitude and longitude, to address the points further.

We believe these adjustments strengthen the manuscript by improving clarity, transparency, and alignment between framing and findings. Importantly, our decision to include WPs as part of the exploration is not contradictory to the results but reflects our aim to demonstrate and quantify the limited value of synoptic-scale indicators in this modelling context. We believe a negative result here is an important contribution, as the WPs are used in operational tools in the UK, such as Fluvial Decider and Coastal Decider.

R2C2: Another concern is how the results are presented. In Figure 4, the UK model is presented as the best model (Generation 6). This would mean the model should outperform regional models at least in 50% of the cases. However, Figure 5a shows us that the UK model has a lower median than NW and NE. That is possible given the variability in the model, as the authors describe; however, the figure may be misleading because to have a real comparison, the authors should compare the UK model with each regional model by selecting the same catchments in both, which appears not to be the case. In fact, Figure 5b shows that the concentration of blue dots is higher in the UK model, which is consistent with Figure 4. The author describes the Simpson's paradox as the problem; however, they forgot that they are responsible for having a fair comparison and splitting of the data. Therefore, they should avoid the paradox, which, from the differences between Figures 4 and 5, is not the case.

I suggest a major revision of the paper, given that they need to reframe the paper to the actual results and check that all the results are presented in a fair way to avoid misleading.

We thank the reviewer for these insightful comments regarding the comparison between the national (UK) and regional models. We fully agree that fair and transparent evaluation requires identical data sets for comparison, consistent metrics, and careful interpretation of performance differences between the UK and regional models. In response, we have made major methodological and presentation revisions to ensure a fair comparison and clear communication of results, as detailed below. We acknowledge that the previous version compared UK and regional model performance across partially different catchment sets, which could have led to misleading interpretations. To address this, we:

1. Recomputed all comparisons using ensuring matched catchments, such that the UK model is evaluated only on the subset of catchments included in each regional model (same locations, same test period).
2. Replaced Figures 7a and 7b with a new multi-panel figure showing, for each region in new Figure 5a (page 13):
    - (a) side-by-side boxplots comparing regional and UK-matched performance distributions (R^2)
    - (b) a spatial map showing the catchment-level difference (ΔR^2)
3. Removed all references to Simpson's paradox, as this was an imprecise explanation of the observed discrepancies. Instead, we now clearly state that differences arose due to aggregated versus disaggregated catchment sets.

Minor comments:

R2C3: Line 24 Check reference (?)

Amended.

R2C4: Line 29-30 What about events with no high intensity but longer duration?

Response: We thank the reviewer for this observation. While our text referenced key simplified mechanisms of flood generation such as intense rainfall and saturated soils, we agree that long-duration, lower-intensity rainfall events also play a critical role, especially in large, lowland, or more permeable catchments where flood response integrates over longer accumulation periods. To address this, we have revised the text to acknowledge this mechanism more explicitly and clarify that while high intensity rainfall is often emphasized, flood generation may also result from prolonged rainfall events that gradually saturate the soil and exceed drainage capacity.

R2C5 and R2C6: Line 43 Any idea why they have been used before?

Line 45 How have other studies been done before? You should present the baseline to have a clear benefit of your approach.

Response to R2C5 and R2C6: In response to the comments regarding line 43 and 45 together: We thank the reviewer for this question. These Met Office 30 synoptic scale WPs have been used in past studies usually to characterize broadscale atmospheric conditions that influence precipitation and drought, or to classify regimes relevant to hydrological modelling. For example, the work by Richardson et al (2018, 2019, 2020). More recently, they have been incorporated into decision-support tools such as the Met Office's Decider which has applications such as Fluvial Decider (Richardson et al, 2020) which uses WPs to assess flood risk probabilistically, based on the WPs association with UK regional precipitation distributions. However, despite their use in such frameworks, WPs have rarely been integrated into data-driven flood magnitude models, in large-sample machine learning contexts. This is possibly due to the added complexity of multi-scale feature interactions, limited precedent for encoding WPs in ML models, and a focus on more direct hydrometeorological forcings such as the prediction of precipitation. We have revised the manuscript to clarify both the rationale for their past use and the gap that our study addresses by formally evaluating their added value in a predictive ML framework.

R2C7: Line 87 What about the look at table approach implemented in RF? How can this go against your results?

Response: We thank the reviewer for raising this point. By "look-up table behaviour," we refer to the tendency of Random Forests to approximate predictions by memorizing specific combinations of feature values observed during training. The model can act like a stored table of rules, rather than learning generalizable relationships, especially when input patterns are highly repetitive or when training and testing distributions are very similar. However, in our case, we believe this effect is limited for two reasons. Firstly, due to event rarity and spatial diversity. Our focus on flood events above the 99th percentile helps ensure that the input patterns are highly variable, with relatively few repeated combinations. This reduces the chance of the model relying on what the reviewer is referring to as memorised look up rules. Furthermore, we employed temporal validation. The performance on the held-out data in time indicates the model is not memorising event specific combinations but capturing transferable patterns in time. We do agree this is a valid concern and could be more robustly considered in the modelling approach and is critical to consider when interpreting the model's decision structure. We emphasize that while RFs can overfit by forming overly specific partitions of feature space, our use of temporal validation and focus on extreme, spatially diverse events reduces this risk. We do not select a spatial-temporal cross validation due to inherent data sparsity when dealing with extreme flood events in a subset of the full catchment set (just the UKBN2 near-natural catchments). We have re-written the Machine learning model structure section, and highlighted that our decision is due to achieving a balance between robustness and data efficiency (page 8).

R2C8: Line 98-99 Why do you need to cluster the model if the RF architecture already does clustering? Why do you think your clustering can do it better than RF?

Response: We thank the reviewer for this insightful comment and acknowledge that our original phrasing may have created confusion. To clarify, we did not perform any additional clustering or data-driven grouping beyond what the Random Forest algorithm inherently does through recursive partitioning of the feature space. What we did do was to train separate models on predefined hydroclimatic regions (e.g., NW, NE, SE, etc.) in addition to a UK-wide model. These "regional models" are not clustered subsets generated by the algorithm, but spatially stratified samples based on established hydrological regions. This design serves a distinct scientific purpose: it enables us to (i) interpret model behaviour, performance and feature importance within consistent hydroclimatic contexts, (ii) assess whether regional models trained on more homogeneous samples perform better than a pooled UK model, and (iii) explore the degree to which flood-generating processes can be interpreted across scales. Thus, the regional grouping supports hydrological interpretability rather than acting as an alternative or competing clustering approach.

To be clear, we trained RF models across nine spatial samples (the UK national sample plus eight predefined hydroclimatic regions), for each of the seven feature sets, resulting in 63 model configurations in total (see Figure 4, page 20 which shows the heatmap of model run performance results). Each feature set incrementally incorporated additional predictors, enabling systematic evaluation of how physically interpretable features, spatial identifiers, and synoptic indicators contribute to model skill and explainability. We have revised the manuscript to reword this, and to clarify the language, removing any reference of regional grouping which may confuse the reader. We also amend the language throughout the paper, in the Methods and the Results and Discussion sections.

R2C9: Line 151-152 Is this analysis linear? If this is the case, what are the implications of that analysis?

Response: We thank the reviewer for this question. To clarify, the analysis described in this paragraph is not linear and does not involve fitting any regression or parametric model. It is a non-parametric, frequency-based assessment of conditional probabilities, specifically: P (WP | flood) refers to the probability that a given WP occurred, given that a flood magnitude event was identified. This approach is purely descriptive and intended to identify which weather patterns are most frequently associated with extreme flood events, both nationally and within regions. We also consider lagged WPs (up to three days prior) to account for potential lead time influence on flood magnitudes. We appreciate the implications of this analysis are interpretive rather than predictive, and it was our aim to highlight patterns of co-occurrence of synoptic patterns and floods, but we acknowledge this does not quantify causal or linear relationships. We revised the manuscript to explicitly state that this is not a linear or model-based analysis, to avoid any confusion.

R2C10: Line 169 Check reference (year?)

Amended.

R2C11: Table 3 Add all variables considered in each category and the abbreviation used per group.

Response: We thank the reviewer for this helpful suggestion. To improve clarity and reproducibility, we have expanded Table 3 which is now Appendix Table A2 (page25) to include the full list of variables introduced at each successive stage of the feature incorporation framework, and any abbreviations used in the results section or SHAP plots, and brief descriptions and units where appropriate. This overlaps with our changes made in response to R1C8, for Table 1.

R2C12: Line 173-175 This is a result, so it should not be in the methodology.

Amended, removed and replaced in results section. We explain here the reasoning for removing latitude, longitude, WPs, and AWPs from feature set 7 due to results showing that when removed model performance did not change, and other more meaningful static catchment characteristics were then used as given high importance. We then follow on to state the results of the hyper-parameter sensitivity tests that took place from the methods conducted.

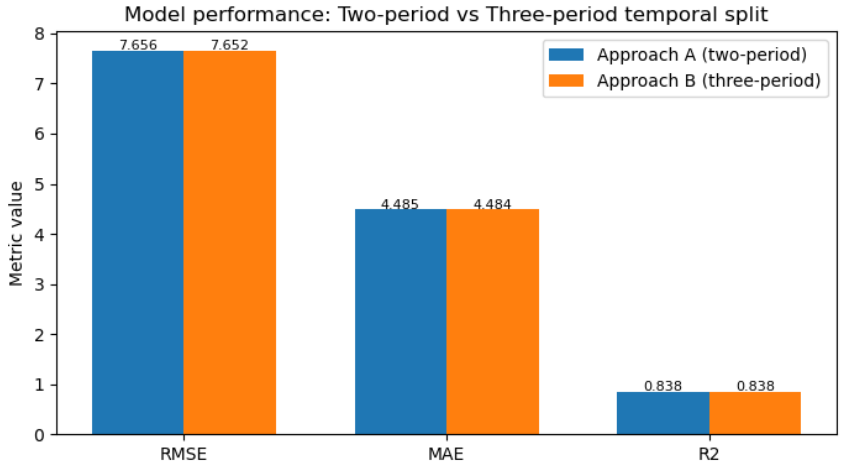
R2C13: Line 179 Why did you use a two-period splitting when in ML, three periods is the common practice to avoid overfitting and leaking information?

R2C14: Line 183-184 This is not completely true because you used this period for the hyperparameter search, so it is not unseen data

Response for RC13 and RC14 together: We thank the reviewer for this valuable point. We used a two-period temporal split, training (1969–2010) and testing (2011–2021) to ensure that the model was validated on future, unseen data in a time-consistent manner. This is common practice in hydrological research, where data length and the rarity of extreme events constrain sample availability, and the desire to 'forecast' future events. Hyperparameter tuning was performed exclusively within the training period sample using RandomizedSearchCV with internal cross validation folds. The test period was never used for model fitting or hyperparameter optimisation. The test set was reserved only for performance assessment and SHAP interpretability. This design avoids forward looking bias and reflects a forecasting scenario. While a three-period split (train–validation–test) is standard in many ML workflows, our setup reflects the structure commonly used in hydrological ML research, particularly when extreme events are sparse. To prevent overfitting and leakage: (1) hyperparameter tuning was performed using RandomizedSearchCV with cross-validation entirely within the training period, and (2) the test set was untouched during both training and tuning and was used only for final evaluation and SHAP interpretation. This approach respects the temporal integrity of the data and avoids forward-looking bias.

However, we acknowledge the reviewer's concern regarding this does not fully isolate the hyperparameter search process from the training period. Therefore, in response, we repeated the analysis using a three-period temporal split: training (1969–2000), validation (2001–2010), and testing (2011–2021). Hyperparameter tuning was performed using the training and validation periods only, ensuring that the test period remained completely unseen. Model performance metrics and feature interpretation results were consistent with the original two-period setup, indicating that data leakage and overfitting were not influencing the outcomes. Although we tested this three-period temporal split, the model's performance remained unchanged compared to the two-period

configuration. We therefore retained the two-period split (training: 1969–2010; testing: 2011–2021) as our main setup, as it maximises the available data for learning hydrological relationships while maintaining a strict temporal separation for evaluation. This design supports the goal of understanding temporal transferability and process-based relationships in EFM predictions, rather than focusing solely on generic predictive optimisation.



Response Figure 1. Comparison of UK national model performance under two temporal data-splitting strategies. Panel A shows results from the original two-period temporal split (training: 1969–2010; testing: 2011–2021), while Panel B shows results from the revised three-period split (training: 1969–2000; validation: 2001–2010; testing: 2011–2021). Performance metrics (R^2, RMSE, and MAE) were nearly identical across both setups, confirming that the two-period design did not introduce data leakage or overfitting. The consistency in results demonstrates the robustness of the modelling framework and supports the use of the two-period configuration, which maximises available training data while maintaining strict temporal independence for evaluation. Similar results were observed across all regional model samples.

We clarify this in the Methodology section.

R2C15: Line 197 The metric can take negative values.

Response: We acknowledge that the statement in the manuscript was incomplete. While R^2 is commonly described as ranging from 0 to 1 in well-performing models, it can take on negative values if the model performs worse than simply predicting the mean of the observed values. We revised the text accordingly to reflect this more precise and technically accurate interpretation. Please see line 530, Appendix C: Supplementary methods and equations. We have also rearranged the paper for clarity, moving the equations to the Appendix.

R2C16: Figure 2 Numbers must be located at the center of the column (x-axis)

Amended.

R2C17: Figure 3 You should uniform the text sizes in the figure.

Amended.

R2C18: Line 293 Could it be that WP30 is associated with the duration or total volume of the event? From figure 1, it is clear that WP30 is related to how big the WP is and the overlapping between the low-pressure area and the UK boundary.

Response: We appreciate the reviewer's thoughtful observation and agree that WP30's spatial extent and persistence may be more closely related to event duration or total rainfall volume, rather than peak magnitude alone. As correctly noted, Figure 1 WP30 is a broad, cyclonic system and this type of system is often associated with widespread and prolonged precipitation across the UK. This may indeed explain why WP30 is frequently

associated with flood days yet not consistently linked to the highest peak magnitudes (Figure 3a). These findings suggest that WP30 could drive lower intensity but longer duration events, especially in larger catchments where prolonged rainfall is a more important factor for flood generation. While this study focused on flood magnitude days in near-natural UK catchments, we acknowledge that evaluating event duration or cumulative rainfall volume would offer additional insights. We have included this interpretation in the revised discussion and suggested it as an avenue for future work. We also added a clarification in the manuscript regarding the frequent association of WP30 with flood days, despite not yielding the highest magnitudes, may indicate that it plays a greater role in driving long-duration or high-volume events, often in larger catchments. Please see the amended discussion, specifically around line 317 (page 17).

R2C19: Line 324 Check reference (?)

Amended.

R2C20: Line 381 Latitude and longitude do not have a hydrological meaning other than specific coordinates in space. For that reason, they are mainly used as an address to identify each catchment and are much more specific than any catchment attributes. If you want to have more hydrological meaning, latitude and longitude should not be used as attributes.

R2C22: Line 423-424 I agree that specific features could help, however, given the use of latitude and longitude, it will be hard the find features that are more specific than those. This is one of the problems of using attributes that are not hydrologically meaningful.

Response for R2C20 and R2C22 together: We thank the reviewer for these insightful comments regarding the inclusion and interpretation of latitude and longitude as model predictors. We fully agree that these variables are not physically hydrological in nature but instead function as spatial identifiers or proxies for location. Their inclusion in the initial feature set (Feature Set 1) was intentional, providing a baseline representation of inter-catchment variability and allowing the model to learn broad geographic patterns that may indirectly reflect hydrological processes (e.g., climatic gradients, physiographic differences). This decision was further supported by the inclusion of weather patterns (WPs) in later feature sets, where latitude and longitude provided essential spatial context for interpreting geographically dependent atmospheric influences. As shown in our conditional probability analysis, identical WPs can produce different flood responses depending on location. Thus, the model required spatial coordinates to properly associate meteorological drivers with regional outcomes. The strong performance of latitude and longitude in early generations therefore likely reflects their implicit spatial encoding rather than true hydrological causality. Similar spatial identifiers have also been used in previous studies (e.g., Slater et al., 2024) to capture inter-catchment variability in national-scale flood models.

When physically meaningful static catchment characteristics were added (e.g., Aridity Index, Baseflow Index, Runoff Ratio) on top of longitude and latitude in Feature Set 4 models, overall model performance changed little indicating that latitude and longitude had already captured much of the spatial variability between catchments. However, the Feature Set 4 additional variables offer stronger physical interpretability, providing hydrological meaning that coordinates lack (e.g., aridity index…..).

To address the reviewer's comment, we performed additional experiments to test the role of spatial and meteorological identifiers. As shown in Response Figure 2, we re-ran the original Generation 6/Feature Set 6 model under four configurations:

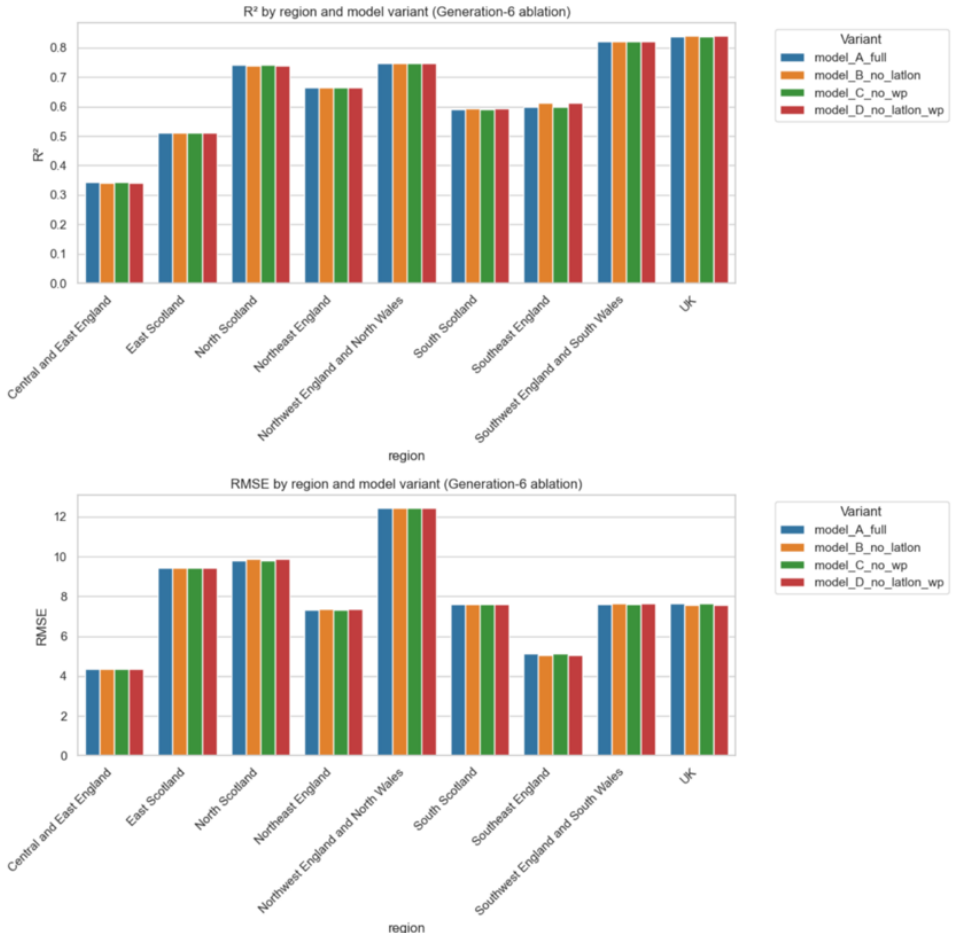
1. full model (all predictors)
2. ablation of latitude and longitude.
3. ablation of weather pattern variables (WPs and AWPs)
4. ablation of both spatial coordinates and weather pattern variables.

Across all regions, R^2 and RMSE values remained effectively unchanged, confirming that model skill was not dependent on latitude, longitude, or WPs. This demonstrates that physically interpretable hydrometeorological and catchment features are sufficient to explain flood magnitude variability. Furthermore, once latitude and longitude were removed, static catchment features such as the Aridity Index emerged as dominant predictors, indicating that the model naturally shifted toward process-based drivers when spatial proxies were unavailable.

These results directly informed the design of the final model (Feature Set 7), which excludes/ablates latitude, longitude, and WPs entirely and applies Variance Inflation Factor (VIF) pruning separately to UK and regional models. This final feature set ensures that all retained predictors represent physically interpretable hydrometeorological and catchment processes. Overall, this analysis confirms that while latitude and longitude can enhance baseline predictive skill by encoding spatial gradients, comparable and more interpretable performance can be achieved without them using more hydrologically meaningful static catchment characteristics. The revised manuscript now clarifies that (i) latitude and longitude functioned as spatial proxies rather than hydrological drivers, (ii) their role was systematically tested and removed in Feature Set 7. Substantial changes can be seen throughout the paper to reflect this.

R2C27: Line 465 Given the use of latitude and longitude in all the generations, and the high performance with the first generation, it is weird that those features do not appear as one of the most important features in SHAP.

Response: Thank you - latitude and longitude are no longer included in the final feature set 7 models, so will not appear in SHAP feature importance plots. All relevant figures have been remade to reflect this (e.g., Figure 6 and 7).



Response Figure 2. Ablation analysis of spatial and weather pattern features (Generation 6). Model performance (R^2 and RMSE) is shown across all regions for four variants: (A) full feature set, (B) excluding latitude and longitude, (C) excluding weather pattern variables (WPs and AWPs), and (D) excluding both latitude/longitude and WPs. Performance remains stable across all configurations, confirming that latitude, longitude, and WPs primarily encode spatial context rather than hydrological processes. The results support their removal in Feature Set 7, which focuses solely on physically interpretable predictors.

We have made amendments at multiple locations throughout the revised manuscript. Appendix Figure B1 (page 26) is a visualisation of the features that were retained and removed in feature set 7 for each of the regional and UK models. Throughout the Results and Discussions section, we revise the use and interpretation of latitude and longitude in the models, clearly, to indicate that that they are not hydrologically meaningful and encapsulate spatial variability (inter-catchment variability) between catchments in the UK and regional models.

R2C21: Line 412 Please describe the intra/inter concept before using it.

Response: Thank you for this helpful suggestion. We agree that the distinction between intra-catchment and inter-catchment variability should be clearly defined beforehand. We have revised the text to introduce and define these terms explicitly. Please see lines 205 -207 (bottom on page 8).

R2C23: Line 429 This goes against the findings you already mentioned and figures 4 and 5b.

Response: Figures 4a, 4b, 5a,5b have been updated to include final feature set 7 and are now Figure 4 and Figure 5 (page 12 and 13), and we show re-calculated UK vs regional catchment level R2 comparison. We have also revised the methodology to reflect our fair comparison.

R2C24: Figure 5b Could you replace the regional figure with the difference between the UK and the regional one?

Response: Yes, please see Figure 5 (page 13). We have also provided figures in the Appendix, Figures B3 and B4.

R2C25: Line 437 Check reference. I agree that it is an important issue; however, you are responsible for the framework to avoid that. How are you calculating the median for the local and global metrics? You should always be calculating the metric per catchment and then computing the median independent of the model, and all over the same period and the same group of catchments.

Response: Reference has been fixed. We appreciate the reviewer's request for improved clarity on how performance metrics were calculated and aggregated. In the revised manuscript Methods section and we explicitly describe that:

- Overall model R^2 value were calculated one per model using all test set predictions pooled across catchments within the UK or regional sample, for the overall model performance, displayed in Figure 4.
- Additionally, R^2 values were computed at the individual catchment level using test set predictions. Median and mean metrics were then computed from these per-catchment values, ensuring consistent comparison across models, regions, and time. Only catchments with ≥10 test events were included to strengthen statistical robustness. This was for the UK and regional comparison of performance at the catchment scale within each model (per region and UK subset of matched data) - See Figure 5.

R2C26: Line 447-449 Could this issue be just part of the overfitting, given the low number of events? I think more work must be done to clarify how the R2 is calculated. Maybe this conclusion is just an artifact of the computing method.

We also acknowledge that limited sample sizes for extreme events can produce noisy R^2 estimates and potential overfitting artifacts. We have therefore clarified that some variability in performance across catchments likely reflects data scarcity rather than true model generalization differences. These caveats are now discussed explicitly in the Results and Discussion section.

Overall, we have made major revisions through recalculating all metrics on matched samples, replaced Figures 5a and b with matched-catchment comparisons, removed the misleading Simpson's paradox reference, clarified per-catchment metric computation, acknowledged potential data limitations, and refined the interpretability discussion to properly contextualize UK and regional SHAP differences. Together, these revisions ensure that all results are fair, reproducible, and conceptually consistent and we have responded to the reviewers' comments thoroughly.

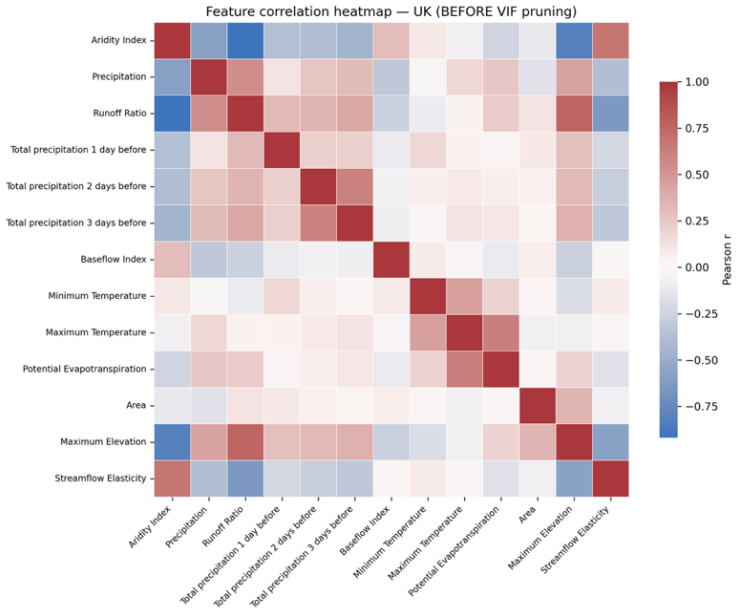
R2C28: Line 486-489 Many of these variables have other collinear variables, so you should try to prune the model with more independent variables. This way, you will have stronger relationships with the attributes.

Response: We appreciate this valuable suggestion. We agree that multicollinearity among input features introduces redundancy and may weaken the interpretability of feature importance scores. In our current framework, we did not explicitly perform collinearity pruning, as the Random Forest (RF) algorithm is often relatively robust to multicollinearity due to its ensemble structure and random feature sampling. For example, as discussed in Linder et al., 2022. However, we recognize that removing or grouping highly collinear features could improve both interpretability and reduce the noise in SHAP-derived importance rankings. In response to the reviewer's comment, we have now included the addition of feature set 7, and throughout the entire revised manuscript we refer now to seven feature sets, rather than six model generations. The seventh feature set has had longitude, latitude, WPs and AWPs ablated, and then a pruning technique applied. We conducted VIF pruning.

We make amendments to Figure 4a and Figure 4b which is now Figure 4 to include this feature set 7, and we show in Appendix Figure B1 which features are retained and dropped in each UK and regional model for feature set 7 - this is because feature set 7 is the only feature set unique to each of the UK and regional models, as VIF pruning is applied to each of the UK and regional data samples prior to running each of the models - taking into account the collinearity of each dataset sample (e.g., all flood magnitude data for the UK national model, and for example the subset of CEE extreme flood magnitude data sample).

We see in our results that removing collinear features does slightly reduce the performance of some models compared to their feature set 6 iterations. For example, UK R-squared result decreased from 0.84 to 0.83, NS from 0.75 to 0.66 (although the 0.66 is not statistically significant change with reference to the baseline model feature set 1), CEE from 0.37 to 0.33 and SW from the 0.83 to 0.82. We explain that this small reduction in performance when removing highly correlated variables as due to there being less information for the model to use to explain the target variable, as it is forced to rely on fewer features. We agree this does push the final feature set 7 model results towards a more interpretable set of SHAP results, and more robust conclusions. After pruning, the model becomes simpler, more interpretable, and more physically grounded, a trade-off between raw accuracy and scientific interpretability.

Correlation heatmap below is provided as a visual understanding - and then we applied VIF pruning after looking at this.



Response Figure 3. Feature correlation heatmap for the UK national dataset. The matrix shows Pearson correlation coefficients between all continuous predictors retained after removing non-hydrological variables (latitude, longitude, and weather pattern indicators). Warm colours indicate positive correlations, cool colours negative correlations. Strong correlations are observed among precipitation and temperature-based predictors, motivating the application of the Variance Inflation Factor (VIF) pruning technique.

R2C29: Line 501 Precipitation is well known in hydrology as one of the most important variables, so I would not say that this is interesting.

Response: We thank the reviewer for this observation. We agree and have revised the manuscript text accordingly.

R2C30: Line 505 That the UK model has different importances does not mean it does not capture the local variability of the importance. The UK's importance is just overall more important. You can think of the regional models as specific branches of a big tree. In that case, each branch has different importance because they are independent. Therefore, this comment is unfair to the UK model.

We thank the reviewer for the thoughtful analogy and fully agree that differences in SHAP feature importance between UK and regional models do not imply that the UK model fails to capture local variability. Instead, both report global feature importance computed across different sample domains:

- The UK model reflects importance aggregated across all catchments, while
- The regional models reflect importance within specific hydroclimatic contexts.

We have revised text throughout the Results and Discussion section. We also acknowledge that while SHAP values offer insights into model behaviour, they represent associations rather than causal relationships.

R2C31: Line 517 There are ways to quantify uncertainty. In fact, RF has already un uncertainty quantification that you did not use (ensemble). Therefore, more effort should be made to consider it.

Response: We agree with the reviewer that Random Forests offer a built-in mechanism for estimating predictive uncertainty. We thank the reviewer for highlighting this important point. In the revised manuscript, Predictive uncertainty was estimated using the distribution of predictions from the Random Forest (RF) ensemble. For each event, the ensemble mean (Pred_Mean) and standard deviation (Pred_SD) across all trees were computed, providing a direct measure of epistemic uncertainty. To allow comparison across catchments and regions with differing flood magnitudes, a dimensionless coefficient of variation (CV = Pred_SD / |Pred_Mean|) was also derived. We have added Methods section 2.5 Uncertainty quantification which outlines these methods. We include an additional figure in the Appendix, Figure B2, to show the Coefficient of Variation results.

R2C32: Line 518-530 Why should researchers spend time refining WP if they had zero importance? Maybe it could be more beneficial to refine catchment attributes that you have already proved are important.

Response: We thank the reviewer for this question. Our SHAP analysis no longer shows the WPs as they were ablated in the final feature set 7 models.

R2C33: Line 543-544 However, you defined the framework, why did you not change the percentile to 95% or another value to have more data? Do you think that the important features or relationships would change if you use other percentiles?

Response: We appreciate the reviewer raising this point. Indeed, the choice of the 99th percentile threshold was a deliberate one, intended to focus the analysis on the most extreme and impactful flood magnitude events. We agree that systematically investigating how feature importance evolves across percentile thresholds would be a valuable next step. Applying the full framework to moderate high-flow events would require retraining and reinterpreting all models to ensure comparable rigour. We plan to pursue this in future work as an extension to the current study. The 99th percentile threshold was intentionally selected to isolate the most extreme and impactful flood events, aligning with the study's focus on rare extremes rather than moderate high flows. This choice ensures that the analysis concentrates on the tail behaviour of the flood distribution, where both hydrological and meteorological drivers can behave differently from more frequent events.

Thank you very much! We thank you for your time and efforts in improving this work and have made substantial changes to address your useful comments.