Overall response

We thank the reviewer for their careful evaluation and for recognising the practical relevance of using impact reports as the forecast target within a medium-range, global framework. The main concerns relate to the scope of the conclusions and the interpretation of the results, rather than to the methods. We address them below by stating more precisely what the evidence shows, defining the prediction target explicitly, and adding the information needed to interpret the scores operationally.

Point-by-point responses

***** Reviewer's comments are indicated in bold and blue , whilst author's responses are provided below each comment. Added references provided in the responses are listed at the end of the document. *****

Major comments

Evidence boundary for the global feasibility claim.

The manuscript's strongest claim is that it provides empirical evidence for the feasibility of global, medium-range flash-flood occurrence forecasts. This is an important claim, but the evidential boundary is not fully clear. The quantitative verification is performed over the CONUS, while the extra-CONUS evidence appears to consist of a single qualitative case study, the 2024 Valencia event. The TA1–TA3 sensitivity experiments are informative, but they remain internal experiments within the same broad geographical domain and reporting system used for model development. It is therefore difficult to assess whether the results demonstrate global forecast feasibility in a general sense, or rather a CONUS-based feasibility demonstration with one out-of-domain illustration. This distinction matters because the abstract and conclusions use broad language suggesting that global, medium-range flash-flood occurrence forecasts are achievable (e.g. L12–14; conclusions around L607–614).

We thank the reviewer for recognising the importance of this claim. However, we maintain that the results provide empirical evidence for the feasibility of global, medium-range forecasts of flash-flood occurrence. While the quantitative verification is confined to the CONUS and the evidence outside it consists of one case study, the spatially constrained experiments are central to showing how a model trained over a limited region can be used to produce global forecasts. TA2 and TA3 withhold the labels of one sub-domain during training and verify the model over the region it has not learnt from. This reproduces the situation a regionally trained model faces when deployed elsewhere, so these experiments provide quantitative evidence on transferability. The Valencia case then extends this evidence to a different continent and reporting system. We acknowledge that both CONUS sub-domains share the NOAA reporting system, so TA2 and TA3 cannot capture differences in reporting practice between countries. A quantitative global verification is currently limited by the lack of impact databases of comparable density and quality

outside the CONUS. Datasets such as Groundsource (Mayo et al., 2026) may make this possible, and we will identify it as future work. We will revise the abstract and conclusions to state explicitly which evidence supports the feasibility claim: the CONUS verification, the transferability experiments and the out-of-domain case study.

The final sentence of the abstract (L12–14) will read: "Hence, this study provides empirical evidence, from the CONUS and an out-of-domain case study, that global, medium-range forecasts of flash flood occurrence are feasible with simple data-driven approaches and readily available data." We will also replace "closing one of the most pressing and long-standing gaps in modern hydrology" (abstract) and "these results close one of the most persistent gaps" (L614) with "offering a pathway towards addressing one of the most pressing and long-standing gaps in modern hydrology".

Interpretation of the prediction target: reported occurrence versus physical occurrence.

The model is trained against impact-report occurrence, defined as at least one flash-flood report within an ERA5 grid box and 24-hour period. This is a useful target, but it is not the same as physical flash-flood occurrence. The distinction matters because impact reports are affected by population density, exposure, reporting practices, and timing or location errors, as acknowledged in the data description.

This issue also affects the interpretation of the SHAP results. While the dominant role of rainfall is physically plausible, the counterintuitive behaviour of LAI and SDFOR suggests that the model may also be learning reporting structure, seasonality, or spatial exposure embedded in the database. The interpretation of the output as "flash-flood occurrence probability" therefore seems somewhat broader than the target variable itself supports, particularly when the model is discussed in the context of transfer beyond the training region.

Relatedly, the scale of the forecast product is easy to misinterpret. The Introduction motivates the study partly through the inability of existing systems to resolve small flashy catchments, whereas the proposed product is an ERA5-grid-box, 24-hour impact-occurrence probability forecast. This is a useful product definition, but it is different from resolving sub-grid catchment runoff processes or outlet-scale flash-flood dynamics.

We agree that impact reports are not equivalent to physical flash-flood occurrence. Reporting is affected by population density, exposure, reporting practices and timing or location errors, and under-reporting in particular means that some flash floods are not recorded. Impact reports are, however, the only observational basis currently available for training and verifying flash-flood occurrence forecasts at this scale. We therefore retain "probability of flash flood occurrence" as the name of the forecast product, and we will state explicitly in the manuscript that the model is trained and verified against reported occurrence, and that the forecasts inherit the limitations of the reporting. A quantitative assessment of how closely reported occurrence reflects physical occurrence was not possible in this study, because no independent dataset of comparable quality to the NOAA Storm Events Database was available. The recently released Groundsource dataset (Mayo et al., 2026) now makes such an assessment possible, for example by comparing

the two databases over the CONUS to estimate the extent of under-reporting. We will identify this comparison as future work.

On the SHAP results, we agree that the behaviour of LAI and SDFOR may partly reflect reporting structure, seasonality or exposure. As detailed in our response to Reviewer 1 (major comment 2), these associations cannot all be attributed to reporting effects. LAI may partly reflect seasonality, SDFOR describes sub-grid orographic variability rather than catchment slope, and the association with soil saturation has a plausible physical basis. We will state more clearly that the SHAP results describe what the model has learnt from the reported target, not causal relationships, and we will discuss the implications for transfer to regions with different reporting practices.

On scale, we agree that the product does not resolve sub-grid catchment runoff or outlet-scale dynamics. It estimates the probability that at least one flash flood is reported within a grid box and day, so it complements, rather than replaces (which indeed, it was never the goal of this study), catchment-scale systems. We will revise the Introduction to separate the motivation (the coverage and lead-time limits of existing systems) from the product definition, consistent with our response to Reviewer 1 (major comment 1).

Operational interpretation of skill and reliability under severe class imbalance.

The verification results suggest useful discrimination, especially in terms of AUC-ROC, but the operational meaning of the reported skill is less clear. The AUC-PR values are low in absolute terms but difficult to interpret without the positive-event base rate; the forecast frequency bias indicates over-forecasting, and reliability appears to hold mainly within low probability ranges, below about 10% at day 1 and about 2% by day 5 (around L338–350). For such a severely imbalanced warning problem, these details are important for interpreting the practical value of the forecasts. Without the grid-box-day sample size and the positive-event base rate, it is difficult to assess how informative the reported AUC-PR values are. The broad statements about skilful and reliable forecasts are therefore difficult to interpret in operational terms, especially where false alarms would matter for warning decisions.

We agree that the operational meaning of the scores should be clearer, and we will add the information requested to better interpret the AUC-PR values.

On reliability, we agree that the forecasts are reliable only below about 10% at day 1 and 2% at day 5, and we will state this limit wherever reliability is mentioned. We note, however, that for rare and localised events such as flash floods, the occurrence probabilities are expected to be low. High-probability bins are therefore sparsely populated, and reliability within the low-probability range carries more weight than it would for more frequent events. We will add the frequency of forecasts in each probability bin to the reliability diagrams to show what proportion of forecasts falls within the reliable range. Because grid-box days without a report are treated as non-events (that was the decision made in this analysis), the observed frequencies in the reliability diagrams are likely to be biased low by under-reporting. We will state this assumption and its implications explicitly in the Discussion.

On frequency bias, as discussed in our response to Reviewer 1 (major comment 3), the acceptability of over-forecasting depends on the intended action and its cost (Mylne, 2002), and part of it is likely to reflect under-reporting. We will replace general statements about "skilful and reliable" forecasts with these specific statements and discuss their implications for warning decisions where false alarms matter.

Ambiguity in the nested cross-validation workflow.

The description of the nested cross-validation workflow is ambiguous. The main text states that the inner loop is used for hyperparameter tuning and the outer loop for generalization assessment, but the Fig. 9 caption says that trial performance was measured over the outer test subset. This wording seems inconsistent with the usual nested-CV design, where the outer test fold should only be used for the outer-loop assessment. This may simply be a wording issue, but it is important because the nested design is central to avoiding data leakage and overoptimistic performance estimates.

We thank the reviewer for identifying this inconsistency, which is more of a sentence construction error. In the caption of Figure 9, the sentence "each trial trains on inner training folds and validates on inner validation folds, with performance measured as the mean of AUC-ROC and AUC-PR over the outer test subset." should read instead "each trial trains on inner training folds and validates on inner validation folds. Model generalisation is subsequently assessed over the outer test subset by computing the mean the of AUC-ROC and AUC-PR.". We will correct the caption accordingly.

Precision of novelty and contribution claims.

The manuscript makes broad claims about novelty and contribution, particularly through the "first empirical evidence" framing. The study's contribution is important, but the exact scope of this "first" claim is not fully clear. The novelty appears to lie in the combination of impact-report targets, global predictors, medium-range lead times, and a continuous-domain forecasting framework, rather than in any single component alone. A clearer distinction between what is new in this study and what is already covered by existing regional, machine-learning, or global-rainfall-based flash-flood guidance work would help place the manuscript more precisely.

We agree that the scope of the "first" claim should be stated precisely and, as the reviewer suggests, the novelty lies in the combination of components rather than in any single one. Machine-learning flash-flood models exist but operate at catchment or regional scales, with lead times rarely beyond 24 hours (Santos et al., 2025). EFAS provides flash-flood guidance up to day 5, but only over Europe (Raynaud et al., 2015). FFGS aims at global coverage through regional systems whose coverage remains patchy (Georgakakos et al., 2022). GloFAS covers a continuous global domain but cannot resolve flash floods. Since the paper was submitted, there are also the Google's flash-flood forecasts (Zlydenko et al., 2026) which have also adopted impact reports as the training target, extracted from news articles (Mayo et al., 2026) and combined with globally available hydro-meteorological inputs. They are, however, restricted to urban areas (population density above 100 people per km^2), and to lead times of up to 24 hours,

and areas that have some history of flash flooding in the past. To our knowledge, no previous study has combined impact-report targets, globally available predictors, medium-range lead times (up to day 5) and a continuous-domain framework covering both urban and rural areas. We will revise the “first empirical evidence” statements in the abstract and conclusions to name this combination explicitly, and add a short paragraph to the Introduction distinguishing the study from the systems listed above.

Minor comments

1. **Figure, caption, and cross-reference consistency.**

18 is difficult to follow because the caption, text, and plotted panels do not appear to match. The figure contains panels a–g, whereas the caption describes only a–e. The text appears to refer to panels b–c as rainfall exceedance probabilities, panel d as the reanalysis prediction, and panels e–g as forecasts. The day-5 initialization/lead-time label also appears unclear.

In the TA1 description, the retained fractions are given as 90%, 50%, and 90% (L280–282), whereas Fig. 10 describes reductions of 10%, 50%, and 90%. The last retained fraction appears inconsistent and may be 10%.

The SHAP discussion refers to Fig. 13e–g, but Fig. 13 appears to contain only panels a–d.

The text appears to refer to Fig. 7a as an exceedance-probability panel, but Fig. 7a is the 50-year return-period threshold map. The intended reference seems to be Fig. 7b.

Some internal section references also appear inconsistent, for example the references to Section 4.1/4.2 and to results in Section 3.4.

1. **Additional reporting details for interpretation and reproducibility.**

The probability threshold selected by maximizing F1 is not clearly reported. Since binary forecasts and frequency bias depend on the thresholding procedure, the selected threshold value would be useful.

The normalization of the SHAP feature-importance percentages is unclear. The listed values exceed 100% when summed, so it is not obvious whether they are relative to the most important feature, normalized in another way, or reported on a different scale.

The workflow is described as repeated cross-validation, but Fig. 9 gives n_repeats = 1. This may be technically implemented through RepeatedStratifiedKFold, but the wording could be confusing.

The grid-box-day sample size and positive-event base rate would be useful to report together, especially for interpreting AUC-PR in the severely imbalanced setting.

1. **Wording and numerical consistency.**

The neural-network training-time comparison is difficult to reconcile across the main text and Supplement. It is not always clear whether the reported values refer to individual trials, one outer fold, or the full tuning process.

Some broad rhetorical phrases, such as "closing one of the most long-standing gaps", seem closely related to the evidence-boundary issue raised in the major comments and may benefit from more precise wording.

In the abstract, "flash occurrence" should presumably read "flash flood occurrence".

The Valencia damage estimate of "16.5 billion" lacks a currency symbol.

All minor comments will be addressed in the revised manuscript, as summarised below.

We will correct the figures, caption, and cross-reference inconsistencies. The caption of Figure 18 will be aligned with its panels, with the forecast initialisation and lead-time labels clarified. The retained fractions for TA1 will be corrected to 90%, 50% and 10% (L280–282). References to Figure 13e–g and Figure 7a will be corrected to Figure 13b–d and Figure 7b, and internal section references will be checked throughout.

We will also add the reporting details requested:

- the probability threshold selected by maximising the F1-score;
- the normalisation used for the SHAP feature-importance percentages;
- the grid-box-day sample size and positive-event base rate, reported together (see our response to major comment 3);
- a clarification that RepeatedStratifiedKFold was used with n_repeats = 1, i.e. a single stratified 5-fold split, so the procedure will be described more clearly.

Neural-network training times will be reported consistently in the main text and the Supplement, specifying whether values refer to a single trial, one outer fold or the full tuning process. Finally, we will replace broad rhetorical phrases such as "closing one of the most long-standing gaps" with wording that matches the evidence (see major comment 1), correct "flash occurrence" to "flash flood occurrence" in the abstract, and add the currency to the Valencia damage estimate.

References

Georgakakos, K. P.; Modrick, T. M.; Shamir, E.; Campbell, R.; Cheng, Z.; Jubach, R.; et al. (2022). The flash flood guidance system implementation worldwide: a successful multidecadal research-to-operations effort. Bulletin of the American Meteorological Society, 103, E665–E679. DOI: https://doi.org/10.1175/BAMS-D-20-0241.1

Mayo, R.; Zlydenko, O.; Bootbool, M.; Fronman, S.; Gilon, O.; Hassidim, A.; et al. (2026). Groundsource: A Dataset of Flood Events from News. EarthArXiv [preprint]. DOI: https://doi.org/10.31223/X5RR2K

Mylne, K. R. (2002). Decision-making from probability forecasts based on forecast value. Meteorological Applications, 9, 307–315. DOI: https://doi.org/10.1017/S1350482702003043

Raynaud, D.; Thielen, J.; Salamon, P.; Burek, P.; Anquetin, S.; Alfieri, L. (2015). A dynamic runoff co-efficient to improve flash flood early warning in Europe: evaluation on the 2013 central European floods in Germany. Meteorological Applications, 22. DOI: https://doi.org/10.1002/met.1469

Santos, L.; Satolo, L.; Oyarzabal, R.; Escobar-Silva, E.; Diniz, M.; Negri, R.; et al. (2025). Machine Learning-based Hydrological Models for Flash Floods: A Systematic Literature Review. Smart Construction and Sustainable Cities, 3. DOI: https://doi.org/10.1007/s44268-025-00071-9

Zlydenko, O.; Fester, H.; Fronman, S.; Gauch, M.; Gilon, O.; Hassidim, A.; et al. (2026). AI learns to forecast global flood events from historical news. EarthArXiv [preprint]. DOI: https://doi.org/10.31223/X5177S