ANONYMOUS REVIEWER #1:

GENERAL COMMENT

The manuscript explores the recently introduced approach to define temperature extremes as three-dimensional phenomena for the Mediterranean region (Rome, Italy) and analyses driving mechanisms of identified Extreme Warm Events (EWEs). In my opinion, the manuscript is well-suited for Weather and Climate Dynamics journal, as it extends current knowledge of vertical characteristics of temperature extremes from both statistical and dynamical perspective. The manuscript, however, contains several problematic parts, which should be clarified before consideration its publication.

We sincerely thank the Reviewer for her/his positive and encouraging general evaluation of our manuscript. We are grateful for the recognition of our work's relevance to *Weather and Climate Dynamics* and its contribution to advancing the understanding of temperature extremes from a three-dimensional, process-based perspective. We have carefully considered all the specific comments and concerns raised by the Reviewer, revisiting the manuscript accordingly.

Please find below a point-by-point reply to the Reviewer's comments. The complete, original Reviewer's comments are given in black, and our responses are listed in blue. The revised text is shown in italics and the Line number refers to the revised version of the manuscript.

In the revised version of the manuscript, modifications are marked in red. Please note that, in the revised manuscript, the original figures (Figure 1, Figure 2, Figure 3, Figure 5, Figure 6, Figure 7) have been updated and replaced with expanded figures in direct response to the Reviewer's requests.

MAJOR COMMENTS

Q1: Classification of Extreme Weather Events vs. Extreme Weather Regimes: Authors classified identified EWEs into 5 categories based on thresholds for 1000, 500, and 300 hPa levels. This is not an issue, we have used the analogous method, but the authors also introduce additional classification of Extreme Weather Regimes (EWRs) using K-means clustering. If I understand correctly, EWEs categories were used solely for Figure 2 and 3 (statistical perspective), while the remainder of the manuscript used EWRs. I am missing the link between these two approaches, for example which of the EWRs interprets Surface only EWE? Judging from Figure A2, this EWE falls into Low Troposphere, Standard Deep and partly to Upper Troposphere (!) clusters. This also points to a drawback a threshold-based methods. To sum up, in my opinion, classification of EWEs is redundant and potentially misleading, as EWE categories are not an input for EWR clustering. Authors may consider redrawing Figure 2 and 3 with EWR categories instead of EWE and omitting EWE classification from the manuscript. K-means would probably provide a more objective method to distinguish between vertical types of temperature extremes.

A1: We thank the Reviewer for this insightful observation and for giving us the opportunity to clarify the conceptual and methodological distinction between EWEs and EWRs. While both approaches investigate the vertical distribution of thermal anomalies, they provide two fundamentally different and complementary goals within our study:

- EWE classification provides an objective identification of extreme events based strictly on standard climatological percentiles of temperature series (>90th percentile for at least 4 consecutive days) at specific pressure levels (1000, 500, and 300 hPa). Therefore, Figures 2 and 3 evaluate the statistical occurrence, trends, and seasonal distribution of thermal extremes

at single- or multi-level thresholds. Omitting them would remove the direct comparability of our findings with traditional surface-bound HW definitions.
- EWR classification captures the full three-dimensional atmospheric dynamics and vertical coupling of the column. The K-means algorithm operates directly on the temperature anomaly profiles of the identified extreme days.

Moreover, the Reviewer correctly points out that a threshold-defined *Surface-only EWE* can occasionally map into different EWR clusters. In our approach, this is a key physical insight and not a drawback. In fact, a fixed percentile threshold at 1000 hPa merely tells us that the surface is extremely warm, but it cannot differentiate whether that surface heat is driven purely by local near-surface processes (*Low Troposphere EWR*) or represents the tail of a broader, developing synoptic-scale subsidence (*Standard Deep EWR*). The K-means clustering (used to identify EWRs) resolves this exact ambiguity by looking at the whole tropospheric profile.

To clarify these aspects, we have modified the Introduction as follows:

LL. 90-105: *"In this study, it is worth noticing that a clear distinction is provided among three related but conceptually distinct phenomena:*

- *"Heatwave" (HW): persistent periods of high near-surface air temperature exceeding a specific percentile threshold for a minimum multi-day period, consistent with standard surface climatological definitions (World Meteorological Organization and World Health Organization, 2015);*
- *"Extreme Warm Event" (EWE): episodes during which daily mean air temperature exceed the level-specific 90th percentile threshold at a given isobaric level (from the surface up to 300 hPa). This term encompasses both classical surface-based HWs and vertically limited warm anomalies that are not necessarily experienced also at the surface;*
- *"Extreme Warm Regime" (EWR): three-dimensional atmospheric configuration formed by the vertical co-occurrence of EWEs at multiple pressure levels. While an EWE identifies events at a given pressure level, an EWR characterizes the coherent vertical profile of the warming, distinguishing deep-column coupled extremes from surface-confined or upper-tropospheric isolated layers."*

Q2: Number of EWR categories: Judging form Figure 5, the "Standard deep" and "Low Troposphere" EWR categories are very similar. They have also quite similar vertical profile (Figure 4) and precursor thermal animacies (to a lesser extent; Figure 7). I understand that from Figure A1b that the five clusters seems to be the most reasonable option but have authors tried using four or six EWRs?

A2: We thank the Reviewer for this observation about cluster separation. In our analysis, to avoid subjective bias in selecting the number of EWRs (i.e., the number of clusters K), we relied strictly on two objective, data-driven validation metrics: (i) the Elbow Method (Inertia) that measures the total within-cluster sum of squares (variance), helping to identify the "elbow" point where adding more clusters yields diminishing returns in explaining total variance, and (ii) the Silhouette Analysis that evaluates how well-separated and cohesive clusters are by comparing the average distance (typically Euclidean distance) of a point to its assigned cluster against its average distance to the nearest neighbouring cluster. As shown in Figure A1, K=5 represents a clear peak in the Silhouette Analysis (Figure A1b) alongside a distinct bend in the Elbow curve (Figure A1a). This confirms that 5 clusters optimize internal cohesion and inter-cluster separation without overfitting.
In addition, preliminary tests were also performed using K=4 and K=6. Assuming K=4, the algorithm merges the Standard Deep and Low Troposphere categories into a single cluster, but this does not allow for capturing the distinct physical mechanism driving lower-tropospheric heat extremes. With

K=6, the algorithm splits one of the existing deep-column regimes into two sub-clusters with minimal differences in thermal amplitude, reducing cluster robustness and lowering the silhouette score.
We have expanded Sect. 2.5 to include a detailed explanation of the purpose of the Elbow and Silhouette methods, as well as a brief summary of the sensitivity tests conducted for K=4 and K=6, as follows:

LL. 287-299: *"To determine the optimal number of EWRs (i.e., the number of clusters K) objectively and avoid subjective selection bias, two complementary validation metrics were evaluated: the Elbow Method and Silhouette Analysis (see Fig. A1). The Elbow Method assesses the within-cluster sum of squares (inertia) as a function of K. The optimal number of clusters corresponds to the point where the rate of decrease sharply changes (the "elbow"), indicating that adding further clusters yields diminishing returns in explaining total variance. The Silhouette Analysis measures cluster cohesion and separation by comparing the mean distance of an observation to other points within its own cluster against the mean distance to points in the nearest neighbouring cluster. As illustrated in Fig. A1b, K=5 corresponds to a local maximum in the Silhouette score, coinciding with a clear inflection in the inertia curve (Fig. A1a). Sensitivity tests using K=4 and K=6 were also evaluated. Selecting K=4 the Standard Deep and Low Troposphere regimes were merged, masking the distinct surface-bound, bottom-up thermodynamics characteristic of lower-tropospheric events. Conversely, K=6 splits vertically coupled clusters without revealing new physical processes, resulting in a lower overall silhouette score. Therefore, K=5 provides an optimal balance between statistical separation and physical interpretability."*

Q3: Potential seasonal differences: Authors identify EWEs thorough the year (Figure 3), which is fine. But although they present temperature trends and EWEs statistical characteristics in different seasons, EWRs were determined from all events regardless a season. I am not an expert on the Mediterranean climate, but I can imagine that the stormier, moister and colder part of the year may affect development of EWE differently than the hot and dry part of the year. I am wondering if the number of EWR categories and their driving mechanisms would be different if EWEs would be partitioned between hot and cool part of the year.

A3: We thank the Reviewer for raising this point regarding potential seasonal dependencies in the identification, vertical structure, and physical mechanisms of EWRs. We fully agree that climatological conditions shifts between summer, winter, and the transition seasons, theoretically altering the optimal number of clusters. To address this comment and also a similar one from another Reviewer, we have conducted a comprehensive seasonal analysis by partitioning all identified EWE days into the four standard meteorological seasons: Winter (DJF), Spring (MAM), Summer (JJA), and Autumn (SON). First of all, following your suggestion, we independently evaluated the optimal number of clusters (K) for each seasonal subset using the Elbow Method and Silhouette Analysis. The dual-metric evaluation yielded an optimal configuration of K = 5 for Spring, Autumn, and Winter, perfectly matching our all-year baseline. For Summer, the metrics indicated a primary local optimum at K = 4. However, to maintain cross-temporal consistency and allow direct physical comparison across all seasons without masking shallow lower-tropospheric processes, K = 5 was retained for all seasonal clusterings. This analysis confirms that the five identified EWRs represent robust, recurring atmospheric modes rather than statistical artifacts of seasonal aggregation.

We updated the Methods section to describe the seasonal partitioning methodology, the independent validation framework for each season, and the rationale for retaining K = 5 for cross-seasonal comparability. In the Results, we added a dedicated paragraph detailing the seasonal dynamics, vertical structures, and spatial stability of both coupled and decoupled EWRs across the four seasons. Finally, in the Appendix A we included four new seasonal composite figures (Figs. A3-A6) depicting Z_{500} and ΔT_{850} for Spring (MAM), Summer (JJA), Winter (DJF), and Autumn (SON).

LL. 300-306: *"Moreover, to evaluate potential seasonal dependencies in the vertical structure and physical mechanisms of EWEs, a seasonal analysis was also performed. Identified EWE days were partitioned into the four standard meteorological seasons: Winter (DJF), Spring (MAM), Summer (JJA), and Autumn (SON). The dual-metric validation framework (Elbow Method and Silhouette Analysis) was applied independently to each seasonal subset. The optimal cluster count yielded K=5 for Spring, Autumn, and Winter, whereas Summer exhibited an optimal metric configuration at K=4. To ensure methodological consistency across all temporal domains, facilitate direct comparison between seasonal regimes and the all-year baseline, and avoid masking distinct lower-tropospheric processes, K=5 was retained for all seasonal clustering analyses."*

LL. 542-561: *"To further assess whether the driving mechanisms and spatial characteristics of EWRs depend on the seasonal cycle, independent seasonal composite maps were constructed for the four meteorological seasons (Figs. A3-A6). This seasonal analysis reveals that the core synoptic configurations and vertical decoupling mechanisms identified in the annual baseline remain physically robust, confirming that the identified EWRs represent recurring atmospheric patterns rather than artifacts of seasonal aggregation. In particular:*

- *Coupled Regimes (Extreme Deep and Standard Deep EWRs): across all seasons, both coupled regimes are consistently governed by high-amplitude upper-level ridges driving strong lower-tropospheric warm advection ($\Delta T_{850} > 4\,°C$) and subsidence heating. During winter (Figs. A6a and A6b) and spring (Figs. A3a and A3b), these regimes manifest as prominent blocking patterns or high-amplitude subtropical ridges steering tropical air masses towards Central Europe. In summer (Figs. A4a and A4b), geopotential gradients are broader and less intense, but positive thermal anomalies remain strong and geographically focused over the Central Mediterranean.*
- *Decoupled Regimes (Mid Troposphere, Low Troposphere, and Upper Troposphere): the intensity and characteristics of vertical decoupling exhibit distinct seasonal signatures. During summer and autumn, the Mid Troposphere EWR (Figs. A4c and A5c) shows marked thermal decoupling, where elevated warm anomalies coincide with near-neutral or slightly negative lower-tropospheric anomalies over the Mediterranean basin, linked to transient troughs and baroclinic activity. Conversely, the Low Troposphere EWR (Figs. A4d and A5d) consistently retains strong surface-bound warming ($\Delta T_{850} > 2$ °C) under relatively zonal or weak upper-level flow, emphasizing boundary-layer thermodynamic processes irrespective of season. Finally, the Upper Troposphere regime (Figs. A3a-A6e) is persistently associated with high-tropospheric ridge axes and tropopause displacements, reaching its strongest spatial coherence during summer and autumn (Figs. A4e and A5e)."*

MINOR COMMENTS

Q4: Figure 1: In the "Detection of extreme warm events" panel, 90th percentile threshold, the red broken percentile line is drawn as horizontal. I believe that a proper representation of the threshold should be curved, as the threshold follows the annual temperature cycle.

A4: We thank the Reviewer for this observation. We agree that representing the 90th-percentile threshold as a horizontal line may incorrectly suggest a constant threshold throughout the year. We have modified Figure 1 by representing the 90th-percentile threshold (THR_p) as a daily-varying curve that follows the annual temperature cycle, while retaining the daily mean temperature series for clarity.

Q5: Lines 179–180: “This moving-window approach reduces the influence of short-lived weather systems while providing a smooth representation of the seasonal cycle.” This sentence is confusing for me. If I understand the method correctly, the moving window was used only for a calculation of the thresholds. The short-lived EWEs are excluded due to a persistence criterion (four days; Line 185). Consider reformulating the sentence.

A5: We agree that the original sentence was potentially confusing. In our methodology, the 21-day moving window is used to calculate the daily climatological mean and the corresponding 90th-percentile temperature threshold, thereby accounting for the seasonal cycle. The exclusion of short-lived events is instead achieved through the persistence criterion of at least four consecutive days. We have therefore reformulated the sentence as follows:

LL. 236-237: *“This moving-window approach is applied to calculate the daily climatological mean and the corresponding 90th-percentile temperature threshold, allowing to capture the seasonal variability of temperature.”*

Q6: Table 1: I would omit the subjective meteorological interpretation, also because authors did objective analysis of EWRs below.

A6: As suggested, the subjective meteorological interpretation has been removed from Table 1.

Q7: Figure 2 and 3: Please enlarge plot axis titles and texts. Also legends are hard to read, consider using a larger font.

A7: As suggested, Figures 2 and 3 have been updated using larger fonts for axis titles, labels, and legend.

Q8: Figure 6: In my opinion, there is not much added value in this figure. The upper panel (EWR frequency) could be integrated into revised Figure 2 see my first comment and the remaining panels could be omitted as the changes between time periods are relatively small. Also how interpretable are upper tropospheric positive temperature anomalies in 1996–2000 vs negative upper tropospheric temperature anomalies in 2001–2005?

A8: We thank the Reviewer for this comment. However, we believe it would be preferable to retain the lower panels of Figure 6, as they carry one of the central physical conclusions of our manuscript. Furthermore, we clarify the interpretation of the upper-tropospheric regime anomalies. The 4-panel layout reveals an important dynamic: while the average duration (EWR duration, ~1.5–3.5 days) and the column-average intensity (average temperature anomaly, e.g., ~4.5 °C for Extreme Deep) of individual EWRs remain remarkably stable across all 5-year sub-periods from 1991 to 2025, the frequency (EWR frequency) exhibits a marked increase, especially post-2015 for Deep EWRs. This stability shows that the surge in overall CHI over Rome is driven by the increased recurrence rate of these events rather than an intrinsic lengthening or intensification of single episodes. Removing the duration, anomaly, and CHI panels would lose this critical proof and hide the distinction between frequency trends and intensity stability.

Also, we would like to clarify that the negative bar in 2001-2005 (and 2006-2010) refers specifically to the Upper Troposphere EWR (cyan bar) and represents the column-averaged temperature anomaly across 1000-300 hPa. As shown in the top panel (EWR frequency), the Upper Troposphere regime occurs very rarely. In 2001-2005 (i.e., in a period dominated by intense lower-tropospheric HWs, such as Summer 2003), the few isolated days classified under this Upper Troposphere regime had warm air aloft accompanied by below-average temperatures in the lower and mid-troposphere. When averaged vertically across the entire 1000-300 hPa column, the resulting $\overline{\Delta T}$ yields a slightly negative value (around -0.5 °C). On the contrary, in 1996-2000, the few Upper Troposphere days coincided with a warmer mid-tropospheric condition, resulting in a slightly positive column average (around +1.6 °C).

This aspect has been clarified in the revised manuscript as follows:

LL. 585-589: *"Lower anomaly intensities characterize the Low Troposphere EWR, whereas the High Troposphere EWR displays greater temporal variability, including slightly negative column-averaged anomalies during 2001-2005 and 2006-2010. This behaviour is primarily driven by its very low sample size and its physical structure: since warm anomalies in this regime are strictly confined to 300 hPa, neutral or slightly cooler-than-average conditions in the lower and mid-troposphere can result in a marginally negative value when averaged across the entire vertical column."*

Q9: Figure 7: Very hard to read, please enlarge labels and a legend.

A9: As suggested, Figure 7 has been updated using larger fonts for axis titles, labels, and legend.

Q10: Table 4: It reads "19/01/1997-22/01/1997", while Figure 8a starts with 19/01/07, probably a typo.

A10: We thank the Reviewer for pointing this out. We confirm that this was indeed a typo in Table 4 and in the text of the manuscript. The event started on 19/01/2007 and finished on 22/01/2007. Table 4 and the main text have been modified accordingly .

Q11: Figure 8: Duration of x-axis varies between individual cases, which may mislead the reader about the case persistence. Consider aligning the x-axis between cases.

A11: We are fully aware that the duration of the selected events varies across the different categories (ranging from 4 days for the Extreme Deep and Standard Deep EWRs up to 10 days for the Mid Troposphere EWR). However, this specific time scale for each panel was intentionally chosen to display the full evolution of the tropospheric conditions throughout the entire duration of each individual event. To avoid any potential misunderstanding, we have clarified this rationale in the revised caption of Fig. 8:

LL. 741-743, caption Fig. 8: *"Note that in each panel the onset time of the event (t_0) corresponds to the start of the time axis in each panel and that the duration displayed on the x-axis corresponds to the full lifecycle of each specific event, ranging from 4 days (panels a, b, c, and d) up to 10 days (panels e and f)."*

Q12: References: Line 268 (Draxel and Rolph, 2010), the name should be Draxler (typo) and the year (2010) do not match the year in the references (2026); Line 571 (Lhotka et al., 2018) is not in the reference list.

A12: We thank the Reviewer for identifying these typos and missing references. We have corrected both instances in the revised manuscript.