

Editor

1) Your short summary section includes the abbreviation CMIP6. For the next revision upload, please adjust the section and provide at least one written-out version of the abbreviation to make it better understandable for non-experts. Please consider the character-limitation of max. 500 characters (including spaces) for the short summary section.

Reply: We have removed a few words and spelled out CMIP6 in the short summary.

2) Please ensure that the colour schemes used in your maps and charts allow readers with colour vision deficiencies to correctly interpret your findings. Please check your figures using the Coblis – Color Blindness Simulator (<https://www.color-blindness.com/coblis-color-blindness-simulator/>) and revise the colour schemes accordingly. --> Figs. 10, 17, A2, A3

Reply: We have updated figure 17 (old numeration). Using the Color Blindness Simulator , we checked figures 10 and A2, where we think the differences should be large enough to differentiate between the used colors. Figure A3 was close to impossible to find sufficient distinct hue's to differentiate between the lines (especially top left). We hope that the addition of asset with each case separately (i.e. split out each individual line in A3), is enough to aid the reader who cannot differentiate between the used lines.

3) Please place all appendix figures and tables within the appendix section (Fig. A3, Table A3, and Table A4 are placed after the appendix section). 4) In line 140 on page 7, you are referring to "supplementary materials". Please refer to either "additional" materials or "materials in the Appendix" to avoid confusion with a separate supplement file.

Reply: We have incorporated this remark

Reviewer 1

General comments

This manuscript describes a catalogue of Strong Nonlinear Surprises (SNS) in ocean, sea ice and atmospheric variables in CMIP6. The authors expanded on the methodology of a previous assessment on CMIP5 by Drijfhout et al. (2015) by automating the detection of SNS and including an algorithm to combine grid cells into spatially connected regions with SNS. They have a set of 6 categories of SNS, including abrupt changes and state transitions.

The developed method substantially improves the previous method, specifically by automating and including a spatial algorithm. The algorithm performs very well, and the authors are able to successfully capture large SNS in the data. The results are of great interest and highly valuable to the community. The results lead to new insights and have a high potential to stimulate further research and discussion within the field on abrupt dynamics in the climate system. The manuscript could benefit from a clearer description of the methods, careful framing of the results, and a reorganized and more substantive discussion.

Reply: We thank the reviewers for their kind words and constructive comments on our manuscript.

Major comments

1.

Could the authors please clarify in the methods section how exactly the regions are determined. Specific points to consider here are the following.

Thresholding is used to select different regions. How are these initial regions created/selected? Is the percentage threshold based on the very first and last values of the timeseries or over a smoothed timeseries/average over n years? This could make a difference for variables with high variability. For the third region finding approach, what is the reasoning for multiplying the percentage scores? In the third phase, formal criteria are applied to the selected regions. Does this merge regions of the same region-finding method, or does it merge any regions regardless of the region-finding approach? If so, will this lead to “smoothing” out of SNS events? Also, what is the point of having higher thresholds in case the different types of regions are merged? Why does it not work to only use the lowest threshold of $T = 85\%$?

Reply:

- The goal of this step is: isolate regions of roughly $\sim 1\text{e}6 \text{ km}^2$, that we can then search for SNS using the formal criteria. To be able to select regions that are not too large (which could lead to smoothing out of the SNS, making them undetectable), we use more than one threshold for each region finding method.
- For example, when we take the start and end-difference (the first method), we first use the threshold of 1 psu. We produce a Boolean map if the start-end difference exceeds this threshold. Using an unsupervised clustering algorithm (reference also added below), we build continuous regions of the Boolean map. This yields one or more clusters on the map with regions higher than the 1 psu criterion. Now, we only keep those clusters that also exceed the area-requirement of $1\text{e}6 \text{ km}^2$. It is rare that the first threshold (99.99%) yields any regions that are sufficiently large. Let's assume we did not find any cluster at this iteration. If we lower the threshold, and run the algorithm again (e.g. with a threshold of 0.9 psu difference), the clusters are generally larger than in previous step. It may now happen that a clustered region is sufficiently large. We save this region, but also check again if for a lower threshold (e.g. 0.8 psu) yields any *new* clusters. We emphasize that it should be new, since we exclude the first selected cluster from the Boolean mask (since we already identified that region). We keep repeating these steps, so at the lowest percentile (85%), we have most likely already identified a set of regions (order of 10 regions / method).
- If we would have instead only applied the 85%-percentile threshold, we would have gotten for example, two regions that are $\sim 5\text{-}10 \times 10^6 \text{ km}^2$. These two regions might then wash out any SNS, which is not in line with the goal of this step. This is why we iterate the thresholds to get smaller regions ($>10^6 \text{ km}^2$) such that we are sufficiently fine-grained to identify SNS.
- In the documentation of the tool, we have provided step-by-step information of how the algorithm identifies and selects regions (https://github.com/JoranAngevaere/optim_esm_tools/blob/master/notebooks/advanced_methods.ipynb).
- We clarified this in the manuscript by:
 - Providing the example of the region finding method using plots as in the linked step-by-step documentation of the tool. Please see appendix A1.
 - Restructuring the text that we first enumerate the four methods, and then explaining their four respective thresholds (near lines 113 and following).
 - The 3rd method is very similar to the 2nd method, and therefore finds very similar regions. It puts a little more emphasis on regions where both the max-jump and the standard-deviation have very high percentiles. The fact that the differences

are small with respect to the 2nd method only means we have to evaluate the following step (applying SNS criteria) a few more times, which is only a computational expense. Therefore, we kept these two methods in the workflow despite their similarity. This is now mentioned in the text.

- The merging phase is applied to regions irrespective of their selection method. This will not allow for smoothing of SNS events in the merging phase, since we only merge regions if the combination of two regions also fulfills the SNS criteria (which are also very strict for that reason).
- All metrics are calculated based on the 10-year running mean, we clarified this.

2.

It is not fully clear what choices the authors made in arriving at the 6 different SNS categories and how to interpret them, i.e. in what ways are they similar or different. What is, for example, the difference in interpretation between categories i and ii? Type ii is towards the end of the time series, and it can therefore be less robustly tested whether the change is persistent. Should this then be interpreted differently from a “real” abrupt change event i? Please give a short explanation of what the authors regard as a state transition/new state (criteria iv to vi).

In addition, the manuscript would benefit from more robust reasoning for the different categories and differentiation between abrupt changes and state transitions. Categories iii to vi are concerned with state transitions instead of abrupt shifts. However, when looking at the detected time series, the SNS often seem abrupt (e.g. sea ice “A” and “a” both change abruptly relative to the timescale of their normal dynamics). What is the motivation for separating these? With regards to the criteria of category iii, can the authors explain why they decided on this criterion instead of using vi with an extra requirement of a minimum surface area? Currently, the results sections are divided into abrupt shifts and state transitions for the same systems. Without a clear reasoning on the difference between the two, perhaps the authors can merge the sections for each physical system instead of having this distinction.

Reply:

- Criteria i-vi do not describe 6 different SNS categories but 6 different sets of criteria used to detect the various SNS. While criteria iii, iv, v are indeed tailor-made to describe SNS for 3 different variables, the other three are not. The difference between criteria i and ii is only determined by the timing of the abrupt event relative to length of the timeseries. Criteria i can never be met when abrupt changes occur

at the end of the timeseries, and for these cases criteria ii are developed. We will clarify this in the text and rename and reorder the criteria to avoid confusion with SNS categories.

- We will rename the A-I categories to comply with the order of discussion in section 3 and treat abrupt changes and state transitions for each category more closely together.
- The main difference between e.g., a and A cases, while both may look abrupt is that state transitions often develop over timescales longer than 1 decade and may not feature a strong enough change over 1 decade to be qualified as “abrupt”, while vice versa abrupt shifts over 1 decade may not qualify as a state transition if the change between before and after the abrupt change is too small to qualify as a state transition, to be differentiated from a state change that can be explained by a state modification by changing forcing without invoking strong internal feedbacks, apart from the feedbacks associated with the abrupt shift. We clarify this better in the text. When an abrupt shift does involve a state transition it will always be qualified as a state transition.
- Criterion iii is specifically tailored to sea-ice collapse identification, while criterion vi identifies state transitions in most other variables of interest (see A1.6). Criterion vi heavily relies on the behavior of the Pi-control dataset, and especially the standard-deviation of the time series. This makes it ill-suited for sea-ice, that often features a constant 100% sea-ice for extended regions, meaning that large regions have standard-deviations of 0%. This effectively means that if there is warming, you will always find large regions that show something extreme with respect to the Pi-control. Therefore, we expect criterion vi, when applied to sea-ice, to find nearly all models and SSPs to have an SNS in sea-ice (even with the area requirement of 5 million square kilometer).
- In addition: applying vi to sea-ice would always find a transition near the shifting sea-ice edge. We do not find this a “surprise” or SNS. That is why we demanded to apply this to the polar ocean as a whole for sea ice.
- Finally, we will treat the abrupt/state transition categories (e.g. Category A and a) them more together, see also response above to the same point of the Ref made for the Methods section.

We implemented this in the text by:

- Renaming the criteria Ab1-St4 (see lines 153 and onward).
- Further motivating them in the lines below that (159 and onward).

3.

Throughout the sections discussing the SNS results, the authors make statements about the mechanisms or forcings of the identified SNS without discussing how they arrived at this conclusion. Can the authors please substantiate the claims they make on this, whether it is based on analyzing the data of multiple variables at the SNS or on literature. We suggest that claims like “forced by”, “caused by”, “leads to”, “driven by” need to be backed up with either references or a note on what is observed in related variables around the SNS.

An (incomplete) list of points where this was done is shown below, and the manuscript would benefit from a thorough check on the whole results section on whether the claims are substantiated.

- Lines 160-162. What is the source for the fact that sea ice loss is caused by the sea ice-albedo feedback in these simulations?
- Lines 205-208. In what way does it lead to a new climate/what type of climate?
- Line 213: “forced by global warming”
- Line 217: Why is this likely driven by the onset of deep convection? Similar question for the explanation around line 220, does the onset of deep convection show in the data? Is it clear what process causes what?
- Line 230: Can the authors show this mechanism or have a reference? If not, mention that this is a proposed mechanism. What is meant by the last sentence of this paragraph?
- Line 288: “the mixed layer collapse is caused by a polar halocline”
- Line 295. Why is freshening a requisite for mixed layer collapse to occur? For example, in Figure 8 it looks like the mixed layer decreases slightly before the freshening starts.

Reply: We will support these claims with further references. We assumed that when such mechanisms were already discussed in Drijfhout et al. 2015 and/or other TP-reviews we did not need to support those claims, but we will make the manuscript more self-contained now.

We implemented this in the text by:

- Adding citations, rewording claims, and/or elaborating at the different items listed above.

4.

It would be good if the computation of the CDFs was added to the methods section, instead of only being explained and discussed in the discussion. The results of the global CDFs could then be placed near the end of the results section. This would improve the structure and readability a lot.

Reply: We have incorporated this suggestion.

Figures 16 and 17 are informative showing the distributions of global warming at which the SNS occurred. In the second panel of Figure 16, it shows that there are very few simulations above 6 degrees of warming. The authors currently use a cut-off of 11 degrees, but maybe this should be lowered to 6 degrees. The high temperature region draws a lot of attention while not being informative due to the very high uncertainty. Moreover, the color palette puts a very strong focus on the SSP585 scenario due to the bright color. In Figure 17, the CDFs of all categories are shown. However, some categories contain just one model simulation. This makes the CDF highly uncertain. Maybe only the CDFs with more than e.g. 5 detected SNS could be shown or separate those with more simulation by a different line style.

Reply:

- We capped figure 16 at 5 models (bottom panel) which corresponds to 6.7 C warming. Similarly, we adjusted the lower bound to -0.26 C warming.
- Consistent with that change, we would propose to make the lines with fewer than <5 models dotted and explain this clearly in the caption.

Furthermore, in the introduction (line 62), it is stated that PDFs are used to give the likelihood of maximum change. Can the authors explain or provide a reference to why a single simulation can statistically give a likelihood? In Figure 3, the global warming level at the point of maximum change in the SNS are used instead of PDFs. What is the reasoning for not using the same method in both cases? For Figure 3, one could take the global warming level at e.g. the midpoint of the PDF instead.

Reply:

- Our use of “likelihood” does not refer to the probability that a model produces an SNS, but rather to the conditional distribution of where the maximum change occurs *given that an SNS is present in that simulation*. Specifically, our procedure

applies multiple plausible analysis windows (smoothing and change timescales) to the same time series, which yields a set of estimates for the global warming level associated with the maximum change. From this ensemble of estimates we construct a probability density function (PDF). Thus, the PDF quantifies the conditional likelihood of the timing of the maximum change under varying methodological choices, not the probability of the SNS itself. We have clarified this wording in the manuscript to:

“Using this method, we construct for each SNS and each model a probability density function (PDF) that quantifies the conditional likelihood of the global warming level at which the maximum change occurs. This PDF is derived from applying multiple plausible smoothing and change windows to the same time series, to reflect the methodical uncertainty. We then average these conditional PDFs across all models in each category to obtain a category-level distribution. Importantly, these PDFs do not describe the probability that an SNS occurs in a model, but rather the distribution of the warming level at which the maximum change is detected, conditional on the presence of an SNS.”

- For figure 3, we reasoned that a simple approach would help the readability of the figure. However, we do see the merit to use, as suggested, the same method for as we used in figure 17 so that it's consistent. This has a relatively small effect on the pi-diagrams which was to be expected since the methods were slightly different.

We implemented this in the text by:

- Using the wording as explained above.
- Change figure 3 to be based on the outcome of figure 17.

5.

The discussion contains important points, but it requires an improved structure. A large part of the discussion section is currently occupied with methodology and new results (how to obtain the CDFs and the global warming levels). The manuscript would benefit from moving this part to the methodology and results sections (as described in major comment 4).

Reply: We agree to restructure the Discussion and moving parts discussing the methodology to the methods section. We have implemented this directly.

The discussion is currently very brief (when the CDF results are not considered). It would be good if the authors linked back to some of the points they mention in the introduction, like the distinction between abrupt shifts and tipping points, and how their results fit into this. In addition, it would be valuable if some discussion on the individual physical subsystems was added, placing their results in the wider literature context including a discussion on future research directions for specific systems.

Reply: We will extend the discussion, also considering the comment above, although we cannot go into depth about the physical mechanisms of each category. However, we can and will give more focus on SNS categories not discussed before and or categories with very different thresholds compared to previous assessments.

We implemented this in the text by:

- Highlighting the difference with Drijfhout et al 2015, especially on category A/a (Northern Hemisphere sea-ice collapse); category D (AMOC collapse), and category B/b (Northern Hemisphere mixed-layer collapse) in the discussion section.
- Elaborating on model bias and how that propagates in this work.

Lastly, the main conclusion the authors draw is that the number of SNS events rises until a global warming level of 6 is reached where it stabilizes, even though few simulations reach such high temperatures. It is a little unclear how to interpret this rise since there are less simulations. We would recommend rephrasing this conclusion such that it is better supported by the previous discussion.

Reply: In the region between 4 and 6 C, the number of models quickly drops from ~54 to ~10. We propose to reword the conclusion statement to: “The frequency of SNS events increases steeply at global warming levels below 2 C and remains relatively high between 2–4 C. While the frequency of SNS rises further between 4–6C this is based on a rapidly decreasing number of models (dropping from ~54 to ~10 in this interval), which correspondingly decreases the robustness of the inferred frequency.”

We implemented this rewording directly into the text.

Specific comments

The authors clearly describe how they distinguish between the terms “abrupt changes/shifts” and “SNS”. They also include a short discussion about the potential harmful effects of the tipping points concept. The strength of the statements regarding the

tipping point controversy does not reflect the content of the paper. We believe it is important to discuss the distinction between tipping points and abrupt shifts, but the way it is currently framed might distract from the goal of the paper.

Reply: We adjusted the introduction section accordingly and removed some of the strong statements.

The authors mention in the introduction that they search for events that are “truly surprising”. What is this exactly according to the authors?

Reply: As detailed in our response to the distinction between abrupt and gradual changes what we mean is that especially state changes in transient climate runs with changing forcing are trivial and we need to develop more stringent criteria to separate state changes from state transitions that involve a large additional effect from internal feedbacks. If this is the case, they can no longer simply be explained by the change in forcing. That is what we mean by “truly surprising”. This will be clarified in the text.

We’ve implemented this directly into the text near line 54.

It is generally well-argued in the introduction that the goal is to detect large and abrupt changes in the data. However, the authors also argue they want to limit the total detected amount of these events. What is the reasoning behind that? Instead of being guided by the quantity of SNS in the data, the goal now seems to only find the largest changes instead of all large events. This needs more justification.

Reply: We will rephrase this and argue we will be extracting the most extreme cases.

We have removed the related wording near line 54.

In the introduction, the authors reference the use of machine learning (line 52). How did the authors make use of machine learning? In the methods section, there is no reference to a machine learning method.

Reply: We added a reference for the used method, which is an unsupervised clustering algorithm. Specifically on line 125, we add the sentence:

We employ the unsupervised clustering algorithm HDBSCAN [reference] to cluster grid cells that surpass the threshold, forming continuous regions.

We added the reference on line 125.

The paragraph at line 85 lists all scanned variables. Why is only one atmospheric variable used? In the introduction it seems that atmospheric variables are also a focus point (also at line 500), which does not come back strongly in the rest of the manuscript.

Reply: One could argue that hfds is as well an atmospheric as ocean variable, but we will rephrase. We are not interested here in atmospheric variables per se, which would ask for a different approach as the atmosphere is a far more noisy system as the ocean and sea-ice and is much more dominated by shorter timescales than 10-year or longer; the focus of this catalogue. What we mean is that the focus is on the interaction between ocean-atmosphere and sea-ice and the variables involved, hence our focus on atmospheric surface temperature and heat fluxes, although we also include precipitation whenever relevant.

We added “We focused on these ocean, sea-ice, and atmospheric variables **involved in the coupling to ocean and sea ice** as they were the largest contributor to the SNS in Drijfhout et al. (2015),...”

Why do the authors combine historical data with SSP scenarios? Is this to gather enough statistics for e.g. the Diptest? If so, please mention this in the methods.

Reply: We combine the historical and scenario datasets indeed to increase the effectiveness of the metrics used in the 6 classification criteria. The diptest is the most prominent and clearest example. Other metrics also benefit from the full time series from 1850-2100 to gain statistics, e.g. when the standard-deviations are calculated and utilized as a proxy for variability.

We added this wording near line 103.

Why is global warming calculated with respect to the average temperature from 1850-1880 instead of the preindustrial temperatures from preindustrial control simulations?

Reply: In practice, because we had the time-series organized by historical+scenario, which meant that we had the values for 1850-1880 readily available. Nevertheless, we checked that for the models with an SNS, the difference between the last 30 years of the pi-control and 1850-1880: $\text{mean}(1850-1880) - \text{mean}(1820-1850) = 0.03 \pm 0.15 \text{ C}$, which is a clearly negligible for the results presented here.

We did not further incorporate this in the manuscript.

Around line 95 the authors mention they only look at yearly averages (except for mixed-layer depth). Some other variables, like sea ice extent, also depend heavily on the season. Summer and winter sea ice likely disappears at different forcing levels. Could the authors analyze summer and winter sea ice separately? By averaging year-round, abrupt changes in summer sea ice are likely missed.

Reply: In principle, we aimed to use yearly averages where possible, and, also for sea-ice, we do find abrupt shifts with this metric. For mixed layer depth we choose to make an exception and use the maximal value of the year since that is most relevant for deep convection. Note that yearly-averaged sea-ice cover is dominated by winter sea-ice and as we focus on disappearance it is just the winter sea-ice that has to collapse to near zero in our analysis. Summer sea-ice disappears much faster, but this is much more a reversible and linear decrease than winter-sea ice disappearance. We nevertheless agree that the seasonal aspect of sea-ice disappearance is worth further investigating, which we include in a follow-up paper, currently being drafted.

We did not add specific wording to the manuscript for this point.

The category of each detected SNS is often denoted by a letter (either lowercase or uppercase). This is difficult to keep track of. We suggest writing it out instead throughout the whole manuscript since it is difficult to remember every category (e.g. “Abrupt shift in NH sea ice” instead of “category a”).

Reply: We will do both, keeping both the category-labels and descriptions.

We’ve added throughout the manuscript. Naturally, the labels like “category a” still exists but are now written out much more often.

In figure 3, the abbreviations are not yet explained. The figure shows different locations for abrupt changes versus state transitions (for example, MLD and sea ice). Could this be clarified? Furthermore, the order in Fig 3 does not correspond to the order in which the systems are discussed in the results section. It would help to align this for clarity.

Reply:

- We will add the abbreviations to the caption.
- We ordered the sections by physical processes, grouping the northern- and southern- hemispheres close together when related processes are discussed. Nevertheless, we will implement the order consistent with figure 3.

We added the abbreviations to the caption, we also changed the order of the sections.

It is not clear what the difference is between section 3.1 and section 3.2. According to the formal criteria they are indeed divided into different categories, but are they really physically different from each other? When looking at the time series in Figure A3, the loss of sea ice is also abrupt, even though they are treated as state transitions instead of abrupt shifts. How big is the overlap between models in sections 3.1 and 3.2?

Reply: See our comment on the same point raised in the methods section. We will restructure section 3 accordingly.

We have implemented the reply above.

At line 199-200, the authors mention that the thresholds are reached earlier in CMIP6 with reference to Figure 5. However, this figure does not relate to this statement. It would be good to mention the new temperature range in CMIP6 for this comparison since this is not explicitly mentioned.

Reply: We agree with the reviewer that we should also mention the numbers here. Currently it is in our discussion, but we will also use it here. We indeed do not provide the right reference (Figure 5 is just one example of a category "A"). It should be figure 17.

We have added the numbers and the right reference.

Line 258: Over what region is this temperature impact measured? Over the area where the mixed layer collapses?

Reply: Yes, over the area of mixed-layer collapse.

We did not further elaborate in the manuscript on this point

Line 262: Please add an explicit reference with whom the authors agree.

Reply: Swingedouw et al. (2021). We rephrase the text accordingly.

We have incorporated the reference.

3.4 and 3.5: In both sections 3.4 and 3.5, changes in the subpolar gyre are discussed. In the first paragraph of 3.4, the authors explain that they do not find any abrupt shifts in SPG convection, but later they do discuss such changes. This is confusing and requires clarification. Furthermore, the comparison with the results of Swingedouw et al. (2021) is framed in 3.4 as if the results do not match, while in 3.5 many of the same models are found to exhibit SNS, only as state transitions instead of abrupt changes (see also major comment 2). Because of the large differences between methods and definitions, we suggest that the authors do not make this statement as strongly. Especially regarding the large area threshold used, smaller scale abrupt shifts (on the order of e.g. the Labrador sea) cannot be found, making a precise comparison nearly impossible. Why did the authors not consider a smaller area threshold for this system, given the scale of the processes relevant for convection?

Reply:

- We mean that, according to our criteria, the mixed-layer changes do not classify as abrupt, but do classify as state transitions. We will clarify this in the text.
- The region-finding method is not fixed to one region (such as the Labrador sea). With the flexible region-finding algorithm we find regions that span more than just the Labrador Sea. The 10^6 km² criterion is already reasonably low, given the flexibility that we added to the algorithms to find arbitrary shapes. Two out of three of the abrupt (category b) models also extend into the Labrador sea.

We have mostly rewritten these sections, and together with the further extended explanation of the region finding methods, these comments should now be addressed in the updated manuscript.

Line 291: What are these larger regions? How are these obtained?

Reply: What we mean is that when we plot the changes over the whole SPG, regions of maximum SSS change do not coincide with regions of maximum mixed-layer depth change, indicating a role for temperature change and changes (differences) in deep densities.

In the rewriting of this section, we removed this sentence.

Line 356: In the text, a comparison is made between this manuscript and other articles with reference to Figure 11. However, this figure does not contain any comparisons; this should be added to the figure.

Reply: we plan a new figure that compares the related categories to McKay. This will be placed in the section together with the discussion on figures 16 and 17.

We added figure a new figure 11, which does make this comparison/

Line 397: Can a reference be added?

Reply: Yes, Stewart et al. (2023).

Line 403-404. Why are the transitions associated with model bias? In what sense does the double ICTS become less pronounced?

Reply: The double ITCZ is a model phenomenon not observed in reality, thus a model bias. The ITCZ becomes less pronounced because the double-banded structure in rainfall disappears favoring a single-banded structure. We see this transition as a model artifact as the double-banded structure should not be there at the first place.

We did not implement changes to specifically address this question.

In the first paragraph of the discussion, it is mentioned that there is a large increase in number of SNS between this assessment and Drijfhout et al. (2015). How is this statement supported? Both assessments used different methodologies and criteria. Using an automated algorithm could likely have increased the number of detected SNS. It would be interesting to see how many events would be detected if some of the CMIP5 variables were re-analyzed with the new methodology (although this would require substantial work and therefore is not a request to the authors).

Reply: We certainly do not claim the finding of more cases is due to CMIP6 model improvement over CMIP5 and is likely to(also) impacted by the changed methodology. We will rephrase this part to avoid the confusion flagged by the reviewer.

We removed this wording and made the discussion more balanced from lines 551 onwards.

Line 475: What is meant by the small bump? It is not clear where in the figure this is visible (there is however a small bump at 0 degrees?).

Reply: We agree with the referee that we should change this to the bump at near 0 degrees.

We have rectified this in the manuscript.

In the discussion, global warming thresholds are given for each category of SNS. Could this be summarized in a table?

Reply: We agree a table is useful here, which we will provide.

This table is new table 1.

Figure 17: What is meant by “maximally changing”? Why does the temperature of the bottom figure range from 0 to 5, while the upper one ranges from 0 to 17?

Reply: We will rephrase this to CDF of SNS occurrence (as function of temperature). We will also elaborate in the caption. The bottom panel is dedicated to the abrupt shifts, occurring at lower warming levels than the state transitions. Hence the different ranges.

We extended the meaning of this each time it is mentioned in the text.

The authors mention at the end of the introduction that they will compare their results to the assessment of Terpstra et al. (2025). However, in the discussion they do not compare much of the results apart from stating that the CMIP6 thresholds are lower than in CMIP5. The authors could also make a comparison between frequency/thresholds between this manuscript and Terpstra et al. (2025). Even though both use different scenarios, and indeed one-on-one comparison is not possible, would it be possible to go into a bit more detail in the comparison?

Reply: We will indeed reserve more space and elaborate a bit more on the comparison with the Terpstra paper in the Discussion section, also as we moved methodological issues to section 2 as requested by the reviewer.

The discussion is added on lines 613 onwards.

Although not strictly necessary, it would be very interesting to have figures with both the time series and spatial extent (like e.g. figures 4, 5, 6) available for all SNS in an online supplement/repository if it does not require too much effort from the authors.

Reply: We agree, also, given the request from referee 2, we will provide an online github repository with figures and data on each case. We discussed with the editor and the editor agreed that this would be a useful addition.

We have made this publicly available, <https://doi.org/10.5281/zenodo.17867115>, throughout the manuscript, this is cited.

Technical corrections

Reply: We thank the reviewer for finding these technical issues. Unless otherwise stated, we will implement the corrections below.

Add consistent numbering format (e.g. line 324 “seven” and “3”)

Line 22: “Ref” should be the actual reference

Line 56: Remove extra brackets around citation

Line 68: Sentence not starting with a capital letter.

Lines 88-89: clarify what the difference is between msftyz and msftmz since now they both have the same full name (or state they are the same)

Line 96: maybe mention nominal resolution explicitly of Gaussian N90 grid for non-expert readers.

Line 99: TAS is written upper case here, but with lower case at line 90.

Line 131: “Generally, i and v_i are generic criteria”. These are types/categories, not criteria.

Reply: They are criteria, but we will rewrite how to discuss them as we explained above.

Line 132 – 133: missing words in this sentence

Line 173: What does “Its similarity” point to? The abrupt change or abrupt shift in the previous sentence?

Line 188: This only occurs in one model, so remove “typically”

Line 206: The abbreviation of ppt is mentioned here but afterwards it is only used in the figures. Maybe this sentence can be removed.

Line 252: Suggestion: SSS decreasing the surface density → freshening

Line 254-256: Check the grammar of this sentence. What do the authors mean exactly by “in terms of atmospheric cooling”?

Line 266-267: NorESM2-MM and NorESM2-LM are mentioned in these two lines. Shouldn't these both be NorESM2-MM?

Line 272: remove comma between number and unit.

Line 284: “unlikely whether” is not clear, maybe rephrase this sentence.

Line 291: “looking to” → “looking at”

Lines 308-313: There is some repetition in mentioned locations of the transitions in different models.

Line 412: “...also work in Nature” → “...also are present in nature”

Page 4, footnote 1: “that” → “than”

In figure 12, the regions of SNS are shown for both SST and SSH. Is it correct that for both variables the regions are exactly the same?

Reply: We only observe the SNS in SST, for the reasons explained in the caption, SSH in this model was not checked for SNS.

Figure 16: Unit of degree Celsius is not displayed correctly in the pdf.

We have implemented these corrections.

Reviewer 2

The manuscript submitted by Angevaar and Drijfhout propose a cataloguing protocol for identifying strong nonlinear surprises (SNS) in the CMIP6 database. The manuscript describes the method relatively simply (most of the details are in the supplementary material) and then describes successively the 11 identified categories. This work is clearly useful to the community, as it proposes an objective and unified tool to work on nonlinear events of the climate system, and gives a first overview of the main findings. Yet, because of the global and general approach, it necessarily remains a bit too general, and has difficulties to escape the classical caveat of a relatively lengthy qualitative description from which the reader can finish with a clear idea (me at least). Other than that, the paper is very well written. Figures are clear and clean. They are sometimes perhaps a bit too simple, see my comments below.

Because I think this paper may eventually be an important milestone for the TP community, I recommend major revisions following some of the suggestions below.

Reply: We thank the reviewers for their kind words and constructive comments on our manuscript.

1 General comment on the SNS 2 types of SNS are introduced (abstract, introduction and methodology). Do the authors consider these 2 are exhaustive nonlinear surprises? If not, what other cases may be considered? How were these selected? Of yes, how and why??

Related to that: how do you ensure that “Slower transitions” are abrupt or Decadal, as claimed in the abstract?

Reply: In the abstract we state that we focus on abrupt changes over decadal timescales, and slower transitions that develop over several decades. We do not claim that these two types of SNS are the only ones that occur. We will clarify this in the introduction, as well as linking to the previous work of Drijfhout 2015, which formed the starting point for the selection of types of SNS.

We have extended the discussion on this topic in the manuscript on several places, most clearly in subsection 2.3.

2 General comment on the main message of the text, that also appears at the beginning of the discussion section to the end of the abstract: a bit catastrophist. How many events per simulated year? how dependent are these conclusions to the detection tool itself? (did you test it in CMIP5?) How realistic are these findings, given the models' biases in terms of mean state, and spread in terms of climate sensitivity? How to estimate the risk?.

Reply: The reviewer poses here several questions which are partly impossible to answer with the present state of knowledge. We will rephrase the text on these points, delete, e.g. the last sentence of the abstract and nuance our statement of more SNS found by emphasizing the role of the new detection tool used here. We will also emphasize that these events simulated by models are in that sense uncertain that the models contain several biases that will affect occurrence of those events, although there is some indication that at least for a few tipping elements models may be too stable.

We've implemented the removal mentioned above, and increased the discussion on the other points in the text where appropriate.

3 Cataloguing: it is interesting that specific events in specific models are described. Yet although some of them are illustrated, most of the time by a single timeseries, many others are not, and in general, the behavior of other related variables discussed in the text is not shown. This makes the various paragraphs describing the various SNS difficult to read and to be convincing, the reader supposed to believe the authors and what they write. models I am not sure how to handle that as once again, I believe that a bit of in depth discussion of each of the SNS is interesting. Otherwise we would have not more than a methodology paper. Perhaps adding a few related variables on the figures illustrating the various cases would help?

Reply:

We will address this in two ways:

- The other reviewer had similar remarks in this direction. To avoid making the already lengthy manuscript even more bulky, we propose to provide an online public repository of all the cases that we have, such that readers can find a complete set of SNS we associate with a given category.
- Point the reader to the material in the appendix, for example table A2-A4 and figures A2-3, especially A3.
- Since we provided the online material (<https://doi.org/10.5281/zenodo.17826419>) we deemed table A4 to be superfluous and removed it from the manuscript.

In short, we added material in <https://doi.org/10.5281/zenodo.17867115>, and we removed table A4. We increased the number of forward references to Table A2, A3 and <https://doi.org/10.5281/zenodo.17867115>. We explicitly enumerate the different sources/examples/all available SNS in line 216-225.

4 Also it is not easy for the reader to follow which models show which type of SNS and whether some models seem to show more or a cascade of SNS. The authors sometimes discuss such cascades here and there in the text (SSS related to MD for example). But this is not systematic and not recalled anywhere synthetically. Would it be possible to improve that? Something in the line of Fig. 16 and the related comment in the discussion section, but for models perhaps?

Reply: We agree with the referee on this point, we will table and discuss possible cascades and give this extra attention.

These have been added to the manuscript in the discussion of those cases where it is relevant.

Minor comments

52: I don't see much machine learning in the detection protocol. Please rephrase.

Reply: We will add a reference for the used method, which is an unsupervised clustering algorithm. Specifically on old line 106, we add the sentence: We employ the unsupervised clustering algorithm HDBSCAN [reference] to cluster grid cells that surpass the threshold, forming continuous regions.

We incorporated the reference and the wording as above.

55-56 "truly": I don't see any assessment of realism or robustness of the findings. Please clarify what you mean by "truly" here

Reply: What we mean with "truly surprising" is that state changes in transient climate runs with changing forcing are trivial and we need to develop more stringent criteria to separate state changes from state transitions that involve a large additional effect from internal

feedbacks. If this is the case, they can no longer simply be explained by the change in forcing. That is what we mean with “truly surprising”. This will be clarified in the text.

We have elaborated this in line 54.

69: capital C missing (cataloguing)

Reply: We have fixed this.

96: historical run (why singular?) combined with the scenarios: this step requires a bit more detail. How did you pair members? Could the protocol find some artificial SNS because of the way scenario members are generated in 2014 and paired with historical members?

Reply: We combine the historical and scenario datasets indeed to increase the effectiveness of the metrics used in the 6 classification criteria. The diptest is the most prominent and clearest example. Other metrics also benefit from the full time series from 1850-2100 to gain statistics, e.g. when the standard-deviations are calculated and utilized as a proxy for variability. We paired the members based on their available properties in the dataset-metadata. We verified that no major discontinuities occur at the historical–scenario boundary (2014) and exclude pairs that do have such transitions as they indicate a mismatch between historical and scenario runs. We will further elaborate this in the text.

We added much more detail in section 2.1

Around l. 105 and following: regions selection:

I suggest using bullet points to describe the four approach for more clarity in the reading

Reply: We will implement this, also the other reviewer made suggestions to improve clarity in this section.

We have implemented this.

A bit of explanation of why criteria 2-4 (why may look a bit redundant or similar at first sight) would be welcome

Reply: We will add some additional explanation here.

We’ve dedicated much more discussion on this in section 2.3.

125 and following. It is a pity that 6 criteria have to be listed with 3 specific lines for 3 variables. I suggest to rank and perhaps to change a bit the order in which the criteria are presented: (iii) should appear as a sub-criteria of (ii) I guess and (iv) and (v) as sub-criteria of (vi) no? Or list (vi) before (iv) and (v) so that the list goes from the more general to the more specific. Also, I find that (iv) and (v) are not really justified as (vi) does not say anything about the structure of the data. The fact of the data have a different structure is of the developer's business, and should not appear in this list of general criteria I think.

Clarifying this list either through reducing it or ranking it would strongly enhance the impact of the method.

Reply: While criteria iii, iv, v are tailor-made to describe SNS for 3 different variables, the other three are not. The difference between criteria i and ii is only determined by the timing of the abrupt event relative to length of the timeseries. Criteria i can never be met when abrupt changes occur at the end of the timeseries, and for these cases criteria ii are developed. We will clarify this in the text and rename and reorder the criteria. We will further detail and motivate while, besides criteria vi, we have developed criteria iii, iv and v. They are not simple subcategories of vi, as the differences imply more than just an increase of the thresholds used in vi. We will also rename the A-I categories to comply with the order of discussion in section 3 and treat abrupt changes and state transitions for each category more closely together.

As with the previous point, we've dedicated much more discussion on this in section 2.3.

151: how is this normalization performed?

Reply: The other reviewer has suggested to use the data underlying figure 17 in this figure, which we will implement and discuss accordingly in this section.

We removed the old version, normalization is now consistent with figure 17.

156 and following: I would rewrite the number of events concerned by each category in the title of each subsection (A= etc)

Reply: We will implement this suggestion by the reviewer.

We have added this to the subsection titles.

161: polar amplification -> perhaps cite XX to be more complete?

Reply: Yes, we will cite e.g., Davy and Griewank (2023).

We've added the citation.

164-165: please clarify this sentence and explain better which types of events are in category A and which are in a. It is the first time these letters are mentioned in the text I think. Furthermore, the beginning of the paragraph discussed the cases in general and I am not sure why CanESM5 is a specificity here (see l 165 "In this model").

Reply: As per suggestion by Ref 1, we will (especially here) use more verbose category names, so that it's clearer. We should clarify how we chose example figure 4 (CanESM5).

We have restructured these lines, with changes to the title as well as giving the category names more clearly to address this point.

L 182: logical link between this "outlier" and what precedes is unclear to me. The previous sentence was describing positive feedbacks favoring sea ice melt. The 2 examples that follow rather concern impacts of sea ice SNS don't they?

Reply: We will rephrase and make this a separate paragraph. What we mean is that sea-ice change itself does not classify with our criteria as an SNS, but the surface air temperature does, while its change is driven by the sea-ice change.

We've implemented the restructuring/rephrasing as in the reply.

200 and following: do you want to speak of SNS cascades?

Reply: We will discuss and table cascades in the revised paper for those cases where SNS-cascades are relevant that are discussed after line 200.

We've implanted this near lines 256.

215 and following (but applying to most cases): is there something in the way sea ice is represented, or mean state bias, that could explain the specific behavior of these few models?

Reply: There is not a simple answer to that, in general open ocean convection is not much observed in the Southern Ocean and appears a result of biased low-resolution models, also not capturing dense water formation on the shelves and their downward and equatorward propagation in the open ocean. On the other hand, the Weddell polynia that intermittently pops up in the observation does support that open ocean convection could play a role. All climate models have severe biases in the Southern Ocean and ocean stratification, a link between bias and the genesis of new open ocean convection sites in climate models is not easily made. We will give this item more discussion/consideration in the revised manuscript.

This extended discussion is now added to section 3.6.

227 and following: I wonder if this example is well placed. Shouldn't it appear rather in a section focusing on MLD SNS?

Reply: 2.1. It is indeed an MLD SNS but clearly driven by a sea-ice SNS. This is a clear example of a cascade and has been discussed as such in the revised manuscript.

We elaborated on the connection between these processes in the text L431-435.

245: wasn't criteria (v) defined specifically for this purpose?

Reply: We shall clarify that this section is for the abrupt-type, while criterion v is aimed at gradual mixed-layer depth state transitions.

We've clarified the distinction between abrupt and state transitions.

253-254: this allusion to Swingedouw et al 2021 is largely a repetition of what precedes I think. Remove the sentence?

Reply: We agree that we can remove this sentence.

We've implemented this.

L 257: not clear to me why the GISS model is suddenly specifically cited here.

Reply: Because in Swingedouw et al (2021) and other papers MLD SNS are detected by temperature changes over the convection area. With that method you miss this model. We will clarify this in the text.

We've implemented this.

310-311 "this site is not particularly known for deep convection": but how is it in this model? This relates to my general comment on the models mean state.

Reply: This is closely related to the concerns raised by the referee for line 215. Our reasoning there also allies here, and we indeed will give model bias more attention for these cases.

We've rewritten this paragraph to reflect this, near line 451.

Around l. 315: I acknowledge the discussion on the models mean state systematic biases. I think this is very useful and could appear on other instances in the manuscript

321: Comparison to Swingedouw et al 2021: we don't really know who to believe. All this is relative to the detection tool

Reply: Indeed, that are no first principles from which an SNS can be defined, so detection is always subject to the set of criteria adopted, and when those differ, the SNS cases you find also differ. We will stress this point more in the text (p.s. there is no reference to Swingedouw at line 321, but in general we can say a bit more about this point when comparing our results with Swingedouw et al.)

We've extensively reworked this section to address these concerns.

325: add a "probably" in this sentence

Reply: We will rephrase this sentence. The Amoc did not yet collapse or was not provided.

We've rephrased

330 is (or was) missing before identified

Reply: We will add this

We've added "is"

346= "in CMIP6" Are you really able to generalize that much? "in some CMIP6 models" rather?

Reply: We will clarify with "in the CMIP6 multimodel mean compared to cmip5 multimodel mean".

We've added this wording instead.

L 400 and following: the monsoon system is often described as a high potential for TP. Could the ITCZ transition be a precursor of linked to such TP?

Reply: This could well be, but this catalogue does not really address possibly much faster monsoon changes, or such changes need other variables than yearly means. Monsoon tipping and abrupt changes in extreme weather statistics are deferred to future work.

We've made it clear in lines 179-185 that these types are beyond the scope of this work.

L 410: is this bias only present in 1 member of the IPSL model? If not, why would the transition be linked to biases then?

Reply: This SNS is seen in one member because it develops after year 2100 and only one member is extended.

We did not alter the text in relation to this comment.

L 427 and following: given all the differences described here, plus the argument on the size of the region (which maybe should be added here): can one really speak of an increase? I think this is a bit overstated

Reply: We do not (intent to) claim the finding of more cases is due to CMIP6 model improvement over CMIP5 and is likely to(also) impacted by the changed methodology. We will rephrase this part to avoid any confusion.

We rephrased this part of the discussion accordingly.

L 441 and following: I suggest repeating which physical component of the climate system the lettered categories relate to.

Reply: We will implement this suggestion

We've implemented this.

L493-494: sorry I don't understand (or don't know) this notation, please explain.

Reply: This notation gives the uncertainty for a non-symmetric error, however, we will use a clearer notation and put these numbers in a table with a clear caption on their properties.

We added table 1 instead.