Reviewer 2

Mignot et al. use self organising maps to evaluate the behaviour of a multi-model ensemble in the Senegalo-Mauritanian upwelling region, with the aim of producing accurate projections of future climate changes in the region. Their algorithm aims to select models that yield a specific desired quantity - in this case, a multi model mean. They then project the selected models through the future to assess changes in the region.

There is clearly a great deal of potential in the technical work in this paper. The idea of using Self Organising Maps as a dimension reduction and interpretation technique is a good one, and appears to work very well. It can clearly add a great deal of value to the analysis of a large multi-model ensemble in this region. However, I feel a degree of restructuring, clarification of the aims of the paper, and editing for overall clarity is required before the scientific content can be properly assessed.

We would like to thank the reviewer for his/her careful reading of the manuscript and his/her constructive and challenging comments. We have restructured the manuscript as suggested, and clarified the methodology as much as we could. This has greatly improved the manuscript. We have also paid specific attention to editing issues for which we apologize. We detail below on all the modifications that have been implemented the text.

I feel the paper would most benefit from restructuring so that the objectives, details and then method of assessment of the algorithm were more clearly laid out earlier in the paper, and the reader were more carefully led through that process. As it stands, intense technical detail follows very broad overview statements, and important details about the analysis are left until later in the paper, so the reader is left confused and searching for appropriate context into which to place technical detail. Some choices in the analysis feel arbitrary, and it is unclear whether this is because they are indeed arbitrary, or that they are inadequately described.

The most obvious candidate for restructuring is the start of section 3, describing the methods used for classification of the models. This section dives straight into a detailed description of self organising maps (SOMs), without a discussion of precisely what the algorithm aims to achieve, how that can be assessed, and why SOMs were chosen as opposed to any other dimension reduction technique. As it is, the paper reads as “we decided to use SOMS and this is what you can do with them”, rather than “we are trying to solve a specific problem and here is how SOMs can help”. One suggestion would be to take the description of the methods from the start of section 6 (Discussion and Conclusion), expand upon it and place it at the start of section 3.

The SOMs appear to successfully cluster the model field into regions with different dynamics. This seems useful and interesting. How does it help solve the specific problem? I think it would be useful to set out near the beginning of the paper the exact strategy that will be used, and how to tell if it is successful or not.

Thank you for these detailed and constructive remarks. We restructured the paper by taking the remarks of the reviewer exposed in the above paragraphs into account. First, at the end of the introduction we now mention that our method is based on classification. At the beginning of section 3, we now justify the use of a classification method on the one hand and then choice of the SOM+HAC on the other hand.
Moreover, Sub-section 3.4 is now section 4. In fact the MCA method used in this section is different from the SOM and quite new in geophysics; it therefore deserves a dedicated section in which we give more details on the functioning of the MCA.
We also cut some long paragraphs in smaller paragraphs in order to facilitate the understanding of the text.

Regarding the relevance of the SOM in particular, note that this question was also raised by reviewer 1. We are grateful to both reviewers for requiring this classification. As answered above, the SOM model has been used to determine a vector quantization of the dataset: i.e. to determine referent vectors that are a representative summary of the learning dataset. The vector quantization compresses the total database into a quite small (with respect to the size of the database) number of referent vectors such as each data is not too different of its nearest referent according to a distance (The Euclidean distance in the present case). The exact number of referents, that is the number of neurons, does not really matter because this number will be reduced by the HAC. Doing so allows us to take the non-linearities of the dataset into account in the analysis. This explanation now appears at the beginning of section 3.

For example, It seems clear that the assessment algorithm (starting line 232) can be used to rank the models in terms of their closeness to observations and dynamics in particular regions.

One downside however, is that it does not give the modeller an intuition into how far the model is from “good” behaviour in absolute terms. We simply get an averaged “skill score” from 28% to 79%, but without an idea of how this might relate to more traditional measures of skill. So how close is the best model and how far is the worst model from reality? We have only a score (useful as that is) to guide us.

In this paper we give a global index that is the mean of 7 indices associated with the seven Region-clusters. This mean index shows the ability of a given model to represent the global area. But for each Model and each Region-Cluster, we give the ratio that represents how that model represents that Region-cluster. These indices are visible on the colorbar of figure 3.So each modelling group can evaluate how its numerical schemes represent the dynamic of the observations.
We also give a visual interpretation of the fitting of the different CMIP5 models with respect to observations in Figure 4. The best models are the closest to the observations with respect to the khi2 distance.
In general, we agree with the reviewer that there is a huge amount of information to take out from this method in general and from Fig. 3 and 4 in particular. Each modelling group could use this information to better understand the reasons for their model’s difference to observation. And specialists of the Senegal-Mauritania region can use it to better understand the reasons for the

weak representation of this region in climate models. In this paper we propose an illustration of the method application. After publication, the method will be free of use for more various applications. We thus added a sentence in the text as well as in the conclusion to stress that point. We are grateful to the reviewer for having stressed that point.

The paper makes the claim that it offers an objective method for the assessment of the behaviour of models with regards historical observations. I struggle to accept this, given the number of subjective choices made with regards to the way the analysis is conducted. Subjective judgements will always need to be made in the analysis of climate model output - this is inevitable, and perfectly reasonable as long as labelled as such. The paper only examines a subset of model fields for example, and a subjective choice as to which of those fields to select has been made.

The reviewer is right in several aspects:

- The study focuses on the ability of CMIP5 models to reproduce the ocean seasonal variability in the Senegalo-Mauritanian upwelling region only. The models which represent this region at best do not necessarily represent other regions at best. Our study is not devoted to the comparison of the CMIP5 models in general but to their ability to reproduce the Senegal-Mauritanian upwelling area only. We now mention that point explicitly in the conclusion. Furthermore, a full representation of the geophysical phenomenon should involve more variables, as explained in Sylla et al (2019) and this first study is more a test-case than a full analysis.

- Concerning the use of statistics, we are aware that statistics are only a support to understand or interpret what is hidden in the dataset, mainly if the number of data and observations is large. This is the present case because we want to compare 47 different models with respect to a set of observations, with a focus on the dynamical behavior of each dataset (multidimensional analysis in a 12-dimensional space, the monthly SST anomalies). We built a method to solve that problem. Other methods could possibly lead to different results depending of what we looked for: The number of possible statistical studies is huge. In that sense, the choice of the method contains some subjectivity.

Nevertheless, the method we propose is not subjective ; it allows to rank the models according to the reduction of information we made (the seven dynamical region-clusters, after the SOM). This is in some way a classical problem in geophysics, where we need to classify, organize the information. Here it is done using relatively novel statistical tools (SOM+HAC). Finally, the MCA is a qualitative (but rational) method to summarize and visualize on a graphic the “similarities” of the models, the observations and the region-clusters.

For these reasons, we consider that the title is not misleading: our method is a step *towards* an objective assessment. Yet, considering the reviewer’s remark, we have modified the sentence claiming for an “objective method” in the abstract and this term is now better justified in the conclusion section.

A core problem that needs to be addressed in the paper can be illustrated by considering the section starting on line 285:
“As indicated in the introduction, the main objective of the methodology is to select an ensemble of models that represents at best the upwelling behavior with respect to the observations and to use this ensemble to predict the impact of climate change in the Senegalo-Mauritanian upwelling with some confidence. The problem is now to determine a subset of models that can adequately represent the observations, as the number of models is small enough we choose to cluster them by HAC according to their projections onto the seven axes provided by the MCA, and select the optimal jump in the hierarchical tree (Jain and Dubes, 1998).”
I cannot see a description of what it means for a subset of models to “adequately represent the observations”.

We agree with the reviewer that the phrase “adequately represent the observations” is misleading.
Through MAC+HAC, we group the models into Model-clusters, using the khi2 distance, according to their proximity to the observations and their internal similarity. Model group 4 appears as the one closest to observations with respect to that distance. In Figure 4, we see the projection of the individual models on the first two axes of the MCA. The fact that only two axes are shown here can introduce some bias in the visualization and this figure must be considered with some caution. We associated a multi-model with the Model-group 4 (close to the observations),whose outputs are the mean of the outputs of the models constituting the Model-group 4. We agree with the fact that we cannot prove that this is the best. An exhaustive research in order to find the best subset is nevertheless prohibitive due to the enormous number of possible combinations. The phrase “a subset of models that can adequately represent the observations” was changed into “a subset of models which has a better skill than Model-All”.

I also cannot see an adequate description for what the “optimal jump in the hierarchical tree” of Jain and Dubes (1998) is, or what it might mean for the ensemble members. The clustering of the models in figure 4 looks reasonable by eye, but there are a large number of other ways that the models could be clustered that might be equally as reasonable. The authors claim that their algorithm selects a number of ensemble members that best represent an ensemble mean. I don’t believe that they provide sufficient justification for why the ensemble mean should be selected for, or that the ensemble members their algorithm selects members in a way that is superior to a subjective selection.

We recall that the HAC (hierarchical ascending clustering) is a bottom-up algorithm for dataset clustering. The key operation in hierarchical bottom-up clustering is to repeatedly combine the two nearest (according to a certain distance) clusters into a larger cluster. The HAC starts from individuals and combines them according to their similarity (with respect to the chosen distance) to obtain new clusters.The process is repeated up to get one cluster only (the full dataset). This algorithm is visualised by a tree-like diagram, the so-called connection tree : the

connections between the clusters are represented by the branches of the connection tree (see figure below) according to their proximities. Due to the bottom-up algorithm, the construction of clusters is therefore objective with respect to the chosen distance. The objects are finally categorized into a hierarchy similar to a tree-like diagram which is called a dendrogram (see figure). A major problem then arises: When do I stop combining clusters and consider that I have optimal clusters?
The problem is semi-qualitative. It depends on the dataset under study. Most of the time, it is a compromise between a sufficient number of clusters to explain the complexity of the dataset and a relatively small number of clusters in such a way that every cluster can be handled and explained.

In the present study, we decided to deepen the statistical aspect of the problem and to choose an "optimal" model according to the data provided by the MCA algorithm (the data are the rows of the matrix $\mathbf{R} = [Rm_i]$ representing the 7 component vector-skill of the models). The HAC clusters the models according to their similarities (based on the Khi2 distance). The HAC used in the MCA analysis yields the following dendrogram (not shown in the paper):

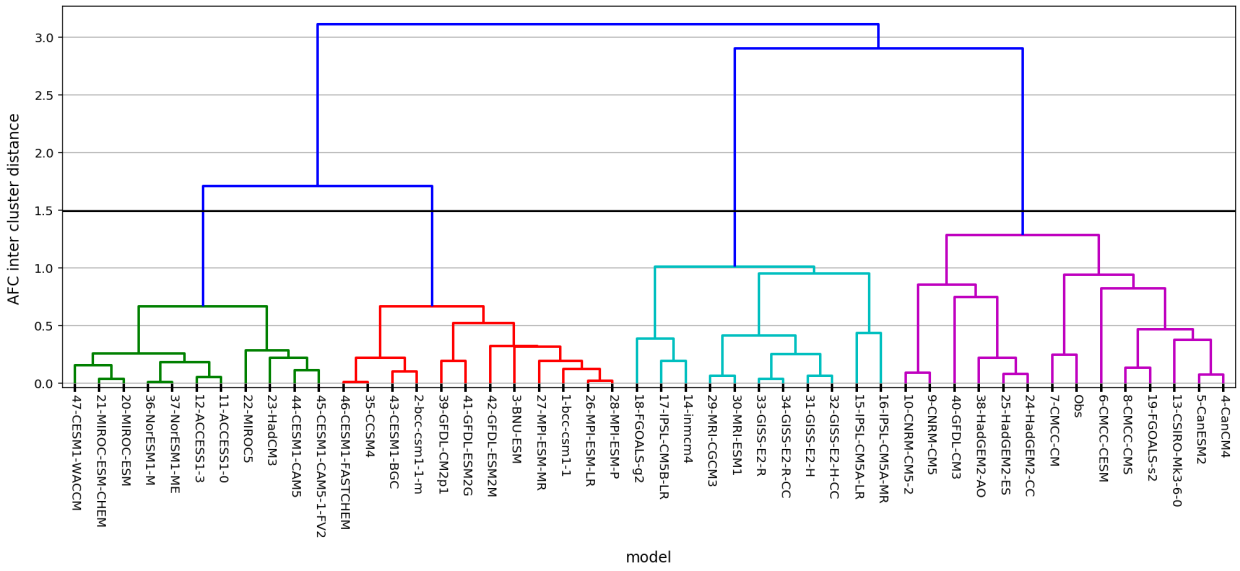

Figure: HAC dendrogram

On the horizontal line, we have displayed the 47 CMIP5 models, each model being associated with its 7 component skill-vector. As the dendrogram represents a hierarchy of clusters, the numbers on the y axis give the distance between two clusters; Clearly there is an optimal jump on this graph: for 4 clusters we obtain well separated Model-groups that are very different. The horizontal black line materializes this optimal jump on the figure (level 1.5 in the vertical axis).. The purpose of this explanation is to highlight the rationality of the selection. We reckon that there is subjectivity in the choice of the approach of the statistical tool, and also in the use of the geophysical knowledge of the region. But by themselves, these tools rely on rational criteria.

. This is presented as a “model weighting” paper, and while that might be possible with this algorithm, I do not believe that is where the strength of the analysis lies.

It was not our intention to present this study as a “model weighting” one. Although model weighting strategies are indeed presented in the introduction, the text only refers to model selection. We have carefully proofread the text so as to make sure to remove this misleading message
We indeed agree with the fact that we do not determine a weighted ensemble model but an ensemble model that better represents the observations than Model-all, which is the mean of all the CMIP5 models we have considered, and also better than the other Model-groups. Our model selection is based on the distance separating the models to the observations .
This combination is provided by the MCA which deals with the 7 component skill-vector associated with each model (and permits to determine a distance to the observation field also associated with a 7 component skill vector) which is more informative than the average skill which has one component only.
Our paper is a “model selection” one.

The paper would be better re-cast as a model analysis paper, using an interesting and useful algorithm to explore the dynamical deficiencies of the models in the region, and informing climate modellers of those deficiencies. I think if the authors wish it to be a model weighting paper, then more emphasis needs to be given to the meaning and justification of the weighting scheme. Further, the authors should develop placing the weighting scheme in the context of established work on the meaning of multi-model ensembles.

We agree with the reviewer that our methodology provides rich and objective information about climate models performance in a specific region. This is one outcome of our study (mainly section 3) and we have strengthened this message in the text and in the conclusion.
Nevertheless, We do not agree with the suggestion of the reviewer that the paper is a model analysis paper. Section 4 indeed provides a way (through the MCA) to use a 7 component skill vector to obtain an efficient combination of climate models leading to an efficient multi-model. (Efficient means that its skill is better than the one of Model-all).
Another question may arise, which is far beyond the objective of the present paper: Is model weighting the best strategy to obtain the most efficient multi-model? Or should we envisage statistical combinations based on multi-parameter analyzes as those developed in the present study?.
In our view, the major contribution of our paper can be summarized into the following sentences included in section 7 (discussion and conclusion):

“The extraction of information embedded in the vector-skill whose 7 components are the skills associated with the 7 sub-regions and the resulting efficient multi-model combination imply the use of advanced statistical tools such as the MCA. Moreover, the study of the vector skill also permits to separate information provided on large offshore ocean circulation from those

occurring in the upwelling region leading to diagnose the deficiencies of some climate models with respect to the modelling of physical processes. Another contribution of the MCA is the visualization of the 47 models and the observations on the plane constituted by the first two MCA axes, which represents 70% of the information embedded in the data. The similarities of the climate models with respect to the observations and the region-clusters are well evidenced. The 'mean' skill associated with each climate model and proposed in this study is easy to use but is far less informative than the vector-skill whose 7 components are the skills associated with the 7 sub-regions. "

We would like to thank again the reviewer for these comments that helped us improve the paper.