Response to reviewer comments

Comment 1.1 *I find the premises and motivation for the paper unconvincing. The authors argue that one of the most interesting aspects of temperature extremes is their persistence, as this is closely linked to their impacts and thus needs to be studied more (ll. 13-27). They further identify two key gaps in the literature: (i) the focus has been on short-lived temperature extremes but it is necessary to look at temperature extremes over longer periods (ll. 28 and following); and (ii) in the past, work has been conducted mostly on arbitrary regions (ll. 42 and following). I understand these to be the main limitations this study seeks to address. Concerning point (i), what I find to be missing is any evidence that looking at 3-week temperature extremes as opposed to e.g. 5-day temperature extremes actually provides a better picture of the impacts (or even of the meteorological drivers) of the temperature extremes. The 5-day or similar threshold generally captures past high-impact events and allows to identify coherent sets of meteorological drivers, and I assume that these are amongst the reasons why it has enjoyed such popularity in the literature. Can the authors actually identify a set of high-impact events that is overlooked by a 5-day minimum persistence criterion but captured by their own definition? Similarly, can they make a case for the fact that using a 5-day minimum persistence criterion confounds the meteorological drivers of longer-lasting events? After all, the main conclusions on the dynamical drivers of the temperature extremes that the authors find in many cases seem to support previous findings in the literature. A further point that I detail further in one of my other comments is that a lower 5-day threshold on heatwaves does not prevent including much longer-lived events in the analysis. Additionally, in many cases, impacts are related to duration of temperature extremes in a non-intuitive fashion (e.g. Xu et al., 2016), which again seems to go against the argument of the authors for looking at 3-week periods.*

Answer: Thank you for this important input. We realise that the motivation needs a clarification and we will rework our introduction and motivation for looking at persistent temperature extremes. We choose to define warm and cold spells to look at longer time windows than is classically done in the literature for the following reasons:

1. From an observational perspective, high or low temperature conditions sometimes persist for weeks (e.g., the October 2022 and the mid-December 2022 to mid-January 2023 warm spells in Western Europe).

2. Short windows windows tend to focus on the period of most extreme temperature within warm and cold spells. Admittedly, a 3-week warm spell would likely be detected with a shorter (say, 5-day) window, but we want to look at such events in their entirety, specifically their build-up and persistence over periods of potentially several weeks. This allows us to highlight some mechanisms that are maybe less obvious for short events, like recurrent Rossby waves.

3. While most 3-week extremes do include short periods of very extreme temperatures, only about half of 5-day extreme events occur within 3-week warm/cold spells (see Figure R1). The two approaches are thus not exactly interchangeable and with a longer window we are more confident that we capture really persistent events.

4. 5-day windows make the regionalisation more challenging. There is mechanically less synchronicity between extreme events in different locations over 5-day periods than over

3-week periods. This leads to a much more complex regionalisation (many more regions) while we want to reduce the dimension of the problem to provide a simple, physically-meaningful regionalisation.

We should also note that the fact that our results generally agree with the existing literature isn't a sufficient reason to discard them. We do not argue that we discovered new mechanisms (on that point we should certainly reformulate certain passages of the manuscript that suggest the contrary), but that we provide a comprehensive, hemispheric-wide perspective on the distribution and drivers of persistent temperature extremes. The regionalisation is an important output of our study (something which has previously not been analysed in this extent in the literature, to our knowledge).

Finally, there is evidence that choosing a longer time window to define events does provide a better representation of impacts. It is true that when it comes to human health, most studies focused on short periods (roughly 3-7 days, cf. Xu et al., 2016). But the impact of warm and cold spells on the energy sector and vegetation, for instance, clearly scales with their duration in a non-linear way (see e.g., Añel et al. 2017; doi:10.3390/atmos8110209, or the many studies looking at the impacts of long summer heatwaves on vegetation). There is hence interest in S2S prediction of warm and cold spells on these time-scales (see e.g., Van Straten et al. 2022 https://doi.org/10.1175/MWR-D-21-0201.1 who focus on four week warm periods).
Regarding meteorological drivers, it is less clear that selecting a 3-week time scale is necessarily more relevant than selecting a 5-day time scale – after all, our results are in agreement with the literature (the opposite would be surprising). Still, the longer time window provides for more robustness in the regionalisation (compared to a 5-day window, see point 4 above) and thus helps associate more robust drivers to specific regions.

Comment 1.2 *Concerning point (ii), there are several papers that have proposed regional partitions of temperature extremes based on somewhat objective meteorological criteria, some of which are cited by the authors later in the paper (e.g. Stefanon et al., 2012 and others in Sect. 6.2). Moreover, there are several studies that have chosen specific regions motivated by non-meteorological but perfectly sensible criteria, such as taking an impacts perspective (e.g. Lowe et al., 2015), maximizing data availability (e.g. Hirschi et al., 2011) or favouring ease of comparison with previous work (notably SREX regions, e.g. Perkins-Kirkpatrick and Lewis, 2020). The authors do mention that arbitrary regions may make sense from an impact perspective; I would argue that defining the regions from an impact perspective makes them distinctly non-arbitrary. I do see a value in defining regions on a hemispheric rather than continental level, as the authors do here, but I find the statement on l. 45 to be a gross misrepresentation of the literature.*

Answer: We agree that "arbitrary" is the wrong word here – we used it from a physical driver perspective (i.e., regions chosen based on impacts instead of based on the coherence of physical drivers). In a revised manuscript, we should reformulate by saying that previous studies have mainly looked at regions based on impacts or observed extremes, while we are interested in a purely meteorologically-driven regionalisation. We will make sure to include additional papers when discussing our motivation such as the ones you suggest.

Comment 1.3 *In Sect. 2.2.2 the authors define their analysis window. I understand the logic*

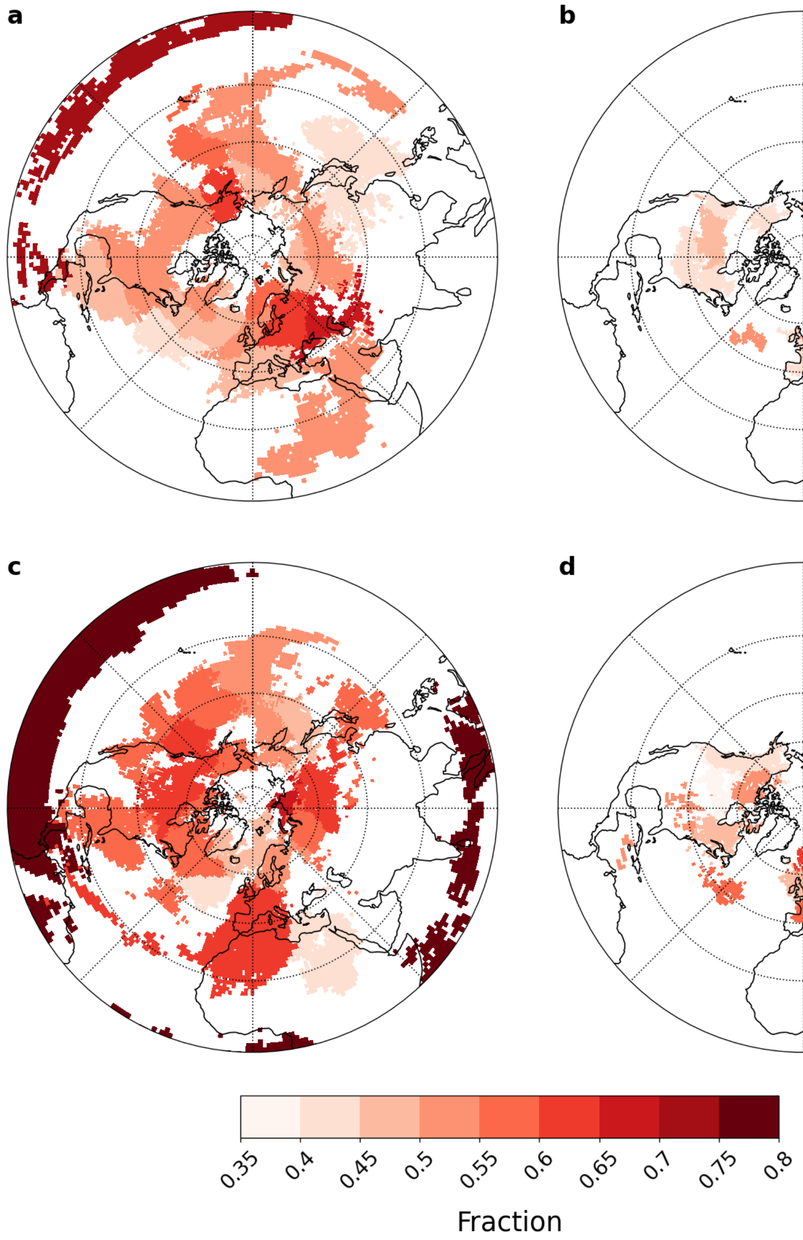

Figure R1: Fraction of 5-day warm and cold spells occurring within 21-day warm or cold spells, for (a) DJF cold, (b) JJA cold, (c) DJF warm and (d) JJA warm spells. In each case spells are defined relative to the 95th or 5th percentiles of the corresponding 5- or 21-day temperature distribution.

of wanting to study the onset phase until the peak of the warm/cold spells, but defining the first day as the one where the regional temperature anomaly changes sign may introduce an unrealistically long build-up phase for some events. For example, I am not sure that I buy the argument of a wintertime warm spell taking on average more than 2.5 weeks to build up, as suggested by table 1.

I also think this makes the reasoning on ll. 91-92 somewhat circular. The authors define a *very generous window to analyse the build-up of temperature anomalies, and then argue for the importance of analyzing temperature deviations over a long time period based on their own definition of the build-up period.*

I would be more convinced if some analysis could show the typical evolution of regional temperatures from the first day of the analysis window and make the case that this reflects a consistent, continuous build-up as opposed to e.g. a period of oscillations around weak anomalies of the same sign and then a rapid increase of absolute anomalies closer to the peak date. Presumably, both from a physical and a more impacts-based perspective, the interest lies in the build-up and persistence of somewhat large temperature anomalies, not in small anomalies, even if persistently of the same sign.

Answer: The boundaries of warm and cold spells could be defined in several ways. Here, we followed Röthlisberger and Papritz (`https://doi.org/10.1038/s41561-023-01126-1`) in defining the beginning of a spell as the last time before the spell when the daily temperature anomaly remained of the same sign (positive for warm spells and negative for cold spells). While for single events, this may not always be the most relevant choice, it certainly makes sense from a statistical perspective: on average, we want to capture the period during which temperatures consistently depart from their climatological average. Again, there are variations across events, but the goal is to capture their average behaviour. We show on Figure R2 the example of four different regions in Western Eurasia, which illustrate the fact that temperature anomalies, on average, take several weeks to build up, persistently deviating from their climatological mean. Note that we do not argue that the 21-day window is necessarily relevant for all regions or all events. We only take a simplified perspective in order to draw robust inferences across regions and events.

One important reason why we define spell boundaries the way we do is that we are looking at S2S timescales, and we are specifically interested in the drivers behind the onset and build-up of the extreme events, not just their extreme part. This is why it makes sense to go back to the last time when the daily temperature anomaly changed sign.

As to whether the reasoning is circular, our initial formulation was clearly misleading. We do not argue that finding analysis windows of about 3 weeks makes the case for looking at 3-week temperature anomalies (since the analysis window is constrained by the length of the event we consider). Instead, the analysis window lengths we find are consistent with mechanisms developing over S2S timescales, rather than very short-term drivers that would suddenly swing temperatures towards extreme anomalies. In that sense, the reasoning is not circular, and we will make sure to clarify it in a revised version.

Comment 1.4 *A separate issue I have with the methodology is that the authors weigh differently each event, and could be giving a disproportionately large weight to events with a long build-up time, regardless of whether these are particularly extreme events or not. What is the range of the analysis windows for individual events and is there a correlation between event severity and duration of the analysis windows within the single regions?*

Answer: We are not sure what you mean by "weigh differently each event". Do you mean to say that longer events would get more weight in the anomaly calculations? There is certainly a large variability in event duration for a single region. The coefficient of variation (standard deviation divided by mean) of event durations is around 0.5 on average. Longer events, however, do not get more weight in the anomaly calculations since anomaly maps are calculated for each event separately before being averaged (this will be made clear in a revised version). And again,

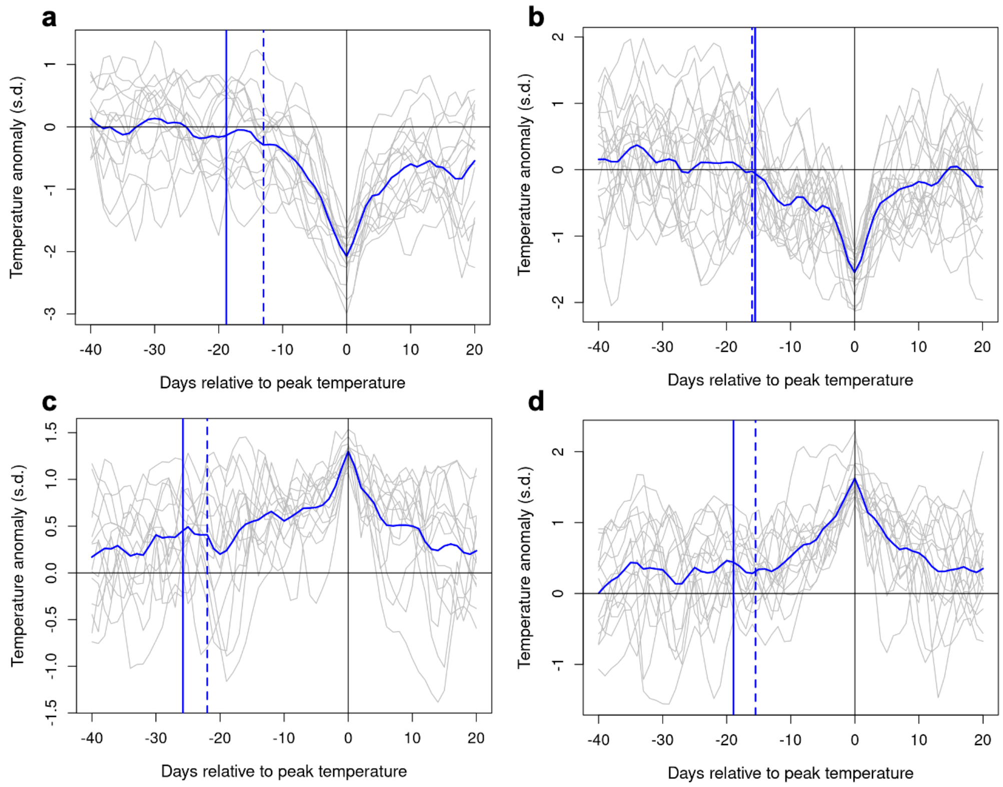

Figure R2: Time series of region-average daily normalised temperature anomalies (unitless) around 3-week warm and cold spells, centred on the day of peak temperature anomaly, for (a) DJF cold spells in region 19 (Northeastern Europe), (b) JJA cold spells in region 6 (Western Russia), (c) DJF warm spells in region 16 (Scandinavia) and (d) JJA warm spells in region 15 (Southwestern Europe). Gray lines represent individual events, and the thick blue lines are the mean. See manuscript Figure 3 for the correspondence between region numbers and their locations.

the point is not to look at specific events, but to highlight the robust signals across events. Finally, we do not find a systematic link between the length of the analysis window and the event severity (measured by either the peak event anomaly or the 3-week event average temperature anomaly).

Comment 1.5 *Further, I am confused by the description of the data preprocessing. On l. 65 the authors mention normalization of daily temperature using the mean and standard deviation. I assume that the former part entails subtracting the mean. On ll. 137 and following they mention subtracting the long-term average. Is the procedure such that the authors first normalize the data, including subtracting a mean value, then average it regionally and then subtract again a mean value?*

Answer: Thank your for pointing this out we will clarify this point in a revised version. We do normalise daily temperature series (l. 65) by removing the mean and dividing by the standard deviation, both mean and standard deviation being calculated on 30-day, 7-year moving windows (to remove both seasonality effects and long-term trends). However, this normalisation is only used for the modeling part. In the event anomaly maps (l. 137), we proceed differently: we calculate event anomalies by removing, for each event, the long-term average calculated for the same period of the year as the observed event. For instance, if we observe a warm spell during the period of September 1-15, 2000, the anomaly maps in a given variable for this events are calculated by removing the long-term average of the variable for all September 1-15 periods (1979-2020). You might argue that for temperature or Z500 this does not take long-term trends into account. That's certainly true, however we did not find very significant differences if we removed trends from temperature and Z500 series beforehand when calculating event anomalies. We will make sure to make this more explicit in the revised version (also related to your next comment).

Comment 1.6 *Finally, as far as I can tell the way statistical significance is computed is never explained in detail. For example, in Fig. 5 and following how is the 90% confidence level computed, and do the authors account for multiple testing?*

Answer: Thank you for pointing this out. We mention at l. 140 that we assess statistical significance based on a bootstrap analysis, but we certainly need to give more details about the procedure. For each region, season and type of spell, we randomly generate 1000 sets of events with the same distribution, in terms of duration and timing (i.e., starting day of the year) as observed events. We then calculate anomaly fields for these random sets of events, and determine from these empirical p-values for the anomaly fields associated with the actual events. All the significance maps include a correction for multiple testing (we apply the false discovery rate correction of Wilks (2016).

Comment 1.7 *While some of the regions defined in the paper are relatively intuitive, others look puzzling to say the least. For example, is it really the case that cold spells in the middle-east are part of a coherent region with cold spells in the south-western Sahel (region 2 (or 3, I can't really tell the colours apart) in Fig. 2a)? Similarly, Fig. 2d seems to suggest that heatwaves in Eastern Europe/Western Russia actually belong to three (or even four) separate clusters – something that I do not recall ever having seen in the literature. I am in general not against introducing new definitions to the literature; indeed, it is part of scientific progress. However, when new definitions are presented – in this case of temperature extreme regions – which appear at odds with the "conventional" ones from the previous literature, some more robust justification and contextualization would appear necessary. I find Sect. 6.2 somewhat dismissive, by providing references to three previous regional definitions of temperature anomalies but not going any further in this direction.*

Answer: Our regionalisation results are of course dependent on the number of regions we select, and the results should be interpreted as showing regions where warm or cold spells are related to similar large-scale circulation patterns. Our analysis focuses on the extratropics, not the tropics, where regionalisation results (region 2 in Figure 3a which you refer to in your comment) should be taken with a grain of salt.

Our results are instead meant to highlight spatial connectivity in a systematic way (specific events may behave differently) to best analyse the physical drivers. In practice, we find that neighbouring regions generally experience warm/cold spells at different times. There is certainly some overlap (regions 3 and 8 in Figure 4-d, for instance, share about 40% of their warm spells,

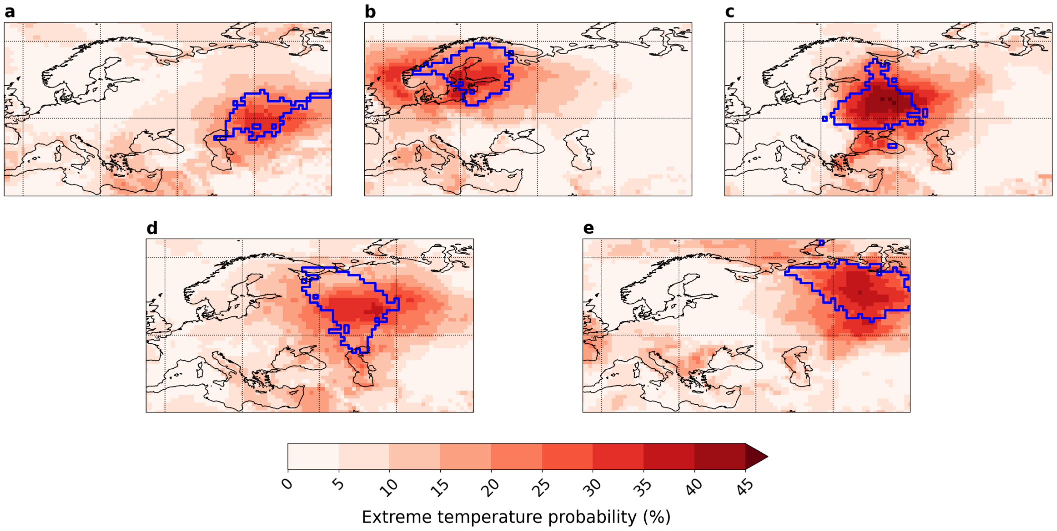

Figure R3: Average probability of exceeding the 95th percentile of 3-week mean temperature anomalies during JJA warm spells in 5 different regions (shown by the blue contours).

which is the highest percentage between any pair of regions in either season), but that is to be expected (see Figure R3 for the example of Europe during summer). In most cases, events in neighbouring regions overlap by less than 20%. The point of the regionalisation is to group together similar locations, but as with any clustering methods, categorising points at the edges and determining the optimal number of clusters are difficult things to do.

Second, the point of our analysis is to provide a different, data-driven and physically interpretable perspective on warm and cold spells, and not to stick to "conventional" regions (though our results are consistent with other regional-scale analyses). The way that we identify regions is also rather different from what is usually done, as we explain in the introduction. However, we should spend more time discussing the consistency of our results with previous analyses. We need to search for more studies, and would be grateful if you have any specific ones in mind to point out to us.

Comment 1.8 *A separate comment on the regionalization results is the conflation of land and sea regions. I assume that the composites shown in Figs. 5 and following are centred on the centroid of each region. The composites then show an odd combination of averages over land and sea regions. Variables such as cyclone frequency may be largely governed by specific regions spanning the storm tracks, and much less relevant for others. I suspect that more coherent results may be obtained if, especially for surface and near-surface variables, the authors tried to composite separately land and sea points.*

Answer: You are right that in Figures 5, 6, 7 and 9, we do mix land and ocean points. However, (i) the whole point of the analysis in this paper is to highlight similarities and differences in warm/cold spell dynamics across space. We purposefully do not discuss the behaviour of specific regions. This would require an additional paper. While the differences you mention (e.g., relative location to storm tracks) are certaintly important locally, this is not what we focus on here. (ii) Compositing land and ocean points separately is not practicable (because the location of e.g.

ocean grid points relative to the rest of the region is not consistent across regions). (iii) While the magnitude of anomalies in e.g. cyclone frequency may be very different between ocean and land regions, our analysis would nevertheless highlight areas where these are of the same sign (which is what we are interested in, after all. (iv) For cold spells the difference between land and ocean is not especially relevant, because cold spells are advection-driven. Differences are then instead to be found in the magnitude of the anomalies (temperature, advection, etc.) but not in their pattern. Regarding warm spells in winter, this could admittedly be a limitation, which we should discuss in a revised version.

Comment 1.9 *Sect. 5 is an odd section, which seems to be midway between a results and a discussion section. Much of the section is speculative and imprecise, and generally seems to make strong statements without a clear support from the analysis presented in the study. An example is the subsection on Recurrent Rossby Wave Packets. The authors dedicate a subsection to this without ever explaining clearly whether they define in some objective way RRWPs. It is unclear to me how the reference they make to Fig. 12 would support their claims. The subsection is largely based on speculation and analogies with previous results in the literature, rather than an objective analysis. Similarly, to support the statement made on ll. 349-350, one would need to at least show precipitation anomalies in Fig. A6.*

Answer: In section 5, we do not claim to quantify the role of all the spell drivers we identified. This may be why you find the section to be "imprecise". A complete and quantitative analysis would go beyond the scope of our paper (which centers on the regionalisation and the common features between warm and cold spells across regions), and will be the subject of future research. That being said, we should certainly expand some of the points in this section. As noted by the other reviewer, we clearly need to clarify the definition of recurrent Rossby wave packets that we use, to better support the results shown on Figure 12. We meant to include the precipitation deficit figure, but forgot to put it next to the sensible heat flux anomaly map (Fig. A6). This is why a figure S5 was incorrectly referenced at l.348 of the manuscript. We include it here in this response (see Figure R4).

Comment 1.10 *Finally, in the author's intention this section should discuss persistence of temperature extremes, but no definition or quantification of persistence is ever presented. The only place where the authors do define a timescale is in choosing the averaging period for the temperature anomalies and computing the "analysis window length". However, neither the persistence of the warm/cold spells (which may or may not be similar to the analysis window length if one is concerned here with the persistence of somewhat large anomalies) nor the persistence of the atmospheric circulation features associated with these is ever explicitly addressed. I struggle to see how one may make statements on the circulation features explaining the persistence of a given surface feature when persistence is never defined, nor is the duration of either feature quantified.*

Answer: We agree that we need to provide more details on our definition of persistence. Here, we take a fixed timescale to define "persistent" warm and cold spells. The choice of timescale, as we explained above, is motivated (i) by impacts considerations and (ii) by observed warm and cold spells during which high or low temperatures indeed persisted for several weeks in a row. What we could show however to make a better case about the perspectice we take, and make it clearer to readers, is some metric of daily temperature persistence during these warm and cold

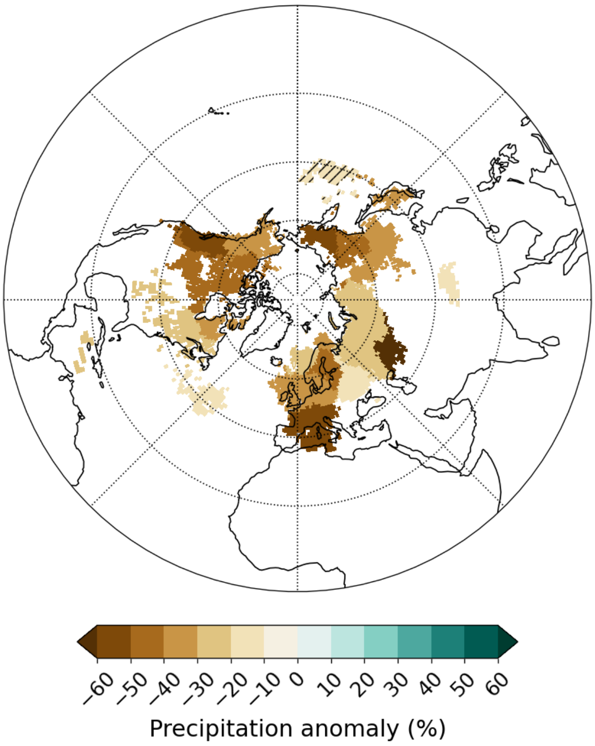

Figure R4: Region-averaged precipitation anomalies (in %) during JJA warm spells. Hatching indicates the absence of significance at 5%.

events. For instance the number of days above some high (1 standard deviation) or extreme (90th percentile) threshold to show that the large temperature anomalies are indeed ”persistent”. By the way, we just published as a preprint a review manuscript on the various definitions of persistence in the literature, which may be of interest (see https://doi.org/10.5194/egusphere-2023-111).

Comment 1.11 *I find the contextualisation of the findings lacking. While in Sect. 6.2 the authors attempt to make a link to previous work which has looked at coherent regions of extreme surface temperature occurrences, what is completely missing is a discussion section making a link to the wealth of literature looking at drivers of different regional cold or warm spells. Admittedly, there is some of this in Sect. 5. However, discussing in more detail the literature related to the mechanisms is a key point needed to contextualise the current analysis relative to the literature, and I would have expected the bulk of the discussion section to be dedicated to this.*

Answer: We are sorry if you find the discussion of previous results insufficient. We did make an extensive literature search and referenced many studies that have looked at (persistent) warm and cold extremes, and in fact cited dozens of papers on this topic. It is true that we could expand section 5 a bit more (it is already 2 full pages long, so the discussion of previous studies is arguably not ”completely missing”). However, as we argued above, the details of the regionalisation itself are not the main point of this manuscript – we do not discuss individual regions in detail, focusing instead on the similarities in terms of synoptic-scale anomalies and drivers across regions and seasons. We will expand section 6.2 as well to provide more context for our regionalisation results in light of previously identified dynamical drivers.

Comment 1.12 *ll. 14-15 Perhaps the authors could explain what spatially and temporally compounding events would be in the context of temperature extremes. Presumably concurrent, geographically remote heatwaves or cold spells or successions of heat waves or cold spells at the same location within a given season? Moreover, compound events are never mentioned or discussed in the analysis presented in the paper, so it seems largely irrelevant to discuss them in the introduction.*

Answer: This sentence is indeed irrelevant to the paper and should be removed in a revised version.

Comment 1.13 *ll. 36-37 I find this statement misleading. First of all, there are indeed papers that have treated temperature extremes on multi-week to monthly timescales (I can think of Galfi and Lucarini (2021) off the top of my head, but a careful literature search would likely bring up others). Second, while many papers do impose a 5-day or similar threshold for heatwave duration, most heatwaves end up being much more persistent than that. It is thus misleading to state that these papers "focused on short-lived events, on the order of a few days only". Indeed, using a few days in duration as a lower threshold can easily lead to identifying multi-week heatwaves, see e.g. Vogel et al. (2020) or Fig. 10 in Grotjahn et al. (2016). This also links to my previous major comment on the motivation for looking at 3-week temperature deviations.*

Answer: It is true that a few papers have looked at multi-week warm or cold spells explicitly. You are also right that multi-week events can also often be detected by looking at short windows (like 5 days) (while the converse it not true, cf. our reply to your first comment). We shall reformulate accordingly.

Comment 1.14 *ll. 63 Since the authors themselves later in the paper mention the challenges of working with 42 years of data, an obvious question is why they have chosen not to take advantage of the ERA5 back-extension, which has been available in preliminary form for over two years and in its final form since mid-2022. The major issues with the preliminary version were tropical cyclones, which is not something that would have affected the analysis presented here.*

Answer: It is true that the ERA5 back extension represents a wealth of new data that upcoming studies should make the most of. However, when the final version came out, we were already very advanced in our analysis, and we decided to stick to the 42-year data.

Comment 1.15 *ll. 108 In Fig. 1 many of the patches with DR>0.4 are fragmented. How does the PAM behave when given data with scattered "holes" in it? Does this affect the robustness of the final clustering?*

Answer: The clustering algorithm does not take any geographical information as input, so whether the DR map is fragmented or not does not matter. It is true that some regions end up including some fragments (i.e., a few lone points scattered away from the region's main group of points) but these represent at most a few percent of a region's points and do not impact the

	Cold	**Warm**
DJF	10-16 (12)	8-23 (12)
JJA	5-13 (9)	8-24 (12)

Table 1: Analysis window length (in days) for 2-week warm and cold spells: inter-region range and median.

	Cold	**Warm**
DJF	12-20 (16.5)	11-25 (18)
JJA	8-17 (13)	12-27 (17)

Table 2: Analysis window length (in days) for 4-week warm and cold spells: inter-region range and median.

results. Such fragments are unavoidable in our regionalisation since by choosing a rather small number of regions forces the algorithm to add isolated points to one of the regions. We explain at ll. 117-121 how we remove these isolated points to better visualise the regions.

Comment 1.16 *ll. 190-191 It would be interesting for the readers to see these numbers for 2 and 4 week averaging times, e.g as an Appendix table.*

Answer: Here are the numbers for 2 and 4 week events:

Comment 1.17 *Sect. 4.1 How do the authors determine this grouping? Is it through some objective criterion or through a subjective analysis of the individual regions?*

Answer: We obtained this grouping by subjective analysis based on the anomaly maps, and we should make it explicit in the revision. Note however that a simple clustering applied to the Z500 anomaly maps yields a similar partition.

Comment 1.18 *A few typos (although I appreciated how well-written the paper generally was):*

l. 211 near-surface diabatic
ll. 272, 328 Incorrect figure reference?
l. 328 (difference in intensity/spatial extent?)
ll. 347-348 (Figures 11) and S5)

Answer: Thanks, we will correct these typos in the revised manuscript.

Comment 1.19 *ll. 296-297 I am not sure I agree with this statement, and indeed the following two sentences seem to counter it. I would suggest adding a "summertime" to the first sentence to avoid confusion.*

Answer: You are right (zonal configurations are important in winter) and we will modify the sentence accordingly.

Comment 1.20 *l. 310 and following. The current formulation seems to suggest that there is plenty of work on the role of upstream blocking but that the authors are the first to highlight the*

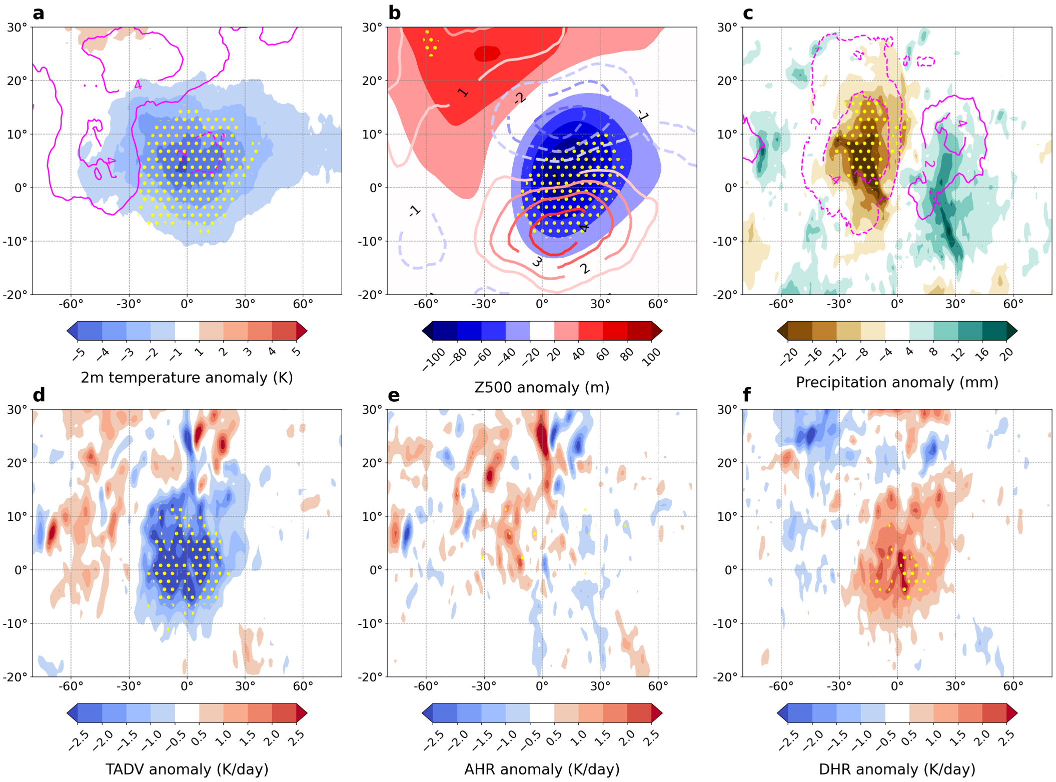

Figure R5: Modified Figure 5 (see manuscript).

role of downstream blocking/processes for the occurrence of upstream temperature extremes. While upstream blocking has certainly received more attention, there are studies which also consider the role of downstream features (e.g. Takaya and Nakamura, 2005; or Lehman and Coumou, 2015, who discuss the role of downstream storm track anomalies in the context of eastern North American extremes).

Answer: You are correct to point out that the initial formulation was misleading, and we do not claim to be the first to highlight the role of downstream blocking. Our point is that our regionalisation captures the varied influences of blocking (up- or downstream) in a coherent way and provides a useful hemispheric perspective on the role of blocking. Takaya and Nakamura discuss a nice example for Eastern Asia which we will include as reference. As far as we can see, however, Lehman and Coumou only perform pixel-wise regressions and do not discuss teleconnections between surface extremes and circulation features.

Comment 1.21 *Fig. 5 and following: using green dots on a red background is not ideal for many readers (including this reviewer).*

Answer: Thank you for pointing it out. It is hard to find a color that stands out well on red, blue, green and brown. We could use yellow (see Figure R5).

Comment 1.22 *Fig. 12 If the authors only look upstream of the regions (30 W of the westernmost point as stated in the figure caption) how can they diagnose downstream blocking as again stated in the caption?*

Answer: We only look 30°W upstream for the R-metric anomalies, not for the blocking (the caption states that "Hatching indicates significant positive blocking frequencies up- or downstream of the region". We will clarify the caption accordingly.

Comment 1.23 *On a separate point, it may be worth explaining in the methods what the R-metric is and how it is being used here.*

Answer: You are right; we realise now that we should have spent more time explaining the R-metric and the corresponding analysis. We will do so if invited to submit a revised version.