Response to reviewer comments

Comment 1.1 *The authors have conducted additional analyses and implemented several textual edits. They have responded to a number of my comments, but I find their response to my main comment, namely whether they can demonstrate the relevance of picking a 3-week timescale, very weak. Indeed, several of the citations they have added to the text actually work against their argument (see Comment 1.1b below), and they never replied to my question of whether it is possible to identify a set of high-impact events that is overlooked by a 5-day minimum persistence criterion but captured by their own definition, or whether they could make a clear case for events whose meteorological drivers would be confounded if using a 5-day minimum persistence criterion.*

Answer: Thank you for your detailed comments. It appears that your major concern regarding our paper remains our choice of the 3-week timescale. You argue that the events thus identified are likely the same as those which would be identified with a shorter (e.g., 5-day) timescale. We realise our initial response had not been clear enough in that regard. We therefore conducted an additional analysis, namely to calculate the fraction of 3-week events we identify which do not include 3- or 5-day heat/cold events. The latter are defined as in the literature, *i.e.* by looking at consecutive runs of 3 or 5 days with extreme temperatures (extreme being defined on a daily basis). See Figures R1 and R2. Note that requiring strictly consecutive daily extremes or allowing for a 1-day gap in the series makes little difference (Figures R3 and R4).

Results show that a significant fraction of the 21-day spells we identify do not overlap with the more classic 3- or 5-day extremes. This is especially true for summer, especially summer cold spells (Figure R1-b), while in winter, only 0-20% of 21-day spells do not include short periods of extreme temperatures.

There is therefore an undeniable overlap between our persistent spells and the more traditional, short extreme hot/cold spell definition. However, our argument isn't about identifying pointwise events so much as identifying the temporal extent of events (see our answer to your Comment 1.3). We do not really see the point of showing that we capture many events with a three-week timescale that are different from the events that are captured when working with 5-day timescales. The whole point of our argument is that when working with 3- or 5-day timescales, one will not necessarily know if the event lasted longer than 3 or 5 days. And the key here is our focus on the longer-lasting events. This is precisely why we adopt the 3-week timescale. We are not interested in whether the longer-lasting events include short periods of extreme daily temperatures. The question is whether these periods indeed correspond to persistent warm/cold anomalies.

We replaced Figure A1 in the manuscript with the present Figure R1, and expanded the text as follows. First, we replaced the reference to Figure A1 in the introduction by "*In addition, a sizeable fraction of persistent spells do not include short periods of extreme temperature anomalies, especially in summer (Figure A1).*". Second, we added the following sentence to the regionalisation discussion of section 5.2: "*Figure A1 shows that while there is certainly a lot of overlap between our 21-day spells and more traditional 3-day extremes, the two are not the same, notably during summer (Figure A1-b,d), which makes a direct comparison difficult.*"

Comment 1.2 *I only partly agree with the authors' argument that "while persistent spells frequently include short periods of extreme temperature anomalies, the opposite is not true".*

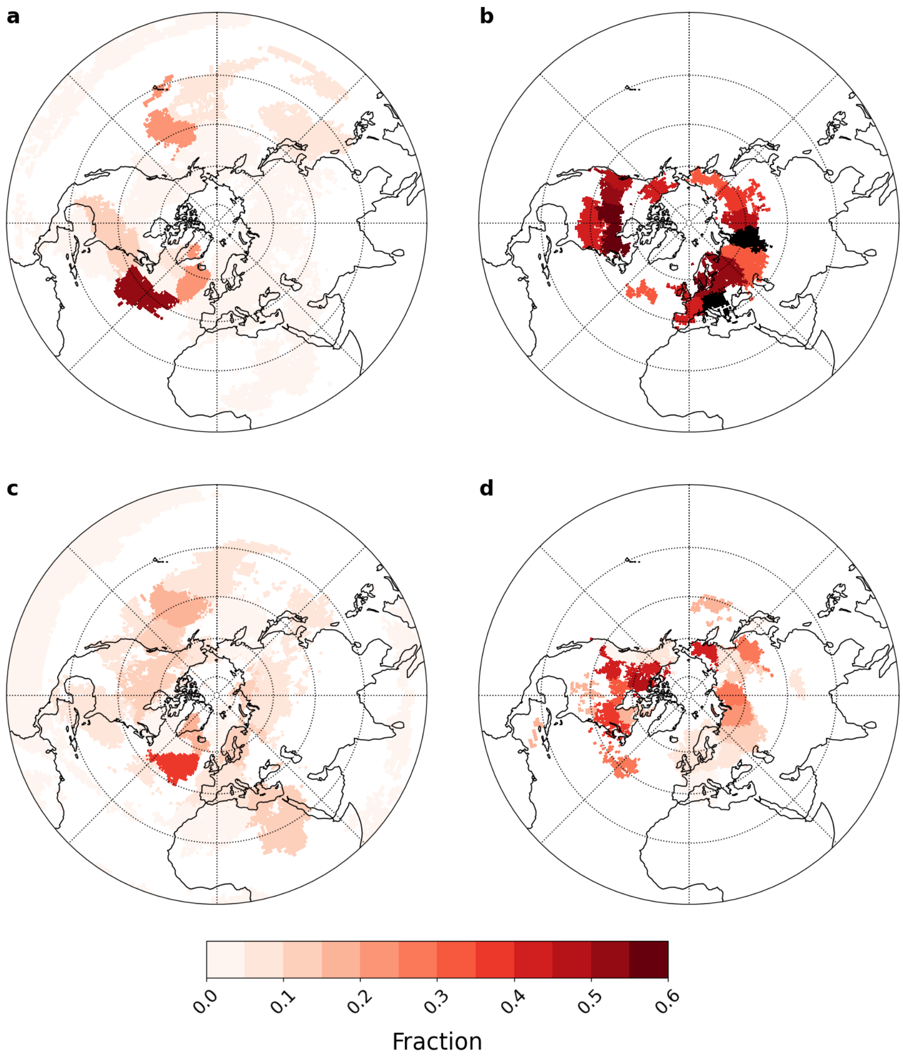

Figure R1: Fraction of 21-day persistent warm/cold spells which do not include a 3-day extreme warm/cold period, for (a) DJF cold spells, (b) JJA cold spells, (c) DJF warm spells and (d) JJA warm spells. 3-day extreme periods are defined based on the daily $5^{\text{th}}/95^{\text{th}}$ temperature percentiles.

Assume we call 3-week events group "A", and 5 or 3-day events group "B". If A and B include a similar number of events, then if group A events frequently include group B events, it is a tautology that group B events must frequently be part of group A events. The authors' statement can only be true if one has many more short periods of extreme temperature anomalies (group B) than persistent spells (group A). If I understood correctly, in their example the authors define extremes using 5-day mean values, thus providing something of a self-fulfilling prophecy (as I assume that this makes the number of events in B larger than the number of events in A – I would recommend indicating sample sizes in the caption to Fig. A1). However, the definition typically used in the literature is 5 (or some other similar number) of (near-)consecutive days above a single-day persistence threshold (this is true of all three papers the authors cite on l. 45. For example, Jiménez-Esteve and Domeisen (2022), define heatwaves using consecutive exceedances

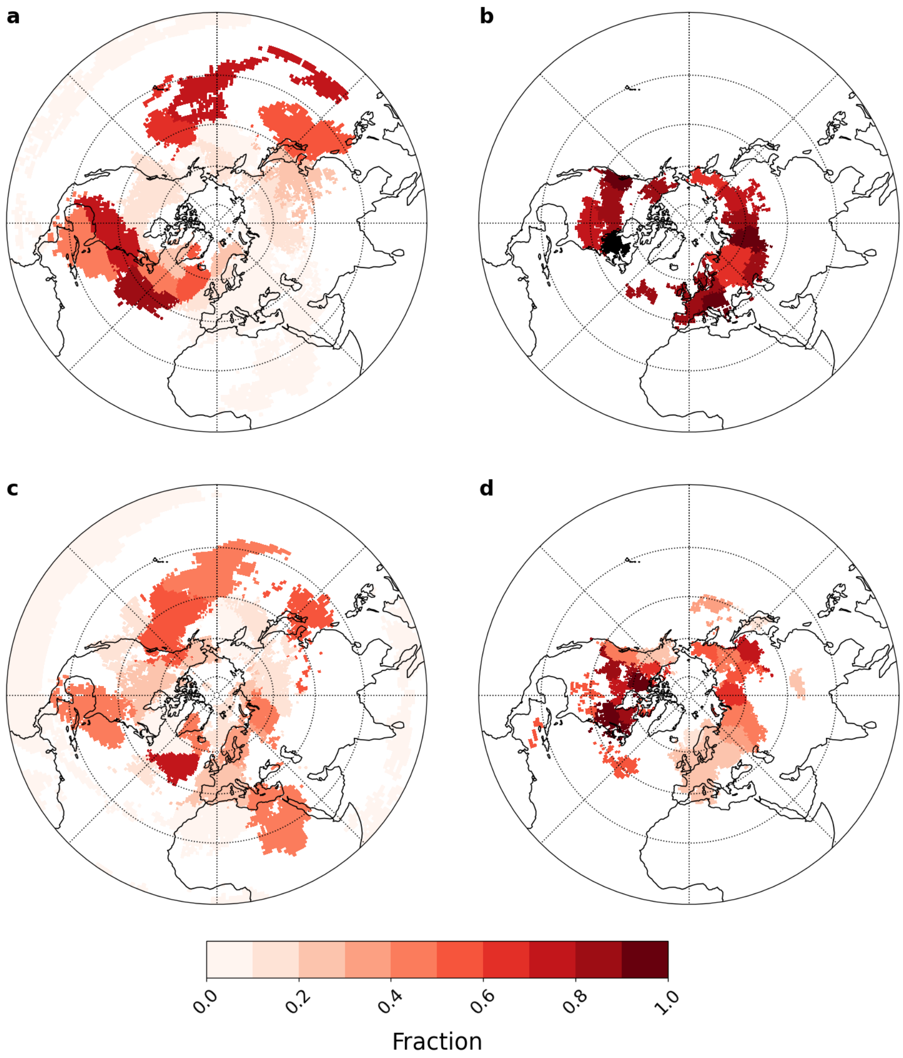

Figure R2: Same as Figure R1, but for 5-day extreme periods.

of single day 95th percentiles). The authors' example thus does not reflect the conventional definition used in the literature. I am doubtful that if the authors were to use single-day thresholds as in the papers they cite, and say a 5-day consecutive exceedance criterion, the results they show in Fig. A1 and that they describe in the text would still hold.

Answer: We fully agree that our argument only holds when there are more events in group B than in group A. But with our event definition, this is precisely the case (something we hadn't said explicitly, though). Since there are many more independent 5-day periods than 3-week periods, one will necessarily find more extreme warm or cold spells using a 5-day window than with a three-week window. In any case, we removed this figure and replaced it with Figure R1.

Comment 1.3 *Besides this, I still do not see a clear-cut case for why one should be interested in persistent above-average temperatures that may never be extreme in terms of daily values. In my original review, I asked whether the authors could identify a set of high-impact events that*

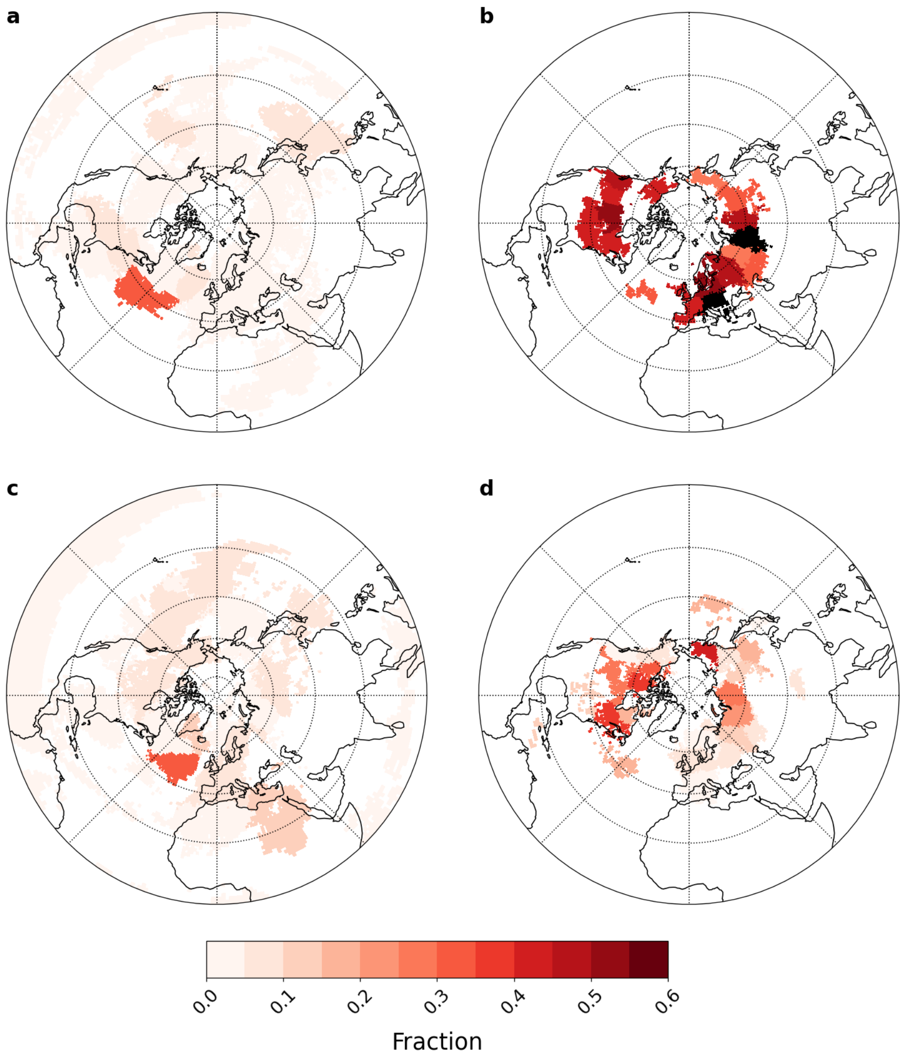

Figure R3: Same as Figure R1, but where a 1-day gap is allowed when defining the 3-day extreme periods.

is overlooked by a 5-day minimum persistence criterion but captured by their own definition. Similarly, I asked whether they could make a clear case for events whose meteorological drivers would be confounded if using a 5-day minimum persistence criterion. So far, the only argument concerning these points that the authors have put forth is that choosing a longer window is useful to identify regions where spells tend to co-occur. That may be true, but does it really provide an advantage over a simple lagged co-occurrence analysis as often performed for shorter events? On this same point, several of the citations that the authors have added to the revised text to support their claim of relevance of prolonged spells, actually use short thresholds to define extremes (e.g. Añel et al. (2017) use three days running means to define their case-study cold spell). Similarly, Chapman et al. (2020) chose to highlight in their abstract 5 and 7-day persistence periods, much closer to the conventional 3-5 day definitions than the 3-week definition proposed here. This supports the idea that, even if there is an impact dependence related to event duration, this is successfully captured by conventional shorter definitions of

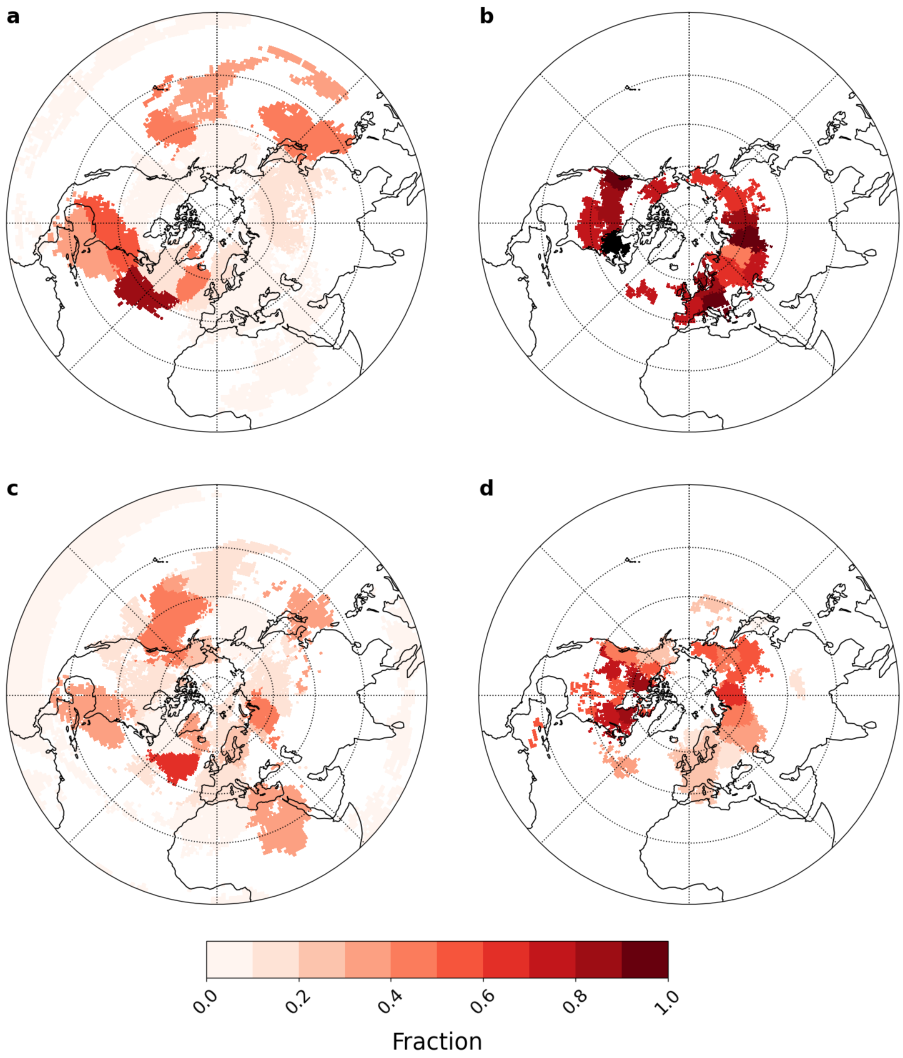

Figure R4: Same as Figure R2, but where a 1-day gap is allowed when defining the 5-day extreme periods.

extreme temperatures. Returning to my question above, it is probably relatively easy to find high-impact persistent warm or cold periods. However, I would expect that it is hard to find high-impact persistent warm or cold periods which would not be detected by conventional 3 or 5 day heatwave/cold spell definitions (these would be "persistent spells which do not include short periods of extreme temperature anomalies" to paraphrase the authors' response).

Answer: First, the fact that persistent warm or cold spells lead to significant impacts is clearly demonstrated in the literature, as supported by the studies we cited. Daily temperatures do not need to be consistently extreme to cause high impacts, if the temperature anomaly lasts long enough. Classic heatwave impacts focus on daily temperature extremes, but at S2S timescales, persistent warm or cold temperatures can still lead to major impacts, not because of their extremeness on a daily basis, but because of their persistence. Note that Añel et al. (2017) do not define cold/heatwaves with a 2-day threshold. They provide a literature review of case

studies, including ones for which the cold/heatwaves lasted several weeks. Chapman et al. (2020) also specifically discuss the case of long-lasting heat events (15 days and longer). It is true that few papers systematically analyse long warm or cold spells and their impacts (hence also the novelty of our study). In any case, whether persistent warm/cold spells lead to high impacts or not (independent of their including short periods of extreme daily temperatures), we are interested in characterising their persistence. Our paper is not focused on demonstrating the impacts of long warm/cold spells with new data. We added a mention in the introduction of this year's persistent heat in North America that led to massive impacts on human and natural systems.
Second, we agree, as stated previously, that many persistent cold or warm spells do include short time periods (2-5 days) with extreme daily temperature. We do not claim to be looking at a completely different category of events. However, we argue that to consider the persistence of warm or cold anomalies, one should look beyond the classic 2-, 3- or 5-day timescale and consider S2S timescales. The question isn't whether the events are also detected with a 3 or 5-day timescale, but what is the timescale that best captures the full extent of the events.

Comment 1.4 *I thank the authors for having contextualized their statement. A small addition I would suggest is to explicitly state in the text what they explain in their reply, namely that the focus here is on a purely meteorologically-driven regionalization.*

Answer: Thank you for this comment. We rephrased the beginning of the last paragraph of the introduction as follows: "*Here, we introduce a simple, meteorologically-driven regionalisation method for sub-seasonal warm and cold spells.*"

Comment 1.5 *Fig. A2 is a very nice complement to the original analysis, and the wording in the text now provides an honest appraisal of possible issues with the chosen definition. I believe it makes the point that indeed the authors' definition does artificially extend the build-up period, but I leave the editor and the authors to evaluate whether this may be an issue or not. Wanting to reflect the physical build-up period when a rapid temperature change occurs, I would by eye set a mean threshold between 5 and 10+ days after the shown mean values (I struggle to interpret the median vertical lines given that only the mean temperature anomaly is shown in the figure).*

Answer: You are right to point out that the mean values (solid vertical lines) are less representative of the true beginning of the build-up period as the median (dashed vertical lines). We didn't show the median temperature series because they were virtually the same as the mean series.

Comment 1.6 *Thank you for the clarification, this makes more sense now. To avoid any possible confusion you could perhaps reiterate that the compositing is done for every event separately in the first or second sentence of Sect. 2.2.3.*

Answer: Thank you, good point. We rephrased as follows: "*Anomalies are calculated for each event within each season and each region separately. We average the different fields during the corresponding event time window, ...*"

Comment 1.7 *Since the authors now have a new "caveats" section they could discuss there (or* *in the conclusions) this point and explicitly mention the ERA5 back extension.*

Answer: We added the following to the "Limitations" section: "*Last, we note that ERA5 is now available back to 1940 (83 years vs. the 42-year dataset used in this study). Considering this back-extension would allow to identify many more warm and cold spells and draw more robust conclusions.*"

Comment 1.8 *l. 60 Misspelled reference to Röthlisberger, also recurs later in the text.*

Answer: We couldn't find where the name was misspelled. This might be a compilation issue with the "ö" sign in latexdiff.

Comment 1.9 *Figure A8. Describe the stippling in the figure caption.*

Answer: Thanks for noticing. We added the following to the caption: "*Hatching in (a) indicates the lack of statistical significance of the region-average anomaly at the 95% confidence level. Stippling in (b) indicates statistical significance of the region-average anomaly at the 95% confidence level.*"

Comment 1.10 *The revised study has two "Table 1"s.*

Answer: Fixed, thanks.