Dear Reviewer,

Thank you for this constructive review. Please find our responses below (**in bold**).

Best regards,

Anna Luisa Hemshorn de Sánchez (on behalf of all authors)

This manuscript presents a large-sample, pan-European analysis of streamflow elasticities to precipitation and sensitivities to temperature, including mean and extreme flows. The dataset is impressive in scope, and the topic is timely and relevant for understanding hydroclimatic variability and change across Europe. The spatial patterns identified are interesting and potentially valuable for both science and water management.

We thank the reviewer for highlighting the timeliness and interesting results of our manuscript.

However, in its current form, the manuscript suffers from some conceptual, methodological, and presentation weaknesses that limit the robustness and interpretability of the conclusions. In particular, issues arise regarding (i) the definition and interpretation of "resilience", (ii) statistical robustness of regression and Random Forest analyses, (iii) insufficient data description, and (iv) weak integration between the main text and the Supplementary Material. Addressing these issues would substantially strengthen the paper.

I have reported below the major and minor comments the authors could consider to improve the manuscript:

We address these points individually below and thereby strengthen the manuscript. The main changes that we will implement in the revised manuscript are the following:

- **State clearly what we mean by resilience to extreme events and highlight the difference between event-scale and state-dependent causes, and set a minimum station density for the E-OBS climate data to reduce related uncertainties.**
- **Work on a clearer visual differentiation between statistically significant and insignificant catchments and explicitly state the fraction of catchments with statistically significant temperature coefficients.**
- **Further stress that there is interdependence among several of the predictors of the random forest analysis and that the results are mostly associative and revise the wording to reduce the emphasis on causality.**
- **Give more detailed information on the data used for this study.**
- **Better integrate the results of the Supplement while optimizing the main text.**

1. The manuscript interprets elasticities of annual maximum flows to mean annual precipitation as a measure of resilience to extreme flows (Section 3.2 and Figure 5). While the figure is informative and the spatial patterns are interesting, this interpretation is conceptually problematic.

Annual maximum flows are typically generated by event-scale precipitation (sub-daily to multi-day), whereas mean annual precipitation reflects a yearly integrated climatic state. Consequently, the elasticity metric used here does not directly represent resilience to extreme precipitation events, but rather the sensitivity of flood magnitudes to interannual hydroclimatic wetness and antecedent catchment conditions.

We agree that the presented elasticity metric does not reflect the resilience to extreme precipitation (P) events (which we also do not claim in the manuscript). Our purpose is to show the sensitivity of mean and extreme streamflow (Q) to interannual mean precipitation and temperature. This is relevant because, while annual maximum flows are triggered by event-scale precipitation (or snowmelt), direct comparison of annual maxima of P and Q yields a weak relationship (median R^2 of 0.16) and low elasticities (Fig. R1) because in most of Europe annual precipitation maxima and annual flow maxima usually occur in different seasons (e.g., Berghuijs et al., 2019). Indeed, maximum flow can still depend on annual mean P, indicative of the general wetness state in that year. We will now better emphasize this in section 3.1 and thereby also better explain what we mean by resilience in section 3.2.

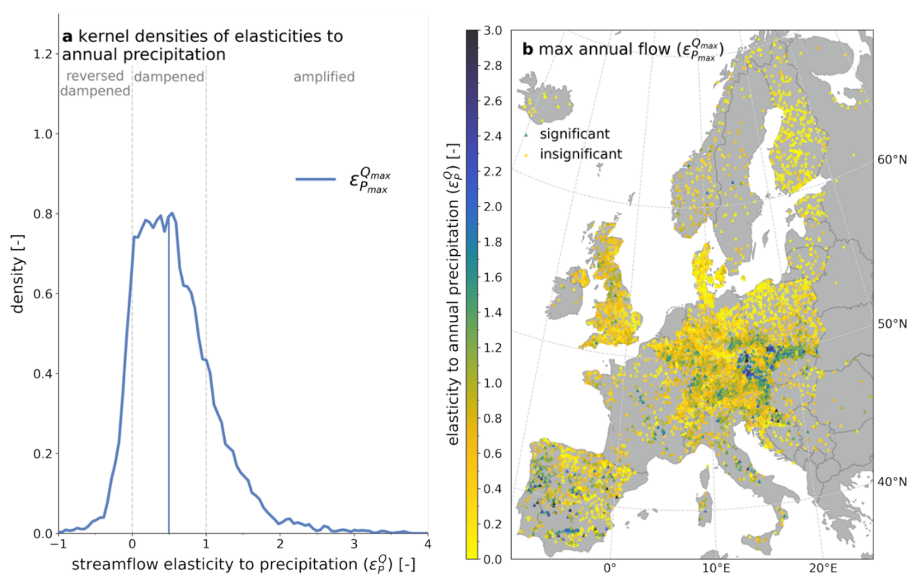

Figure R1: Frequency distribution (a) and spatial distribution (b) of annual maximum streamflow elasticity to the annual maximum precipitation.

This interpretation is, in fact, supported by the Supplementary Material. There, the authors test two hypotheses to explain the similarity between elasticities of mean and maximum flows: (1)

a correlation between mean and maximum precipitation, and (2) the control of antecedent wetness and landscape state on flood response. Their analysis (Figure S4) shows that the correlation between mean and maximum precipitation is weak and spatially scattered at the European scale, suggesting that hypothesis (2) dominates. This indicates that the derived elasticities primarily reflect state-dependent flood amplification rather than resilience to event-scale extreme precipitation.

While the metric itself is not meaningless, the terminology "resilience to extreme flows" risks overstating what is actually quantified. I therefore recommend explicitly acknowledging the time-scale mismatch, aligning the main-text interpretation more closely with the Supplementary findings, and softening or rephrasing the resilience terminology accordingly.

We agree that it is good to bring forward the implications of using mean annual (or seasonal) precipitation. We will do this by incorporating more of the interpretation from the Supplement (S3) into the main text of the revised manuscript. We think that the concept of "resilience of extreme flows" is not exclusively reserved for resilience of extreme flows to event-scale extreme precipitation but could also describe the resilience of extreme flows to the general wetness state of the landscape (mean annual P) or to the annual mean temperature. The difference in timescale will be discussed in the revised manuscript.

In addition, the use of mean annual precipitation derived from gridded datasets such as E-OBS introduces further uncertainty, particularly in southern Europe, where station density is very low and precipitation extremes are known to be less reliably represented (see here: https://climatedataguide.ucar.edu/climate-data/e-obs-high-resolution-gridded-meanmaxmin-temperature-precipitation-and-sea-level). This adds another layer of uncertainty when relating annual precipitation metrics to extreme flows in these regions.

To reduce the uncertainty related to low station density, we will set a minimum station density of E-OBS as a criterion for the catchment selection for the analysis of the revised manuscript. We will state this procedure and remaining uncertainties more clearly in the revised manuscript. We only use the E-OBS data to get annual mean P and T (instead of annual extreme P)_which also reduces the impact of this uncertainty.

2. The use of a multiple linear regression framework to estimate elasticities and temperature sensitivities is appropriate. However, statistical significance in such models must be assessed at the parameter level, not merely at the level of model fit.

This is especially relevant for the temperature coefficient, which explains a very small fraction of variance ($R^2 \approx 0.03$ when used alone). As far as I understood, while the authors state that parameter-level p-values are used, the manuscript does not clearly show how often temperature coefficients are statistically significant, nor this is clear from the figures and whether inclusion of temperature significantly improves the model relative to precipitation-only models.

Without this information, temperature sensitivities risk being over-interpreted. I suggest, to explicitly report the fraction of catchments with statistically significant temperature coefficients, and clarify the added explanatory value of temperature relative to precipitation-only regressions.

Indeed, we did calculate statistical significance at the parameter level using p-values. Based on this we make the differentiation of significant and insignificant catchments in the maps (Figures 2b-d, 4b-d, 4f-h, 5b-d, 5f-h and 5j-l). We acknowledge that the differentiation is visually not that clear. In the revised manuscript we will work on a clearer visual differentiation and state explicitly the fraction of catchments with statistically significant temperature coefficients.

We also want to emphasize that while sometimes at the station level the relationships are statistically insignificant, the exposed broadscale regional patterns of low and statistical insignificant streamflow elasticities to precipitation and streamflow sensitivities to temperature suggest systematic regional differences in catchment functioning even when the individual stations are uncertain. Having regions where streamflow is not sensitive to precipitation or temperature is relevant to know and will logically be statistical insignificant.

3. The Random Forest analysis is used to infer which catchment characteristics "shape" elasticities. However, several methodological aspects are insufficiently documented or justified:

- The paper does not clearly present training vs. testing performance, nor any assessment of robustness across multiple splits or cross-validation.

  The performance of the training sample will be stated in the revised manuscript in addition to the already stated testing performance. In the revised manuscript we will add a cross-validation.

- Reported R^2 values (≈0.30–0.51) indicate that a substantial fraction of variability remains unexplained, which is understandable but limits interpretability.

  The implications of the low R^2 values will be highlighted further in the revised manuscript.

- Feature importance is derived from impurity-based metrics, which are known to be biased in the presence of correlated predictors, a major issue given the strong interdependence among climate, soil, and landscape variables. While predictor independence is not required for RF prediction, it strongly affects feature importance interpretation, which here is framed in physical terms. In this respect, I suggest to clarify that random forest results should be interpreted as associative rather than causal, discuss limitations of impurity-based importance under collinearity, and provide clearer information on model validation and robustness.

We agree that flows relate to landscape features but that correlations with individual landscape features are not always causal. This problem is not unique to this manuscript but present in most empirical studies on flow behaviour and landscape characteristics.

In the revised manuscript, we will further indicate that there is interdependence among several of the predictors and that RF results are mostly associative. At the same time, to reduce the interdependence of the predictors, we excluded parameters with a higher collinearity than p>0.8 as mentioned in the manuscript in lines 175-176. In the revised manuscript we will revise the wording to reduce the emphasis on causality.

4. The manuscript relies on numerous datasets (streamflow, precipitation, temperature, catchment attributes, human influence metrics), many of which are introduced with minimal or no description. Key information is often missing, including:

- Histograms with area of the basins.

We will add histograms of the basin area to the revised supplement and state their size range (5th percentile, median, 95th percentile) in the main text. The spatial distribution of the catchment area is shown in Fig. S7 of the Supplement.

- data sources and spatial resolution for climate variables,

In the revised manuscript we will add information on the data sources and the spatial resolution of the climate variables from E-OBS.

- temporal aggregation and handling of missing data,

In section 2.1 and sections 2.2 we explain the temporal aggregation to hydrological years and seasons based on monthly values from EStreams. We also mention that we used missing data to filter for a minimum of 15 years of valid data (lines 86-88). In the revised manuscript we will add the information on minimum valid monthly data to calculate the value for each hydrological year or season. Further, we will differentiate better between the processing of

streamflow data and climate variables in terms of temporal aggregation and handling of missing data in the revised manuscript.

- spatial aggregation and handling of small catchment precipitation representativenss vs precipitation and temperature spatial resolutions.

  As also mentioned in the reply to comment 1, we will set a minimum station density of E-OBS as a criterion for the catchment selection for the analysis of the revised manuscript. This will then be set in context with the smallest catchment size.

- whether predictors are static or time-varying,

  Most predictors are classified as static in the EStreams dataset. Some predictors, like the land cover that were time-varying in EStreams were averaged over time for the analysis. We will add this information to the revised manuscript.

- uncertainty and limitations of human influence datasets.

  We agree there are limitations and uncertainty in the data on human influence datasets. We will better acknowledge that in the revised manuscript.

I think that referring readers to EStreams alone is not sufficient, given the interpretive nature of the study. Adding a concise but explicit data description section or table summarizing sources, resolution, temporal coverage, and key preprocessing steps would clarify better the paper and its potential effect. Also, consider adding potential limitations of the datasets.

In the revised manuscript we will expand the section describing the data used by including sources, resolution and temporal coverage. We will highlight key processing steps more clearly in the revised manuscript. For the predictors we will also add information on the data sources to Table 1 of the revised manuscript.

5. Several essential analyses (regression diagnostics, correlation structures, robustness checks, RF diagnostics) are relegated to the Supplementary Material, but their implications are not consistently summarized or integrated into the main text. Conversely, some sections of the main text are overly long and repetitive. I think that this creates a disconnection that makes it difficult for readers to fully evaluate the robustness of the results without repeatedly consulting the Supplement.

In the revised manuscript we will better integrate the results of the SI while optimizing the main text.

Minor comments

- Environmental zones are introduced without prior definition.

  In the revised manuscript we will briefly introduce and add the reference to the study that defined the environmental zones.

- Chiew et al. 2006 are not properly defined. Please add *a* and *b* as one study refers to Australia and one is global.

  In the revised manuscript we will add a and b to the two different references.

- Lines 50-52. Can you be more precise with "narrow climate change" here?

  In case you are referring to "narrower climate range" in lines 160-163, we will rephrase "European scale" to "European range of climates" in the revised manuscript to clarify what we mean.

- The distinction between statistically significant and insignificant elasticities is visually unclear in several figures.

  See reply to comment 2.

- Lines 273–275: Additional recent literature could be acknowledged, including Nanda et al. (2023), https://journals.ametsoc.org/view/journals/hydr/26/7/JHM-D-24-0143.1.pdf

  Note that the URL links to another paper. However, if with Nanda et al. (2023) you are referring to this study (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4635449), it is a modelling study on the Lower Mekong River Basin that focuses on improving streamflow simulations by comparing four different configurations: an open loop run, calibration with soil moisture data, calibration with discharge data, and a combination of both. While this is related to the general theme of studying streamflow, it is not that closely related to our work.

- Line 298: See also Massari et al. (2024) for a more recent contribution: https://www.sciencedirect.com/science/article/pii/S002216942300954X.

  In the revised manuscript we will refer to this article.

- Lines 350-353. Possible effect of soil crusting?

  This could be the case, but other drivers might also be causing this.

- Line 405. “The most”?

  In the revised manuscript we will rephrase to “The more humid basins” to “humid basins”.

- The manuscript is considerably longer than typical journal standards and could be substantially shortened without loss of scientific content.

  In the revised manuscript we will optimize the wording to make it shorter where we can and thereby stay within the length appropriate for HESS. The total word count, including the abstract, main text, and declarations sections is now 7472 and the main text contains 7 figures.

References

Berghuijs, W. R., Harrigan, S., Molnar, P., Slater, L. J., and Kirchner, J. W.: The Relative Importance of Different Flood-Generating Mechanisms Across Europe, Water Resour Res, 55, 4582–4593, https://doi.org/10.1029/2019WR024841, 2019.