Response to reviewers for Ashcroft et al. "*AI based seasonal large ensembles for fluvial flood risk: Evaluation over the Elbe basin*"

We thank the reviewers and the editor for their careful reading of the manuscript and for their constructive and insightful comments. Below, each reviewer or editor comment is reproduced in black followed by our response in orange. Section references refer to the revised manuscript unless otherwise stated.

Reviewer 1

This manuscript provides a novel framework for supplementing the short period of record of historical precipitation and river flow data through an AI-based methodology that creates hundreds of ensemble members from which to assess rare extreme events. This methodology represents a creative and sophisticated approach to addressing the problem that low-likelihood, high-impact events are rarely seen in our short historical period. I am impressed by the breadth of knowledge presented on the methodology in this study and think this application will constitute a considerable advancement on the academic literature of this topic.

While I think this approach can be a meaningful contribution to the literature, there are several areas where the manuscript's contents can either be better explained or address topics that are currently lacking. I will detail my main concerns below, with a list of minor comments after.

Major comments

-How do we know that the precipitation outputs of PrecipHENS are physically plausible?

We know, and the authors do an excellent job of showing, that the PrecipHENS precipitation outputs are statistically plausible in reference to the historical data, but the modeling framework presented here is entirely based on an AI-based emulator of ERA5, so it is important to know whether or not these outputs represent potential physically-realistic scenarios. In short, how do we know that PrecipHENS generates thousands of "right answers" instead of thousands of "wrong answers". With a physics-based framework like UNSEEN, we can trust that they are "right answers" because the underlying model is a physics-based dynamical model, but that is not the case for

PrecipHENS. I acknowledge that this is a very difficult question to answer with any degree of certainty, so I think the authors should at least acknowledge that this remains an open question related to their method if they are unable to fully answer this question.

We agree that this is a key open question. We have added explicit text to the Discussion acknowledging that, because SFNO is an AI-based emulator rather than a physics-based dynamical model, it cannot provide the same formal physical guarantees as approaches such as UNSEEN (lines 747-750).

To provide some reassurance that PrecipHENS does not generate systematically unphysical patterns, we have refined the PCA-based analysis in Section 3.1.4, Figures 10-12. Originally, the analysis used a historical event from 2003 (within the SFNO training period); we now use a storm from December 2023, which is outside of the SFNO training period. PrecipHENS is still able to produce events with coherent spatial structures similar to this out-of-sample storm. While this does not demonstrate that *all* generated events are physically correct, it does show that the system can produce coherent, spatially realistic precipitation structures that are distinct from its training examples. In changing this example we have therefore updated Figure 18 to use the closest members to the new date, though this has not impacted the discussion around the Ffgure.

We have expanded the Discussion to emphasise that assessing physical realism in long AI-generated weather sequences remains a developing area, with two key considerations (lines 790-796):

- Most AI models are trained and validated on short-range forecasts; much less work has examined if they maintain physical consistency over weeks to months, which is the timescale relevant to our application.
- As AI-generated datasets become more commonly used in natural-hazard risk applications, there is a growing need for dedicated meteorological assessment from specialist research groups. We highlight this need explicitly and encourage future collaborations between AI-model builders, hydro-meteorologists, and the climate-hazard community.

-What do the data distributions look like for each of the three cases (Historical, Benchmark, PrecipHENS)?

The authors show many sophisticated statistical analyses to illustrate that PrecipHENS passes the tests for G1-4; however, a more basic depiction of the precipitation data generated for each of these cases is missing. It would greatly improve the manuscript to see how the distributions (e.g., PDFs) vary across the Historical, Benchmark, and PrecipHENS, especially regarding the tails. This is important because we want to know if the precipitation data is being drawn from the same distributions or not, which has implications for the differences between Benchmark and PrecipHENS shown

throughout the paper. One possibility here is that I can see the underlying distribution as a way to potentially show that PrecipHENS is generating hundreds of "right answers" (from the comment above), especially if there is no discernable difference between the Historical and PrecipHENS PDFs.

Additionally, it would be helpful to see a precipitation plot version of Figure C4 (which I really like!) to get a better understanding of the similarity of actual precipitation values between the Historical and PrecipHENS. This also can help with an understanding of how trustworthy the precipitation output from PrecipHENS is.

We agree that examining the full precipitation distributions is a useful diagnostic, particularly for understanding differences in the bulk and tail behaviour across datasets.

Upper-tail behaviour

The behaviour of the upper tail is already explicitly assessed through the fitted GEV parameters (Figure 8) and the corresponding return-level estimates (Figure 9). These diagnostics capture the extremes of the precipitation distribution.

Bulk and mean behaviour

The spatial distribution of mean precipitation is shown in Figure 4, and the bootstrapped mean precipitation in Figure B1 further explores central tendencies across the domain. As an additional illustration, Figure R1 shows the full distribution for Dresden, considering only "wet" days (≥ 1mm), consistent with the manuscript. For Dresden, this corresponds to 46.8% of Historical days, 46.6% of Benchmark days and 44.6% of PrecipHENS days. The main deviations apparent in the PDFs reflect the dry tendency in the bulk of the PrecipHENS distribution: consistent with what is already visible in Figures 4 and B1. For this reason, we do not propose including Figure R1 in the main manuscript, as it represents a single--site illustration rather than a regional assessment. However, we provide it here for completeness and transparency.

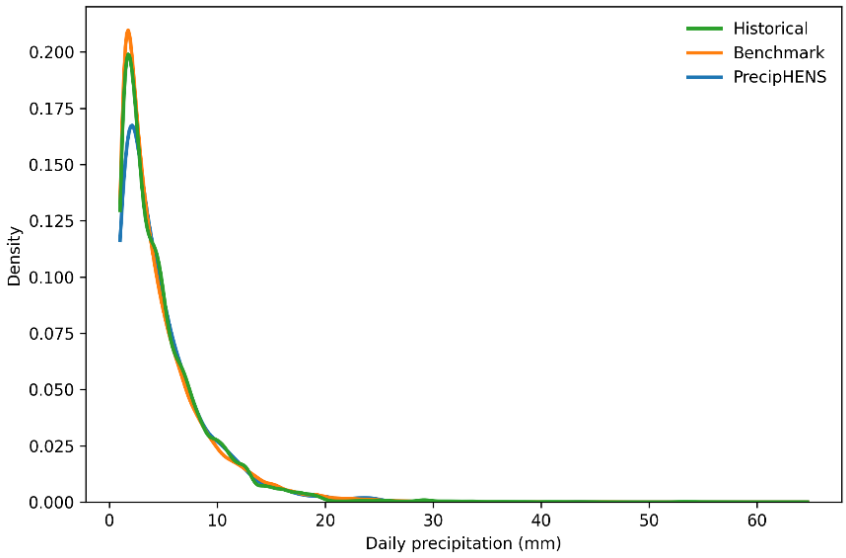

Figure R1 Kernel density estimates of daily precipitation (≥ 1mm) for the Historical, Benchmark and PrecipHENS datasets at Dresden.

We also produced a precipitation analogue to Figure C4 (Figure R2). However, because precipitation is substantially noisier than river flow, the resulting timeseries is dominated by high-frequency variability, making it difficult to discern any interpretable patterns. We therefore do not propose including this figure in the manuscript but provide it here to demonstrate that this option was explored.

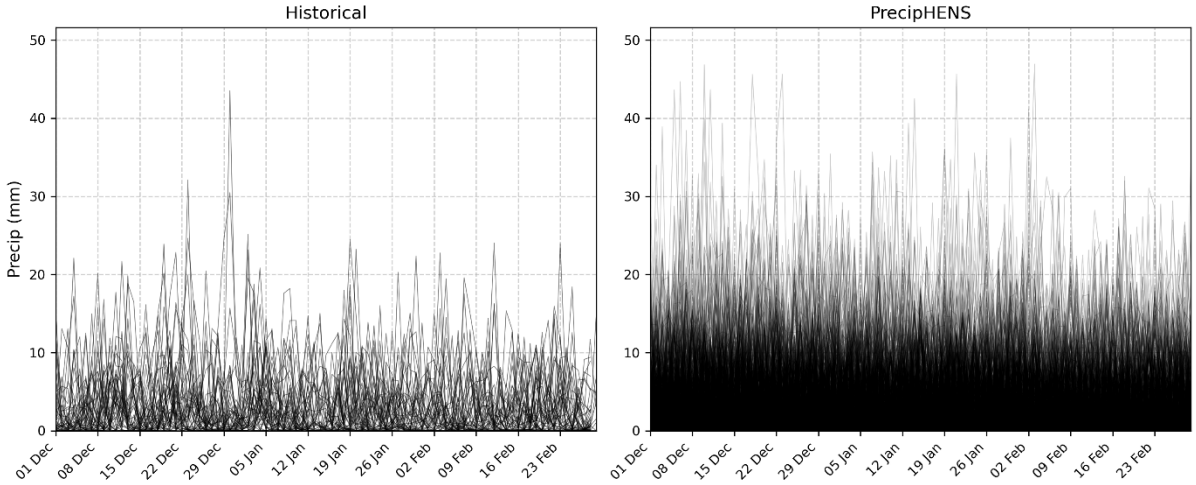

Figure R2 Full-season precipitation timeseries at the Dresden reference location. The historical data (left) are the historical precipitation series from all 43 seasons, while the PrecipHENS data (right) are each of the 1008 weather simulations.

-What are the big picture goals this approach is trying to accomplish?

A clearer representation of the problem from the beginning of the manuscript would help to ensure there is no misunderstanding related of the capabilities of this methodology. For example, this method is entirely based on historical data, so it is very relevant to an understanding of extreme events today (or in the next few years, let's say). But as extreme events are occurring with greater frequency and magnitude, it will likely

become out-of-date (i.e., an underestimation) for calculation of extremes a decade from now or longer out into the future. This is an important point about this method that is not currently addressed by the manuscript.

This can also help to contextualize the illustrated dry bias of PrecipHENS. While it may not be entirely generalizable for the Elbe River individually, extreme precipitation in general is increasing in the future with climate warming. Is it a problem that the PrecipHENS approach is showing a dry bias in reference to the historical data when we know with a fair degree of certainty that the historical data is likely the lower end of what we expect in the near future? This can especially be seen in the longer return period events in Figure 9 that clearly have a dry bias (far fewer points falling above the reference line than below).

We thank the reviewer for prompting clarification of the intended scope and limitations of the framework. We have revised the Introduction and Discussion to make three points explicit.

Intended scope (present day risk)

We have updated the Discussion to state explicitly that SFNO is trained on ERA5 data from 1979-2015. We acknowledge that, taken alone, this date range may appear outdated relative to today's climate. We now clarify in the manuscript that SFNO is not intended to memorise historical sequences, but instead learns operators governing large-scale atmospheric evolution rather than any specific weather sequences of the training period (Bonev *et al.*, 2023) (lines 197-199). Nevertheless, we emphasise in the Discussion that PrecipHENS is fundamentally tied to the recent-historical climate on which it is trained, and therefore represents the present-day risk rather than future non-stationary changes in extremes. We now state this limitation explicitly and highlight that assessing long-term climate-change signals is beyond the intended scope of the method (lines 751-757).

Stationary limitation

In the Discussion, we now acknowledge explicitly that PrecipHENS inherits the assumption of stationarity from its training data. As the reviewer notes, future extreme precipitation is expected to intensify under climate warming; thus the method will become increasingly misaligned with future conditions if used for long-horizon climate-change applications. We highlight this limitation and state that the approach is intended for representing present-day variability and tail risk rather than future non-stationary extremes (lines 751-757).

We also emphasise in the Section 1.4 that PrecipHENS should be understood primarily as a framework for generating large ensembles of plausible weather sequences for present-day flood risk estimation, rather than as a definitive or optimised product. The

core contribution of this work is to demonstrate the potential of AI weather models to generate seasonal-scale precipitation and temperature fields suitable for flood-risk estimation.

Interpretation of dry bias

With regards to the dry bias in PrecipHENS over the Elbe region, we would like to clarify that this is not a systematic bias globally. Based on preliminary follow-on analyses, we have evidence that in some regions PrecipHENS exhibits a dry tendency whilst in others it shows a wet tendency, and that this bias can also vary seasonally. Because the precipitation is an output of the diagnostic model that has 20 input variables, a rigorous understanding of the precipitation bias should involve those input variables themselves, which were not retained in this proof-of-concept study due to storage constraints. We have mentioned the future collaborative work required with meteorologists to further understand and unpick the multivariate biases present in such data (lines 790-791).

While Figure 9 may visually suggest a dry bias at the longest return periods, the interpretation of these deviations is complicated by substantial uncertainty in the historical return -level estimates. For completeness, we note that for all return -periods shown there are more points above the 1:1 line than below (RP100 return level for the 1-day accumulated precipitation features 508 points above the 1:1 line compared to 260 below); however, we agree that uncertainty dominates interpretation at the most extreme return levels.

The deviation from the 1:1 line in Figure 9 should not be interpreted as a model bias alone. The short historic period means there is high uncertainty in the RP fit, so our goal with Figure 9 is not wholly to have complete agreement with the 1-1 historical line. The motivation of Figure 9 was comparable figure set-ups in relevant studies (e.g., Figure 8 in Quinn et al. (2019)), and it is difficult to show the historical uncertainty on Figure 9 which includes all grid points. We have created an additional figure (Figure R3 here and Figure B4 in the revised manuscript) for the location at Dresden, used throughout the paper, to highlight the 1-day precipitation extreme exceedances that includes uncertainty bounds. In this case, the precipitation estimates associated with high return periods are higher in both the benchmark and PrecipHENS than the historical data (i.e. the Dresden grid cell would be one of the points above the reference line in Figure 9), however the entirety of the PrecipHENS 95% confidence interval sits within the associated historical interval. We have included this Figure in the appendix of the paper to assist in visualising the uncertainty of the historical estimates which we have clarified in lines 521-525.

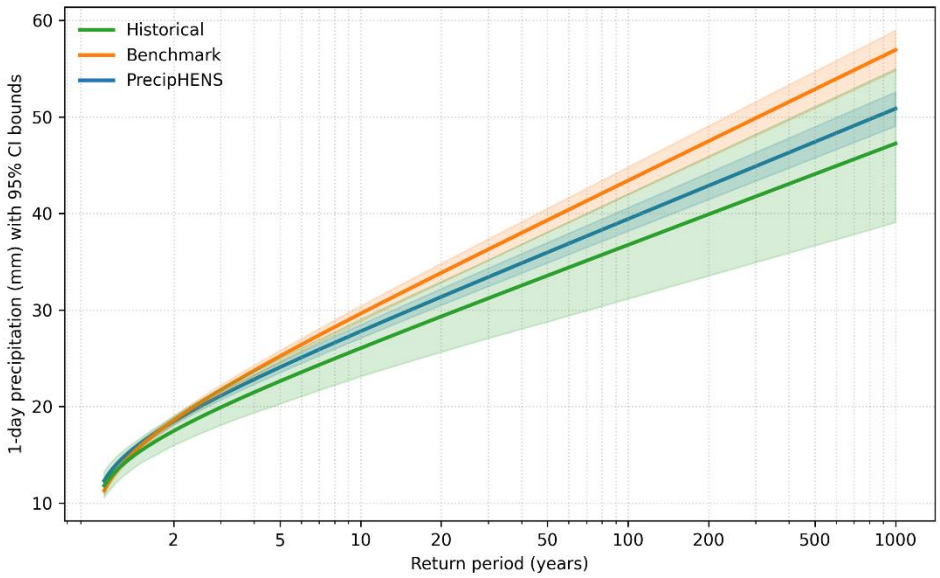

Figure R3: The estimated return period curve (with associated 95% confidence interval) obtained by a Gumbel distributional fit across the historical, benchmark and PrecipHENS data for the Dresden location.

-Add motivation for why the Benchmark method is used

The Benchmark method is essentially a resampling of the historical period to give it a much longer period of record, but with the same underlying density of extreme events. Why is it necessary to construct this Benchmark method as a point of comparison with PrecipHENS instead of performing a comparison with only the Historical Data? This type of motivation in Section 2.2.1 would improve this manuscript.

The benchmark is not simply a resampling of the historical storms: it resamples the storm-arrival process but also perturbs the magnitude, spatial structure and peak location of the storm. This produces a long, diverse synthetic time series that preserves the observed dependence structure and large-scale storm morphology through the conditional multivariate extreme-value model, while generating sufficient variability in storm footprints to support robust statistical characterisation of extremes. Because flood-risk estimation is performed at regional scales, where the spatial coherence of precipitation fields is a critical driver of losses, this dependence-preserving behaviour is essential, and is one of the reasons the approach is widely accepted in the flood-risk community as fit for purpose.

We use the benchmark for two primary reasons. First, it enables us to judge how closely different summary statistics of PrecipHENS match a distribution that is already validated and routinely used in operational flood-risk estimation. Since the observational record contains only a limited number of spatially coherent extreme storms, a benchmark that reproduces regional-scale dependencies provides a more robust reference point for evaluating model performance. This effectively establishes an expectation for the magnitude and nature of deviations between the historical record and PrecipHENS. We see this across Figures 4–9, where the benchmark aligns closely

with the historical distribution, providing confidence that deviations observed in PrecipHENS are meaningful.

Second, the perturbation schemes of both PrecipHENS and the benchmark allow us to quantify the expected variability in a large cohort of spatial storm footprints: again, a key requirement for regional-scale flood-risk modelling. The benchmark therefore provides a meaningful reference against which to compare the footprint variability generated by PrecipHENS, as illustrated across Figures 10-12.

We have added text to Sections 1.1 and 2.5 (e.g. lines 366-371) explaining that the benchmark extends the historical dependence structure while introducing new realisations, thereby creating a controlled baseline for assessing the magnitude and nature of differences between PrecipHENS and historically consistent extremes.

-Selection of initial conditions

Are all initial conditions for PrecipHENS taken from 9-16 Nov. 2023 as lines 243-245 appears to suggest? The description here is a bit unclear on these details related to the model. If initial conditions are indeed all derived from this 7-day period, I would argue that this is unnecessarily restricting the variability of the initial conditions, which would be a potential issue with the model and its results. If this were the case, the model would be considerably more robust with a more diverse set of initial conditions for atmospheric patterns than simple variations on a 7-day timeframe. I am hoping that I am understanding this poorly and there is more diversity in the set of initial conditions run with the model.

In short, all PrecipHENS ensemble members were initialised from lagged start dates between 9–16 November 2023. This choice was motivated by the need to remain outside the training period, allow synoptic scale uncoupling via a burn-in phase, and remain within computational constraints for this proof-of-concept study. We agree that a wider range of initial conditions would be desirable for an operational or optimised implementation and have added this to the Discussion (lines 745-746).

Outside of training period

It was important for this study to ensure that the ensemble was created from initial data outside of the training period of PrecipHENS to assess the ability to replicate climatological patterns and extrapolation. Given that SFNO had training from 1979–2015, validation from 2018, and testing from 2020, there were essentially 6 options for initial years to choose from.

Burn-in period and synoptic-scale uncoupling

It was important for this analysis to avoid repeating the historical ERA5 weather, i.e. as SFNO is a forecasting model, we would not want the simulations to be predictive over the time period we are analysing (the winter season). We therefore employed a burn-in

period of 9–16 November to 30 November, inclusive, to allow the simulation to uncouple from the initial conditions at the synoptic-scale, whilst retaining the 'of interest' part of the simulation being 1 December to 28 February. A priori, we were unsure how the long-rollouts of SFNO would behave with the lead-time (in terms of both the stability of the simulation over an excess of 3 months and reversion to long-term climatology), and so to limit the potential impact of this, we wanted the initial conditions to be as close together as possible. We had the computational budget for 8 initial lagged dates for our {8 (dates) x 9 (perturbations) x 14 (checkpoints)} design, and so that led to the 9–16 November.

Pragmatism within the already limited scope

We acknowledge that this proof-of-concept was about presenting a framework for developing upon, rather than an optimised product/solution, hence why the ensemble here was only for a winter season, rather than a full annual seasonal approach. Given the limited scope in those two respects, needing to cover all 6 of available initial years was not a priority. For an optimised solution following on from this study, a wider range of initial conditions would be suggested.

Large-scale uncoupling in the ensemble

Although we anticipated synoptic-scale uncoupling during the burn-in period, the preservation of large-scale patterns was unknown *a priori*. We have briefly explored the monthly SOI (southern oscillation index) arising from the PrecipHENS ensemble from this study, see Figure R4. This shows that the ensemble has a good spread in ENSO state despite being formed from initial dates limited to November 2023. There is evidence that the weakly negative state from ERA5 in November 2023 has persisted in terms of the mean of the ensemble (the mean of the PrecipHENS ensemble for each month is below zero). These initial results do indicate that a more diverse range of initial dates would be required for an optimised solution.

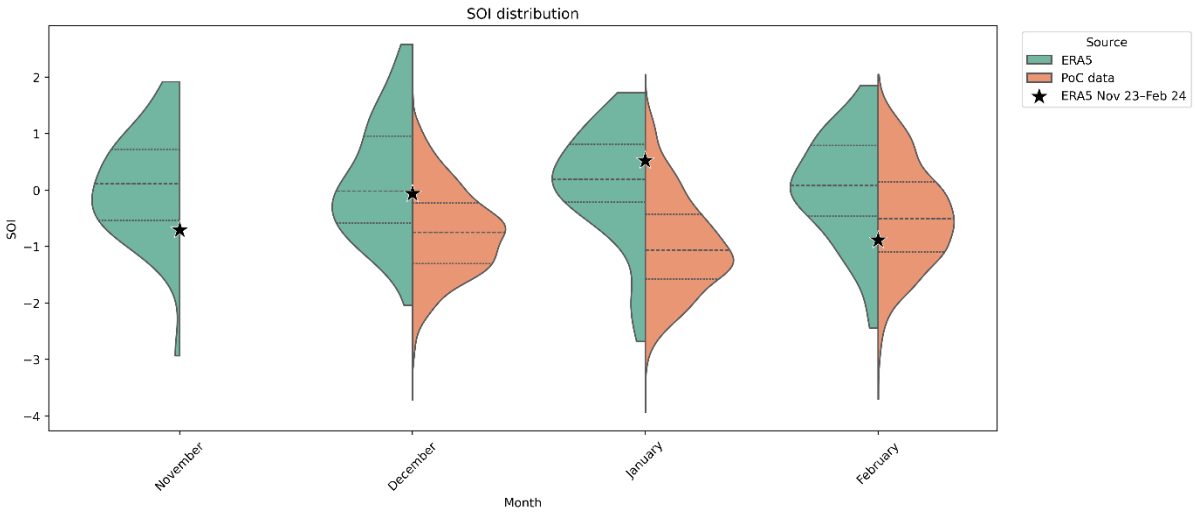

Figure R4 The SOI distribution (monthly) for ERA5 1980–2024 (green, left violin) and the PrecipHENS ensemble (orange, right violin). The monthly values for 2023 are highlighted with stars, noting that the ensemble was created using initial data from a set of lagged dates in November 2023.

Minor comments

-Line 191: What is the historical record here? Stations? ERA5? Something else?

Updated to "historical data" which is defined in section 2.1.2 (line 381).

-Figure B2: I recommend using a different color scale for this figure because it is temperature. The figure currently gives the incorrect assumption that PrecipHENS is biased cold. I suggest either flipping the bias color bar direction so brown is warmer, or using a blue-to-red color bar with red indicating warmer.

Thank you, we have updated to a blue-to-red colour map (Figure B2).

-Lines 388-390: Why are the PrecipHENS' Gumbel parameters closer to the Historical than the Benchmark even though the Benchmark is based on inference about the extremes? It would be nice to get another sentence or two describing why this counterintuitive behavior exists.

The tendency for the benchmark to overestimate the frequency of the most extreme events is most likely a construct of the underlying resampling method for event frequency that it uses. In the manuscript we have summarised this takeaway with the following text:

"...suggests that the benchmark approach tends to over-represent the tail of precipitation, and that this over-representation accelerates as the tail becomes more extreme. This is likely due to the methodology of the statistical extrapolation, whereby the historical events are repeated and perturbed at the rate of the limited historical record, without consideration of whether the representation in the historical record is uniformly likely. For example, the method is sensitive to the presence of a rare, extreme

event that is in the historical input record and is likely over-sampling the rarest, most extreme observed events." (lines 526-532)

-Figure 9: Did the authors consider showing the 500-year and 1000-year return periods on this plot as well? Since there are 1000+ years of data, this approach would also be applicable to these long return periods.

We thank the reviewer for the suggestion and have provided this figure below (Figure R6). We have found that, although these long return periods can be computed from the dataset, the resulting figure becomes substantially more clustered due to the overlapping points. In addition, the observed points become increasingly uncertain with increasing return periods. Importantly, the overall patterns and conclusions remain essentially unchanged from those in the manuscripts. Thus, for clarity we retain the original figure for visual clarity.

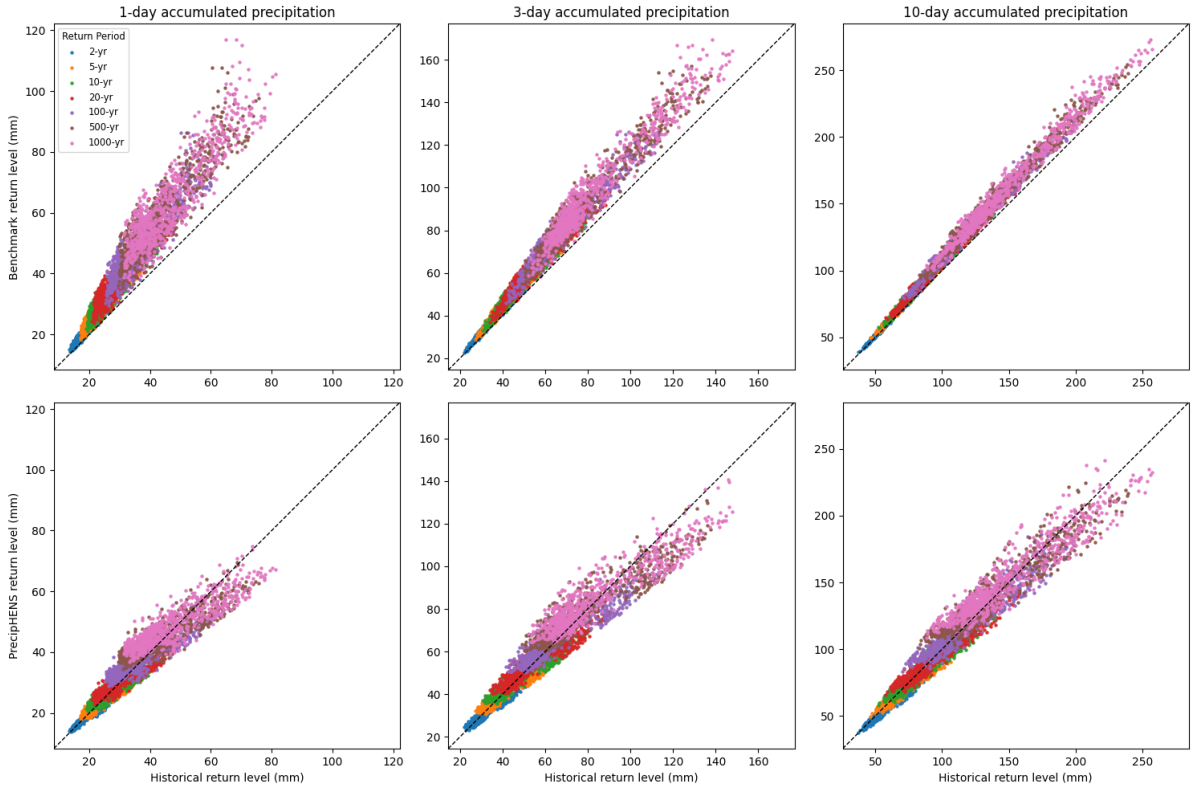

Figure R6: Equivalent of Figure 9 in the manuscript but with additional return periods of 500 and 1000 years.

-Lines 454-457: What does the calculated proportion of non-overlapping 1x1 grid cells in PCA space that contain at least one event from each model mean exactly? I understand that it shows us greater diversity of events in PrecipHENS, but there could be another sentence or two clarifying how this method is actually constructed and why it's done this way.

The 2-dimensional space representing the support for the first two principal components was divided into 1x1 grid cells, where the measurement unit of a grid cell is

the “dimensionless” principal component loading (i.e., the figure spans approximately [-25, 175] in the PC1 dimension and [-125, 125] in the PC2 dimension so this space was split into 1x1 grids). For each of these grids, a binary outcome was taken for whether the respective dataset (from historical, benchmark and PrecipHENS) had at least one extreme point within that projected grid cell (i.e. was there at least one point in the figure in that grid). The proportion of grid cells with a true outcome was used for the figures quoted.

This summary was intended as an illustrative numerical summary of relative spread to complement the subjective image and was not intended to be a formal measure of physical plausibility. To avoid overstating their significance, we have removed these numerical values from the abstract and explicitly state the numbers are for illustrative purposes in the main body of the text (lines 578-582).

We have clarified in the manuscript that the PCA space does not represent the full set of plausible atmospheric extremes, and occupying a broader region of this space should be interpreted as greater structural diversity rather than a guarantee of physical realism (lines 578-582).

-Figure 11: What dataset is used to construct the data shown in this figure?

This is from the historical data, we have ensured this is clear in both the caption and text (lines 582-584, Figure 11).

-Line 583: I believe the authors mean Figure 18 (not Figure 17).

Thank you, we have updated this (line 707).

Reviewer 2

This manuscript introduces and validates a novel AI-driven framework, PrecipHENS, which integrates a Spherical Fourier Neural Operator (SFNO)-based model combined with an Adaptive Fourier Neural Operator (AFNO)-based diagnostic precipitation model to generate seasonal-scale, high-resolution weather ensembles for flood risk assessment. Through a case study in the Elbe River basin, this manuscript systematically compares the PrecipHENS outputs against a multivariate extreme-value statistical model (the benchmark) across multiple dimensions such as climatology preservation, spatiotemporal coherence, extreme event representation, and methodological generalizability. The results demonstrate that PrecipHENS can realistically reproduce historical climatic features and generate extreme precipitation events with greater spatial diversity, and the produced meteorological data can be translated into plausible river flow simulations via a hydrological model named GR4J. This study successfully demonstrates the feasibility and considerable potential of AI weather generation models for directly constructing extreme precipitation and flood risk

event sets, representing a significant methodological novelty with important application prospects.

The paper is well-written with sound experimental design, although there are a few points requiring clarification and minor adjustments. Comments are listed below.

Major comments

The authors note that this study applied the lumped hydrological model, GR4J, on individual watersheds without explicit channel routing. This is a reasonable simplification given the core focus on validating the meteorological input. However, the specific limitations this choice imposes on the applicability of the conclusions should be more explicitly discussed. It should be stated that the current framework's output is flow event sets at sub-basin outlets, and its direct use for fluvial flood risk assessment has constraints, particularly concerning flood peak propagation and superposition.

We thank the reviewer for this useful point. We agree that it is important to note that the absence of a dedicated channel routing module (e.g., Muskingum-Cunge or kinematic wave) introduces constraints. The nested approach described at lines 295–304 means that we do not explicitly model the propagation of flood waves and volumes through the river basin. Instead, we focus on capturing correlations in the hydrological response to forcing, which are found (lines 677–679) to be reproduced well. The GR4J models conserve mass at sub-basin scale, however the nested approach does not guarantee continuity at the larger basin scale. The approach would not represent features of flood events where conservation effects in the flood routing were the dominant control on risk, which might potentially occur in situations where large volumes of water enter floodplain storage or in a sequence of prolonged flood events.

In the case of Europe wide event set generation such constraints are a necessary compromise when considering the computational expense of flow routing. Our revised manuscript includes the following paragraph.

"While GR4J effectively captures the rainfall-runoff transformation at the sub-basin scale, the lack of explicit channel routing means that the temporal translation and attenuation of flood waves are not specifically modelled. The fluvial simulations are therefore physically independent between sub-basins and do not capture downstream interactions or continuity along the river network. The framework is designed to characterise the hydrological response to meteorological forcing and to reproduce the observed temporal and spatial correlation of flood events across locations, rather than to simulated routed hydrographs. Consequently, the conclusions of this study are most applicable to understanding how meteorological variability translates into coherent sub-basin flood responses, and less applicable in settings where flood-wave routing, floodplain storage, or prolonged multi-peak events dominate flood behaviour." (lines 769-776)

All datasets in this study, including the 44-year historical reference used by the statistical benchmark model and the AI model trained on data from 1979-2016, rely on relatively short climatic records. A key point that could be more explicitly discussed is that the short record length fundamentally limits the reliable characterization and statistical estimation of low frequency and high return-period extreme events. This implies not only that the historical data itself may lack some genuine extreme patterns, but also that both the benchmark and PrecipHENS methods, which involve extrapolation or learning from this short series, carry inherent and considerable uncertainty for events far exceeding the record length (e.g., high return periods of 50 or 100 years). This should be more clearly elevated as a common foundational limitation affecting all approaches.

We thank the reviewer for highlighting the important issue of record length. We agree that the limited duration of the historical periods means that the observational data cannot fully sample the range of extreme spatial patterns that may occur in the present-day climate. We have clarified this motivation more explicitly in the Introduction (lines 51-53, 75-76, 107-108).

We also now discuss how this constraint influences the benchmark and PrecipHENS methods in different ways. The benchmark model inherits its dependence structure directly from the historical extremes, so its representation of rare spatial patterns is shaped by the limited variety present in the input record. In contrast, PrecipHENS learns the dynamical operators from the ERA5 training periods and is therefore able to generate a wider set of spatially distinct extremes (as demonstrated in the PCA analysis), but its behaviour remains conditioned on the same underlying historical climate state. We now emphasise that both approaches involve extrapolation beyond the observational period, but the nature of the extrapolation differs across the methods. We now elevate this point in the Discussion (lines 762-768) as a foundational limitation common to the historical record, the benchmark, and PrecipHENS, rather than a limitation specific to any single method.

We also reiterate that in the revised manuscript that this study presents a framework rather than an optimised product, and that our aim is to demonstrate the feasibility and behaviour of AI-generated seasonal-scale weather ensembles under current climate conditions, rather than to provide a definitive extreme-value characterisation for long return periods (lines 129-131, 773-775).

Similar to (2), the manuscript mentions that the SFNO model was trained on data before 2016. Given climate non-stationarity, readers may be concerned about the implications of this data timeliness for future risk analysis. It is recommended that the authors briefly discuss this as a recognized limitation.

We have made this limitation more explicit in the manuscript. For further information we refer the reviewer to the response to the question "**What are the big picture goals this approach is trying to accomplish?**" from reviewer one.

Minor comments

P2, L47, "Each event set comprises of spatially …", remove "of".

We have updated this line (line 58).

P8, L200, "…with such length following a Poisson distribution with the unit being the number of years.", revising to "…with the length (in years) following a Poisson distribution."

We have updated this line (line 391).

Reviewer 3

This manuscript introduces PrecipHENS, a novel AI-driven framework that combines a deep-learning-based weather emulator (SFNO + AFNO) with the conceptual hydrological model GR4J to estimate fluvial flood risk in the Elbe River basin. By generating an ensemble of more than 1,000 synthetic winter seasons, the study seeks to overcome the limitations of relatively short historical records when estimating tail risk. The proposed approach is compared against a traditional statistical method based on Heffernan and Tawn (2004), with performance evaluated using four criteria: whether each method reproduces the observed precipitation climatology; spatial and temporal coherence; representation of extremes; and methodological robustness.

Overall, this manuscript fits well within the scope of Natural Hazards and Earth System Sciences. However, a few questions require clarification, and some revisions are needed.

Specific comments

The text refers to "Figure 17" when describing river flow simulations at Dresden. Figure 17 displays Gumbel parameters; the correct reference is Figure 18. Please correct this reference.

Thank you, we have updated this reference.

The authors highlight that PrecipHENS spans approximately 81% of the PCA space, compared to only 19% for the historical record, and interpret this as evidence that the framework generates novel extremes. Could the authors clarify how the physical plausibility of these novel events is ensured? Specifically, do these outlier events correspond to coherent atmospheric structures?

We thank the reviewer for raising this point. The percentage quoted for the proportion of PCA space occupied by each dataset were intended as an illustrative numerical summary of relative spread to complement the subjective image and was not intended to be a formal measure of physical plausibility. To avoid overstating their significance, we have removed these numerical values from the abstract and explicitly state the numbers are for illustrative purposes in the main body of the text.

We have clarified in the manuscript that occupying a broader region of PCA space should be interpreted as greater structural diversity rather than as evidence of physical realism (lines 577-581). A more detailed discussion of physical plausibility and atmospheric coherence is provided in our response to Reviewer 1 ("**How do we know that the precipitation outputs of PrecipHENS are physically plausible?**"), and corresponding text has been added to the Discussion (lines 747-755).

In addition to the plausibility of the extreme weather event structures, their plausibility in terms of larger atmospheric conditions is another avenue for future work. We discuss this further in response to the below comments on both large-scale conditions, such as ENSO, and additional atmospheric variables (lines 756-761).

Are all 1,008 ensemble members derived from initial conditions sampled strictly from the 9–16 November 2023 window? If so, have the authors tested initializing from different years (e.g., neutral, El Niño, and La Niña years) to ensure that the ensemble spans the full range of climatological variability?

We refer the reviewer to the response to reviewer one on "**Selection of initial conditions**".

The current evaluation focuses primarily on end-of-pipe outputs (i.e., precipitation), without assessing the intermediate atmospheric variables generated by SFNO, such as geopotential height, pressure fields, or wind patterns. Given that the Discussion explicitly acknowledges that "remaining uncertainties concern how well the underlying SFNO emulator reproduces large-scale atmospheric circulation patterns", could the authors include a brief evaluation or at least a visualization of these generated atmospheric fields?

Due to both the focus in flood risk management being specifically on precipitation, and that the benchmark approach we are evaluating against is built for a single meteorological variable, we have not evaluated further atmospheric variables for this study.

We agree that expert meteorological assessment of AI-generated weather remains limited for long rollouts of emulators like SFNO. These models are typically trained and validated for short-range forecasts (e.g., Bonev *et al.*, 2023; Mahesh *et al.*, 2024, 2025) where a varied range of atmospheric variables are assessed. Recently, evaluations at a

seasonal scale have been conducted for SFNO-HENS indicating the model is capable of producing realistic propagation of the Madden-Jullian Oscillation (Peings *et al.*, 2026).

As AI weather generators become more widely applied in natural hazard risk analysis, collaboration with academic groups specialising in synoptic meteorology will be essential to evaluate physical consistency and diagnose potential systematic artefacts, and so we would welcome this further work to be instigated following this proof-of-concept. We intend to add to the Discussion section on this point (lines 789-796).

Editor

Thank you for the submission of your manuscript, "AI based seasonal large ensembles for fluvial flood risk: Evaluation over the Elbe basin," to NHESS.
As you know, three reviewers have now provided detailed reviews, which you have replied to in thoughtful detail. Reviewer #1 recommended a major revision, while the other two recommended minor revisions. Therefore, I'd like to invite you to submit a revised version of your manuscript, with all those comments appropriately addressed. In addition, I also carefully read through the manuscript as it looks very interesting to me. I totally agree with Reviewer #1 that this manuscript can be a meaningful contribution to the literature, but it requires a major revision for better clarification and presentation on several key points to further enhance the impact of this novel paper. So, I'd also like to offer a few minor comments after reading the paper, most of which are related to the manuscript's presentation.

We thank the Editor for the careful reading of our manuscript and for the constructive guidance provided alongside the reviewer comments. We have revised the manuscript substantially to address these points, with particular emphasis on clarity of scope, narrative structure, and accurate terminology. Below, we summarise how each of the Editor's comments has been addressed in the revised manuscript.

(1) Title: I would suggest adding a few words here to better highlight the study's major focus, i.e., seasonal large ensembles of WHAT? Of precipitation, or weather scenarios, etc.

Thank you for this helpful suggestion. In response, we have revised the title to more clearly specify the nature of the seasonal large ensembles considered in this study, namely *AI-based seasonal weather ensembles*, and to emphasise their role as input to fluvial flood risk estimation.

The revised title is:
"Evaluation of AI-based seasonal weather ensembles as input for fluvial flood risk estimation: A case study over the Elbe basin."

We believe this wording better highlights the study's main focus while remaining appropriately general with respect to the underlying meteorological drivers.

(2) Introduction: this section needs to be carefully reorganized to better explain the big picture (as noted by Reviewer #1) and the gap/scope of this work. Subsections 1.1 and 1.2 do not look like an appropriate way to present the Introduction, and in my view belong more appropriately in the Methodology section. Similarly, Lines 77–83 and Section 1.1.3 definitely should go to the Method section. As an introduction, it needs to identify the gaps in the research field, and the new opportunities provided by AI for this research topic.

The Introduction has been substantially reorganised to improve presentation of the big picture, research gap, and scope of the work. Material previously placed in Sections 1.1–1.2 that described technical construction details has been relocated to the Methodology section, while the revised Introduction now focuses on the following: (i) the importance and challenge of event-set generation for flood-risk estimation, (ii) limitations of existing physical and statistical approaches, and (iii) the new opportunities and open questions associated with AI-based weather models. The revised Introduction explicitly frames the contribution as an evaluation of AI weather models for risk-relevant use cases, rather than a description of the PrecipHENS framework itself.

(3) Accurate expression: the authors need to carefully revise the manuscript for accurate wording. Here, I just provide a few examples to highlight this concern: (a) "G3. Realistic representations of extreme events" (Line 108) is too general, and it actually refers to the "tail behaviour of extreme events" or the "statistical distribution of extremes". Please be accurate and specific; (b) According to the context, the subsection title "1.2 River flow generation" should be replaced with something like "Hydrologic evaluation", which means examining the use of PrecipHENS in terms of hydrological modelling. The subsection title "1.1.3 Precipitation goals" also looks weird to me.

Throughout the manuscript, we have carefully revised the wording to improve precision and ensure consistent use of accurate terminology. For example, (G3) has been revised to "Statistical tail behaviour of extreme events", and the remaining goal definitions have been updated in a similar manner to better reflect their intended scope. We have also reviewed and refined all subsection titles so that they more clearly represent their content and undertaken a systematic pass through the manuscript to remove awkward, ambiguous, or potentially misleading phrasing.

(4) Methodology: this section also needs better organization. As noted by Reviewer #1, I would also like to encourage the authors to rethink the role of the "benchmark approach". The core of this manuscript is the novel PrecipHENS framework, while that

benchmark statistical method is just a comparison to PrecipHENS. So, the narrative logic in the Methodology section should focus first on PrecipHENS, and then introduce the benchmark method and hydrological evaluation method as follow-up parts to evaluate the validity of PrecipHENS. In addition, all data descriptions need to be introduced naturally and logically around the methodology of PrecipHENS, so that readers can understand the data and the PrecipHENS framework more clearly and easily.

The Methodology section has been reorganised to reflect the central role of PrecipHENS as the core contribution of the manuscript. The framework is now introduced first, followed by the hydrological evaluation and then the statistical benchmark as a diagnostic reference. The benchmark is explicitly positioned as a validation baseline rather than a competing alternative, and its limitations (notably the lack of temperature and inability to support rainfall–runoff modelling, along with its dependence on historic storm structures) are now stated clearly. Whilst we retain a dedicated data section, it has been moved to a more natural position in the manuscript, where it directly supports and clarifies the description of the PrecipHENS framework.

(5) Fig. 2: for consistency, in this flowchart the term “historical record” should be replaced by “historical data” (as stated by the authors in Line 150).

Thank you, we have updated this figure (now Figure 1).

(6) Lines 175–176: I think there should be a clarification of why the winter season was chosen as the study focus of this work. I am not familiar with Europe’s hydroclimate, but for a manuscript on extreme precipitation, I would expect this season to be consistent with the typical storm/flood season in the selected study region. The authors could explain this in either the Introduction or the Method section.

Additional clarification has been added to explain the choice of the winter season for the Elbe basin case study (lines 242-247). While this rationale was already implicit in the original manuscript, the revised text now states it explicitly, noting that winter represents the dominant season for basin-scale fluvial flooding in the region, driven by prolonged precipitation and, in some areas, rainfall–snowmelt interaction. This makes winter a particularly relevant and demanding test case for evaluating coupled precipitation–temperature realism and hydrological response.

We believe these revisions substantially strengthen the manuscript and clarify both the capabilities and limitations of the proposed framework, and we thank the reviewers again for their constructive feedback.

References

Bonev, B. *et al.* (2023) 'Spherical Fourier Neural Operators: Learning Stable Dynamics on the Sphere', *Proceedings of the 40th International Conference on Machine Learning*, 202, pp. 2806–2823.

Mahesh, A. *et al.* (2024) 'Huge Ensembles Part II: Properties of a Huge Ensemble of Hindcasts Generated with Spherical Fourier Neural Operators'. arXiv. Available at: https://doi.org/10.48550/arXiv.2408.01581.

Mahesh, A. *et al.* (2025) 'Huge Ensembles Part I: Design of Ensemble Weather Forecasts using Spherical Fourier Neural Operators'. arXiv. Available at: https://doi.org/10.48550/arXiv.2408.03100.

Peings, Y. *et al.* (2026) 'Subseasonal forecasting and MJO teleconnections in machine learning weather prediction models', *Journal of Geophysical Research: Atmospheres*, 131(3), p. e2025JD044910.

Quinn, N. *et al.* (2019) 'The Spatial Dependence of Flood Hazard and Risk in the United States', *Water Resources Research*, 55(3), pp. 1890–1911. Available at: https://doi.org/10.1029/2018WR024205.