Dear Reviewer,

Thank you very much for your positive and constructive comments. The points raised focus on several key aspects of the study, and thus we believe that in addressing them the clarity and usefulness of the article has been greatly improved. Please find below our responses to the individual comments and a description of the corresponding changes to the manuscript made to address each point. For each point, the initial comment is given in bold, while modifications to the text are italicised and shown in blue.

1. **The conclusion section does mention the caveat of neglecting the role of time dimension. The SST case has a dominant ENSO signal that is mostly a time oscillation of a fixed spatial pattern, while in the geopotential height case there are traveling wave signals with spatially changing patterns. For the later case, can the time dimension be included in the analysis so that a more physically interpretable mode can be found? Can including time dimension directly in the d-dimensional data produce different results from applying temporal regularization?**

   In general, explicitly including information about time-dependence in the dimension reduction method may produce better results in terms of extracting more physically interpretable modes, but in practice whether or not this is actually achieved will depend on the details of how the time dimension is incorporated. It is also correct that this may produce different results from applying some forms of temporal regularization, although as regularizing terms can be freely constructed the comparison between the two approaches has to be done on a case-by-case basis. Extending the definitions of the various methods to explicitly model time-dependence is an option that we had intended to include in the paragraph beginning on line 425 of the original manuscript, but we believe that the description given there is too narrow. To address this, we have extended the opening sentence to clarify that we include models that include some sort of explicit time-dependence, which now reads

   The idea of imposing temporal regularization via assumed dynamics for the latent weights suggests that another approach to better target particular features is to start from an appropriately defined generative model, or otherwise explicitly incorporating appropriate time-dependence when constructing a reduction method.

2. **From the description in section 2, it is a little hard to conceptualize the key differences between the AA and CC methods, and their advantages over the k-mean clustering method. Could you list the cost function and constraints for each method in a succinct manner and highlight their differences?**

   We agree that the description of the various methods given in Sect. 2 does not provide a sufficiently clear summary of the key distinctions between the possible choices, and puts too much emphasis on numerical details. To address this, we have moved the discussion of the numerical implementation beginning on line 191 to a separate appendix, and instead now conclude Sect. 2 with the following brief summary of the different methods that highlights the relevant choices that enter into the definition of each:

   The various choices of cost function and constraints defining the above methods are summarized in Table 1. While all of the methods fit within the broader class of matrix factorizations, the different choices of cost functions and constraints lead to important differences in the low-dimensional representation of the data produced by each method. For instance, in contrast to PCA, the basis vectors produced by k-means, convex coding, or AA are in general not orthogonal. In some circumstances, this non-orthogonality may be advantageous when the structure necessary to ensure orthogonal basis vectors (e.g., via appropriate cancellations) obscures important features or makes interpretation of the full PCA basis vectors difficult. A k-means clustering may, for

Table 1: Summary of the definitions of each of the four methods compared in this study. Each method is defined by a choice of cost function to be minimized together with a set of constraints placed on the factors Z and W (or Z and C for AA). The choices for each impact that nature of the features that each method extracts from a particular dataset.

Method	Cost function	Constraints	Targeted features
PCA	$\frac{1}{2T}\lVert X - ZW^T\rVert_F^2$	$\mathrm{rank}(ZW^T) \leq k$	Directions of maximum variance
k-means	$\frac{1}{2T}\lVert X - ZW^T\rVert_F^2$	$Z_{ti} \in \{0,1\}$, $Z\mathbf{1}_k = \mathbf{1}_T$	Data centroids
Convex coding	$\frac{1}{2T}\lVert X - ZW^T\rVert_F^2 + \lambda_W \Phi_W(W)$	$Z \succeq 0$, $Z\mathbf{1}_k = \mathbf{1}_T$	Basis convex hull
AA	$\frac{1}{2T}\lVert X - ZCX\rVert_F^2$	$C \succeq 0$, $C\mathbf{1}_T = \mathbf{1}_k$, $Z \succeq 0$, $Z\mathbf{1}_k = \mathbf{1}_T$	Data convex hull

example, provide a much more natural reduction of the data when multiple distinct, well-defined clusters are present. The cost function and choice of constraints that define convex coding and archetypal analysis imply that the optimal basis vectors produced by these methods are such that their convex hull (i.e., the set of all linear combinations of the basis vectors with weights summing to one) best fits the data. Consequently, both are well-suited for describing data where points can be usefully characterized in terms of their relationship to a set of extreme values, be they spatial patterns of large positive or negative anomalies in a geophysical field or particular combinations of spectral components in a frequency domain representation of a signal. PCA and k-means may be less useful in such cases, as neither yields a decomposition of the data in terms of points at or outside of the boundaries of the observations. AA differs from more general convex encodings in imposing the stricter requirement that the dictionary elements, i.e., the archetypes, lie on the boundary of the data. It is, in this sense, conservative, in that the features extracted by AA lie on the convex hull of the data and so correspond to a set of extremes that are nevertheless consistent with the observed data. In the absence of any regularization ($\lambda_W \to 0$), the general convex codings that we consider admit basis vectors that lie well outside of the observed data. By doing so, the method finds a set of basis vectors whose convex hull better reconstructs the data than in AA, at the cost of representing it in terms of points that may not be physically realistic. This behaviour, and the impact of the different choices of cost function and constraints, is sketched in Fig. 1.

3. **Comparing Figures 9 and 13, the behavior of CC and AA finding basis with more extreme departures from mean than k-mean clustering is the same for both the SST and Z cases. Does one case prefer bases with larger departures from mean than the other? Or, is the magnitude of bases functions less important than their alignment with the actual physical modes?**

   We tend to agree that, as far as the discussion in the article goes, the magnitude of the basis functions is less important than the alignment of the various basis choices, although it is certainly not immaterial. There are two points that we should have explained more clearly in the original manuscript. Firstly, with respect to the common behavior of AA and CC choosing a dictionary representing larger departures from the mean state than k-means, this is expected from the definitions of the methods: AA and CC are designed to find points that lie at the "boundary" of the observed data, in the case of AA, or outside of it when using a general convex coding. As a result, both methods will select bases further from the mean than k-means, which selects the cluster means as a dictionary. The extent to which CC chooses more extreme basis points than AA will depend on the particular characteristics of the dataset, and so it is difficult to compare the extent to which one case may show larger departures from the mean than another. However, it is right to note that such large departures are obtained with CC, and so to try better emphasize this point we have included this observation in the summary paragraph concluding Sect. 2, where it is pointed out that the resulting basis vectors may not be physically realistic. For the purposes of the case studies, we would argue that the main distinction is the fact that, for the geopotental height case, the bases do not correspond to the PCA basis (or the physical modes) and so are more difficult to make use of. To try to clarify this point, we have expanded the third paragraph of Sect. 3.2 to highlight some of the similarities and differences. Beginning on line 339 of the original manuscript, these changes read:

   The relative dispositions of each of the states with respect to each other, as measured by Euclidean distance, are visualized in Fig. 13 on the basis of a two-dimensional metric MDS. In some respects, the performance of the different methods is similar to that seen in the SST example; for example, as expected on general grounds the ordering of the methods with respect to achieved RMSE (not shown) is the same as in Fig. 10. Similarly, AA and the unregularized convex coding select, by design, basis vectors that lie on or outside of the convex hull of the data, with the latter having the freedom to choose a basis corresponding to much larger departures from the mean so as to reduce the reconstruction error; note that the precise degree to which the basis vectors lie outside the convex hull will depend on the particular data at hand, but the behavior is otherwise generic. However, unlike the SST case, ...

4. **For the Z case, both AA and CC methods are not aligned with the PCA bases functions. Do you have any insights of which method is superior in this case? Or, are they all not finding the physical modes because of missing the time dimension?**

   Yes, neither AA nor CC are finding the expected modes as they do not account for the important role of persistence in defining these modes. We agree that the discussion of which method is superior, or at least more useful, in the case of the geopotential height anomalies was not clearly done. In this case, we argue that the regularized convex coding is more appropriate for this case, as it better targets the persistent features characterizing atmospheric extremes, along with k-means. While they also do not explicitly include the time-dimension, they do pick up persistent or quasi-stationary features, and hence give a better extraction of the large-scale anomaly structures. To try to clarify this point, we have added a short summary at the end of Sect. 3.2, beginning at line 386 of the original manuscript, to read:

. . . but could be improved by, for example, making use of a soft clustering algorithm instead. To summarize, in the geopotential height case where extreme events are defined not just by large anomalies but persistent structures, methods such as k-means, or the more heavily regularized convex coding applied here, that better identify such structures provide bases that are more amenable to interpretation in terms of physical extremes and so may provide a more direct starting point for analyses of such events. Unregularized convex coding and archetypal analysis, on the other hand, are less well suited in this respect, as they do not yield a direct assignment of such events to individual states.

Finally, it is worth noting that, as in the SST case study, . . .

5. **For the RMSE for reconstructed data (Fig 10), is there a similar plot for the geopotential height case?**

The equivalent plot for the geopotential height case study is shown in Figure 1. The ordering of the methods

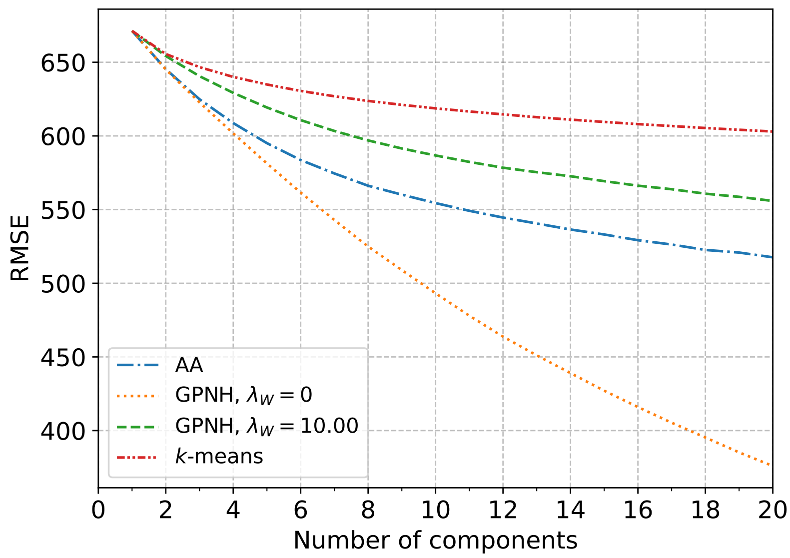

Figure 1: RMSE for reconstructing the leading 167 PCs of geopotential height anomalies resulting from each of the methods.

with respect to the reconstruction RMSE is the same as in the SST case shown in Figure 10, although note that since the clustering in this case is performed on the PCs themselves for efficiency there is not an equivalent line for PCA. This ordering of the methods in terms of reconstruction error is expected on general grounds as briefly discussed following Figure 10 (lines 304 - 311); that is, for a given dataset, PCA yields the globally minimal reconstruction error as guaranteed by the Eckhart-Young theorem, while the unconstrained convex coding can approach the same level of fidelity by choosing basis points sufficiently far outside the convex hull of the data. AA and k-means, being more constrained, in turn have a larger reconstruction error for a given basis size, with AA typically performing better than k-means in this respect by virtue of being able to provide a soft-clustering of the data. As this behavior is the same in both the SST and geopotential height case, we do not include Figure 1 in the article to reduce repetition. However, we do think that we should have better emphasized the expected performance of the different methods in terms of reconstruction error, as this is an important point to consider when, e.g., one is interested in simply achieving a compression of the data with minimal loss of information. To address this point, we have firstly expanded the discussion in the last paragraph of Sect. 3.1 to better explain this point, with the following additions beginning on line 311:

The ordering of the methods in terms of reconstruction error observed in Fig. 10 is expected to be the case more generally. As noted in Sect. 2, PCA provides the globally optimal reconstruction of the data matrix with a given *rank, in the absence of any constraints, and so amounts to a lower bound on the achievable reconstruction error. Of the remaining methods, the additional freedom to locate the basis elements outside of the convex hull of the data when performing an unregularized convex coding allows for a lower reconstruction error than is achievable using archetypes. For a given basis size the hard clustering resulting from k-means generally results in the largest RMSE. For larger values of the regularization λ_W, the optimal basis elements sit within the convex hull of the data and provide a fuzzy representation of the data with a progressively increasing RMSE. In this particular analysis of SST data, the performance of the different methods with respect to reconstruction error is one distinguishing factor that may guide the choice of method; while all four produce similar large-scale spatial patterns, for a given basis size PCA provides the lowest reconstruction error and might be preferred if information loss is a significant concern.*

We also now explicitly note in the text that the same ordering in terms of RMSE is observed in the geopotential height case; this is included in the additions in response to point 3 above.

Additionally, we have also made the following minor changes to the manuscript:

- On line 6 of the original manuscript, in the abstract the use of the acronym "EOF" has been replaced by "empirical orthogonal function".
- Starting on line 21 of the original manuscript, the description of empirical orthogonal functions has been expanded to read: *Perhaps the most familiar example in climate science is provided by empirical orthogonal function (EOF; Lorenz (1956); Hannachi et al. (2007)) or principal component analysis (PCA; Jolliffe (1986)), which identifies directions of maximum variance in the data, or, more generally, the directions maximizing a chosen norm.*
- For clarity, on line 314 it is highlighted that the bases produced by the methods correspond to spatial patterns, and now reads: *As a result, all of the dimension reduction methods that we consider extract similar bases (patterns) ...*