Response to minor revisions

We'd like to thank both reviewers again for their time and considered comments on both drafts of this paper, we feel they have helped significantly strengthen our presentation of the work. Here we respond to the individual minor revisions requested in the second round of review. A small number of typos have also been corrected.

Reviewer 1

1. I appreciate the discussion the authors have added concerning the issue of zonal vs blocked flows in Sect. 3.1. As a further suggestion, I would recommend streamlining the terminology they use. Right now, they adopt both the zonal/blocked and cyclonic/anticyclonic terminology in the same paragraph, which may confuse some readers not familiar with this specific topic.

The language in section 3.1 has now been streamlined to use the language of blocked and zonal flow states almost exclusively.

2. I have missed where the authors discuss how their stability metric could be used to study nonstationary regimes. Please ensure that this point is indeed included in the text somewhere.

This is on Line 734: 'The stability metric introduced in this paper could in fact be a useful tool in future studies focusing on ocean-driven regime variability, or indeed in analysing non-stationary regime frameworks.'

3. Original comment on ll. 68-69. What I meant to say here is that picking "things" that are well-represented to evaluate models and using the fact that they are well-represented as an argument that they are good evaluation metrics comes with the risk of a circular argument. I think the authors now make a strong point for why their regimes are useful in a model evaluation context, but in general they should be careful in using the lack of spatial regime variability (to use the terminology from Fig. 1) as an argument to support the usefulness of their regimes. One may indeed argue that, if there are robust regimes in the "real world" that are poorly spatially reproduced by the models, they can still provide useful model evaluation information, as long as they are not forced into a "temporal variability" framework with the spatial discrepancy being ignored.

We appreciate both this and the original comment, which have helped us to frame and clarify the motivation and application of this work. We believe we now avoid any such circular reasoning in this draft, and more concretely detail the technical challenges posed by non-stationary regime patterns in the introduction, as well as the benefits of stable regimes in terms of clear relationships between model features and regime dynamics in section 5.

4. L. 475 "Th"
Corrected to 'The'.

5. Original comment on l. 408. This is a minor, and mostly linguistic point, and I am perhaps being too picky. I follow the authors' argument, but I still do not understand how this says something about the "correctness" of individual models. If I understand correctly what the figure is showing, the standard deviation here combines information from different models, so at most it says something about CMIP5 or CMIP6 as a bulk as opposed to information about individual models. It also gives us no information about how many models fall within 1 standard deviation of the mean. Perhaps a more precise statement could be: "how the individual models differ in representing the correct level of…"

*The figure shows the standard deviation **in time**, averaged across the models, not the standard deviation **between** models. We have added this small clarification to the text to avoid confusion. As we believe this to be the source of the disagreement on wording, we stand by the current phrasing: whether CMIP5 or CMIP6 models have the 'correct amount' of variability or not is indeed shown by the difference between the model and reanalysis curves in figure 12.*

Reviewer 2

141 is 'non-gaussian' actually referred to the phase space?

*No, strictly speaking it refers to the distribution of points within the phase space, and so now we now write '...produces a phase space **distribution** which is unambiguously non-Gaussian…' to be more precise.*

170 I presume members of the ERA family share some similarities and so do members of the 20CR family. Also some are coupled some are not, maybe provide just a bit of text to comment on this.

We now include a couple of sentences on this on line 171: 'Of these only CERA20C uses a coupled ocean-atmosphere model. Reanalyses produced by the same centre will share some similarities in the features of the assimilating model and in data-assimilation procedures, and so therefore are not totally independent'

Figure 9, caption , after c) put a capital letter.
Figure 13 It appears that the figure is cropped at the bottom
Figure 13 and Figure 14, the labels are not explained, some are intuitive but only if one navigates through the paper carefully. 'Driv' is unclear. Please make sure that the labels are clearly defined. Also in figure 15 Stdev10 is not very rigorous

The figure labels and captions have been clarified and corrected.

First sentence of section 6, 'The two most important aspects…' you may refine the choice of words here
Changed to 'two important aspects'.