Overall response

We thank the reviewer for their careful and constructive evaluation of the manuscript. Their comments have helped us identify areas where the presentation and broader framing of the study can be clarified and strengthened. We also thank the reviewer for recognising the value of medium-range guidance and the use of impact occurrence as a forecast target. Specific proposed manuscript changes are described below. These additions aim to strengthen the clarity, robustness, and broader process relevance of the study while preserving the core analysis and conclusions.

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

1. The flash-flood definition is incompatible with the working spatial scale

Table 1 defines fluvial flash floods as occurring in catchments <500 km² with concentration times of minutes to hours. The model operates on the ERA5 reduced-Gaussian grid (~31 km, i.e., grid-box areas approaching 1000 km²). A single grid box typically contains many flash-flood-relevant catchments, and the model cannot resolve the catchment dynamics it claims to predict. This mismatch undermines both the construct validity of the predictand and the interpretation of "flash flood occurrence" at the grid scale. The authors briefly acknowledge resolution as a limitation in Section 6.4, but the issue deserves more prominent treatment up front: what exactly is being forecast at 31 km, and how does the binarisation procedure (Appendix A, Figure A1) reconcile point reports occurring in sub-grid catchments with a grid-cell yes/no label?

We appreciate the reviewer's concern about spatial scale and would like to clarify how the event definition relates to the forecast grid. The catchment-size criterion of less than 500 km² applies to the flash-flood events represented by the point reports, not to the area of the forecast grid box. A grid box may contain several smaller catchments, and the model predicts the probability that at least one qualifying flash flood will be reported somewhere within that box during a specified 24-hour period. It does not imply that the entire grid box will experience flooding or identify the event's precise location or timing within that period. During binarisation, one or more qualifying reports within the same grid box and day produce a single positive label, regardless of whether they concern one or several catchments. A zero denotes no qualifying report in that box and period. This aggregation preserves reported occurrence while removing finer spatial and temporal detail that is too uncertain to be modelled reliably (Panwar and Sen, 2020). Incomplete reporting remains an observational limitation, separate from the choice of grid-box size. For example, the ecPoint methodology represents an established basis for extracting local-extreme rainfall information from global models without explicitly resolving every small-scale process (Hewson and Pillosu, 2021). This supports the approach adopted in this study, which predicts

the probability of flash-flood occurrence without explicitly simulating flood dynamics within individual catchments.

We will introduce this distinction in the Introduction, while the detailed labelling explanation will remain in Appendix A.

Suggested wording for the Introduction: In this study, flash-flood occurrence refers to at least one reported event within a forecast grid box during a specified 24-hour period. The catchment-size criterion of less than 500 km² used to describe fluvial flash floods in Table 1 applies to the events represented by the point reports, not to the area of the forecast grid box. A single grid box may therefore contain several flash-flood-relevant catchments. The model estimates the probability of occurrence somewhere within that box, without identifying the event's precise location or timing within the 24-hour period. A positive occurrence does not imply that the entire grid box experiences flooding.

Suggested wording for Appendix A (to replace the explanation accompanying the binary assignment in Figure A1c): Reports are aggregated into grid-box days, with a label of 1 assigned when at least one qualifying report is present and 0 otherwise. Multiple reports within the same grid box and day produce a single positive label, whether they concern one or several catchments. This aggregation represents reported occurrence rather than the number of events or the extent of flooding, and removes finer spatial and temporal detail. A zero indicates the absence of qualifying reports.

2. Population-density confounding may not be a peripheral artefact

The authors indicated that the SHAP-derived associations between high LAI / low SDFOR / saturated soils and flash flood occurrence likely reflect the spatial distribution of who reports, not where flash floods occur. But the implications go further than they acknowledge. If the model has learned a latent population-density proxy, then its skill outside CONUS will be highest precisely in densely populated, vegetated, low-relief regions and lowest in the data-sparse mountain and arid regions where flash flood mortality is concentrated globally (Libya 2023, Afghanistan 2024, parts of the Andes). This is the inverse of where additional warning capability is most needed. A useful diagnostic would be to retrain after stratifying by population density — for example, restricting positive labels to grid boxes with population below a threshold — and report whether the SHAP rankings change. If the rankings flip toward steeper terrain and sparser vegetation, the population-confounding hypothesis is confirmed, and the global transferability claims need to be substantially softened.

We understand the reviewer's concern that LAI and SDFOR may indirectly reflect population density, leading the model to capture differences in reporting as well as flash-flood occurrence. However, the SHAP associations alone do not establish that these variables act as population-density proxies, or determine where forecast skill will be highest or lowest outside CONUS.

Population may influence both reporting and the physical environment, and these effects are difficult to separate. Marjerison et al. (2016), for example, found relationships between population, environmental characteristics and flash-flood reports that challenge a simple interpretation based predominantly on reporting bias. The three associations also require

separate consideration. High LAI may partly reflect seasonality, because vegetation and flash-flood occurrence can share seasonal patterns. SDFOR measures sub-grid orographic variability rather than the slope of the catchment generating a flood; low SDFOR at a reported impact location is therefore not necessarily inconsistent with runoff originating in steeper terrain. The positive association with soil saturation has a plausible physical basis, because wet antecedent conditions can favour runoff generation (Ortenzi et al., 2025). These explanations do not exclude reporting effects, but show why the associations cannot all be attributed to population density.

A comparison across population groups could help investigate the reviewer's hypothesis. However, restricting the training sample would also change its climatic and land-surface characteristics. A change in SHAP rankings could therefore reflect several differences between the samples, particularly because attribution depends on relationships among predictors (Aas et al., 2021). Such a change would warrant further investigation, rather than by itself confirm population-related reporting bias.

We propose that a useful first diagnostic would be to assess the existing model in different population groups, including both positive and negative cases within each group and reporting their event frequencies. However, this would assess performance against recorded events, while retaining uncertainty about unreported floods. We will explain how reporting differences, seasonality and physical processes could contribute to these SHAP patterns.

3. The frequency bias

Figure 12c shows FB ≈ 1.2 for reanalysis but ~2.5 for forecasts at all lead times — a 150% overprediction of yes-events. Section 6.1 frames the choice of balanced loss as preferable because weighted loss inflates false alarm rates twentyfold, but FB = 2.5 is not a minor cost; in an operational early-warning context, persistent overprediction at this magnitude erodes trust as severely as the false-alarm problem the authors warn against.

We agree that FB around 2.5 warrants explicit discussion. However, this does not mean that every forecast probability is inflated by a factor of 2.5, as illustrated by the reliability diagrams in Figures 12l–o.

This over-forecasting does not, by itself, establish that the forecasts lack operational value. Their usefulness depends on the intended action, its cost, the losses it could prevent and the forecast lead time (Mylne, 2002). Acceptable false-alarm levels therefore depend on the intended use: low-cost readiness actions may justify accepting more false alarms than costly protective measures. As an illustration, UN OCHA's "readiness approach", presented at the UNDRR Masterclass Series on "AI for Multi-Hazard Early Warning Systems", accepts higher false-alarm rates at medium-range lead times because the forecasts support preparedness rather than protective action. The importance of sufficient lead time for initiating flash-flood preparedness is also illustrated by Giannaros et al. (2022). We will expand the discussion to interpret the verification results in this context.

Published verification against rain-gauge observations shows that ecPoint rainfall forecasts (Hewson and Pillosu, 2021) and the ERA5-ecPoint reanalysis (Pillosu et al., 2026) are reliable. We therefore consider under-reporting in the target dataset a likely major contributor to the over-

forecasting, because some cases classified as false alarms may correspond to unrecorded floods. Our current verification does not, however, quantify the extent of under-reporting in the NOAA dataset or its contribution to FB. This could be investigated using the recently released Groundsource dataset (Mayo et al., 2026), provided that differences in event definition, coverage and location accuracy are taken into account.

We tested a weighted loss function and retained the unweighted loss because it produced substantially fewer false alarms (Section 6.1). This choice is consistent with Carriero et al. (2025), who found that class-imbalance corrections can worsen probability calibration, including for XGBoost. Although their study concerns clinical prediction, it supports evaluating the unweighted approach on its merits. We remain open to calibration improvements in future releases of the forecast dataset, and will discuss this further in the manuscript.

4. Cross-validation does not respect spatial and temporal structure

The repeated stratified k-fold procedure (Figure 9) assumes IID samples. Adjacent grid boxes within the same storm system are strongly spatially autocorrelated, and 24-hourly periods within a multi-day event are temporally autocorrelated. Random folds will place training and test samples from the same physical event in different folds, inflating the apparent skill. A spatial blocked CV (e.g., leave-region-out, or buffered spatial folds) and event-based temporal CV would give more honest performance estimates. This is particularly important because the central transferability claim requires accurate generalisation estimates.

On a related note, Multi-day event expansion may introduce label leakage and double-counting. For a hurricane producing flash floods over five consecutive days, this generates five training rows whose meteorological predictors and impact labels are not independent. This both inflates the apparent class frequency (mildly mitigating the imbalance the loss function is designed to handle) and introduces correlated samples into the training set. The authors should report the distribution of event durations and quantify how many "samples" actually correspond to independent meteorological events.

We thank the reviewer for raising the important point on sample dependence. In our view, however, this has a limited bearing on the results reported in this paper for the following reasons:

a) Multi-day expansion is not, by itself, leakage or double-counting. Flooding reported in impact databases over X number of consecutive days can legitimately correspond to X independent labels, and collapsing these into one would change the target. We applied this disaggregation approach because we found that the impact database may aggregate consecutive distinct occurrences into one single entry, so report duration does not necessarily correspond to the duration of a single rainfall system. We will document this more clearly in Appendix A.
b) The repeated stratified nested cross-validation (Figure 9) was used only to tune hyperparameters and select a candidate model within the training period (2001–2020). The skill scores reported in Figure 12 (Sections 4.1) and Figure 14 (Section 4.3) are computed on an independent verification period (2021–2024) that is fully separated from the training data (events in the verification set cannot also appear during training). If within-fold dependence inflated the cross-validation scores, we would expect a clear

drop in skill over the independent period, but we do not see that. Moreover, blocking can overestimate errors when a model is used within the region it was trained on (Roberts et al., 2017), and it does not by itself guarantee an unbiased assessment (Wadoux et al., 2021). Dependence between samples could, in principle, have influenced which hyperparameter set was selected. However, the cross-validation estimates (AUC-ROC ≈ 0.78–0.82, Figure 11b) agree closely with those obtained on the temporally independent verification period (AUC-ROC = 0.797, Figure 12d), which indicates that any optimism introduced by dependence within the training folds may be small. We will state this explicitly in the revised manuscript.

c) The transferability claim does not rest on random folds. It rests on the spatially constrained experiments (TA2 and TA3, Figure 10). These are a leave-region-out design: models are trained on one sub-domain (west or east of 100°W) and verified over the independent 2021–2024 period across the full domain, including the region withheld during training. The Valencia case provides a further out-of-domain test. This design is consistent with the reviewer's recommendation, and we will make the link explicit in Sections 3.6 and 6.3.

5. No baseline beyond the no-skill diagonal

AUC-ROC of 0.8 is a meaningless number without a reference point. A logistic regression on the same features, climatology of flash-flood frequency by month and grid box, or even raw rainfall return-period exceedance probability used directly as the forecast probability would give a defensible baseline. The Valencia case (Section 5.2) actually provides a hint that the model's added value over raw rainfall exceedance is real (downstream propagation via the adjacent-grid-box feature), but a domain-wide quantitative baseline comparison is needed to support the headline claim that XGBoost adds skill.

We agree with the reviewer. A comparison against a baseline forecast was part of the first author's PhD thesis, from which this paper is extracted (Pillosu, 2026). Chapter 5 of the thesis evaluates rainfall return-period exceedance probabilities, used directly as forecast probabilities, as a baseline for flash-flood occurrence. This analysis was not included in the manuscript to keep its length manageable. We will add it to the Supplementary Material and refer to it in Section 4.1, so that the skill of the XGBoost model can be compared directly with the rainfall-only baseline.

6. "ensemble methods manage class imbalance effectively" justification is incomplete

L.206 cites Ayodele (2023) and Altalhan et al. (2025) for the claim, but neither shows that ensembles solve imbalance — at best, they manage it under appropriate hyperparameter and loss configurations.

We thank the reviewer for this comment. The manuscript refers to ensemble techniques managing, rather than solving, class imbalance (L.206). We agree, however, that "effectively" overstates what the cited studies show without qualification. We will therefore revise the sentence to make clear that ensemble methods can manage class imbalance under appropriate hyperparameter and loss-function configurations. This is the reason this study evaluated both

standard and weighted loss functions (Section 3.1.2) and tuned hyperparameters within a nested cross-validation framework.

Suggested wording: Ensemble learning techniques can manage class imbalance when combined with appropriate hyperparameter and loss-function configurations (Ayodele, 2023; Altalhan et al., 2025).

7. ERA5 is the only meteorological input — but better data exist over CONUS

The CONUS is unusually well-observed by NEXRAD and MRMS, which provide quality-controlled rainfall at ~1 km / 2-min resolution. Using ERA5(-ecPoint) rainfall as the predictor over CONUS conflates two error sources: (i) ERA5's representation error, and (ii) the ML model's inability to map predictors to outcomes. A useful ablation would be to retrain the model with MRMS rainfall as input, holding everything else constant. If MRMS-driven skill is much higher, the bottleneck is the input data, not the model architecture, and operational improvements should be sought there. This would also clarify whether the improved performance with ecPoint reflects genuine sub-grid information or just bias correction.

We would like to clarify the distinction between raw ERA5 and its post-processed version, ERA5-ecPoint. ecPoint is calibrated against rain-gauge observations worldwide and explicitly addresses the sub-grid variability and grid-scale biases underlying the reviewer's concern (Hewson and Pillosu, 2021). Because its corrections are weather-dependent, it represents sub-grid information rather than applying a uniform bias correction, and it outperforms comparable post-processing approaches (Ben Bouallègue et al., 2020; Pillosu et al., 2025). Independent comparisons against higher-resolution systems point in the same direction. Over Italy, Gascón et al. (2024) found ecPoint overall more skilful than raw and post-processed convection-permitting COSMO forecasts at 2.2 km, although the latter added value in complex orography. Hemri et al. (2022) reported similar results over Switzerland. For reanalysis, Pillosu et al. (2026) show that rainfall climatologies derived from ERA5-ecPoint at 31 km, the dataset used here, outperform those from higher-resolution ECMWF reforecasts. Finer resolution does not therefore necessarily improve the representation of extreme rainfall (Hewson, 2024). We will summarise this evidence in Section 2.3.2.

The study aims to develop a medium-range flash-flood forecasting system for a continuous global domain. MRMS and NEXRAD are regional observational products, unavailable at medium-range lead times or outside the United States, so they could not serve as inputs to the intended system. We agree, however, that retraining with MRMS rainfall over CONUS would be a useful diagnostic to separate input-data error from model error. We consider this better suited to a separate study and will identify it as future work in Section 7.

8. Non-stationarity in the impact reports is unaddressed

Figure 2a shows a clear upward trend in geo-located reports from 2001 to 2024. This trend is dominated by changes in reporting (network expansion, smartphone-era citizen contributions, changes in NWS protocols) rather than by changes in flash flood frequency. Training on 2001–2020 and verifying on 2021–2024 means the model may be implicitly learning the reporting trend rather than the hazard. Please:

Quantify the trend in reports per grid box (separately from event counts);

Test whether de-trending the labels changes model skill;

Discuss how non-stationarity in the labels themselves affects the validity of the FB and reliability scores.

We agree that non-stationarity in the reports is an important concern. However, the increase in report counts in Figure 2a does not, by itself, establish that reporting changes dominate the trend, or that the model learns reporting behaviour rather than flash-flood occurrence. Because our labels are binary, additional reports within an already-positive grid-box day do not increase its contribution. We will therefore:

- quantify the trend in positive grid-box days, both domain-wide and per grid box, separately from raw report counts;
- discuss how non-stationarity in the labels affects the interpretation of FB and reliability over the 2021–2024 verification period;
- if a significant trend persists in positive grid-box days, test whether de-trending the labels changes model skill.

Minor Comments

1. The distinguishing criterion for flash floods vs. riverine floods should be clarified explicitly in the Introduction, not only in Table 1. At 31 km, the "rapid rise" / "small steep catchments" defining criteria do not directly apply to the predictand.

2. L.159–160: "The figure demonstrates the model's ability to capture this saturation pattern up to five days in advance." This is a qualitative visual statement. Please report a quantitative metric (e.g., spatial correlation, RMSE of PMSS forecast vs. reanalysis) for the demonstration to be persuasive.

3. L.183 (ecPoint): Is the procedure correctly described as generating 100 sub-grid realisations per ERA5 grid box, distilled to 99 percentiles (1st–99th)? That is essentially a 100× downscaling in the statistical sense. A one-sentence clarification would help readers unfamiliar with ecPoint.

4. LAI as a climatological field cannot capture post-fire or post-deforestation changes in vegetation, which are first-order drivers of debris flow and flash flood risk in the western US (e.g., Front Range) and in Mediterranean Europe. This limitation should be discussed alongside the LAI-as-seasonality-proxy issue raised in Section 6.2.

5. The 1-year and 50-year return period thresholds (Figures 6a, 7a) — please describe the derivation method (annual maxima, partial duration series, distribution fitted) in the main text or Appendix B.

6. Static features that are conspicuously absent: soil hydraulic properties beyond soil type code (saturated hydraulic conductivity, porosity), flow accumulation / upstream contributing area, and HAND (or another channel-network proximity metric). The 2025 Texas Hill Country event at Camp Mystic is a reminder that confluence geometry is first-order. These should at least be discussed as prospective additions in Section 6.5.

7. Figure 8 functions more as a textbook taxonomy than a results figure and could be moved to the supplement. Note also that in the standard ML taxonomy, "ensemble methods" comprise bagging, boosting, and stacking; feed-forward neural networks are not an ensemble method per se. The figure conflates "models tested" with "ensemble families" in a way that is mildly misleading.

8. Section 3.6: The TA2 design (labels in one region, model exposed to the full domain) treats absent labels as true non-events. The authors acknowledge this introduces bias. It is not surprising that TA2-1 (labels only in the data-poor west) collapses to no skill — there are far fewer positive cases there to begin with (Figure 2b shows 35.4% of all reports in SE alone). The TA2-1 result is partly a statement about CONUS report distribution, not just about training strategy. Consider rebalancing the TA2 design to vary label-source region while controlling for label count.

9. Figure 12 shows performance on both the training and verification datasets. The training-dataset overlay is useful for assessing the generalisation gap, but the figure is busy. Consider plotting training scores in the supplement and keeping the main figure focused on out-of-sample performance.

10. Throughout the figures: please replace machine-readable variable names (tp_prob_1, tp_prob_1_adj_gb, SDFOR, PMSS) with human-readable axis labels (e.g., "Probability of 1-yr RP rainfall exceedance"). Figure 13 in particular is opaque on first reading.

11. AUC-PR values (~0.01–0.03) approach the no-skill threshold for the imbalanced base rate. Please add the no-skill PR baseline value (= positive class frequency) explicitly to each PR panel, so readers can see the actual margin of skill.

12. Appendix A, Figure A2 caption uses "(d)" for what appears to be panel (b) in the figure — please check.

13. L.7: "flash occurrence" → "flash flood occurrence".

14. L.285: "the second approach (TA2) simulates the scenario of training a global model over the whole global domain" — should be the full CONUS domain, not the global domain.

The minor comments concern three aspects of the manuscript: the clarity of definitions and methods, the presentation of figures, and the scope of the discussion.

We will correct the typographical and labelling errors (comments 12–14) and replace code-style variable names in all figures with readable labels.

Regarding the figures, we would prefer to keep Figure 8 in the main text (comment 7). In our experience, this overview helps readers less familiar with machine learning to follow the analysis underpinning the results. We agree, however, that a feed-forward neural network is not an ensemble method, and we will correct the caption so that only bagging and boosting are described as ensemble families. We would also prefer to keep the training and verification scores in the same panels of Figure 12 (comment 9), because their direct comparison shows that performance does not drop between the training (2001–2020) and the independent verification (2021–2024) periods. This comparison is central to the paper's conclusions and

supports our responses to the major comments on sample dependence and generalisation. Finally, each PR panel in Figure 12 already includes the no-skill reference line (comment 11). Because the positive-event frequency is very low, the line lies very close to the x-axis and its value is difficult to read, so we will state its numerical value in each panel and in the caption.

We will clarify the following points in the text:

- the flash flood definition, moved into the Introduction, distinguishing the physical event from the grid-box-day predictand (comment 1);
- the statistical, not spatial, nature of the 100 ecPoint realisations (comment 3);
- a reproducible description of how the return-period thresholds are derived (comment 5).

We will also discuss the limitations of the LAI variable (comment 4), and the possibility to add to the model other candidate predictors (comment 6).

In a few cases we propose a narrower change than the one suggested, because we believe it better matches the scope of the study:

- Instead of adding a separate soil-moisture verification (comment 2), we will cite published quantitative evaluations of PMSS forecasts.
- We will keep the sparse-label TA2 design instead of replacing it with a count-matched experiment (comment 8). We will state more explicitly that differences between TA2-1 and TA2-2 reflect both the training strategy and the unequal reporting density between the western and eastern CONUS. This unequal reporting density reflects the real-world conditions a globally deployed model would face, and the experiment is intended to capture it.

References

Aas, K.; Jullum, M.; Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artificial Intelligence, 298, 103502. DOI: https://doi.org/10.1016/j.artint.2021.103502

Ben Bouallègue, Z.; Haiden, T.; Weber, N. J.; Hamill, T. M.; Richardson, D. S. (2020). Accounting for representativeness in the verification of ensemble precipitation forecasts. Monthly Weather Review, 148, 2049–2062. DOI: https://doi.org/10.1175/MWR-D-19-0323.1

Carriero, A.; Luijken, K.; de Hond, A.; Moons, K. G. M.; van Calster, B.; van Smeden, M. (2025). The Harms of Class Imbalance Corrections for Machine Learning Based Prediction Models: A Simulation Study. Statistics in Medicine, 44, e10320. DOI: https://doi.org/10.1002/sim.10320

Gascón, E.; Montani, A.; Hewson, T. D. (2024). Post-processing output from ensembles with and without parametrised convection, to create accurate, blended, high-fidelity rainfall forecasts. Quarterly Journal of the Royal Meteorological Society, 150, 3117–3145. DOI: https://doi.org/10.1002/qj.4753

Giannaros, C.; Dafis, S.; Stefanidis, S.; Giannaros, T. M.; Koletsis, I.; Oikonomou, C. (2022). Hydrometeorological analysis of a flash flood event in an ungauged Mediterranean watershed under an operational forecasting and monitoring context. Meteorological Applications, 29, e2079. DOI: https://doi.org/10.1002/met.2079

Hemri, S.; Hewson, T.; Gascón, E.; Rajczak, J.; Bhend, J.; Spirig, C.; et al. (2022). How do ecPoint precipitation forecasts compare with postprocessed multi-model ensemble predictions over Switzerland? ECMWF Technical Memoranda, 901. DOI: https://doi.org/10.21957/hy89j7svk

Hewson, T. (2024). Capturing extreme rainfall events. ECMWF Newsletter, 178. https://www.ecmwf.int/en/newsletter/178/news/capturing-extreme-rainfall-events

Hewson, T. D.; Pillosu, F. M. (2021). A low-cost post-processing technique improves weather forecasts around the world. Communications Earth & Environment, 2, 132. DOI: https://doi.org/10.1038/s43247-021-00185-9

Marjerison, R. D.; Walter, M. T.; Sullivan, P. J.; Colucci, S. J. (2016). Does Population Affect the Location of Flash Flood Reports? Journal of Applied Meteorology and Climatology, 55, 1953–1963. DOI: https://doi.org/10.1175/jamc-d-15-0329.1

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

Ortenzi, S.; Massari, C.; Ciabatta, L.; Cencetti, C.; Dionigi, M.; Marchesini, I.; Stelluti, M.; Di Matteo, L. (2025). Leveraging SMAP-Based Soil Moisture to Identify Runoff Thresholds on Flash Flood–Prone Basins in the Mediterranean Region. Journal of Hydrometeorology, 26, 975–990. DOI: https://doi.org/10.1175/jhm-d-24-0143.1

Panwar, V.; Sen, S. (2020). Disaster Damage Records of EM-DAT and DesInventar: A Systematic Comparison. Economics of Disasters and Climate Change, 4, 295–317. DOI: https://doi.org/10.1007/s41885-019-00052-0

Pillosu, F. M. (2026). Outrunning flash floods: improving global medium-range forecasts for better preparedness. PhD thesis, University of Reading. DOI: https://doi.org/10.48683/1926.00128953

Pillosu, F. M.; Hewson, T. D.; Gascón, E.; Vučković, M.; Prudhomme, C.; Cloke, H. (2026). Bridging the scale gap: enhancing point-scale rainfall estimates by post-processing ERA5. Bulletin of the American Meteorological Society, early online release. DOI: https://doi.org/10.1175/BAMS-D-25-0019.1

Pillosu, F. M.; Hewson, T. D.; Stephens, E.; Prudhomme, C.; Cloke, H. (2025). Does a multivariate approach enhance univariate grid-to-point post-processed rainfall forecasts? A comparative analysis. ECMWF Technical Memoranda, 924. DOI: https://doi.org/10.21957/eb6c62b7af

Roberts, D. R.; Bahn, V.; Ciuti, S.; Boyce, M. S.; Elith, J.; Guillera-Arroita, G.; et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40, 913–929. DOI: https://doi.org/10.1111/ecog.02881

Wadoux, A. M. J. C.; Heuvelink, G. B. M.; de Bruin, S.; Brus, D. J. (2021). Spatial cross-validation is not the right way to evaluate map accuracy. Ecological Modelling, 457, 109692. DOI: https://doi.org/10.1016/j.ecolmodel.2021.109692