Review of the manuscript "**A global description of the three-dimensional structure of temperature anomalies during heat extremes**" written by Lau et al.
(EGUSphere-2026-5697)

Recommendation: Major revision

Summary
This manuscript presents a global, systematic description of the three-dimensional structure of temperature anomalies during surface heat extremes, using 58 ERA5 model levels over the period 1951–2023 (73 years). In our opinion, it is the type of study with real potential to systematically advance our understanding of the three-dimensional structure of temperature extremes. The descriptive approach based on contours above a local 95th percentile, vertical stacking, and ellipse geometry is innovative and well reproducible. The topic is timely and falls within the scope of the journal.

On the other hand, part of the main conclusions carries uncertainties that are insufficiently discussed or not discussed at all. Three fundamental interpretation issues (conditional statistics, the selection design of the sample, and the effect of level binning over topography) is not discussed. I therefore recommend major revision, with the conviction that none of the main conclusions needs to be retracted, as the requested changes concern interpretation and statistical support, not the essence of the approach.

Comments relevant to all four conclusions

1) Sample design and independence: the TXx criterion and temporal replication

The TXx criterion conditions the sample on annual maxima of Tmax ("winner-take-all"). A strong heatwave in a year with an even stronger heatwave tend to be removed from the sample, while a considerably weaker heatwave in a year without "competition" passes. The intensity distribution in the sample therefore reflects "rank within the year" rather than physical intensity, with consequences for all correlations and conclusions. In addition, multi-day events contribute to several autocorrelated stacks. We suggest reporting the number of independent events, and possibly also performing sensitivity tests concerning (i) different TXx criteria (e.g., the annual 99th percentile of Tmax), and (ii) using one value per event (an average for all heat wave days). We note that TXx is not even listed among the "four main parameters" in Sect. 4.1.

2) Conditional statistics (survivorship bias) and a broken cross-reference

Intensity and shape metrics are computed only at levels where a cell exists. According to Fig. A1, only approx. 40–60% of stacks are available at 500 hPa and 30–50% at 350 hPa.

Therefore, mid- or upper-tropospheric results (correlations, shape, orientation) are based on a systematically selected minority and the authors themselves acknowledge that this selection favours more intense events (Sect. 2.2, 3.3). Results 3.1 refer to a discussion of this bias in Sect. 4.1 ("As discussed in Sect. 4.1..."), but that discussion is missing from the manuscript. This discussion should be delivered, including its connection to the interpretation of all upper-level results, and a demonstration of the robustness of the key conclusions with respect to the treatment of missing cells.

3) Quality of upper-air data in the Southern Hemisphere before the satellite era

The sample starts in 1951, when the radiosonde network in the Southern Hemisphere is very sparse and satellite soundings are practically non-existent. Southern Hemispheric upper levels in the first third of the sample are therefore largely a model construct. We suggest a robustness test: compare the key Southern Hemispheric results for the pre- and post-1979 periods or discuss the impact on the detection of heat cells higher in the troposphere.

<u>Points tied to the individual conclusions</u>

Conclusion 1:

- The claim that "tropical cells are larger, especially over the ocean" is undermined by the authors' own admission in Sect. 4.1: the largest tropical cells over the central Pacific are ENSO structures, i.e., a phenomenon the authors themselves label as a limitation of the method.
- Tibet/Andes artifact: the claimed large "mid-tropospheric" cells over high plateaus (Fig. 3c) are most likely an artifact of the fixed-level binning. Over Tibet (surface ≈ 575 hPa), the "mid" bin contains levels just above the surface, so these are near-surface cells labelled as mid-troposphere. The Andes (a narrow mountain range with lowlands alongside) probably provide a cleaner test case – the contrast between large cells over the ridge and small cells in the lee-side lowlands. We suggest a control computation with bins defined with respect to the actual surface pressure for at least one mountain region, or a withdrawal of the interpretation in Sect. 3.1.

Conclusion 2:

- Our main concern with this conclusion is that it rests entirely on the ellipse fit itself, and the reliability of that fit is never tested independently. First, the fit can make cells look more stretched than they really are. The ellipse is fitted to contour vertices, which carry noise from the 0.5° grid. This noise probably tends to elongate the fitted ellipse, so a cell that is in reality close to circular can easily be fitted with

q ≈ 0.6. The authors defend the conclusion by noting that almost no cells have fitted q between 0.9 and 1. However, if the fit inflates elongation, the absence of high fitted values cannot prove that the true shapes are stretched. Second, the near-identical value of q everywhere (0.5–0.6 in the tropics and extratropics, over land and ocean, at all levels) is itself surprising or even suspicious. Different physical mechanisms should produce differently shaped anomalies, so such a uniform number may simply reflect how threshold contours of any smooth field look in this algorithm, rather than the physics of heat extremes. We suggest adding to Sect. 2.3 or 4.2 an explicit statement that the fitted axis ratio can be biased by noise in the contour vertices and softening phases such as „remarkably homogeneous/uniform"

Conclusion 4:

- In the abstract and Conclusion 4, the correlation between heat cell area and intensity in the mid-troposphere is described as "robust". Given that (i) the correlation at 500 hPa is computed only on the subset of stacks in which a cell exists at that level (Fig. A1), a subset that the authors themselves state is biased towards more intense events (Sect. 2.2, 3.3), and (ii) the sample is temporally autocorrelated (multi-day events contribute multiple similar stacks), we ask that "robust" be replaced by a wording that reflects this conditionality.

<u>On the argumentation against the "heat dome":</u>

The argument has three parts but two of which are in our opinion rather weak: (i) "the shapes are not circular" – in the literature the heat dome is primarily a vertical concept (trapping of warm air under a high-pressure system) and the authors' critique also rests on a shape quantification burdened by fit uncertainties. (ii) refusing the term with the tropical case, when the authors themselves state that heat domes are the mid-latitude concept. On the other hand, the authors objection to the implied uniformity of mechanisms of heat domes is legitimate and supported. We recommend refining the conclusion from "advocate against the use of this image and terminology" to a more differentiated formulation (e.g. the dome as a useful simplification of the vertical structure in the mid-latitudes but it holds neither horizontally nor in the tropics).