Response to reviewer comments

Comment 1.1 *The manuscript is well-written and the figures are of high quality. However, I have significant concerns about the methods, the implication of the results, and the comparison with the existing literature. In my opinion, the study does not address the research question stated in the title.*

Answer: Thank you for your comment. We agree that our initial title was misleading: we do not analyse the persistence of warm and cold spells per se, but the characteristics of persistent warm and cold spells (which is different). We would modify the title accordingly if invited to submit a revision. Our replies below should address your concerns as to the methodology and comparison with existing literature.

Comment 1.2 *In the Introduction, the authors identify two gaps in the literature on the processes responsible for warm and cold spells: (i) persistence and (ii) spatial dependence. With respect to the first gap (persistence), what can we learn meteorologically from longer lasting events that we don't already know from the existing literature? There is already a lot known on driving factors for heat and cold waves. I am not arguing that it is not interesting to look at the persistence, but the authors insufficiently motivate why it could be interesting to look at persistent events (apart from impacts). In addition, a minimum duration of three days does not exclude long-lasting events, e.g. in Zschenderlein et al. (2019), a heat wave with a duration of more than 40 days was included in the analysis. It is therefore not true that previous work generally focused on short-lived events only, as the authors claim.*

Answer: We should certainly 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 at least 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.

Additionally, 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).
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.

As to what we can learn from longer-lasting extremes, we openly admit that our results are in full agreement with the previous literature. It would have been surprising to have found something completely new! (on that point we should certainly reformulate certain passages of the manuscript that may suggest the contrary) But this fact isn't a sufficient reason to say that analysing longer events is useless. We argue however that our results 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 to this extent in the literature, to our knowledge). The long timescale is also important to obtain a robust regionalisation.

Comment 1.3 *With respect to the second gap (spatial dependence), the authors argue that the recent literature looked at arbitrary regions and mixed up areas where warm or cold spells are shaped by different physical processes. They are by far not arbitrary in the recent literature, either regions are motivated by impacts (which is perfectly fine and definitely not arbitrary) or by different climates (e.g. humid vs. dry climate). There are also studies looking at the different physical processes in a region, e.g. Sousa et al. (2018, Fig. 5). It is therefore not overlooked that in one region the processes can differ regionally.*

Answer: We agree that "arbitrary" was a very wrong choice of word. We interpreted 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 (but not exclusively) looked at regions based on impacts or observed extremes, while we are interested in a purely meteorologically-driven regionalisation (which admittedly has not been done, at least on a hemispheric scale).

Comment 1.4 *The study promises a lot to provide new insights into the persistence and spatial dependence of warm and cold spells. However, both aspects are not really quantified in the study. I detail my comments below:*
1. Persistence: Sections 4 and 5 are believed to address this research question. However, these sections rather analyse synoptic conditions during warm and cold spells (which is nicely done) and identify three atmospheric drivers (upstream blocking, downstream blocking, subpolar troughs). Many studies already investigated the position of the block relative to the high/low temperature extreme at the surface (e.g. Pfahl and Wernli 2012, Pfahl 2014, Bieli et al., 2015, Santos et al., 2015, Sousa et al. 2018, Zschenderlein et al., 2019). The authors should therefore make the new findings more clear. While I find the synoptic analysis very nice, I don't understand what this has to do with persistence. And this is the main aspect of the paper (at least it is stated

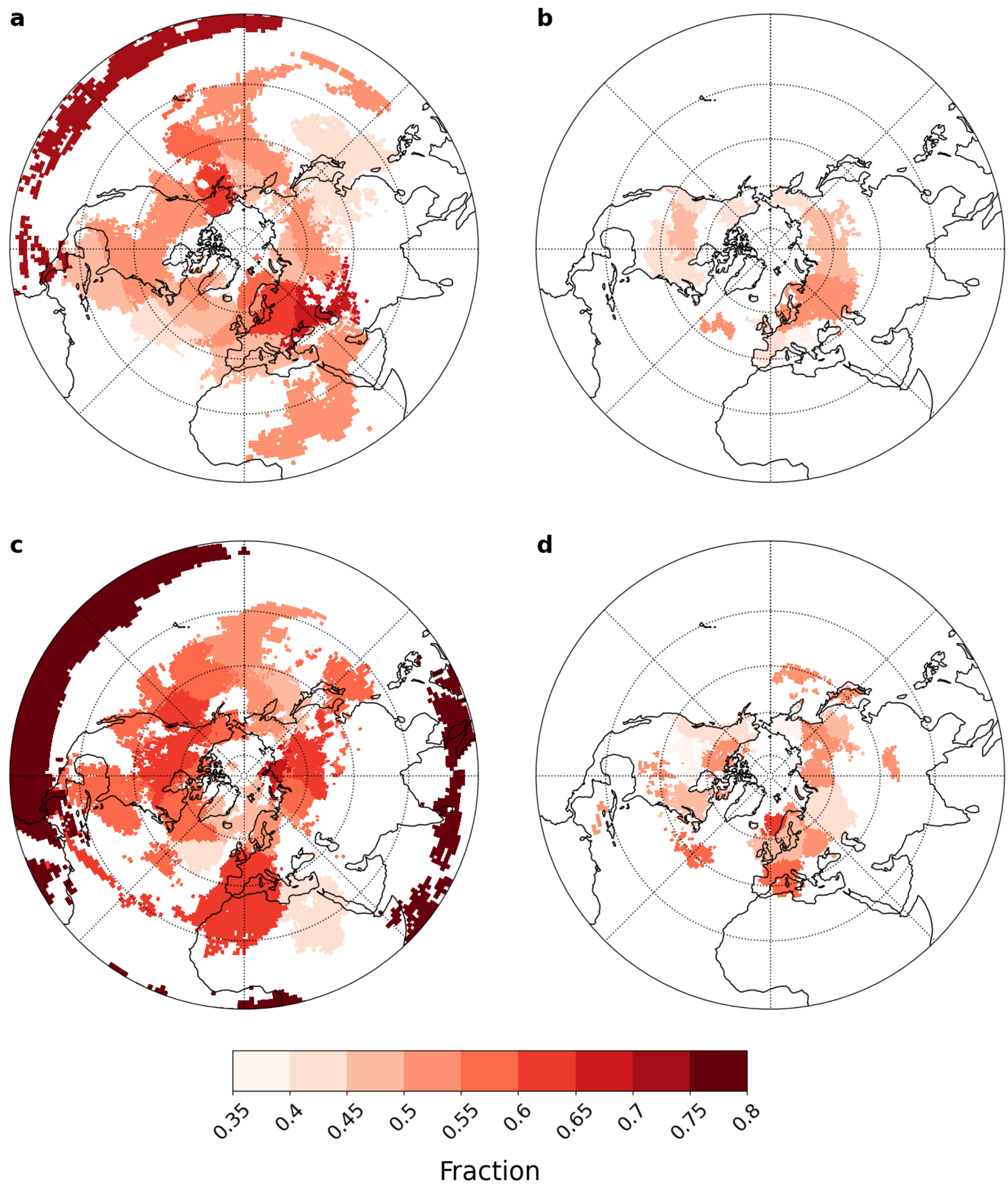

Figure R1: Fraction of 5-day warm and cold spells occurring within 21-day warm or cold spells, for (a) DJF cold, (b) JJA cold, (c) DJF warm and (d) JJA warm spells. In each case spells are defined relative to the 95th or 5th percentiles of the corresponding 5- or 21-day temperature distribution.

in the title). For example, how important is the position of the block relative to the warm/cold spell with respect to the persistence in comparison to a heat/cold spell of about 5 days? I think the authors try to discuss the persistence aspect in Section 5, but not really successful. They are rather speculating about the role of blocks, RRWPs, land-atmosphere feedback, and subpolar troughs (with results from the existing literature and not their own) and not quantifying the importance of these processes with respect to persistence.

Answer: We realise now that the title we chose was misleading since it led readers into thinking that we would be analysing persistence of warm and cold spells in itself rather than the drivers during persistent events and their possible role in the persistence of the warm and cold spells (which is not exactly the same thing). A detailed analysis of persistence would indeed require investigating warm and cold spells across different timescales to understand what determines their length. We will therefore make sure to reformulate our title and introduction so as to avoid misleading readers as to the content of the manuscript.

Comment 1.5 *2. Spatial dependence: I think the authors want to address this research question with the regionalisation approach (Fig. 3). But in the end, they summarise all regions into three categories (and they don't describe how they do it, see also my minor comments). What is the added value of the regionalisation? What can we learn from it? What do the different colors in Fig. 3 imply meteorologically? In my view, Figure 3 is insufficienty discussed.*

Answer: The regionalisation certainly helps identify distinct regions, which could be analysed separately in detail, but this is not the purpose of our manuscript (this will be the topic of future work, hopefully also by others who will make use of our regionalisation results). Instead, in this paper we wish to highlight the common features during warm or cold spells across different regions, so as to identify robust signals, while still analysing the different seasons separately and retaining some spatial variability (for instance when it comes to blocking or recurrent Rossby waves).
We understand that it may be frustrating not to see the case of specific regions be discussed in detail, and many regionalisation papers do take that path (we did just that in a previous paper on the regionalisation of temporal clustering of precipitation extremes; https://doi.org/10.1175/JCLI-D-21-0562.1). With a hemispheric perspective, such a paper would however be too long and we chose instead to highlight similarities across regions. In this sense, the regionalisation is extremely helpful to reduce the dimension of the problem: instead of conducting an analysis for each grid point separately, the regionalisation allows to first reduce the problem to a small, manageable set of regions. In addition, taking this regional approach allows us to shed light on the structure of regional-scale anomalies (e.g., eastern and western halves of regions sometimes behaving differently).
We could nevertheless discuss Figure 3 a bit more (e.g., the distribution of region area, a more in-depth comparison to previous regionalisation attempts, differences/similarities across seasons, etc.)

Comment 1.6 *I am not convinced by the method for the identification of persistent warm and cold spells. While I understand that you incorporate the days prior to the spells to investigate the onset, I do not understand why your analysis window ends with the peak temperature anomaly. What if this anomaly is reached already very early during the warm/cold spell and high/low temperatures still persist after that peak?*

Answer: It is true that the way we define the analysis window is not perfect. The way we do it here emphasises the build-up part of the spell, because we wanted to avoid including the decay part (when the drivers are likely weakening). Additionally, as long as the drivers are in place, temperature anomalies are likely to increase in magnitude (towards either extreme warmth or cold), even weakly, rather than to remain stable. We don't think that going a few days beyond the peak anomaly would make much difference, but in a revised version we will mention this point, and possibly compare results with an analysis window that extends a bit further (e.g., as long as the anomaly does not drop "too much" back towards the climatology)

Comment 1.7 *And for the quantile regression I would exclude grid points in the tropics beforehand, because a low Z500 variability in this area is no surprise.*

Answer: The quantile regression is applied to each grid point separately, so there is no influence of e.g. the tropics onto the mid-latitudes. As to the regionalisation, we remove grid points with low model skill (deviance ratio below 0.4) beforehand, which includes the vast majority of the tropics.
We could have restricted our analysis to, say, polewards of 20° or 30°N. Still, since some of the regions extend slightly more to the south, it is just as well to keep the whole Northern Hemisphere (and in any case, tropical grid points are largely excluded, as explained above).

Comment 1.8 *The results of this study are not set into context with the existing literature. An attempt is made in Sections 5 and 6.2, but it is not clear what new knowledge is introduced by the study.*

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. We will also expand section 6.2 to provide more context for our regionalisation results in light of previously identified dynamical drivers.
Additionally, we will make it clearer what are the new results brought by this study. A main result is the regionalisation itself. Though we do not discuss the case of individual regions separately (this will be the topic of future research), we highlight the coherence of the regionalisation and underline the many similarities and differences in terms of synoptic-scale anomalies and drivers of warm and cold spells across the Northern hemisphere extratropics. Admittedly we do not claim to have identified new, previously unknown drivers of warm and cold spells, and we also need to make clearer that we do not discuss persistence per se but rather the characteristics of persistent warm and cold spells (our initial title was misleading in that regard).

Comment 1.9 *L5-10 (Abstract): All listed processes (blocks important precursors of warm and cold spells, location of the blocks, recurrent Rossby waves, land-atmosphere feedbacks) are already described in many papers (see list in major comment 2) or in a new review paper by Domeisen et al. (2022, "Atmospheric processes" section; the paper was probably not yet published at the time the authors submitted to WCD).*

Answer: Our findings are indeed consistent with the literature (the opposite would be surprising). We will make sure to include the recent reference of Domeisen et al. which was indeed not yet published when we wrote our manuscript.

Comment 1.10 *L65: Does this mean that you first subtract the mean and divide by the standard deviation?*

Answer: Yes, indeed. We will make it explicit in the revision.

Comment 1.11 *L66-68: Why do you use different methods for the trend removal of T2M and Z500?*

Answer: The main reason is that we did not want to normalise the Z500 data; otherwise, tropical variability would have been dominant and we would not have captured the mid-latitude signals. So we simply removed the long-term trend in the Z500 mean.

Comment 1.12 *L70: Is the persistence criterion of 4 days valid for an individual grid point or for the whole cyclone mask?*

Answer: It applies to the cyclone as a system (this is a Eulerian persistence criterion).

Comment 1.13 *L70-80: Should be moved to the methods section.*

Answer: These lines are indeed a mix between data and methods, but we would prefer to focus the methods section on the core methods for this paper (modeling and analysis) rather than on the data pre-processing.

Comment 1.14 *L84: Is the percentile based on daily or 3-weekly values? And please do not use different words for the same method (normalise, rescaled)*

Answer: On 3-week values. The details are given just a bit later in section 2.2.1 so to avoid any confusion we will remove the reference to percentiles at l. 84.

Comment 1.15 *L89: Why non-overlapping 3-week intervals?*

Answer: We want the intervals to be non-overlapping to avoid as much dependency between model input data points as possible.

Comment 1.16 *L105: How sensitive are your results with respect to the DR threshold?*

Answer: We chose this specific threshold (0.4) to capture about as much of the extratropics in all seasons as possible. A lower threshold would lead to more tropical grid points retained in the regionalisation (extratropical regions remaining essentially the same in the regionalisation algorithm). A higher threshold removes points from the extratropics and quickly tends to decrease the "optimal" number of regions.

Comment 1.17 *L112: What is a silhouette coefficient?*

Answer: We should define this in a revised version. The silhouette coefficient is a common metric used to identify an "optimal" cluster number. It essentially compares the mean intra- and inter-cluster distances, with higher values indicating a better clustering (high inter-cluster distance and low intra-cluster distance).

Comment 1.18 *L129: Why not using the back-extension of ERA5?*

Answer: This comment was also made by the other reviewer. 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 1.19 *L135: Is the decay phase of the event not interesting?*

Answer: It certainly is, but not in the context of our paper which focuses on the mechanisms responsible for the onset and maintenance of the warm and cold spells. We can expand the analysis window a bit, but we should avoid including too much of the decay phase since there is no reason why the same drivers should still be in place.

Comment 1.20 *L143: median point – do you mean centroid?*

Answer: No, here we take the actual (geographical) median point, and not the cluster centroid (which in PAM are called medoids, and minimise the distances to all other cluster points).

Comment 1.21 *Figure 1: I don't understand the values on the y-axis. Is it normalised with the 95th percentile? I also do not understand why the grey shaded area goes from approximately day 11 to 31.*

Answer: Figure 1 illustrates an example of a warm spell by showing the corresponding daily normalised temperature series. This is why the y-axis of Figure 1 is in multiples of the standard deviation of daily temperature anomalies (cf. normalisation process in section 2.1). As to the grey shaded area, it corresponds to the identified 3-week warm spell (identification based on a rolling 3-week average of daily temperature anomalies; section 2.2.2). Based on this event (which occurs from day 11 to day 31 here), we calculate an analysis window between the last day with a negative temperature anomaly (day 6, in this case) and the day of peak temperature anomaly during the 3-week event (day 28, in this case). This figure is just an illustration of the method and the actual values of the days is meaningless (the x-axis could start at -11, for instance). We will change the caption to clarify these points.

Comment 1.22 *Section 4.1: How have you identified the three main groups?*

Answer: This is something we should have specified in the manuscript. We identified the main groups visually, after removing the tropical regions.

Comment 1.23 *Figure 5 (and others): Mark the centroid of the regions. Add wind directions or Z500 mean values.*

Answer: By construction, the centroid of the regions is at the origin in these figures. The mean Z500 and wind fields are not as meaningful as the anomaly fields here since we average across regions located at different latitudes/longitudes.

Comment 1.24 *General remark to the composites: I personally think that compositing can be quite dangerous here, because you are averaging over areas with different sizes and shapes. Can you comment on that?*

Answer: You are right and this is certainly a limitation that should be explicited. However, the whole point of the compositing is to highlight the similarities in terms of synoptic-scale anomalies between different regions. In some cases (e.g., blocking or recurrent Rossby waves) we were careful to analyse each region separately (to obtain Figures 4, 10 and 12). Region-specific details would have to be discussed in a different analysis.

Comment 1.25 *L225: Be careful because you are mixing land and ocean grid points (see red colours in Fig. 4).*

Answer: This is true (and the point was also made by the other reviewer. There are certainly limitations to our analysis due to the mixing together of land and ocean grid points, which we should make explicit. However, for cold spells the difference between land and ocean is less of a problem, because cold spells are advection-driven. Differences are to be found in the magnitude of the anomalies (temperature, advection, etc.) but not in their pattern. Additionally, the point of our analysis is to highlight similarities and differences in warm/cold spell dynamics across space.

Comment 1.26 *Figure 6: The scale in Fig. 6c differs from Fig. 5c!*

Answer: We will make sure to use the same scale, thanks.

Comment 1.27 *L250: These results are in line with Zschenderlein and Wernli (2022).*

Answer: We will add the citation here, thanks.

Comment 1.28 *Section 4.4 and Figure 12: I am not able to follow because you haven't described the R-metric sufficiently.*

Answer: This point was also raised by the other reviewer, and we will make sure to add details about the R-metric in a revised version. The current information in section 2.1 is rather sparse. All details are in Röthlisberger et al. (2019), but we should expand a bit for readers unfamiliar with this work. The R-metric is calculated as the magnitude of the envelope of meridionally-averaged 35-65°N 250 hPa 6-hour meridional winds, which were first averaged using a moving 14-day window, and then filtered to retain wavenumbers 4-15 only (the ones relevant for synoptic Rossby waves). A high R-metric thus indicates a strong wavenumber 4-15 component in the series, and to multiple Rossby waves because of the 14-day running mean.

Comment 1.29 *L290: remove the second "is"*

Answer: We will correct that, thanks.

Comment 1.30 *L310: This agrees with a host ... → ok, but where is the new knowledge regarding the persistence?*

Answer: See our answer to your Comment 1.8.

Comment 1.31 *L312-313: This is probably also true for events on the shorter time scales.*

Answer: Yes, certainly.

Comment 1.32 *L318: remove half sentence in brackets at the end of the line.*

Answer: Corrected, thanks.

Comment 1.33 *L348-350: This is not clear to me. Increasing surface heat fluxes would imply diabatic heating, but we see diabatic cooling in Fig. 11f. Radiative cooling is typically dominant in the mid- to upper-troposphere.*

Answer: In Figure 11f, the diabatic heating term is calculated as the residual of the energy balance equation, and is not restricted to radiative cooling. Still, what probably happens is that the radiative cooling term outbalances surface sensible heat fluxes, since it has to balance not only sensible heat fluxes, but also adiabatic warming and horizontal advection. Note that our analysis also covers day and night, and that sensible heat fluxes peak during the day but are weaker at night, while radiative cooling can remain large across the clock. We would however make sure to mention this point in a revised version.

Comment 1.34 *L351-352: What can we learn from this sentence?*

Answer: We mentioned this because a number of previous studies have argued that extreme warm spells were often associated with antecedent precipitation deficits. However, in our case we probably do not look at event that are extreme enough to see the connection.

Comment 1.35 *Section 5.4: You discuss how a subpolar trough develops. But I don't see how you discuss subpolar troughs and their connection to persistence of warm and cold spells.*

Answer: You are right that we do not discuss the development of subpolar troughs, but rather point out that persistent wintertime cold spells at high latitudes are likely associated with such systems (the analysis of synoptic-scale anomalies points in that direction). They connect to persistent cold spells because of the associated circulation anomalies, but also because such systems have a general tendency to persistence, as we discuss in the manuscript (cf. also Woolings et al. (2023); https://doi.org/10.5194/wcd-4-61-2023).

Comment 1.36 *Section 6.1: A matter of taste, but the discussion of Fig. 3 appears quite late in the study. I would prefer to discuss the results earlier.*

Answer: We agree. One possibility would be to merge current sections 5 and 6 into a larger discussion section, and to move section 6.1 to the beginning, before discussing the physical drivers.

Comment 1.37 *L383: Acronym S2S not introduced*

Answer: Corrected, thanks (S2S = "subseasonal-to-seasonal")