Responses to Reviewer #3

Dear Reviewer,

We sincerely thank the reviewer for the positive evaluation of our manuscript and for recognizing the value of the three-dimensional event-based perspective, the phase-dependent analysis of heatwave evolution, and our efforts toward data and code transparency.

Reviewer's overall assessment

This study identifies heatwaves over mainland China during May–September of 1986–2024 as spatiotemporally contiguous three-dimensional events using ERA5-Land daily maximum temperature, and compares event persistence, cumulative severity, affected area, track length, and moving speed across three warming phases (1986–1998, 1999–2011, 2012–2024). The central finding, that recent heatwave change after 2012 is expressed less through a rise in event counts and more through the joint strengthening of event-level characteristics, is well supported by the analyses and is a useful contribution. The event-based framing is timely, the manuscript is clearly organized, and I appreciate that the event catalogue, gridded products, and analysis scripts are archived, which supports reproducibility.

Two earlier referees have already commented in detail on the transparency of the identification procedure, the definition of HWA, terminology, and table and figure presentation. I agree with those comments and will not repeat them. My review focuses on complementary points concerning the phase division, the interpretation of the correlations among event metrics, the trend estimation within phases, and event assignment near the seasonal and phase boundaries. None of these requires new data, and most can be addressed with clarifications, modest additional analysis, or more careful wording. I therefore recommend minor revisions, with General Comments 1 and 2 deserving the most attention.

Response: We sincerely thank the reviewer for the positive assessment of the manuscript and for the constructive comments. We appreciate the recommendation for minor revision and will carefully address all points raised below.

General Comments

R3G1: The Pettitt test identifies a single change point, and the manuscript reports 2012 as that turning point (lines 170–172). The basis for the second boundary at 1998/1999 is less clear. The 3 phases are each exactly 13 years long, which gives the impression that the earlier boundary was chosen at least partly for symmetry, and the supporting argument currently rests on the global warming hiatus literature (Li et al., 2019) rather than on a test applied to this dataset. Please state explicitly how the 1998/1999 boundary was determined, for example through a sequential or multiple change-point test and briefly demonstrate that the main conclusions are robust to reasonable shifts of this boundary (for instance, moving it by one to three years). Related to this, the fitted warming trends in Phases I and II are not statistically significant ($p > 0.05$ in Figure 3), so describing all three periods as distinct warming phases should be worded with care; the essential and well-supported contrast is between the pre-2012 and post-2012 periods.

Response to R3G1: Thank you for this important comment. We agree that the basis for the 1998/1999 boundary was not sufficiently explained in the original manuscript and that the terminology used for the three periods requires greater care.

The 2012 boundary was identified using the Pettitt change-point test applied to the annual mean daily maximum temperature series over mainland China, and therefore represents the statistically detected change point in the present study. By contrast, the boundary between 1998 and 1999 was not identified as a second statistical change point. It was adopted as a literature-informed analytical subdivision corresponding approximately to the late-1990s warming slowdown discussed in previous studies, including Li et al. (2019). This boundary also divides the 1986–2024 study period into three equal 13-year intervals, thereby reducing differences in sample length and facilitating consistent comparisons among periods. We will explain these considerations explicitly and refer to the three intervals as "analysis periods" or "temporal periods", rather than implying that all three are statistically detected warming phases.

We further evaluated the sensitivity of the results to the earlier boundary by shifting the starting year of Period II from 1996 to 2002, corresponding to shifts of up to three years around the original 1998/1999 boundary. Because these alternative divisions produce periods of unequal length, event occurrence was evaluated using the annual mean event count rather than the raw total number of events. Across all seven boundary settings, the annual mean event count in Period III remained 0.7%–4.6% lower than that in Period II. In contrast, the mean HWD, HWS, HWA, and HWL in Period III remained consistently higher than those in Period II, by 8.9%–11.8%, 22.1%–26.0%, 5.1%–7.7%, and 12.2%–15.4%, respectively. These four characteristics were also consistently higher in Period III than in Period I under every alternative boundary.

These results demonstrate that shifting the earlier analytical boundary affects the exact period-specific values but does not alter the central conclusion that recent heatwave change is characterized by stronger event-level processes without a continued increase in event occurrence. We will include the boundary-sensitivity results in the Supplement and briefly summarize them in the revised manuscript. We will also emphasize that the clearest statistically supported temporal contrast is between the pre-2012 and post-2012 periods.

R3G2: HWL is defined as the sum of DD − 1 daily centroid displacements, so a longer-lasting event accumulates a longer track by construction, even if its per-day movement is unchanged. Similarly, HWS integrates temperature exceedance over both days and area, so it embeds HWD and HWA in its definition. The strong positive correlations in Figure 7, and their interpretation in Section 4.2 as evidence that track length carries particular process significance, are therefore partly built into the metric definitions rather than purely empirical findings. I do not think this invalidates the analysis, but the manuscript should acknowledge the construction-based component of these relationships explicitly. To show that mobility carries information beyond duration, I suggest a short supplementary analysis, for example the correlation of per-day displacement (HWL/HWD, which is Speed) or of duration-detrended residuals with the other characteristics, or partial correlations between HWL and HWS or HWA controlling for HWD. Interestingly, the weak correlations of Speed with HWS and HWD already reported in Section 3.3 are relevant here: they suggest that much of the HWL signal is indeed carried by duration. A franker treatment of this point would make the argument of Section 4.2 more convincing, and the conclusion (3) statement that track length better captures the integrated evolution of events should be qualified accordingly.

Response to R3G2: Thank you for this insightful comment. Specifically, because HWL is calculated as the cumulative distance of centroid displacement over the lifetime of an event, longer-duration events can accumulate larger track lengths even when their mean daily displacement remains unchanged. Similarly, HWS integrates temperature exceedance over both time and affected area, and therefore has inherent relationships with HWD and HWA.

We acknowledge that this construction-based component was not sufficiently discussed in the original manuscript. We will therefore revise Sections 3.3 and 4.2 to clarify that the relationships between HWL and the other event characteristics represent associations among integrated event properties and should not be interpreted as independent effects of mobility.

Following the reviewer's suggestion, we additionally performed partial-correlation analyses between HWL and HWS/HWA while controlling for HWD. For all 609 events, the partial correlation between HWL and HWS decreased markedly to $r = 0.146$ ($p < 0.001$), compared with the much stronger bivariate correlation reported in Figure 7. The corresponding partial correlations were $r = 0.242$ ($p = 0.001$), 0.163 ($p = 0.016$), and 0.092 ($p = 0.184$) for Periods I, II, and III, respectively. This substantial weakening indicates that much of the apparent association between HWL and cumulative severity is indeed related to event duration.

In contrast, the relationship between HWL and HWA remained evident after controlling for HWD. The partial correlation was $r = 0.546$ ($p < 0.001$) for all events and remained similar across the three periods, with $r = 0.546$, 0.518, and 0.576, respectively (all $p < 0.001$). These results suggest that, although the relationship of track length with cumulative severity is strongly influenced by event persistence, track length retains a clear association with the spatial extent of contiguous heatwave events that cannot be explained by duration alone. The weaker relationships involving Speed reported in Figure S6 provide complementary evidence that accumulated track length and mean movement rate describe different aspects of event evolution.

Accordingly, we will revise Section 4.2 and Conclusion (3) to avoid implying that track length independently controls other heatwave characteristics. Instead, we will describe HWL as an integrated trajectory metric whose relationships with duration and severity partly reflect its cumulative construction, while its persistent duration-controlled association with affected area indicates additional information on the spatial evolution of contiguous heatwave events. The partial-correlation results will also be provided in the Supplement.

R3G3: Each phase contains only 13 annual values, so fitted slopes are sensitive to single years and endpoints (for example, the very active years 2022 and 2024 fall in Phase III). Please state in the Methods which trend estimator and significance test were used (ordinary least squares, Sen's slope with Mann-Kendall, or other) and justify the choice of $p < 0.1$ as the significance level in Table 2; marking $p < 0.05$ as well would help readers judge the strength of the evidence. Please also clarify how the annual values of the event metrics were constructed, that is, whether each annual value is the mean over all events starting in that year, and how many events enter each annual mean, since annual means over roughly 15 events can be noisy. A brief caution about the short window length would be appropriate in the limitations paragraph.

Response to R3G3: Thank you for this constructive comment. We agree that trend estimates based on only 13 annual values per period should be interpreted with caution, particularly because individual active years and the positions of the endpoints may influence the fitted slopes.

In the revised Methods, we will clarify that the annual trends shown in Figure 5 and Table 2 were estimated using ordinary least squares (OLS) linear regression, with year as the independent variable. The statistical significance of each trend was evaluated using a two-sided t-test of the regression slope coefficient. We retained OLS as the primary estimator because its slope provides a directly interpretable annual rate of change that can be compared consistently among indicators and periods. To evaluate the robustness of these short-window estimates, we additionally conducted Theil–Sen slope estimation, Mann–Kendall testing, and leave-one-year-out analyses. The corresponding results will be reported in the Supplement.

We will also clarify how the annual event metrics were constructed. HWF represents the number of heatwave events initiated in each calendar year. The annual values of HWD, HWS, HWA, HWL, and Speed were calculated as the means of all heatwave events initiated in that year. Across 1986–2024, the number of events contributing to the annual statistics ranged from 9 to 22, with a mean of 15.62 events yr^{-1}. This information will be added to the Methods to make the annual aggregation procedure and its associated sampling variability explicit.

We agree that the original use of a single symbol for $p < 0.10$ did not adequately distinguish conventional statistical significance from weaker evidence. Rather than treating $p < 0.10$ as conventional statistical significance, the revised Table 2 will distinguish $p < 0.05$ as statistically significant and $0.05 \leqslant p < 0.10$ as marginal evidence. Exact p values and the non-parametric robustness results will be provided in the Supplement. Under the OLS analysis, the Period III trends in HWD, HWS, and track length were significant at $p < 0.05$, with p values of 0.0416, 0.0270, and 0.0438, respectively. The positive trend in HWA was marginal ($p = 0.0839$), whereas the trends in HWF and Speed were not statistically significant.

The leave-one-year-out analysis further showed that the fitted slopes of HWD, HWS, HWA, and track length in Period III remained positive after the exclusion of any single year, including 2022 or 2024. This indicates that the positive direction of these trends was not driven by either of these active years alone. However, the exact slope magnitudes and significance levels changed under some exclusions, confirming that the statistical strength of the period-specific trends is sensitive to the short 13-year window. We will therefore revise the Results and Discussion to distinguish directional consistency from statistical significance and add a limitation statement noting that the period-specific slopes should be regarded as supporting evidence rather than as stand-alone proof of the reported period contrasts.

Specific Comments

R3S1: (Lines 135–136). At 0.1° resolution the grid-cell area varies substantially with latitude across the study domain, from roughly 120 km^2 in the far south to below 80 km^2 in the north. Please confirm that Ai in Table 1 and the 1.6% area criterion are computed from true (latitude-dependent) cell areas rather than from cell counts, and state this in the Methods.

Response to R3S1: Thank you for pointing out this important issue. We confirm that the area calculations use the actual latitude-dependent area of each 0.1° grid cell rather than a constant cell area or the number of grid cells. The term in Table 1 therefore represents the actual geographic area of grid cell so the calculation of HWS accounts for the variation in grid-cell area across mainland China.

The same area calculation was used when applying the 1.6% minimum-area criterion. Specifically, the area of each spatially connected hot-temperature patch was calculated by summing the actual areas of all grid cells contained in that patch and was then compared with the threshold corresponding to 1.6% of the land area of mainland China. We will add this clarification to the Methods section.

R3S2: (Lines 160–167). Please report the p values of the Phase I and Phase II temperature trends in the text, not only in Figure 3, and temper the wording accordingly (see GC1).

Response to R3S2: Thank you for this suggestion. In the revised manuscript, we will report the exact p values of the fitted temperature trends for Periods I and II in the main text and clarify that, although both fitted slopes are positive, neither trend is statistically significant.

Consistent with our response to General Comment 1, we will also revise the wording in this section to avoid describing all three intervals as statistically distinct warming phases. We will instead refer to them as analysis periods with different background temperature conditions and emphasize that the clearest statistically supported contrast is between the pre-2012 and post-2012 periods.

R3S3: (Lines 253–259 and Figure 7). The text states that the relationships are nonlinear and that the increases in HWD and HWS steepen when HWL exceeds approximately 1×10^4 km. Please state what functional form was fitted to produce the curves in Figure 7, how the confidence bands were obtained, and how the 1×10^4 km value was identified (visually or through a formal breakpoint fit). Only a small number of events appear to exceed this value, so the steepening should be described with appropriate caution.

Response to R3S3: Thank you for this constructive comment. We agree that the fitting procedure and the interpretation of the upper tail of Figure 7 require clearer explanation.

To make the fitting procedure explicit and avoid arbitrary selection of polynomial order, we re-evaluated the relationships on the original data scale using first-, second-, and third-order polynomial regression models. For each relationship and analysis period, the polynomial order was selected independently according to the minimum corrected Akaike information criterion (AICc). Consequently, the selected functional form was not uniformly nonlinear across all panels: linear models were selected for HWS and HWD in Period I and for HWD in Period II, whereas higher-order polynomial models were selected for the remaining panels. We will state this procedure explicitly in the Methods and revise the general wording that previously described all relationships as nonlinear.

The shaded regions in Figure 7 represent the 95% confidence intervals for the fitted mean response, calculated using the default confidence-interval output of MATLAB predict for the selected fitlm model. This information will also be added to the Methods and the figure caption.

We further confirm that the approximate value of 1×10^4 km was not obtained from a formal breakpoint or segmented-regression analysis. It was an approximate visual reference based on the distribution of the observations and the apparent shape of the fitted relationships. Only 14 of the 609 events (2.30%) had HWL values greater than 1×10^4 km. We therefore agree that this value should not be interpreted as a statistically defined threshold. Accordingly, we will remove the statement that HWD and HWS increase more rapidly once HWL exceeds approximately 1×10^4 km. The revised text will instead describe the upper-range pattern cautiously and acknowledge the greater uncertainty associated with the limited number of long-track events.

R3S4: (Lines 269–275). The inverted U-shaped relationship between Speed and the other characteristics is one of the more interesting results and deserves one or two sentences of physical interpretation. A plausible reading is that very fast-moving events pass through a region too quickly to accumulate severity, while near-stationary events accumulate heat locally but produce short tracks. Even a cautious remark along these lines would help readers, and it also connects naturally to Luo et al. (2024) on slower-moving large heatwaves.

Response to R3S4: Thank you for this insightful suggestion. We agree that the inverted U-shaped relationships merit a brief physical interpretation and will incorporate the interpretation proposed by the reviewer into the revised Discussion. Very fast-moving heatwave events may pass through a given region too quickly to maintain prolonged local heat accumulation, thereby limiting their persistence and cumulative severity. At the other end of the spectrum, nearly stationary events may sustain strong local heat conditions but undergo little spatial displacement, resulting in relatively short tracks and a more spatially confined event footprint. Events with intermediate movement speeds may combine sufficient local residence time with continued spatial propagation and may therefore be associated with larger event-scale duration, cumulative severity, and affected area. We will relate this interpretation to the slower movement of large heatwaves reported by Luo et al. (2024). We will also make clear that this explanation is a plausible interpretation of the observed statistical relationships rather than a causal mechanism directly demonstrated by the present analysis. Because Speed is defined from both HWL and HWD, part of its statistical relationship with duration is mathematically coupled. We therefore interpret the inverted-U pattern descriptively rather than as direct evidence of a causal mechanism.

R3S5: (Lines 303–316 and Figure 9). The finding that May shows the highest mean HWS, HWA, HWL, and Speed is somewhat counterintuitive, since July is climatologically the hottest month. Two clarifications would help. First, because a calendar-day-relative threshold is used, May events represent large anomalies relative to a cooler climatology, which should be said explicitly so that readers do not interpret May events as the hottest in an absolute sense. Second, please check whether the May maxima are driven by a small number of very large events (the Phase I case event of May 1995 is itself a 30-day event starting on 5 May); showing medians alongside means in Figure 9, or reporting the event counts per month, would indicate how robust the monthly contrasts are.

Response to R3S5: Thank you for this insightful comment. We agree that the higher mean values of HWS, HWA, HWL, and Speed in May require careful interpretation because July is climatologically the hottest month over mainland China.

First, we clarify that the heatwave events in this study were identified using calendar-day-specific 90th percentile thresholds rather than absolute temperature thresholds. Therefore, the larger May values should be interpreted as characteristics of events identified relative to the May climatological background, rather than as evidence that May experiences higher absolute temperatures than July. We will add this clarification to the revised manuscript.

Second, following the reviewer's suggestion, we examined the monthly event counts and compared the mean and median values of the event-level characteristics. The numbers of identified heatwave events during May–September were 128, 137, 119, 109, and 116, respectively. The number of May events was therefore comparable to that of July, indicating that the relatively high May values were not caused by an insufficient sample size.

Furthermore, median values will be shown together with mean values in the revised Figure 9 to evaluate whether the monthly contrasts were dominated by a small number of extreme events. The median values generally showed similar seasonal variations to the mean values for HWD, HWS, HWA, HWL, and Speed, suggesting that the larger May values were not solely driven by a few extreme events. However, we acknowledge that individual long-duration events may still influence monthly averages, and therefore we will revise the interpretation of the monthly differences more cautiously.

Recommendation

This is good research just needs some minor revisions. The dataset, method, and central message are sound, and the requested changes concern justification, interpretation, and presentation rather than the core analysis.

Response: Thank you for your positive assessment and constructive suggestions. We have carefully considered all comments and completed the additional analyses and clarifications needed to address them.

The corresponding revisions will be incorporated into the manuscript and Supplement during the subsequent revision stage.

Closing note: We again thank the reviewer for the constructive suggestions, which have helped us improve the clarity, robustness, and interpretation of the study. The corresponding revisions will be incorporated into the manuscript and Supplement during the subsequent revision stage.