Anonymous Reviewer #1

Responses are in red.

General Comments:

The manuscript reports a Bayesian mixture of sparse regression models for clustering using the Dirichlet process mixture model for feature selection. The purpose is to be able to find interpretable clusters that can allow for climate model downscaling based on key features. This is a potentially useful approach to the discovery of unique features in highly complex datasets for use in statistical downscaling. The authors thoroughly describe the model setup and theory, but they spend less time on the intended purposes of using this approach to better understand and discover real-world covariates to be used in statistical downscaling. The purpose of sparse regression for enhancing the interpretability is clearly stated by the authors. However, the results from the experiments suggest that many challenges with simultaneous clustering and discovery exist, including temporal averaging impact (annual vs. monthly) and regional spatial and temporal impacts on final feature selections and clusters. These challenges should be discussed in greater detail. After reading through the model theory and description, I had high hopes that this mixture of sparse regression models would lead to great insights into the variability of precipitation along the Western U.S. However, this was not quite the case, as the author's interpretation of the results was general and did not show features that could be clearly defined and understood. Greater insight needs to be expressed in the manuscript by detailing how these cluster results could be applied in downscaling, which would assist the reader in understanding the broader point of the paper. It is not clear how downscaling will benefit from this approach. The manuscript is of publication quality, but needs a major revision starting with answering the specific questions below and providing additional descriptions and interpretation of the results as they apply to regional characteristics and the potential use in downscaling.

The importance of statistical downscaling in translating climate model simulations to projections at stakeholder relevant scales is widely accepted in the climate science or adaptation community. However, the value of statistical downscaling has been questioned from two primary perspectives, which are interrelated: First, while statistical relations developed based on historical and current climate may hold as long as conditions remain relatively "stationary", to what extent will they generalize in the future with changes in radiative forcing? We note here that stationarity in this context does not merely refer time-varying trends, but a fundamental change in the data-driven relationships between the variables of interest and the covariates. Second, even under "stationary" conditions, can one validate that are statistical downscaling based hindcasts are "right for the right reasons", or is any accuracy a possible spurious relation that may not generalize? The latter question ultimately also relates to the choice of covariates, and in turn to physical interpretations. Thus, the credibility, physical interpretability of statistical downscaling, and the ability of these results to generalize, depends crucially on the covariate selection step, but simple validation based on accuracy, or even apparently intuitive explanations, may not be adequate to ensure the ability to generalize or guard against spurious relations. The covariate selection step may therefore need to be examined both in isolation, and then in the context of predictions and interpretability, through well-designed metrics and based on Occam's razor considerations.

As we have clearly acknowledged in the manuscript, both in the introduction and the conclusion sections, the application of our methods on climate dataset, especially for prediction, are preliminary and exploratory in nature. The primary purpose of this manuscript is to develop methods for covariate discovery, in the context of statistical downscaling of precipitation, which can generalize to conditions well beyond those for which they are developed. This problem by itself is important enough, and challenging enough, to examine within one focused manuscript. However, to demonstrate the effectiveness of our statistical model, we have evaluated their application in the Northwest and Western United States. We have discovered at least two distinct clusters in each of the regions, where each cluster shows dependence of precipitation on different sets of covariates. We acknowledge in the manuscript that this does not address statistical downscaling in its entirety. Nevertheless, this addresses the crucial step of covariate discovery, which may need to be examined both in its own right and in conjunction with prediction models that use the discovered covariates. The development of linear or nonlinear prediction models, based on the covariates discovered by our method, is our plan for the future. In conjunction with such predictions, we hope to develop better strategies for falsifiability of the predictions under conditions that may resemble the kind of non-stationarity expected in the future.

The reviewer's point is indeed very well taken, and one on which we have spent much thought. However, in the interests of not making the manuscript too complex and to avoid the risk of sounding overly speculative, we have not included our entire thought process in the discussion. We will be happy to consider including a more detailed discussion in the manuscript under a "future research" section if the reviewer so recommends.

Meanwhile, the following sentence has been appended at the end of Section 4.2.1: “The clusters discovered here, and the corresponding covariates, can be utilized to develop individual non-linear prediction models per cluster.”

Specific Comments:

Section 4.1.2, Page 617, Line 14-20: This argument needs to be justified more clearly. See Bader et al., 2008

We assume that the reviewer is referring to section 1. We have added the sentence: “Uncertainties in sub-grid scale cloud-microphysics and ocean eddy processes, as well as poor understanding of the effects of carbon cycle and other biogeochemical processes, on climate systems still limits the ability of the physics-based climate models to reliably project future climate, especially at regional scale.” in place of “This may suggest that purely physics based models may have reached their limit.” We have also added a reference to Bader et al. (2008).

Section 4.1.2 With the degrading NMI value for such a small value of K, given ideal case (i.e. simulated synthetic data), what does this say about how this model perform with weather and climate data, which is likely to have varying values of K depending on season, location, year, co-dependence of climate indices?

DPMs automatically find the number of clusters K; thus, it should be able to adapt to varying values of K depending on season, location, etc. However, DPMs prevent the model to “learn” an unnecessarily large value of K, if a smaller K is sufficient to describe the model, thus managing complexity. We have acknowledged in Section 4.1.2 that the performance of the method degrades as the number of components *K* grows larger. We believe it is reasonable to expect that there will only be a limited number of distinct relationships between average rainfall and their covariates when we apply our method at the regional scale. These distinct relationships may even relate to a finite number of rainfall generation processes. However, even in situations where a large number of relationships exist within a particular region, our method may not be able to identify all of the distinct methods, but it can nevertheless be expected to outperform the use of a single model. The single model will attempt to learn a relationship that is the average of all distinct relations, which our approach will still attempt to distinguish among major categories of relationships even though some of them may be lumped together. In addition, as observed in Figure 2, our proposed models, i.e., unconstrained and constrained sparse regression (DPM with sparse regressors) degrade less than the nonparametric regression (DPM with regression) model. In the future, we plan to investigate and handle this issue further by introducing “diversity priors” on the indicator variables $\boldsymbol{Z}$ that will encourage more clusters to counterbalance the tendency of DPM to merge clusters together.

Section 4.1.2 By averaging precipitation to an annual value, you have reduced your ability to interpret results looking at both atmospheric and climate features. Why was annual averaging chosen? The temporal averaging (i.e. annual averages) may also limit the new discoveries and insight that can be drawn from the results by smoothing out any variability that is likely to be teleconnected to atmospheric and climate related phenomena.

We agree with the reviewer that by using the annual average of features, we have limited our ability to discover dependence on covariates at a smaller temporal scale. However, the use of annual averages does reduce the amount of noise in the observed rainfall data, which enables us to examine the robustness of our methods with less ambiguity. Given the focus of this manuscript on methods development and evaluation, we decided to examine annual averages. Future studies need to examine values at multiple temporal resolutions.

Section 4.2, Page 638, Line 6 This is unclear.

We have replaced the line “While spatially coherent clusters are more natural, geographical features (e.g. mountains, lakes etc.) of the region must also be taken into consideration while interpreting the results” by the following line – “While spatially coherent clusters are more likely to occur in nature, geographical features such as mountains and lakes and even man-made structures such as large dams and reservoirs may abruptly disturb the spatial smoothness of clusters, since their presence may alter the climate pattern of the nearby areas with respect to the surrounding regions.”

Section 4.2, Page 638, Line 10 Annual/Seasonal average? I thought it was only the annual average for each variable.

We are assuming the reviewer here refers to page 637, Line 10. Yes, we have used both annual and seasonal averages of atmospheric covariates. A list of covariates is provided in Table 1.

Section 4.2.1, Page 637, Line 15: A description of the relaxation procedure is necessary. Quantification of this sensitivity may also be informative of how sensitive your model is to spatial and temporal dependence. This would be an interesting aspect of the evaluation of your model.

We added the following sentences:
"We applied spatial "must-link" constraints among pairs of data-points belonging from the same location. Ideally, if there are n points in a location, we will be required to put $^nC_2$ constraints to cover all pairs of data-points. To reduce complexity, initially we kept only those constraints that connect data-points from consecutive years. However, this reduced set of constraints proved to be too restrictive and all data-points tended to merge into a single cluster. So, we kept removing the constraints in an intuitive manner until more than one cluster emerged for a region. We found more than one cluster for all regions except the southern region. We stopped removing constraints until new clusters stopped emerging for a region." instead of the following existing sentences – "When we applied "must link" constraints over pairs of data-points from the same location in the form of a broken chain (broken for sparsity), all data-points tended to merge into a single cluster. However, as we started to relax the constraints, more than one component started to emerge in most regions except the southern region."

Section 4.2.1, Page 637, Line 19-22: So does this mean that the model did not do its job of finding interpretable results for feature selection and downscaling?

As we have acknowledged both in the introduction and the conclusion section of the paper, the results on the climate dataset are exploratory and preliminary in nature. The purpose of our method is to improve our understanding of the relation between annual average rainfall and various atmospheric covariates to facilitate statistical downscaling of rainfall. We have chosen to focus on two regions where the clusters are easier to visualize. This helps us to examine the effectiveness of our method to generate new insights about the complex relationships among rainfall and atmospheric or other large-scale covariates. For other regions, mixed cluster memberships were observed in and many sites, thus making them harder to visualize.

Section 4.2.1, Page 638, Line 1-10: It is not clear how the authors came to this conclusion. Neither the figures nor text clearly state the basis of their interpretation. It is unclear how the model expressed "dependence" on particular features. Is this based on inference from the selected sparse model? This needs to be made clearer.

The above conclusion is made by inspecting the coefficients of the sparse linear models within each cluster. We have added the following sentence to clarify this in the text: "As mentioned earlier, we obtained one sparse linear model for each of the discovered components within a region. Since a non-zero coefficient in the sparse model implies dependence on the corresponding covariate, we can obtain interesting insights about the dependence of average rainfall on various atmospheric and climate indices from the coefficients of the individual sparse models within each cluster."

Anonymous Reviewer #2

General Comments

This paper is about a Bayesian sparse regression technique and an application to statistical downscaling. The work seems to be interesting, and potentially novel. While it is quite valuable to have interdisciplinary submissions in this field, a significant caveat is that the main contribution needs to be identified in the submission. The downscaling application does not seem to be emphasized in the submission, e.g. it appears to be just an evaluation of their method. (cf. "We have evaluated our method both on synthetic and climate datasets."). If instead the contribution is on the technical (methodological/algorithmic/modeling) side, it needs to be more clearly stated in the exposition. In the Conclusion, the authors state "Our major contribution is to develop an efficient and scalable variational inference algorithm for inference in the fully Bayesian model." In order to facilitate evaluation of this claim, the authors would need to delineate and distinguish their contributions from past work, especially in the technical sections. Currently, sections 2 and 3 do not clearly distinguish between prior work in the technical area, including by the authors, and the claimed technical contributions in the submission. This makes it difficult to evaluate the novelty of the contribution. Finally, if the primary contribution is indeed on the methodological/algorithmic/modeling side, then it would seem much more appropriate to submit the manuscript to a data mining, statistics, AI, or machine learning publication, where it could be refereed by reviewers with the relevant expertise. This reviewer is not aware whether or not similar/parallel submissions have been made by the authors to such venues, but this should be clearly stated by the authors.

We agree with this reviewer that the main contribution in this paper is the development of a non-parametric Bayesian framework for finding multiple underlying sparse regression relationships that can potentially be used to select appropriate covariates for statistical downscaling of rainfall. To our knowledge, this is a novel method that combines Dirichlet Process mixture and Bayesian version of sparse regression to facilitate discovery of multiple sparse linear relationship within a complex dataset. The results on the climate dataset need to be viewed as mostly exploratory and preliminary in nature, which we have readily acknowledged both in the introductory (last paragraph) and in the concluding sections of the paper. This work by the authors is not currently under review in any other venues.

Specific comments:

It would be helpful to expand the related work discussion to a variety of Bayesian techniques for downscaling by Andrew Robertson (IRI, Columbia LDEO) and collaborators.

We thank the reviewer for pointing us to the work of Andrew Robertson and his collaborators. However, most of their work can be categorized under weather generator class of statistical downscaling methods, whereas our method can be classified under regression based methods. We have added a reference to the work of Greene *et. al.* in the second paragraph of the introduction section where we described related work.

Clarifications are needed in the technical section, e.g.:

- "stop when the probabilities E[z] stop changing anymore"
- E[z] is an expectation, not a probability.

In reality, the iterative algorithm stops when indicator variables **Z** stop changing any more, not the expectation E[**Z**]. We This was a mistake in our write-up. We have fixed this error in the revised manuscript, and would like to thank the reviewer for pointing this out.

- How do you quantify that they "stop changing"? Is there a threshold? How do you set the threshold?

We did not use any threshold. Algorithm stops when indicator variables **Z** does not change in 5 consecutive iterations.

- How is 5k chosen for the non-zero components?

We chose *5k* (where *k* is the index of the cluster, *k* =1,…,K) for the non-zero components within the *k*-th cluster so that two clusters are distinctly identifiable in case the indices of non-zero components of the clusters are same. We will clarify it in the revised manuscript.

- It would be good to show experiments for various K values.

We have shown results of experiments for various values of K in Figure 2.

- How does your method compare to other clustering techniques, including non-generative techniques such as k-means++?

Our method is not comparable with parametric unsupervised clustering techniques such as k-means++ owing to at least two reasons:

A) Parametric methods require the number of clusters to be known beforehand. One of the principal advantages of our non-parametric technique is that it can automatically estimate the number of clusters from data.

B) Our method simultaneously learns a sparse regression model within each discovered cluster which is not possible for methods such as k-means++ that performs only clustering.

- Many figures are too small the read the axes.

We will increase the font size of the axes labels in the revised manuscript.

Technical corrections:

The submission needs extensive copy editing to fix numerous grammatical errors. E.g. there are many instances in which verbs do not agree with nouns (e.g. plurality). Due to these errors (along with typos, missing/repeated words), the submission is very tedious to read in its current form, in particular the introductory and non-technical sections. A few corrections are listed below; however the above should be done in a concerted way, e.g. by a copy-editor.

We will go through another pass and correct any grammatical errors we find in the revised manuscript.

GCM is defined twice with two definitions. "Global Climate Model" –> "General Circulation Model."

Both definitions are occasionally used interchangeably in the literature, although since climate models are no longer just confined to the atmosphere, perhaps the use of “Global Climate Models” is more justified. We have changed “General Circulation Model” to “Global Climate Model” in the revised manuscript for uniformity.

"facing the mankind" –> delete "the"
"Variational Bayes inferences" –> "Variational Bayesian inference"

We made the suggested changes in the revised manuscript.