Response to Reviewer 6

We thank the reviewer for the positive assessment and the detailed comments, which will help us to clarify the heatwave definition, the use of the datasets and the construction of the analogue composites and to improve our manuscript overall. Reviewer comments are shown in italics; our responses follow in plain texts.

General comments: The manuscript presents a comprehensive analysis of the June 2026 heatwave in Germany. Overall, I find this a solid study. The manuscript addresses several aspects of this anomalous heat event, combining its spatial and temporal characterization, long-term context, and circulation-based analysis. Although the study is primarily diagnostic, its originality lies in the exceptional nature of the event itself, and its characterization provides useful information for improving our understanding of such extreme events and, as the authors mention, potentially contributing to their anticipation and preparedness. However, several methodological aspects require clarification before acceptance. My comments and suggestions are provided in the attached review.

Comment 1: *I am not fully clear on how the specific days associated with the June 2026 heatwave were defined. The manuscript identifies a five-day period (25–29 June), and some of the event characterization focuses on the three most intense days. However, in the introduction, this heatwave is introduced as with the WMO definition or DWD (which, please note, is not referred to in the references). Different institutions may use different definitions of heatwaves. It would be useful to clarify how these specific dates were selected and how they relate to the heatwave definition applied in the study. Are these dates identified similarly with these definitions?*

Related to this point, the Methods state that the calendar-day P90 threshold is computed "at each grid point (or for each area-averaged regional series)." Are heatwaves independently identified at each grid point and separately from the regional-mean series? If so, these approaches could lead to different characteristics, right? Ideally, heatwave occurrence would first be identified locally at individual grid points and subsequently spatially aggregated, but other approaches are also possible. Please clarify exactly how heatwave events were defined for the station, gridded data and regional Germany federal states.

> **Response:** The heatwave detected in the Germany-average series lasted from 18 to 28 June 2026 (11 days). Within this event, 25–29 June is the period of the provisional national records and of the peak circulation anomaly (DWD, 2026), and it is used as the event peak for the flow-analogue analysis. The WMO/DWD descriptions in the introduction only characterize the event and are not used for detection. The reviewer is correct that heatwaves are detected in two complementary ways, which we be describing more explicitly in the revised version of the manuscript: (i) at each 1 km grid point, with statistics averaged over Germany afterwards (detect-then-aggregate), and (ii) in the area-averaged TX series of each federal state and of Germany (aggregate-then-detect). The station records are used only for annual maxima, standardized exceedances and the GEV analysis. Because averaging damps local extremes, the two approaches give different values (e.g., 13.6 heatwave days in 2026 averaged over all grid points vs. 11 days for the Germany-average series), and we compare results only within the same approach.

Comment 2: *It is not always clear which observational dataset is used for each part of the analysis. My understanding is that the long-term station data are mainly used in Section 3.3, whereas most other analyses rely on the 1-km HYRAS-DE gridded product. I suggest explicitly stating in the Methods which dataset is used for each analysis and why. This is particularly relevant for record-breaking temperatures, since extreme values can be sensitive to spatial interpolation and area averaging. Have the authors compared the HYRAS-DE values with the corresponding station observations during the 2026 event? Such a comparison, even if presented briefly or in the Supplement, would help demonstrate that the gridded dataset adequately captures the magnitude of the observed extremes. Relatedly, it may be worth explaining*

why station observations are not used more extensively in the event-level analysis, given that they provide the direct measurements underlying the record-breaking temperatures.

Response: A new paragraph in Sect. 2.1 will specify the dataset used for each analysis. We compared HYRAS-DE with the station observations: at the 17 non-summit long-term stations, the May–June 2026 maximum of the nearest 1 km grid cell differs from the station maximum by less than 0.3 °C (mean −0.06 °C), and at Neißemünde-Coschen, one of the stations that set the new national record, HYRAS-DE gives 41.6 °C vs. 41.7 °C observed. HYRAS-DE therefore captures the magnitude of the extremes at the station locations. As the stations are among the observations from which HYRAS-DE is interpolated, we describe this agreement as consistency rather than independent validation. Station observations are not used more extensively at the event level because they are too sparse to resolve the spatial structure of the event, which is essential for the regional analysis.

Comment 3: *I understand the practical reason for restricting the analysis to May–June. However, some additional climatological context would be useful. How common are heatwaves in May–June in Germany compared with July–August? Is June already a climatologically important heatwave season, or is the timing of this event itself unusual? The manuscript mentions that early-season heat can have particularly important impacts, but a brief quantitative or literature context regarding the climatological frequency and intensity of May–June heatwaves would strengthen this argument. It would be interesting if the authors make clear how representative this restricted early-summer period is relative to the full summer season.*

Response: We will add a paragraph to the Introduction to support our choice (e.g., between 1951 and 2025, the annual maximum of Germany-average TX occurred in July or August in 58 of 75 years and in June in 13 years, and only 36 of 255 days with Germany-average TX above 30 °C occurred in June (217 in July–August)). Early-summer heatwaves have become more common (years with at least one May–June heatwave: 18% in 1951–1990, 34% in 1991–2025). June is therefore not the main season of extreme heat in Germany, which makes the timing of the 2026 event unusual: its Germany-average peak was higher than on any May–September day of 1951–2025.

Comment 4: *I suggest providing a concise table summarizing the heatwave metrics used in the manuscript, including their definitions and units. This would make the methodology considerably easier to follow. In particular, I would appreciate further justification for the definition of cumulative intensity. Since heatwaves themselves are defined relative to the calendar-day P90 threshold, I wonder why cumulative intensity is calculated relative to the climatological mean rather than relative to the same P90 threshold.*

Response: A new Table will summarize all metrics with definitions, units and the figures in which they are used.

Comment 5: *In Figures 10 and 12, the dynamic component is represented using the mean T850 anomaly of the selected analogues. It would be very informative to also show, or at least quantify, the distribution of T850 anomalies among the individual analogues. The mean alone may conceal substantial variability among analogue events. Showing, for example, boxplots, violin plots, or the percentile range of the analogue temperatures would allow the reader to assess where the 2026 event lies within the historical analogue distribution. This would be particularly useful when comparing the early and late epochs.*
Such an analysis could provide a more direct answer to the question of whether the 2026 event lies completely outside the range of temperatures historically associated with comparable circulation patterns, rather than only showing its difference from their mean.

Response: Thank you for this suggestion. We will revise the aforementioned figures and add the required information/figure in the revised version of the manuscript.

Comment 6: *It is not entirely clear how the epoch-based composites in Figures 12–13 are constructed. Section 2.4 states that, for each event day, the 25 closest Z500 analogues are selected and that the analogue selection is subsequently restricted to the early (1970–2000) and late (2001–2025) periods. However, the Results refer to substantially different numbers of "qualifying analogues" in the two epochs (e.g., n = 38 and n = 87), while Figure 13 appears to show yet another pair of sample sizes. Please clarify.*
There also appears to be an inconsistency in Figure 13 between the sample sizes reported in the figure panels and those given in the caption/text. Please check and correct these values.

> **Response:** For each event day, the 25 closest analogues are selected from 1970–2025 and then split by year into the early (1970–2000) and late (2001–2025) epochs; the epoch composites pool these analogues over the five event days (125 analogues in total). The numbers therefore differ between the epochs because the analogues are selected by similarity, not by a fixed quota. The correct numbers are n = 37 and n = 88, which will be properly mentioned in the text, caption and figure. The comparison will be complemented by equal-size resampling, a permutation test and within-epoch analogue selection.

Minor comments

L 128–130. The expression *"percentile-based algorithm equivalent to the widely used framework..."* reads somewhat awkwardly. I suggest rewriting it.

L124 I would suggest using *"early-summer period"* rather than *"short summer season May–June"*.
Figure 1. It would be useful to include the locations of the 18 long-term stations listed in Table 2. A small elevation/topography inset or background could also facilitate the interpretation of the station distribution, particularly for the mountain stations discussed later in the manuscript.
Figure 4. Please clarify exactly how the P90 shown in this figure is obtained. Is it calculated from the regional-average series? This is not clear from the caption.
Figure 12 / L351–355. The manuscript states that *"the late-epoch analogue composite is systematically warmer than the early-epoch composite on all event days."* This is not immediately evident from Figure 12; for several days, the early-epoch values appear comparable to or even slightly higher than the late-epoch values. Please verify this statement and either provide the corresponding numerical values or revise the interpretation accordingly.

Figure 13. The figure indicates that both epochs are associated with a broadly similar blocking configuration. However, the Z500 anomaly patterns are not strictly identical and show some differences in position and amplitude. Therefore, the statement that the two circulation patterns are *"nearly indistinguishable"* may be somewhat stronger than what is directly supported by the figure.

> **Response:** All the minor comments will be addressed and implemented in the revised version of the manuscript.