Response to Referee #1

We highly appreciate the reviewer for the constructive and detailed comments. We have carefully considered all comments and have outlined our responses and the corresponding revisions to the manuscript below.

Major Comments

Comment 1: Throughout the paper, terms like flood, flash flood, POT floods, and flashiness are central, but the hydrological definition is not completely transparent from the text. In the Methods section where AMF, POTF, POTM, POTFL and MDF are defined, please clearly state: The exact thresholds for POT events (e.g. percentile, exceedance criteria); The minimum inter-event time (if any) used to separate consecutive events; How you handle overlapping events and multi-peak events. It would also help to explicitly say whether your "flash floods" are defined purely from streamflow response time and hydrograph shape, or whether you also enforce a rainfall duration criterion (e.g. rainfall concentrated within ≤6 h). I suggest adding a short paragraph early in the Methods section with a operational definition (1-2 sentences) and a reference.

Response 1: We appreciate these comments. In the revised manuscript, we have added an operational definition of flash floods and a complete description of the POT event extraction procedure. In this study, flash floods are operationally identified as rapid, high-intensity runoff responses of short duration, characterized solely from instantaneous streamflow hydrographs, and no rainfall duration criterion is imposed. The study focuses on small-to-medium catchments (drainage area < 3000 km^2). For the POT method, the discharge threshold at each gauge was iteratively set to yield an average of two events per year. Inter-event independence follows Lang et al. (1999): consecutive peaks must be separated by at least $D > 5 + \log(0.386A)$ days (A in km^2), and the minimum inter-peak discharge must fall below 75 % of the lesser peak. If these criteria are not met, the peaks are merged and only the maximum value is retained.

Comment 2: You use 294 "small and medium-sized catchments" across CONUS, mostly with areas <3000 km^2. The broad selection rules are mentioned, but some important details are not fully clear. I suggest clarifying filtering rules & missing data: What is the minimum record length and completeness required (e.g. at least X years of hourly data, no more than Y % missing)? How are gaps handled (e.g. is a year with missing sub-daily data excluded from the trend analysis)? Do you exclude gauging stations with clear signs of strong regulation (e.g. upstream reservoirs) or use all available?

Response 2: We appreciate these comments. We have revised Section 3 to clarify the filtering protocol, data imputation methods, and the handling of potentially regulated catchments.

Station selection followed a multi-stage protocol centered on the May-September period, when flash flood activity peaks across CONUS. A seasonal year was classified

as complete only if no gap of more than six consecutive hours of missing streamflow occurred during the five-month window, and stations with fewer than 20 such complete seasonal years were subsequently discarded. Once a station passed this threshold, all available data across its full period of record were used for trend analysis to maximize time-series length. Stations affected by persistent drought or containing any full year of missing data were additionally excluded. No interpolation or gap-filling was applied. The completeness criterion itself ensures near-continuous coverage within each qualifying seasonal year, minimizing the influence of data gaps on trend estimation.

No explicit screening for upstream regulation was applied. The restriction to small-to-medium headwater catchments (median area 886 km^2) limits exposure to major reservoir operations, though minor regulation effects at individual stations cannot be entirely excluded.

Comment 3: You use a rich set of precipitation indices (AFHP1h/6h/12h/24h, AMP-N, RX1D, RX6H etc.), which is excellent. However, the connection between these indices and the typical response time of your catchments could be explained more clearly. Please explain why you chose these particular time windows (1 h, 6 h, 12 h, 24 h) and how they relate to: Median concentration time or lag time of the catchments. The typical duration of the POT events you identify in streamflow. For catchments with longer response times, 24 h rainfall might be more relevant. Do you see different patterns there? A few sentences in the Results section (or Discussion) would help.

Response 3: We appreciate these comments. The 1-h, 6-h, 12-h, and 24-h accumulation windows were selected to span the range of concentration times across our study catchments. The shorter windows (1 h and 6 h) target high-intensity bursts that drive flood peaks in smaller catchments with faster responses, while the longer windows (12 h and 24 h) represent cumulative rainfall forcing more relevant to larger catchments with longer lag times. Additionally, the AMP-N index further adapts the accumulation window to each catchment's hydrologic response time via $n = log(0.386A)$. The median MDF across all 294 catchments is 21 h. Given that the MDF encompasses both the rising and recession limbs, and the rising limb alone is typically completed within sub-daily timescales, and the 1 to 12 h precipitation windows capture the rainfall input most directly linked to flood peak generation.

To examine whether larger catchments show stronger associations with longer duration precipitation, we computed Spearman rank correlations between each precipitation index and flood metrics, stratified into four drainage-area classes (<500, 500-1000, 1000-2000, and >2000 km^2). Median correlations remain stable across size classes for all three flood metrics, with ranges across groups typically below 0.05 and no systematic strengthening of longer duration indices in larger catchments (Fig. S1). The relevant revisions have been made in Section 2.1 and Section 5 of the revised manuscript.

Comment 4: The Random Forest (RF) analysis is an important part of the study, but some methodological details are currently too brief. It would be helpful to provide basic RF settings (number of trees, maximum depth, minimum samples per leaf, random seed) and to indicate whether any cross-validation or out-of-bag error estimates were used. In addition, the procedure for converting continuous importance values into discrete dominance classes (e.g., "precipitation-dominated", "temperature-dominated") should be explained—particularly the threshold used and the rationale behind it. Given that many predictors are correlated (e.g., PET-T, various precipitation indices), acknowledging potential biases in RF importance and, if feasible, adding a simple sensitivity check such as permutation importance or removing one of two highly correlated predictors would be valuable. (Or add some discussion)

Response 4: We appreciate these comments. We have expanded Section 2.4 to provide a detailed description of the RF procedure. Hyperparameters were optimized per catchment using the Tree-structured Parzen Estimator (TPE). Data were split into 80% training and 20% validation sets, and the TPE maximized validation R^2 over 100 evaluations. The optimized hyperparameters include n_estimators (50-500), max_depth (None or 10-40), min_samples_split (2-20), and min_samples_leaf (1-10). Neither cross-validation nor out-of-bag estimation was used, as the RF serves driver attribution rather than predictive generalization. The 20% validation set serves only to prevent hyperparameter overfitting.

The original fixed 15% threshold lacked an objective basis. In the revision, variables are ranked by mean decrease in impurity (MDI) and a cumulative importance curve is constructed for each catchment. The inflection point, defined as the rank of maximum absolute second-order difference, marks the transition from dominant to marginal contributors. As reported in Section 4.4 of the revised manuscript, the median inflection point across 294 catchments falls at approximately 22% cumulative importance. The 15% threshold is more stringent than all observed inflection points and thus retains only the most prominent drivers in every catchment.

As noted, inter-correlated predictors share MDI scores, diluting individual variable importance. A discussion has been added to Section 5. Aggregating importances into three broad driver categories before catchment classification renders within-category redistribution inconsequential, while cross-category redistribution is expected to be minor given the distinct physical nature of the three groups.

Comment 5: The "P+T-dominated" category appears especially meaningful but is not yet fully explored. A short summary of the climatic and land-cover characteristics of these catchments, and how they differ from purely precipitation-dominated catchments, would help the reader interpret why both drivers matter simultaneously. The two DRIVE temperature scenarios ("dynamic temperature" vs. "static 1981 temperature") are central to the conclusions, but the construction of these scenarios needs further clarification. Please specify exactly which meteorological variables differ between DRIVE-DT and DRIVE-ST—whether only air temperature is modified, or whether

entire 1981 meteorological fields are reused.

Response 5: We appreciate these comments. We now compare the baseline climatology of P&T-dominated and P-dominated catchments in Section 4.4. The P&T-dominated catchments are drier (median annual precipitation 982 vs. 1184 mm) and cooler (14.5 vs. 15.9 °C) than the P-dominated catchments, yet warming rates are comparable between the two groups. Land cover compositions differ little between the two groups. In these drier catchments, temperature-driven increases in evaporative demand represent a larger fraction of the water budget, making runoff generation more sensitive to warming. During revision, we also corrected a computational error in the original driver categorization. The updated percentages are reported in Section 4.4 and Table S1 without affecting the main conclusions.

We have clarified the DRIVE scenario design in Section 2.3. In DRIVE-ST, only air temperature is replaced by repeating the full 1981 3-hourly air temperature series for each simulation year. All other forcing fields retain their observed 1981-2020 time series and are identical between DRIVE-DT and DRIVE-ST.

Comment 6: The land-cover analysis is important, particularly the conclusion that LULC change has limited hydrological impact in most catchments. Providing more methodological detail would strengthen this argument. For example, please describe how GLC_FCS30 is reclassified into the land-cover categories required by DRIVE and TS-DUH, whether fractional cover or dominant classes are used, and how TS-DUH incorporates land cover (e.g., roughness, routing speed, retention). In the Results, when reporting that most catchments show <3-5% hydrological change, a brief summary of the actual land-cover transitions in the 294 catchments (e.g., forest-cropland, cropland-urban) would help readers assess whether the small hydrological response reflects limited LULC change or limited model sensitivity. For the urban case study (e.g., Atlanta), reporting the baseline impervious fraction and the absolute magnitude of the peak-flow increase would provide useful context.

Response 6: We appreciate these comments. We have expanded Section 2.3 to detail the land cover representation in both models. The GLC_FCS30 land cover maps were reclassified into broader categories compatible with each model's classification scheme via a cross-walking table. DRIVE uses a fractional cover approach, calculating the percentage area of each reclassified land cover type within each 0.125° grid cell from the underlying 30m pixels. These fractions then determine grid-averaged vegetation parameters such as leaf area index, albedo, and roughness length via the VIC parameter library. In TS-DUH, Manning's roughness coefficients (n) for each 90m grid cell are derived as the area-weighted average of n values assigned to the underlying 30m land cover pixels. These n values determine flow velocities across the catchment, thereby capturing the influence of land surface characteristics on flood wave propagation at a finer scale.

In Section 4.6, we have added a quantitative summary of land cover changes between 1985 and 2015. Across all catchments, the total area undergoing any class transition

averaged only 3.3%. The dominant net shifts were a decline in forest cover from 65.2% to 64.4% and an expansion of impervious surfaces from 2.8% to 3.6%, with all other classes changing by less than 0.3%. Given the minimal magnitude of these transitions, the small simulated hydrological response is physically consistent with the limited degree of land cover modification observed.

For the Peachtree Creek case study, we have revised Section 5 to include the baseline impervious fraction and the absolute magnitude of peak-flow change, as suggested.

Comment 7: Because the study applies Mann-Kendall tests and Poisson regressions across nearly 300 catchments and multiple flood and climate indices, it would be appropriate to briefly discuss the potential for false positives arising from multiple hypothesis testing. Even a short note on whether any false-discovery-rate (FDR) considerations were made—or how many significant trends might be expected by chance—would be helpful. Clarifying how sensitive the key conclusions (e.g., "67% of catchments show significant increases") are to the chosen p-value threshold would further improve transparency. For the temperature and LULC scenario results (e.g., the reported 3.6%, 8%, and 20% changes), showing the distribution across catchments (e.g., boxplots or histograms) would allow readers to understand variability and uncertainty rather than rely on single summary statistics.

Response 7: We appreciate these comments. We applied the Benjamini-Hochberg FDR correction to all trend tests and added a discussion in Section 5 of the revised manuscript. Full FDR-corrected results are provided in Table S3. Under the null hypothesis, 294 tests at $\alpha=0.05$ would yield approximately 15 false positives per variable. The heavy precipitation frequency signal is highly robust: 190 of 195 significant AFHP6h upward trends survive FDR correction at $q=0.05$. The POTF signal is attenuated but remains well above chance expectations, with 22 of 51 upward trends retained after correction. Significant trend counts at $p < 0.01$, 0.05, and 0.10 are tabulated in Table S2, confirming that qualitative conclusions are robust to threshold choice.

We agree that distributional context is essential for interpreting these results. We clarify that the reported 3.6% and 8.0% values represent differences in the proportion of catchments with significant increasing trends between the dynamic and static temperature scenarios, rather than flow changes by catchment. Distributions of temperature-induced flow changes for each catchment are presented in Fig.7b, and LULC-induced changes in Fig.8b and 8d. The 20% peak flow increase pertains specifically to the Peachtree Creek catchment, rather than an average across all catchments.

Minor comments

Some expressions in the abstract could be clearer. For example, the phrase "mitigate this effect" is slightly vague—please make clear which effect is meant (e.g., "mitigate the increase in flash-flood frequency associated with heavier precipitation"). The final sentence is long and contains several ideas (spatial-temporal variation, urbanization,

flood-risk management). Splitting it into two sentences would improve readability.
Response 8: We appreciate these comments. We have revised "mitigate this effect" to "mitigate the expected increases in flood frequency and intensity associated with heavier precipitation" for clarity. The final sentence of the abstract has been split into two sentences to improve readability.

The manuscript uses both "land cover" and "land-cover." Please choose one style and apply it consistently. A common approach is to use "land cover" as the noun and "land-cover change" when used as an adjective.
Response 9: We appreciate these comments. We have systematically reviewed the entire manuscript and confirmed that "land cover" is used consistently throughout the text.

Please ensure all abbreviations are defined at their first appearance in the main text—not only in tables. Some that may need checking include POTFL, MDF, AFHP, AMP-N. Key abbreviations should also be briefly defined in figure captions so that each figure can be understood on its own.
Response 10: We appreciate these comments. We have carefully reviewed all abbreviations and ensured that each is defined at its first occurrence in the main text. We have also revised all figure captions to include brief definitions of key abbreviations, so that each figure can be understood independently of the main text or tables.

Please check for consistent unit formatting throughout the manuscript—for example "mm d^{-1}" vs "mm/day" and "°C" vs "degC." Make sure the units for PET and temperature-related indices are given at their first mention and shown clearly in figure axes.
Response 11: We appreciate these comments. We have thoroughly reviewed the entire manuscript and standardized all unit formatting for consistency.

In places where "flood intensity" is used, please specify whether this refers to peak flow, specific discharge, or flood volume.
Response 13: We appreciate these comments. We have added an explicit definition in Section 2.1.

When describing the effect of warming, please consider adding qualifiers such as "in our model experiments" or "under the scenario where only temperature changes."
Response 14: We appreciate these comments. We carefully addressed all of them.