Response to Reviewer 1

We thank the reviewer for a careful and constructive assessment. The comments on the robustness of the dynamic/non-circulation partition, the early-versus-late epoch comparison, the sensitivity of the flow-analogue analysis, and the record and recurrence claims prompts us to substantially strengthen the analysis. In particular, we will repeat the flow-analogue decomposition over a Euro-Atlantic matching domain with a year-block bootstrap and perform a full sensitivity analysis. We will also add a formal extreme-value (GEV) analysis, and correct the inconsistencies identified by the reviewer.

Comment 1: *The dynamic component is estimated from the mean T850 anomaly of the 25 closest Z500-pattern analogues, while the difference between that composite and the observed T850 anomaly is treated as the thermodynamic component. However, matching the pattern of Z500 anomalies does not guarantee that the circulation amplitude, pressure gradients, horizontal temperature advection, subsidence strength, or persistence are comparable. Indeed, the observed Z500 anomaly in Figure 11 appears substantially stronger than the analogue composite. Consequently, the reported 3 to 4 C residual may contain the effect of unmatched circulation intensity, antecedent soil moisture, advection, and other processes; not simply background warming. Could the authors repeat the analysis using analogues constrained by both Z500 pattern correlation and anomaly amplitude, or statistically adjust T850 for differences in Z500 amplitude? At minimum, the contribution associated with circulation mismatch should be quantified before the residual is described as evidence of thermodynamic amplification.*

> **Response:** We agree and we will address this point in several ways. Our similarity metric is the latitude-weighted Euclidean distance between Z500 anomaly fields, which, unlike a pattern correlation, penalizes differences in amplitude as well as in pattern. In the revised version of the manuscript, we will repeat the decomposition over a Euro-Atlantic matching domain (40° W–40° E, 25–75° N) and test its sensitivity to the matching domain, number of analogues, calendar window, distance metric and epoch boundary. We will also describe the residual throughout as a "non-circulation (residual)" component and state explicitly that it is a circulation-constrained approximation containing background warming, any unmatched circulation amplitude and possibly land-surface feedbacks.

Comment 2: *The manuscript states that the 2001–2025 analogue composite is systematically warmer than the 1970–2000 composite on all five event days. However, Figure 12 appears to show the early-period composite warmer than the late-period composite on approximately 26–28 June, with only limited or reversed separation on the remaining days. This visual result does not support the stated systematic warming across all days. The authors should provide the exact daily values, uncertainty intervals, and statistical significance of the early minus late differences. They should also verify whether the plotted curves, legend, or textual interpretation are incorrect. This issue directly affects one of the paper's central conclusions.*

> **Response:** The reviewer is correct, and we thank them for catching this error. The pooled late-minus-early difference over Germany is −0.1 °C with a year-block-bootstrap 95% confidence interval. The result is the same for equal-sized samples, a permutation test and within-epoch analogue selection. Thus, in the revised version of the manuscript we will remove the statement that the late epoch is systematically warmer and state that the epoch contrast is inconclusive; the evidence for a large non-circulation contribution rest on the full-period residual.

Comment 3: *The authors should define what constitutes a qualifying analogue, report similarity-distance distributions, account for unequal epoch lengths and repeated dates, and test whether the late-period excess differs from what would be expected by chance. Independent equal-sized analogue sets should also be selected within each epoch. In addition, Figure 13 visually labels the samples as n = 40 and n = 85, whereas its caption and main text report n = 38 and n = 87.*

> **Response:** A qualifying analogue will be defined explicitly. We will report the distances in physical units. We will also verify the text and figures for inconsistencies.

Comment 4: *Bootstrapping the 25 selected analogue days quantifies variability within that particular set, but it does not account for uncertainty arising from the choice of analogue number, European matching domain, distance metric, calendar window, epoch boundaries, or temporal dependence among analogue days. These choices are especially important because the authors acknowledge that the estimated dynamic–thermodynamic partition depends on them. A comprehensive sensitivity analysis should vary at least the number of analogues, matching domain, ±calendar window, similarity metric, and epoch boundary. A year block bootstrap would also be more appropriate than resampling individual days where analogues from adjacent dates or the same years may not be independent. Confidence intervals should be reported for the thermodynamic residual and early-versus-late difference, not only for the dynamic composite.*

Response: In the revised version of the manuscript, we will replace the day-level bootstrap by a year-block bootstrap (resampling whole years, 1000 resamples) and report 95% confidence intervals for the dynamic component, the residual and the early-minus-late difference.

Comment 5: *The Methods define the standardized exceedance using the standard deviation of the daily temperature maximum series, whereas the Figure 5 caption states that the standard deviation is calculated from the regional annual-maximum series. These are fundamentally different distributions and could produce very different values, including the reported 4.9σ exceedance.*

Response: The Methods wording was wrong: the standardized exceedance uses the 90th percentile and the standard deviation of the annual (May–June) maximum TX series over 1971–2000, i.e., it is a z-score relative to the block-maxima distribution. This is what was computed and is consistent with the reported values. The Methods and the captions of Figs. 4, 5 and 7 will be changed accordingly.

Comment 6: *Cumulative intensity is calculated as exceedance above the climatological mean, although heatwave days are selected using the 90th-percentile threshold. This formulation can mix event duration, seasonal climatology, and threshold exceedance in a way that may influence the rank-1 result. The authors should clarify the exact distributions used for P90 and σ, rename warming ratio as a standardized exceedance or z-score, and demonstrate robustness using alternative accepted heatwave metrics and thresholds, such as exceedance above P90, 95th-percentile definitions, and different minimum-duration criteria. The conclusion that the event was intensity-dominated should be shown to be insensitive to these methodological decisions.*

Response: "Warming ratio" will be replaced by "standardized exceedance (z-score)" throughout. Following Hobday et al. (2016), the cumulative intensity is integrated above the calendar-day climatological mean, i.e., it measures the total heat accumulated above normal conditions; this will be properly explained, and all metrics will be defined in a new Table. In the revised version of the manuscript, we will also recompute the 2026 ranks under five definitions (P90 or P95 threshold, 3- or 5-day minimum duration, cumulative intensity above the mean or above the threshold).

Comment 7: *Trends in heatwave duration and frequency do not by themselves establish that an event of this magnitude must be expected to recur with increasing frequency and intensity. Such a statement requires an extreme-value or formal event-attribution analysis with estimated return periods and probability changes. The record claims also need more careful qualification because most heatwave metrics are restricted to May--June, whereas phrases such as the most intense heat event on record imply comparison with the complete warm season. In addition, the long-term station records require explicit information on homogenization, relocations, instrumentation changes, missing data, and urbanization effects. The station and HYRAS-DE analyses should not be presented as fully independent confirmation because HYRAS-DE itself is derived from station observations.*

Response: We agree with the reviewer's comments/suggestions and in the revised version of the manuscript we will add a GEV analysis). Moreover, the statement that the event "must be expected to recur with increasing frequency and intensity" will replaced by a statement framed in terms of the observed trends and the GEV estimate. We will also state explicitly that the long-term series

are the quality-controlled daily records of the DWD Climate Data Center (Kaspar et al., 2013) and are not homogenized, so relocations, instrument changes and urbanization may affect century-scale comparisons; we therefore interpret them mainly through the rank and standardized magnitude of 2026 relative to each station's own record. We report data completeness (May–June data > 95% complete at 15 of 18 stations) and exclude years with more than six missing May–June days. We agree that HYRAS-DE is not independent of the stations, thus the two analyses will be described as consistent rather than independent lines of evidence.