Dear Reviewers,

Thank you very much for your positive comments and suggestions. We believe that in attempting to address each of the points made the quality of the manuscript has been greatly improved.

Please find attached our responses to the specific comments made in each review, and a description of the corresponding changes to the manuscript made to address each point. In each case, the initial comment is given in bold, while modifications to the text are italicized and shown in blue.

In addition to the modifications in response to specific comments detailed in the attached responses, we have also made some minor changes:

- On line 6 of the original manuscript, in the abstract the use of the acronym "EOF" has been replaced by "empirical orthogonal function".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) . . .*

We attach a copy of the revised manuscript with all changes highlighted.

Yours sincerely,

Dylan Harries and Terence O'Kane

Response to Referee 1

1. **it is helpful to use a table to succinctly summarize the key differences among the techniques. Some of the algorithmic details are unnecessarily elaborated (e.g., pg8) whereas the actual differences are obscured. For example, does K-mean cluster produce orthogonal basis vectors? and are the clusters easier to interpret than principal components of PCA. What are the unique advantages of AA relative to PCA and K-mean cluster?**

   We agree that it is helpful to have a single summary of the key differences between the methods, and that too much focus was placed on the numerical solution of the optimization problem. To address this, the discussion of the numerical implementation of the solution for the AA/CC optimization problem beginning on line 191 of the original manuscript has been moved to a separate appendix. Instead, we now conclude Sect. 2 with a summary of the key differences between the different methods, including a table (Table 1) collecting the different cost and constraint functions, that reads as follows:

   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 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.

   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	$\frac{1}{2T}\|X - ZW^T\|_F^2$	$\text{rank}(ZW^T) \leq k$	Directions of maximum variance
k-means	$\frac{1}{2T}\|X - ZW^T\|_F^2$	$Z_{ti} \in \{0,1\}$, $Z\mathbf{1}_k = \mathbf{1}_T$	Data centroids
Convex coding	$\frac{1}{2T}\|X - ZW^T\|_F^2 + \lambda_W \Phi_W(W)$	$Z \succeq 0$, $Z\mathbf{1}_k = \mathbf{1}_T$	Basis convex hull
AA	$\frac{1}{2T}\|X - ZCX\|_F^2$	$C \succeq 0$, $C\mathbf{1}_T = \mathbf{1}_k$, $Z \succeq 0$, $Z\mathbf{1}_k = \mathbf{1}_T$	Data convex hull

2. **please explain how the case studies were chosen. Are the outcomes of the case studies supposed to inform us about the geophysical variables that one, or several of the approaches are more suitable than others?**

   Yes, this is correct; the two case studies were chosen to highlight instances where some of the methods may not be as suitable or useful as the others, depending on the focus of the analysis. In the case of the SST data, the existence of a single, well-separated mode of variability results in all four methods producing similar bases with

which to represent the data. Consequently, for this application all of the methods are relatively comparable when it comes to defining a set of modes, and are distinguished by other features such as the magnitude of the reconstruction error for a given basis size, as noted in point 3. The second case study is chosen to emphasize the fact that this agreement between the methods is, however, dependent on the features of the SST data, so that for variables where this scale separation is absent, such as geopotential height, the choice of method becomes more important from the point of view of extracting useful modes. To attempt to summarize this motivation for the choice of case studies, the first paragraph of Sect. 3 has been expanded and now reads as follows:

We now turn to a set of case-studies that demonstrate some of the implications of the various differences noted above in realistic applications. We consider two particular examples that highlight the importance of considering the particular physical features of interest when choosing among possible dimension reduction methods. The first example that we consider, an analysis of SST anomalies, is characterized by a large separation of scales between modes of variability together with key physical modes, particularly ENSO, that can be directly related to extreme values of SST anomalies. This means that the basis vectors or spatial patterns extracted by PCA, k-means, and convex coding are for the most part rather similar in structure, and so the choice of method may be guided by other considerations, such as the level of reconstruction error. We then contrast this scenario with the example of an analysis of mid-latitude geopotential height anomalies, where there is neither scale separation nor do the physical modes coincide solely with extreme anomaly values. As a result, methods that are based on constructing a convex coding of the data, without targeting features that capture the dynamical characteristics of the relevant modes, produce representations that are, arguably, more difficult to interpret, and hence may be less suitable than clustering based methods.

3. **For the SST case, I wonder what the take home message is in terms of the difference among the three methods as illustrated in Figs. 3-9? Fig 10 shows that the PCA features lower RMSE than others and yet the conclusion appears to be that these methods are all comparable.**

   We agree that the take home point was not clearly made in the SST case. For the SST data, all four methods are comparable in the sense that the physical space patterns that are identified as basis vectors are similar, as illustrated in Figs. 3 - 9, and so no particular method is obviously preferable for identifying relevant modes. Methods that are based on locating the convex hull of the data in this case perhaps provide a more direct interpretation of extreme events, which can be characterized for SSTs in terms of the magnitudes of the anomalies, but the patterns produced do not differ dramatically from ordinary PCA. As rightly noted, though, the latter also provides the optimal RMSE for a given basis size, and so is distinguished from the other methods in terms of the fidelity of the reconstruction; indeed, this behavior in terms of reconstruction error is generic and we should have highlighted this detail in the first version of the article. To address this point, we now summarize our conclusions from the case study at the end of Sect. 3.1, beginning at line 311 of the original manuscript, as follows:

   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.

4. **For the Z500 anomaly case, what is the recommendation of lambda_w? And what methods offer a clear linkage between the resulting patterns and "physical extremes"?**

   In general, the choice of the regularization parameter will depend on the particular use case, a point which we agree should have been made more clearly. For our purposes, the values $\lambda_W = 0$ and $\lambda_W = 10$ were chosen to provide an illustration of the impact of this parameter on the fitted bases, and in particular to highlight the role of the regularization in producing bases that vary from minimizing the reconstruction error to focusing

on features in the data that are less impacted by noise. In the geopotential height case, we argue that placing more emphasis on the latter provides clearer linkage to the physical extremes, as doing so tends to prefer states that exhibit some degree of persistence and thus matches the characteristics of the physical features. This is also achieved by the use of k-means, whereas AA in its usual formulation takes into account only those points lying on the convex hull of the data and so does not provide a particularly good characterization of the physical extremes in terms of the associated patterns. It should be noted of course that the dimension reductions considered here are a starting point for further analysis, e.g., causally relating the observed weather patterns to known extreme events requires further dynamical analysis, but methods that identify patterns that closely resemble the relevant structures provide clearer starting points for such analyses. To address this, we have expanded the discussion (starting at line 386 of the original manuscript) at the end of Sect. 3.2 to summarize these observations:

. . . 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, . . .

In practice, the regularization λ_W can be chosen using standard model selection criteria depending on the user's objective; for instance, cross-validation could be used to select a value of λ_W that provides the best out-of-sample reconstruction or prediction error. As this was not pointed out in the original manuscript, we have now added the following comment at line 359 to address the problem of choosing λ_W:

In this sense, by appropriate choice of regularization it is possible to interpolate between a representation of the data in terms of points on the convex hull and mean features, depending on how much weight is placed on minimizing the reconstruction error. While here we consider only a few levels of regularization for illustrative purposes, in general the degree to which this is done, i.e., the choice of λ_W, can be guided by standard model selection methods, such as cross-validation.

Response to Referee 2

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 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

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	$\frac{1}{2T}\|X - ZW^T\|_F^2$	$\text{rank}(ZW^T) \leq k$	Directions of maximum variance
k-means	$\frac{1}{2T}\|X - ZW^T\|_F^2$	$Z_{ti} \in \{0,1\}$, $Z\mathbf{1}_k = \mathbf{1}_T$	Data centroids
Convex coding	$\frac{1}{2T}\|X - ZW^T\|_F^2 + \lambda_W \Phi_W(W)$	$Z \succeq 0$, $Z\mathbf{1}_k = \mathbf{1}_T$	Basis convex hull
AA	$\frac{1}{2T}\|X - ZCX\|_F^2$	$C \succeq 0$, $C\mathbf{1}_T = \mathbf{1}_k$, $Z \succeq 0$, $Z\mathbf{1}_k = \mathbf{1}_T$	Data convex hull

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

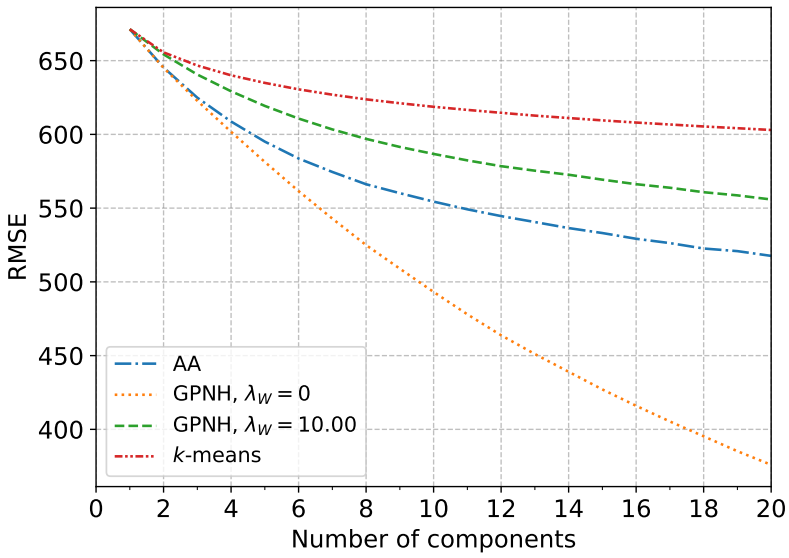

Figure 1: RMSE for reconstructing the leading 167 PCs of geopotential height anomalies resulting from each of the methods.

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.

Applications of matrix factorization methods to climate data

Dylan Harries[1] and Terence J. O'Kane[1]

[1]CSIRO Oceans and Atmosphere, Hobart, Australia

Correspondence: Dylan Harries (Dylan.Harries@csiro.au)

Abstract. An initial dimension reduction forms an integral part of many analyses in climate science. Different methods yield low-dimensional representations that are based on differing aspects of the data. Depending on the features of the data that are relevant for a given study, certain methods may be more suitable than others, for instance yielding bases that can be more easily identified with physically meaningful modes. To illustrate the distinction between particular methods and identify circumstances in which a given method might be preferred, in this paper we present a set of case studies comparing the results obtained using the traditional approaches of ~~EOF~~ empirical orthogonal function analysis and k-means clustering with the more recently introduced methods such as archetypal analysis and convex coding. For data such as global sea surface temperature anomalies, in which there is a clear, dominant mode of variability, all of the methods considered yield rather similar bases with which to represent the data, while differing in reconstruction accuracy for a given basis size. However, in the absence of such a clear scale separation, as in the case of daily geopotential height anomalies, the extracted bases differ much more significantly between the methods. We highlight the importance in such cases of carefully considering the relevant features of interest, and of choosing the method that best targets precisely those features so as to obtain more easily interpretable results.

1 Introduction

An ubiquitous step in climate analyses is the application of an initial dimension reduction method to obtain a low-dimensional representation of the data under study. This is, in part, driven by the purely practical fact that large, high-dimensional datasets are common, and to make analysis feasible some initial reduction in dimension is required. Often, however, we would like to associate some degree of physical significance to the elements of the reduced basis, for instance by identifying separate modes of variability. Given the wide variety of possible dimension reduction methods to choose from, it is important to understand the strengths and limitations associated with each for the purposes of a given analysis.

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. The difficulties inherent in interpreting EOF modes physically have been thoroughly documented (Dommenget and Latif, 2002; Monahan et al., 2009) and partly motivate

various modifications of the basic EOF analysis (Kaiser, 1958; Richman, 1986; Jolliffe et al., 2003; Lee et al., 2006; Mairal et al., 2009; Witten et al., 2009; Jenatton et al., 2010).

Another approach to constructing interpretable representations is based on cluster analysis, which, in its simplest variants, identifies regions of phase space that are repeatedly visited (MacQueen, 1967; Ruspini, 1969; Dunn, 1973; Bezdek et al., 1984). The utility of clustering-based methods is founded on the apparent existence of recurrent flow patterns over a range of time-scales (Michelangeli et al., 1995). As estimating the multidimensional probability density function (PDF) associated with the distribution of states is generally difficult, clustering methods attempt to detect groupings or regions of higher point density, which may in some cases be approximations to peaks in the underlying distribution, and otherwise detect preferred weather patterns or types (Legras et al., 1987; Mo and Ghil, 1988). Extensions to standard clustering algorithms may take into account the fact that such patterns usually exhibit some degree of persistence or quasi-stationarity (Dole and Gordon, 1983; Renwick, 2005). Hierarchical or partitioning based clustering techniques, of which k-means is a popular example, have been widely used to identify spatial patterns associated with regimes or to classify circulation types (see, e.g., Mo and Ghil (1988); Stone (1989); Molteni et al. (1990); Cheng and Wallace (1993); Hannachi and Legras (1995); Kidson (2000); Straus et al. (2007); Fereday et al. (2008); Huth et al. (2008); Pohl and Fauchereau (2012); Neal et al. (2016)). Despite their widespread use, there can be difficulties in interpreting the resulting patterns (Kidson, 1997), while ambiguities in selecting the number of clusters can lead to conflicting characterizations of regimes (Christiansen, 2007). In particular, while k-means is widely used due to its simplicity, if the data does not fall into well-defined, approximately spherical clusters, the method need not provide a particularly useful classification of each sample, in which case an alternative clustering algorithm may be more appropriate.

The output of clustering algorithms such as k-means is an assignment of each data point to a cluster, and a collection of cluster centroids corresponding to the mean within each cluster. The result is a partition of the phase space in which the elements of each partition are taken to be well represented by the cluster mean, as in vector quantization applications (Lloyd, 1982; Forgey, 1965). While this can yield a useful decomposition of the data, the ideal case occurring when the data do indeed form well-defined clusters, a representation in terms of cluster means is not always effective or suitable for all applications. For instance, the k-means centroids need not be realized as observed points within the dataset[1], and, by virtue of their definition, will generally not afford a good representation of edges or extrema of the observed data. When such features are relevant, a possible alternative is to employ methods that construct a basis from points lying on (or outside of) the convex hull of the data. An example of this approach is given by archetypal analysis (AA; Cutler and Breiman (1994); Stone and Cutler (1996); Seth and Eugster (2016)), in which the basis vectors are required to be convex linear combinations of the data points that minimize a suitable measure of the reconstruction error. This has the effect of identifying points corresponding to extrema of features within the data. Unlike partitioning based cluster algorithms such as k-means, AA does not yield a partition of the data into a set of clusters, so that it is not so straightforward to assign observations to a set of regimes. Instead, each observation is represented as a linear combination of the reduced basis of archetypes with appropriate convex weights, and is in this sense similar to PCA. Compared to PCA, however, the archetype basis may (in some cases) be more easily interpreted, as the basis

[1] A simple example would be a naïve analysis of noisy annular data, for which k-means may give rise to centroids lying within the annulus; as noted above, in general the success of all of the methods considered in the following will depend on the underlying topology of the data.

vectors correspond to extreme points of the observed data, or convex combinations thereof, rather than abstract directions in the data space.

Unlike PCA and standard clustering methods, AA has only relatively recently found use in climate studies (Steinschneider and Lall, 2015; Hannachi and Trendafilov, 2017). As a result, there has been little comparison of the results of using AA for dimension reduction versus more traditional approaches. Dimension reduction methods such as PCA, k-means, and AA can all be expressed generically as approximately representing the data in terms of some set of basis vectors (Mørup and Hansen, 2012), possibly with an additional stochastic error term[2],

$$\boldsymbol{x}(t) \approx \hat{\boldsymbol{x}}(t) = \sum_{i=1}^{k} z_i(t)\boldsymbol{w}_i + \boldsymbol{\epsilon}(t), \tag{1}$$

where $\boldsymbol{x}(t)$ is the d-dimensional observation at time t, $\hat{\boldsymbol{x}}(t)$ the corresponding reduced representation, and $\{\boldsymbol{w}_i\}$ is the set of k basis vectors. Each method differs in the definition or criteria used to obtain the basis $\{\boldsymbol{w}_i\}$ and weights $z_i(t)$, and hence the choice of method necessarily plays a key role in the nature of the retained features in the data and the interpretation of subsequent results (Lau et al., 1994). Understanding how the methods differ is important for assessing the appropriateness of each for a given task, and for comparing results between methods. For instance, AA is a natural choice when the salient features are in close correspondence with extremal points, but might be less informative in representing data with a regime-like or clustered structure in which a single cluster must be represented by multiple archetypes. When the data exhibits an elliptical distribution in phase space, modes extracted by PCA are easily interpreted, and k-means clustering can also be expected to perform well, but this need not be the case for more complicated distributions. Thus, naïvely we might expect that a clustering-based approach might be more useful for the purpose of identifying recurrent regimes, whereas AA may be a better choice for locating extremes of the dynamics.

Ultimately though, a fuller understanding might be best obtained by using generalizations of the above methods, or combinations of multiple methods. AA can be regarded as a constrained convex encoding (Lee and Seung, 1997) of the dataset, where the basis vectors are restricted to be linear combinations of observed datapoints. Relaxing this constraint allows for a more flexible reduction in which the archetypes may not be representable as convex combinations of the data (Mørup and Hansen, 2012), although the effectiveness of doing so will depend somewhat on the underlying data generating process. The problem of finding a convex coding of a given dataset can be unified with PCA and k-means clustering by phrasing each as a generic matrix factorization problem (Singh and Gordon, 2008); we review the basic formulation in Sect. 2. In addition to providing a consistent framework for defining each method, it is straightforward to incorporate additional constraints or penalties, e.g., for the purposes of feature selection or to induce sparsity (Jolliffe et al., 2003; Lee et al., 2006; Mairal et al., 2009; Witten et al., 2009; Jenatton et al., 2010; Gerber et al., 2020). Solving the resulting (usually constrained) optimization problem amounts to learning a dictionary with which to represent the data, with different methods producing different dictionaries. By carefully defining the optimization problem, the learned dictionary can be tuned to target particular features in the data. Below,

[2]Probabilistic formulations of PCA and AA have also been proposed (Roweis, 1998; Tipping and Bishop, 1999; Seth and Eugster, 2016), in which the problem is formulated as inference under an appropriate latent variable model; similarly, soft k-means is well-known to be closely related to Gaussian mixture modelling, e.g., MacKay (2003). For simplicity, in the following comparisons we will only consider the deterministic formulations of the methods.

we demonstrate this process by utilitizing a recently introduced regularized convex coding (Gerber et al. (2020)), which allows for feature selection to be performed by varying a regularization parameter. By tuning the imposed regularization to optimize the reconstruction or prediction error, the relative performance of selecting a basis lying on or outside the convex hull can be compared to one that preferentially extracts cluster means.

The purpose of this paper is to explore some of the above issues in the context of climate applications, in the hope that this may provide a useful aid for researchers in constructing their own analyses. In particular, we aim to illustrate some of the strengths and weaknesses of PCA and other dimension reduction methods, using as examples k-means, AA, and general convex coding, by applying each method in a set of case studies. We first apply the methods to an analysis of global sea surface temperature (SST) anomalies, as in Hannachi and Trendafilov (2017). Interannual SST variability is in this case dominated by El Niño-Southern Oscillation (ENSO) activity (Wang et al., 2004), for which there is a large scale separation between this mode and the sub-leading modes of variability, and consequently all methods are effective in detecting this feature of the data. Differences between the methods do arise at smaller scales, however. We then consider a similar analysis of daily 500 hPa geopotential height anomalies. Unlike SST, the time-series of height anomalies generally does not exhibit noticeably large excursions corresponding to weather extremes; in this case, extremes in the data are characterized not by the amplitude of the anomalies but by their temporal persistence. This demonstrates a limitation of all of the considered methods, namely, that (at least in their basic formulation) persistence or quasi-stationarity is not taken into account. Thus, a direct application of AA or convex codings may not be effective in detecting extremes, while clustering methods may be more informative in detecting recurrent weather patterns (to the extent that any such regimes are present in the analyzed fields).

The remainder of this paper is structured as follows. In Sect. 2 we review the dimension reduction methods used. In Sect. 3, we compare the results obtained using each method in a set of case studies to illustrate the distinctions between the methods. Finally, in Sect. 4 we summarize our observations and discuss possible future extensions.

2 Matrix factorizations

In this section, we first describe the dimension reduction methods that we use in our case studies. As noted above, PCA, k-means, and convex coding applied to multidimensional data can all be phrased as matrix factorization problems. Given a collection[3] of d-dimensional datapoints $\boldsymbol{x}(t) \in \mathbb{R}^d$, $t = 1, \ldots, T$, we may conveniently arrange the data into a $T \times d$ design matrix X with rows formed by the data samples. In this notation, the reduced representation Eq. (1) becomes

$$X \approx \hat{X} = ZW^T, \tag{2}$$

where $Z \in \mathbb{R}^{T \times k}$ is a $T \times k$ dimensional matrix with rows giving the weights $z_i(t)$, and $W \in \mathbb{R}^{d \times k}$ contains the basis vectors $\boldsymbol{w}_i$ as columns. Note that here and below we assume that the columns of X have zero mean; if not, this can always be arranged

[3] In the following, we use notation appropriate for a time-series of observations with separate samples indexed by time t, but of course the discussion is not limited to this case.

by first centering the data,

$$X = \left(I_{T\times T} - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T^T \right) \tilde{X},$$

where $\tilde{X}$ denotes the original data, $I_{T\times T}$ the $T \times T$ identity matrix, and $\mathbf{1}_T$ the T-dimensional vector of ones.

The factors W and Z are calculated as the minimizers of a suitably chosen cost function $F(W, Z; X)$, measuring (in some application dependent sense) the quality of the reconstruction $\hat{X}$, subject to the constraint that they are within certain feasible regions Ω_Z and Ω_W,

$$(W, Z) \equiv \underset{Z\in\Omega_Z, W\in\Omega_W}{\arg\min}\ F(W, Z; X). \tag{3}$$

A typical cost function takes the form of a decomposable loss function, measuring the reconstruction error, together with a set of penalty terms imposing any desired regularization, i.e.,

$$F(W, Z; X) = \frac{1}{T}\sum_{t=1}^{T} \ell(W, \boldsymbol{z}(t); \boldsymbol{x}(t)) + \lambda_W \Phi_W(W) + \lambda_Z \Phi_Z(Z). \tag{4}$$

In Eq. (4) we have, for simplicity, supposed that W and Z are independently regularized, with the tunable parameters λ_W and λ_Z governing the amount of regularization. Common choices for the loss function include the ℓ_2-norm[4],

$$\ell(W, \boldsymbol{z}(t); \boldsymbol{x}(t)) \equiv \frac{1}{2}\left\| \boldsymbol{x}(t) - W\boldsymbol{z}(t) \right\|_2^2, \tag{5}$$

in which case the unregularized cost is proportional to the residual sum of squared errors (RSS), the Kullback-Leibler divergence for non-negative data,

$$\ell_{KL}(W, \boldsymbol{z}(t); \boldsymbol{x}(t)) = \sum_{i=1}^{d} \left(x_i(t) \ln \frac{x_i(t)}{(W\boldsymbol{z}(t))_i} - x_i(t) + (W\boldsymbol{z}(t))_i \right), \tag{6}$$

or the wider class of Bregman divergences (Bregman, 1967; Banerjee et al., 2005; Singh and Gordon, 2008). In addition to the soft constraints imposed by the penalty terms $\Phi_W(W)$ and $\Phi_Z(Z)$, the feasible regions $\Omega_Z \subset \mathbb{R}^{T\times k}$ and $\Omega_W \subset \mathbb{R}^{d\times k}$ define a set of hard constraints that must be obeyed by the optimal solutions. The definition of a given method thus comes down to a small number of modelling choices regarding the cost function and feasible regions.

PCA, in its synthesis formulation, is equivalent to minimizing an ℓ_2-loss function with $\lambda_W = \lambda_Z = 0$, i.e.,

$$F_{PCA}(W, Z; X) = \frac{1}{2T}\left\| X - ZW^T \right\|_F^2, \tag{7}$$

where $\|A\|_F$ denotes the Frobenius norm of a matrix, subject to the constraint that the reconstruction $\hat{X}$ has rank at most k. The problem in this case has a global minimum (Eckart and Young, 1936) given by the singular value decomposition (SVD), and the retained basis vectors $\boldsymbol{w}_i$ correspond to the directions of maximum variance. In our numerical case studies, we adopt

[4]The ℓ_p-norm of a vector $\boldsymbol{x} \in \mathbb{R}^d$ is given by $\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{d} x_i^p \right)^{1/p}$ for $p > 0$, while for $p = 0$ the ℓ_0-norm is defined to be the number of non-zero components of $\boldsymbol{x}$.

the convention $Z = U\Sigma/\sqrt{T-1}$, $W = V$ for the PCA factors W and Z, where the SVD of $X = U\Sigma V^T$ for real X. While the existence of a direct numerical solution for the optimal factorization makes PCA very flexible and easy to apply, the basis vectors may be difficult to interpret, for instance if they have many non-zero components. Sparse variants of PCA that attempt to improve on this may be arrived at by introducing a sparsity inducing regularization $\Phi_W(W)$, of which a common choice is the ℓ_1-penalty

$$\Phi_W(W) = \sum_{i=1}^{k} \|\boldsymbol{w}_i\|_1 . \tag{8}$$

The partitioning that results from k-means clustering may also be written in terms of a factorization $\hat{X} = ZW^T$. Whereas PCA by construction yields an orthonormal basis with which to represent the data, the k-means decomposition is in terms of the cluster centroids and iteratively attempts to minimize the within cluster variance (Hastie et al., 2005). This is equivalent to minimizing an ℓ_2-cost function, as in Eq. (7), subject to the constraint that the weights matrix Z has binary elements, $Z_{ti} \in \{0, 1\}$, and rows with unit ℓ_1-norm, $\|\boldsymbol{z}(t)\|_1 = 1$. In other words, the optimization is performed within the feasible region

$$\Omega_Z = \{Z \in \mathbb{R}^{T\times k} | Z_{ti} \in \{0,1\}, \quad Z\mathbf{1}_k = \mathbf{1}_T\}. \tag{9}$$

The corresponding basis matrix W then contains the cluster centroids $\boldsymbol{w}_i$ as columns. Although both PCA and k-means can be seen to minimize the same objective function[5], the additional constraints in k-means clustering mean that finding the exact clustering is NP-hard (Aloise et al., 2009; Mahajan et al., 2012), and heuristic methods must be used instead (e.g., Lloyd (1982); Hartigan and Wong (1979)). These iterative methods are not guaranteed to find a globally optimal clustering, and must be either run multiple times with different initial guesses or combined with more sophisticated global optimization strategies to reduce the chance of finding only a local minimum.

The convex codings that we employ below, like PCA and k-means, are based on minimizing the least squares loss Eq. (7). The least restricted version (Lee and Seung, 1997; Gerber et al., 2020) requires only that the reconstruction lies in the convex hull of the basis, or, in other words, that the weights Z satisfy the constraints

$$\Omega_Z = \{Z \in \mathbb{R}^{T\times k} | Z \succeq 0, \quad Z\mathbf{1}_k = \mathbf{1}_T\}. \tag{10}$$

where the condition $A \succeq 0$ indicates an elementwise inequality. In the absence of any hard constraints on the basis W, the optimization problem to be solved reads

$$(W, Z) = \underset{Z\in\Omega_Z}{\arg\min} \left[\frac{1}{2T} \left\| X - ZW^T \right\|_F^2 + \lambda_W \Phi_W(W) \right], \tag{11}$$

with Ω_Z as in Eq. (10); note that, if these constraints are further tightened to require that the columns of Z be non-negative and orthonormal, $Z^T Z = I_{k\times k}$, solving this optimization problem (in general, for $\lambda_W = 0$) would yield the hard k-means clustering of the data (Li and Ding, 2006). In the absence of any regularization ($\lambda_W = 0$), it follows from the Eckart-Young

[5] See Ding and He (2004) for additional discussion of the relationship between PCA and k-means.

theorem that for a given basis size k the reconstruction error achieved by PCA is always no larger than that achieved by this convex coding, which in turn achieves a residual error that is smaller than that for k-means. Thus, were the goal to simply achieve the optimal least squares reconstruction error for a given basis size, PCA remains the preferred choice. The more constrained decomposition implied by performing a convex coding may yield advantages in terms of interpretability or feature selection. The choice of regularization Φ_W provides further flexibility in this respect, for instance, an ℓ_1-penalty as in Eq. (8) can be used to induce sparsity, while Gerber et al. (2020) suggest a penalty term of the form

$$\Phi_W(W) = \frac{1}{dk(k-1)} \sum_{i,j=1}^{k} \|\boldsymbol{w}_i - \boldsymbol{w}_j\|_2^2, \tag{12}$$

observing that this places an upper bound on the sensitivity of the reconstruction to changes in the data. It is worth noting that, when combined with an ℓ_2-loss function, as the regularization $\lambda_W \to \infty$ all of the basis vectors $\boldsymbol{w}_i$ reduce to the global mean of the dataset.

For $\lambda_W = 0$, the optimal basis vectors $\boldsymbol{w}_i$ produced by solving Eq. (11) will tend to lie on or outside of the convex hull of the data. Standard AA adds a further more conservative constraint to force the $\boldsymbol{w}_i$ to only lie on the convex hull, thus requiring that the archetypes are realizable in terms of convex combinations of the observed data. Under the more relaxed constraints, while the overall residual error might be reduced, one may extract basis vectors that, in reality, never occur and so are difficult to make sense of. In AA, the additional constraint amounts to requiring that $W = X^T C^T$ for some non-negative $C \in \mathbb{R}^{k\times T}$ with unit ℓ_1-norm rows, i.e., $C \in \Omega_C$ with

$$\Omega_C = \{C \in \mathbb{R}^{k\times T} | C \succeq 0, C\mathbf{1}_T = \mathbf{1}_k\}. \tag{13}$$

The corresponding optimization problem reads

$$(C, Z) = \underset{C\in\Omega_C, Z\in\Omega_Z}{\arg\min} \frac{1}{2T} \|X - ZCX\|_F^2, \tag{14}$$

with Ω_Z as in Eq. (10). ~~The impact of these different choices of cost function and constraints are sketched in Fig. 1.~~As for k-means, the optimization problem to be solved in performing either a convex coding or archetypal analysis cannot be solved exactly, and numerical methods must be employed to find the corresponding factorization. We describe one such algorithm for doing so in Appendix A.

~~The optimization problem posed in either the convex coding case or by AA cannot be solved analytically, as in k-means clustering. Moreover, the combined cost function is not convex in the full set of variables W and Z (or C and Z for AA). It is, though, convex in either W or Z separately, when the other is held fixed, and local stationary points may be straightforwardly found by alternating updates of the basis and weights. For instance, using the penalty function Eq. (12), we may update W with fixed Z via~~

$$W \leftarrow X^T Z \left[Z^T Z + \frac{4T\lambda_W}{dk(k-1)} \left(k I_{k\times k} - 1_{k\times k} \right) \right]^{-1}.$$

For fixed W, Z may then be updated by projected gradient descent. An alternative that avoids having to perform direct projection onto the simplex (Mørup and Hansen, 2012) is to reparameterize the latent weights Z as

$$Z_{ti} = \frac{H_{ti}}{\sum_{i=1}^{k} H_{ti}}, \quad H_{ti} \geq 0,$$

which automatically satisfies the stochastic constraint on $\boldsymbol{z}(t)$, and H_{ti} is required only to be non-negative, making the necessary projection trivial. The factors W and H may then be updated via, e.g., a sequence of projected gradient descent update steps of the form

$$\nabla_W F(W,Z;X) \leftarrow -\tfrac{1}{T}(X^T Z) + \tfrac{1}{T} W\left[Z^T Z + \lambda_W G_W\right],$$
$$W \leftarrow W - \eta_W \nabla_W F(W,Z;X),$$
$$\nabla_Z F(W,Z;X) \leftarrow -\tfrac{1}{T} XW + \tfrac{1}{T} ZW^T W,$$
$$H_{ti} \leftarrow \max\left\{Z_{ti} - \eta_H\left[(\nabla_Z F(W,Z;X))_{ti} - \textstyle\sum_{j=1}^{k} Z_{tj}(\nabla_Z F(W,Z;X))_{tj}\right], 0\right\},$$
$$Z_{ti} \leftarrow \tfrac{H_{ti}}{\sum_{i=1}^{k} H_{ti}},$$

where, for simplicity, we have assumed that the derivative of the penalty term Φ_W may be written in the form

$$\frac{\partial \Phi_W}{\partial W} = \frac{1}{T} W G_W,$$

which is the case if, for instance, $\Phi_W \propto \mathrm{Tr}\left[W G_W W^T\right]$ for some symmetric matrix G_W. The step-size parameters η_W and η_H may either be fixed or determined by performing a line-search.This procedure may be iterated until the successive changes in the total cost fall below a given tolerance. As convergence to a global minimum isnot guaranteed, in practice this procedure is repeated for multiple initial guesses for W and Z to try to improve the likelihood of locating the optimal solution. Inspecting the update equations Eq. (A3) the cost per iteration is seen to scale as $O(dTk) + O(k^2T) + O(k^2d) + O(kd) + O(kT)$. In particular, in the usual case of interest with $d, T \gg k$, the leading contribution to the cost is linear in each of d, T, and k, i.e., $O(dTk)$, making the method suitable for large datasets and comparable with k-means and similar decompositions (Mørup and Hansen, 2012; Gerber et al., 2020).

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 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

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

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.

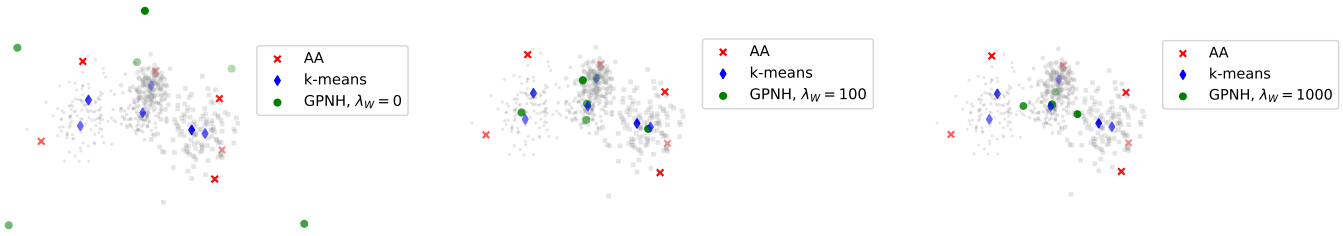

Figure 1. Illustration of the different decompositions obtained using k-means, regularized convex coding (denoted GPNH), and AA, and of the impact of the regularization parameter λ_W when using the penalty term Eq. (12). For increasing (from left to right) $\lambda_W = 0, 100, 1000$, the number of selected features progressively decreases and the basis varies from lying outside the convex hull of the data to the global mean.

3 Case studies

~~Having specified the dimension reduction methods, we now turn to a set of~~ We now turn to a set of case-studies ~~to explore the implications of the different choices made in their definitions~~that demonstrate some of the implications of the various differences noted above in realistic applications. We consider two particular examples that highlight the importance of considering the particular physical features of interest when choosing among possible dimension reduction methods. The first example that we consider, an analysis of SST anomalies, is characterized by a large separation of scales between modes of variability together with key physical modes, particularly ENSO, that can be directly related to extreme values of SST anomalies. This means that the basis vectors or spatial patterns extracted by PCA, k-means, and convex coding are for the most part rather similar in structure, and so the choice of method may be guided by other considerations, such as the level of reconstruction error. We then contrast this scenario with the example of an analysis of mid-latitude geopotential height anomalies, where there is neither scale separation nor do the physical modes coincide solely with extreme anomaly values. As a result, methods that are based on constructing a convex coding of the data, without targeting features that capture the dynamical characteristics of the relevant modes, produce representations that are, arguably, more difficult to interpret, and hence may be less suitable than clustering based methods.

3.1 Sea surface temperature data

Following Hannachi and Trendafilov (2017), we first apply the methods described in Sect. 2 to monthly SST data. The source of the data is the Hadley Centre Sea Surface Temperature dataset (HadISST), version 1.1 (Rayner et al., 2003), consisting of monthly SST values on a $1^\circ \times 1^\circ$ global grid spanning the time period from January 1870 to December 2018. Monthly anomalies are calculated by removing from the full time-series a linear warming trend and additive seasonal component, where the annual cycle is estimated based on the 1981 to 2010 base period. The analysis region is restricted to the region between 45.5°N and 45.5°S.

As the standard and most familiar method, we first perform PCA on the SST anomalies over the time period January 1870 to December 2018 to establish a baseline set of modes. The anomalies at each grid cell are area weighted by the square root of the cosine of the point's latitude. To provide a rough measure of the out-of-sample reconstruction error[6], the EOFs and PCs are evaluated on the first 90% of the anomaly time-series (i.e, January 1870 to February 2004), and the root-mean-square error (RMSE) defined by

$$\mathrm{RMSE}_{\mathrm{train/test}}(k) = \sqrt{\frac{1}{dT}\sum_{i=1}^{d}\sum_{t=1}^{T}[x_i(t) - \hat{x}_i(t)]^2} \tag{15}$$

is computed for the training set and for the test set consisting of the remaining 10% of the data. In Eq. (15), $\hat{x}_i(t)$ is the i^{th} dimension of the reconstruction from k EOFs, and T refers to the size of the training or test set, as appropriate. We consider the results of retaining the first k modes for $k = 1, \ldots, 40$; for reference, the first 40 modes account for approximately 85% of the

[6]Of course, in practice proper estimates of the out-of-sample performance would be obtained by an appropriate cross-validation procedure or similar. However, as here we are primarily interested in the qualitative differences between the different methods in terms of extracting recognizable states, we do not focus on the technical details of model selection or optimizing predictive performance, and simply present these out-of-sample estimates to show the general features of each method.

total variance. The fraction of the total variance in the training set associated with the leading 10 modes is shown in Fig. 2. The most obvious feature of this variance spectrum is the well-known separation between the first and subsequent modes, with the first mode associated with ENSO variability on interannual time-scales and large spatial scales. This is evident in the spatial patterns of the EOFs shown in Fig. 3, in which the first mode shows the canonical ENSO pattern of SST anomalies, while the higher order modes correspond to spatially smaller scale variability.

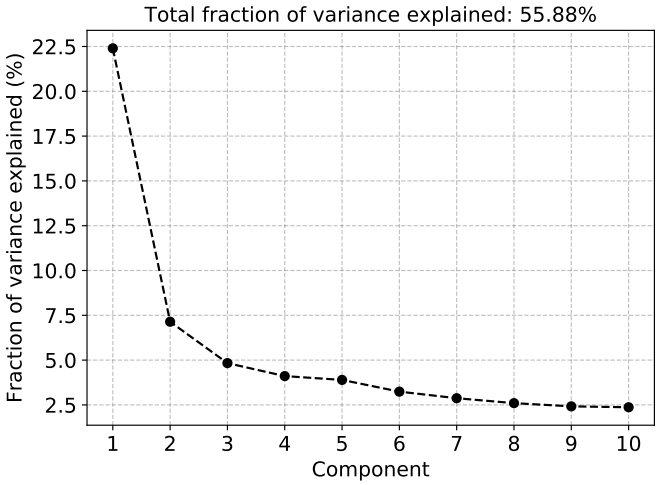

Figure 2. Fraction of variance explained by the first 10 modes obtained from PCA of SST anomalies.

With the above EOF patterns as a point of reference, we now turn to comparing the representation of the dataset produced by each of k-means, archetypal analysis, and convex coding. In each case, the same dataset (i.e., latitude weighted, detrended anomalies) is used as input to the dimension reduction method. The individual data points consist of anomaly maps at each time-step, in other words, the dimension reduction is performed in the state space rather than the sample space (Efimov et al., 1995), such that each dictionary vector corresponds to a particular spatial pattern of SST anomalies. We fit each method using dictionaries of size $k = 1, \ldots, 20$ on the first 90% of the dataset, using the last 10% of samples to provide a simple picture of out-of-sample performance.

We consider the results of a k-means clustering analysis of the data first. For each value of k, the algorithm is restarted 100 times with different initial conditions and the partitioning found to have the lowest reconstruction error is chosen. As is typical of clustering procedures where the number of clusters must be specified a priori, determining the most appropriate choice of k is non-trivial. One commonly used heuristic is to inspect a "scree" plot of the within-cluster sum of squares (Tibshirani et al., 2001),

$$W_k = \sum_{i=1}^{k} \frac{1}{|C_i|} \sum_{t_1, t_2 \in C_i} \|\boldsymbol{x}(t_1) - \boldsymbol{x}(t_2)\|_2^2 , \tag{16}$$

where $|C_i|$ is the size of cluster C_i, as a function of the number of clusters. The preferred number of clusters k^* is identified as the location of an elbow or kink in this curve, if any such feature is present. Alternatively, various indices (see, e.g., Arbelaitz

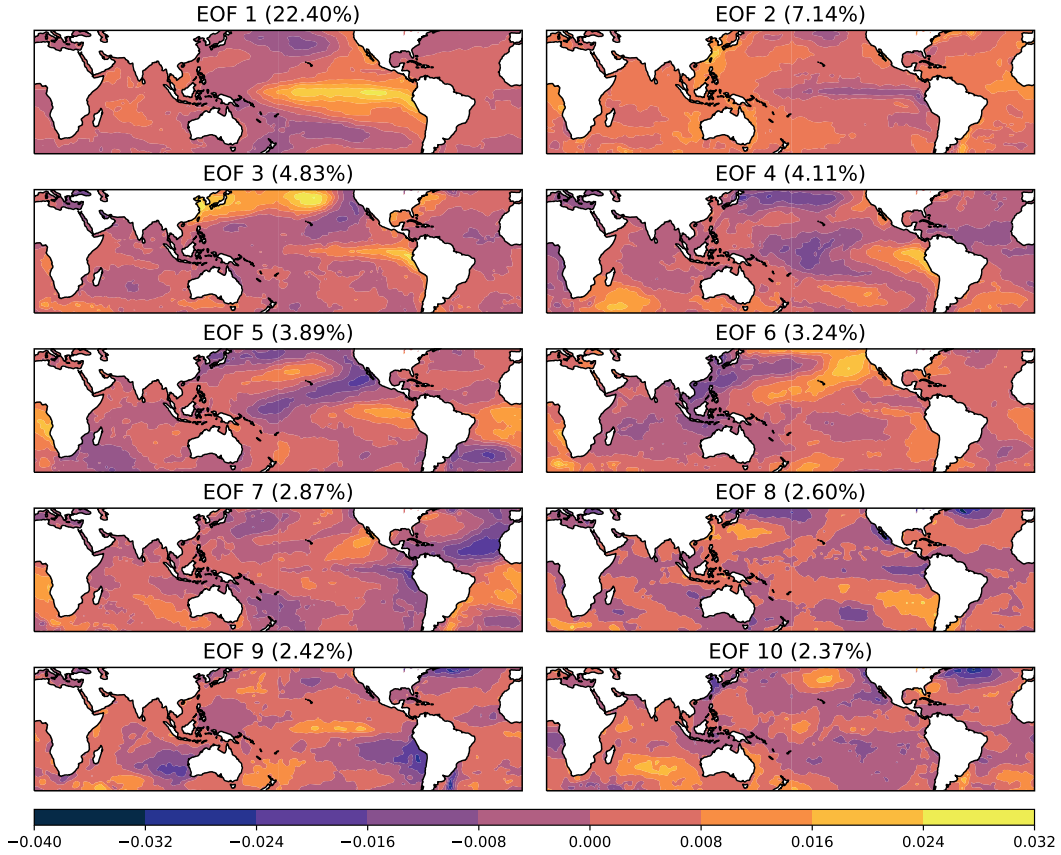

Figure 3. Spatial patterns for the leading 10 modes obtained from PCA of SST anomalies.

et al. (2013) for a review) or Monte Carlo procedures, such as the gap statistic (Tibshirani et al., 2001), have been proposed for assessing whether a given k is suitable. For the present application, plots of the normalized within-cluster sum of squares and the gap statistic computed using 100 Monte Carlo experiments for each k with a null model generated by PCA are shown in Fig. 4. The plot of W_k as a function of k does not show an obvious elbow at any particular $k \leq 20$; similarly, using the gap statistic curve one would conclude that $k = 1$ cluster is preferred[7]. This simply reflects the fact that the anomaly data does not form well-separated clusters (with respect to the assumed Euclidean distance) in the full state-space, making interpreting the different clusters as dynamically distinct regimes difficult, although this does not preclude using the clustering model as a discrete representation of the full dynamics (e.g., Kaiser et al. (2014)).

The partitioning provided by a simple clustering method such as k-means may also still be useful in classifying samples so long as proximity in state space, or similarity more generally, carries meaningful information for an application. In the case of detrended SST anomalies, the magnitudes and spatial distribution of the anomalies are of themselves informative, and

[7]We note that, when using a null model generated by PCA, the gap curve in Fig. 4 might be considered to indicate possible clustering into 5 clusters within the single, large cluster. However, this conclusion depends strongly on the precise null model used to generate the reference data.

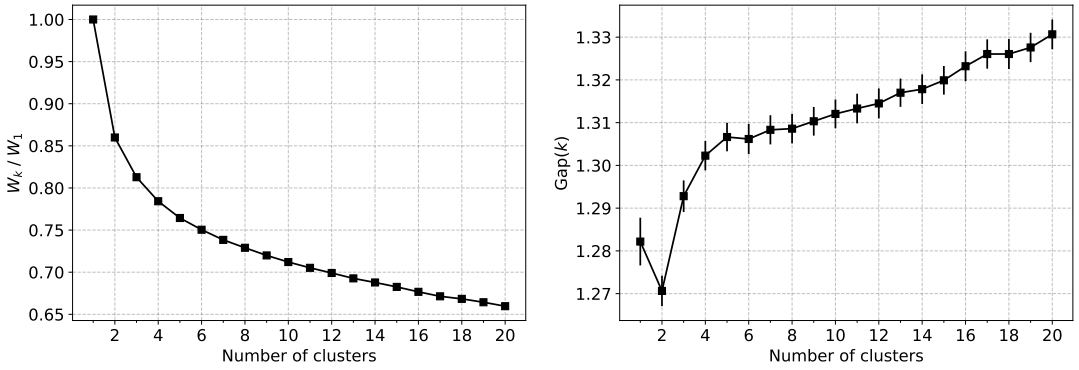

Figure 4. Plots of the normalized within-cluster sum of squares (left), W_k, and the corresponding gap statistic $\mathrm{Gap}(k)$ (right) for k-means clustering of global SST anomalies.

key drivers such as ENSO manifest as relatively large excursions from anomalies associated with higher-frequency variability. Consequently, a k-means clustering of the anomalies does result in a partition in which the cluster centroids, or a subset thereof, can be identified with physical modes, as the algorithm finds several clusters consisting of these extremes of the point cloud while the remainder partition the bulk of the data. The simplest, if somewhat trivial, case for which this can be seen is for $k = 3$ clusters, shown in Fig. 5. Two of the plotted centroids are recognizable as canonical El Niño and La Niña patterns, while

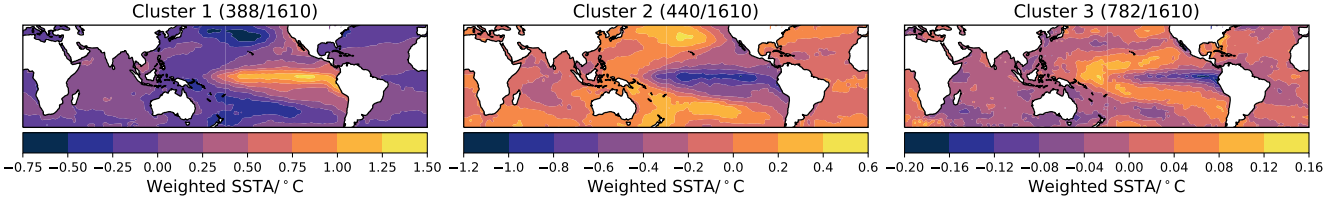

Figure 5. Spatial patterns of the cluster centroids obtained from a k-means clustering of SST anomalies with $k = 3$. The number of months assigned to each cluster is shown above each map.

the last corresponds to a near-climatological state. The former two clusters are dominated by months at the ends of the first principal axis (i.e., along the leading EOF). The fact that two clusters are required to represent both phases of the leading EOF is due to the fact that the overall sign of the cluster centroids, which correspond to points in the data space, is meaningful, unlike in PCA; note that the same is true for the convex coding methods to be considered below. To obtain some sense of the relative distributions of the points within each of the clusters, in Fig. 6 we show the projection of the data into two-dimensions generated by metric multidimensional scaling (MDS). The points assigned to the first and second cluster are those furthest from the mean state, while the remaining cluster accounts for the bulk region. The difference between the first and second centroids is closely aligned with the leading EOF. Similar behavior results when a larger number of clusters is specified.

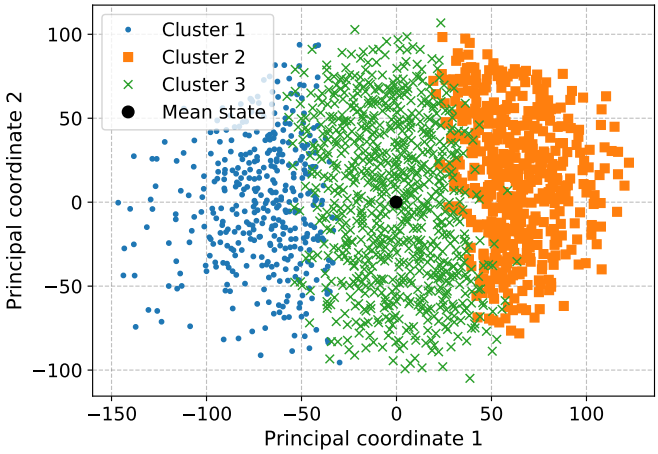

Figure 6. Two-dimensional projection of HadISST SST anomalies obtained by metric MDS with a Euclidean distance measure. The assignment of each point to the clusters produced by k-means clustering with $k = 3$ in Fig. 5 is also shown.

The fact that, due to the dominating role of ENSO variability, a k-means clustering results in several clusters corresponding to large magnitude anomalies in turn implies that the corresponding centroids will be relatively close to the convex hull of the anomaly point cloud. Consequently, these centroids will also closely resemble a subset of the archetypes derived from an archetypal analysis of the same data, at least qualitatively. Relaxing the requirement that the dictionary vectors lie on the convex hull, one may still expect that a convex coding applied to the data will identify features along the same directions, albeit with inflated magnitude so as to reduce the resulting reconstruction error. To verify this, we perform a standard archetypal analysis and a regularized convex coding of the detrended anomalies, in each case restarting the optimization algorithm 100 times and choosing the best encoding from the multiple starts. The regularization in Eq. (12) is used; to illustrate the effect of this regularization, we show results for $\lambda_W = 0$ (i.e., no regularization) and $\lambda_W = 10^3$. The fitted archetypes for $k = 3$ are shown in Fig. 7, and the corresponding dictionary vectors for the regularized convex coding are shown in Fig. 8. The patterns

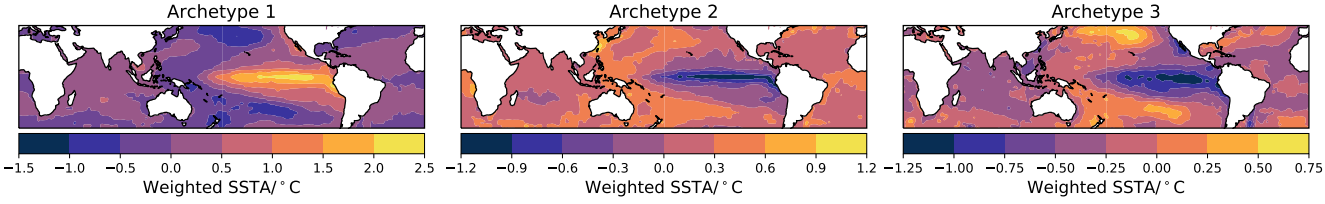

Figure 7. Spatial patterns of the archetypes found from AA with $k = 3$ archetypes.

obtained using the two methods in Fig. 7 and Fig. 8 are very similar; in both cases, El Niño and La Niña patterns are found together with a third configuration containing a cold tongue in the equatorial Pacific with positive anomalies to the north and south, as noted by Hannachi and Trendafilov (2017). However, the magnitude of the anomalies is substantially larger when

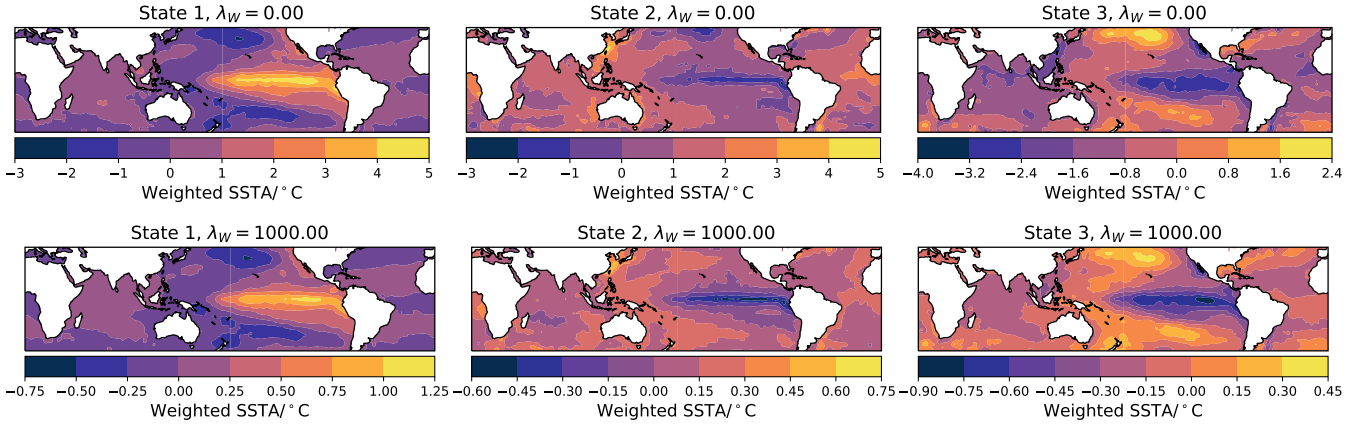

Figure 8. Spatial patterns for the convex coding dictionary vectors for $\lambda_W = 0$ (top) and $\lambda_W = 10^3$.

using an unregularized convex coding as the fitted dictionary vectors are chosen to sit well outside the observed point cloud. While the states correspond to very similar relative signs of anomalies, it is arguably more difficult to interpret the individual dictionary vectors found by the unregularized convex coding physically, as the large anomalies represent far more extreme states than are observed in the data. On the other hand, this also permits a much smaller reconstruction error for a given basis size, and so may be preferable when a higher fidelity reconstruction is required[8]. The effect of the regularization Eq. (12) is to penalize over-dispersion of the dictionary vectors, and so select features that are insensitive to small variations in the input data. For sufficiently large λ_W, the resulting states lie on or within the convex hull of the data, and as a result are more comparable to the states found by AA or k-means. This is illustrated in Fig. 9, from which it can be seen that the basis vectors produced by the regularized convex coding are closer (in Euclidean distance) to those produced by AA than the unregularized basis vectors.

This trade-off between reproducing the data with small errors and constraining the basis vectors to be close to the observed data is also evident in Fig. 10, where the reconstruction RMSE within the training and held-out test datasets for each of the methods are plotted as a function of the dimension of the reduced order representation. In the absence of any regularization, the RMSE produced by a convex coding of the data is close to the globally optimal result obtained from PCA. For $\lambda_W = 10^3$, the obtained RMSE is larger and similar to that for AA, while k-means leads to the largest errors due to the very coarse representation of each datapoint by the centroid of its assigned cluster. It follows that, for a given number of basis vectors, an unregularized convex coding yields performance approaching that of PCA in terms of reconstruction errors, while AA and k-means in turn are more constrained and hence reproduce the data with somewhat larger errors. 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

[8]Similar remarks can be made for the AA/PCH-δ model proposed in Mørup and Hansen (2012).

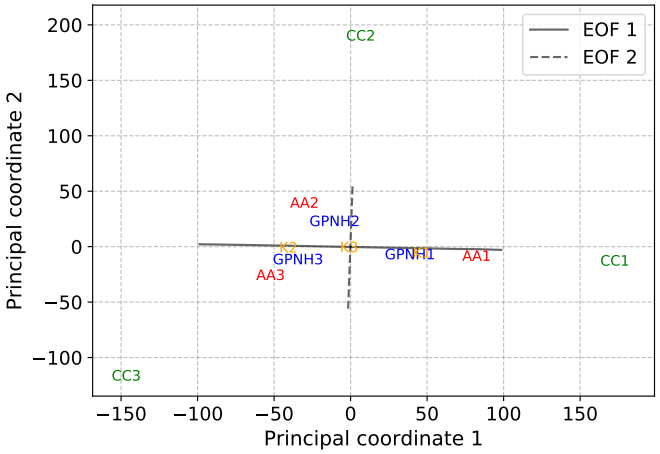

Figure 9. Two-dimensional projection of basis vectors obtained by AA, convex coding, and k-means based on a metric MDS analysis with Euclidean dissimilarities. The projected locations of the basis vectors for each method are indicated by the labels AA, CC, GPNH, and K for AA, convex coding with $\lambda_W = 0$, convex coding with $\lambda_W = 1000$, and k-means, respectively. For reference, the images under the MDS projection of line segments lying along the directions of the first and second EOFs, with lengths proportional to the variance explained by each, are also shown.

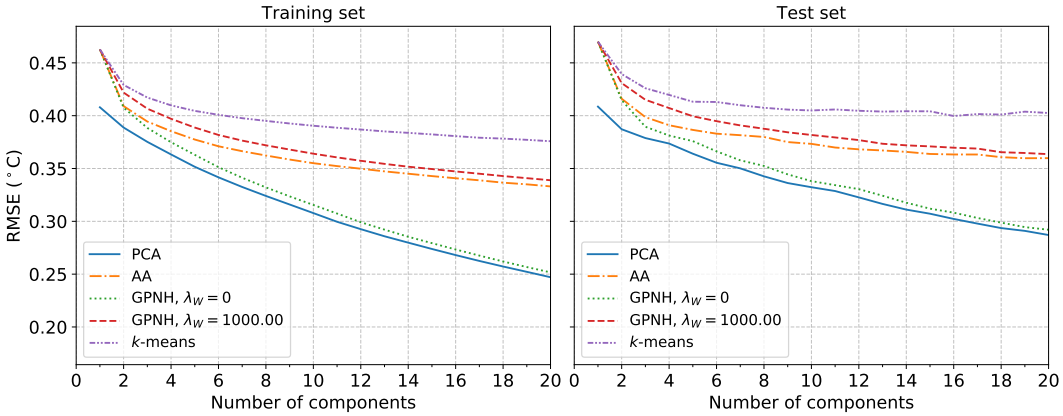

Figure 10. Training and test set RMSE for the reconstruction of SST anomalies resulting from each of the methods.

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.

3.2 Geopotential height anomalies

SST anomalies are an example of a dataset in which a dominant mode is well separated in scale from subleading modes of variability. As a result, all of the dimension reduction methods that we consider extract similar bases (patterns) with which to represent the data, and these can be identified with well-known physical modes. Moreover, physically interesting events such as extremes correspond to large magnitude anomalies relative to the mean state, i.e., at the boundary of the point cloud, and so can be directly extracted by those methods that look for dictionary vectors in the convex hull of the data. This is not true for many variables of interest, however, and so we now compare the behavior of the methods when applied to data that do not exhibit these features.

We consider Northern Hemisphere (NH) daily mean anomalies of 500 hPa geopotential height, $Z'_{g500\text{hPa}}$, between 1 January 1958 and 31 December 2018. The geopotential height data used are obtained from the Japanese 55-year Reanalysis (JRA-55, Kobayashi et al. (2015); Harada et al. (2016)). Anomalies are formed by subtracting the climatological daily mean based on the 1 January 1981 to 31 December 2010 reference period. Unlike monthly SST, there is no clear scale separation in the resulting time series; in Fig. 11 we show the leading EOFs and the percentage of the total variance explained by each. Additionally, the

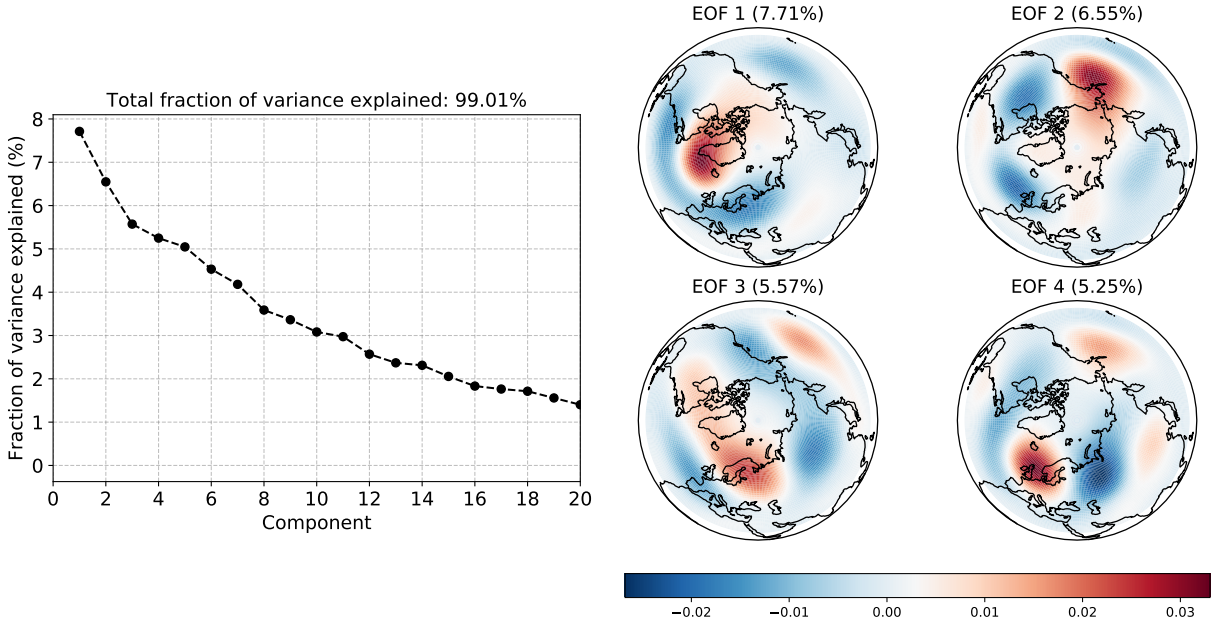

Figure 11. Fraction of the total variance associated with each of the first 20 of 167 retained EOF modes of daily NH 500 hPa geopotential anomalies (left), and the spatial patterns associated with the leading 4 modes. The 167 retained modes account for approximately 99% of the variance.

characterization of regimes is more complicated. In particular, physically significant features in the height field are expected to be quasi-stationary or persistent (Michelangeli et al., 1995; Dole and Gordon, 1983; Mo and Ghil, 1988), and are not solely distinguished by their location in the state space. Such features may therefore be better identified as modes of the state space PDF, arising either due to longer residency times or frequent recurrence of particular states. Where this gives rise to regions of higher density in state space, clustering algorithms such as k-means might be expected to reasonably well represent the corresponding regimes. Convex reduction methods, on the other hand, default to being less sensitive to recurrence; as in the preceding SST example, the fitted dictionary vectors will correspond to points near the boundary of observed data. In doing so, AA and methods based on convex coding preferentially represent the data in terms of deep, but potentially infrequent or highly transient, lows or highs. While this remains an adequate representation simply for reconstructing the full observations, the resulting basis may be more difficult to relate to traditionally identified metastable atmospheric regimes, for instance.

In the case of NH geopotential height anomalies, the cluster centroids, archetypes, and dictionary basis vectors that result from applying k-means, AA, and convex coding to the leading 167 PCs[9] of $Z'_{g500\text{hPa}}$ are shown for $k = 4$ clusters in Fig. 12. Note that inspection of the scree plots (not shown) does not indicate a strong preference for a given number of clusters or states for $k \leq 20$, and we choose $k = 4$ as a simple example. 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. ~~Unlike the SST~~ 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, the $Z'_{g500\text{hPa}}$ data do not exhibit a dominant axis of variability, i.e., there is no clear scale separation between modes. In Fig. 13 this manifests as the absence of a clearly preferred axis along which the basis vectors are distributed[10], c.f., Fig. 9. As the variance in the data is not dominated by a single principal axis, the basis vectors extracted by each of the methods are more evenly distributed around the point cloud, in turn leading to greater differences between the fitted bases.

While all methods identify a feature that is strongly reminiscent of the NAO, there is somewhat more variation in the remaining representative states, in contrast to the case of SST anomalies. In particular, the centroids identified by k-means are no longer in close correspondence with the representative states constructed by either AA or an unregularized convex coding. The k-means centroids are characterized by relatively small magnitude anomalies with positive amplitude at longitudes associated with blocking (Pelly and Hoskins, 2003); in particular, the location of the anticyclonic anomaly in cluster 3 in Fig. 12 closely coincides with the center of action of the EU1 (Barnston and Livezey, 1987) or SCA (Bueh and Nakamura, 2007) pattern associated with Scandinavian blocking. The archetypes, in comparison, appear to exhibit a more pronounced wave-

[9]Clustering on the PCs was done so as to reduce the overall cost of the methods; we have checked that, for small numbers of clusters, the spatial patterns that result are very similar.

[10]Similar behavior is evident for different numbers of states, e.g., for $k = 3$ the MDS projection results in a triangular arrangement of points with the climatological point located close to the centroid of the resulting shape.

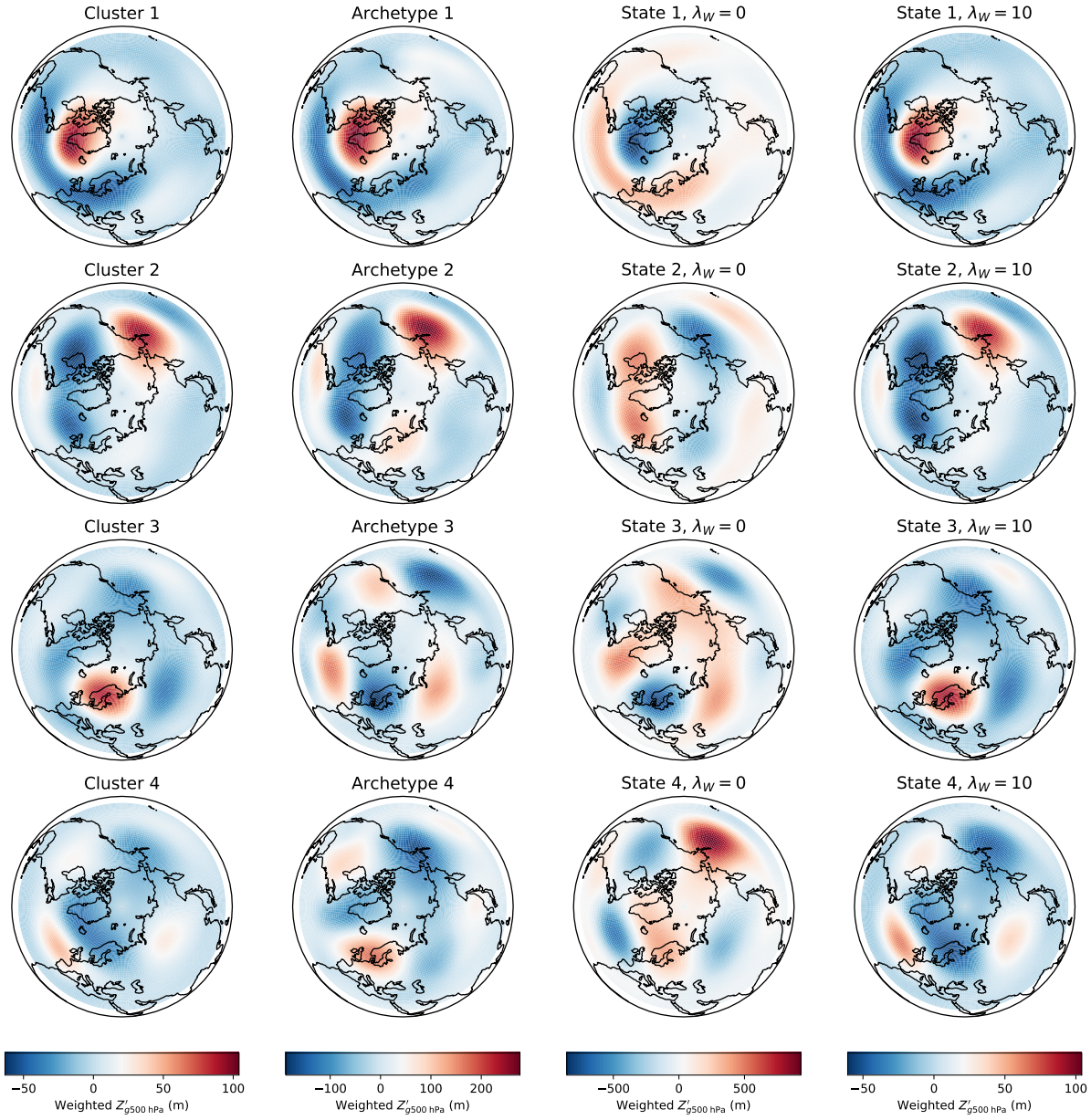

Figure 12. Spatial patterns of geopotential height anomalies corresponding to the k-means centroids (first column), archetypes (second column), convex coding basis vectors with $\lambda_W = 0$ (third column), and convex coding basis vectors with $\lambda_W = 10$.

train structure, in addition to corresponding to generally larger magnitude anomalies. As expected, the basis obtained via an unregularized convex coding represents far larger anomalies again, defining representative states that are much more extreme than the bulk of the observed daily anomaly fields. The role of the regularization in feature selection is clearly illustrated by

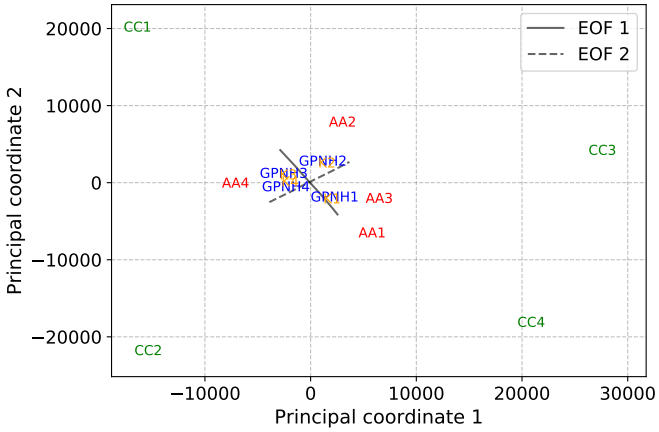

Figure 13. Two dimensional projection of spatial patterns of geopotential height anomalies obtained using the various dimension reduction methods by metric MDS. The results of transforming line segments lying along the directions of the first and second EOFs, as in Fig. 9, are also shown.

comparing the two right-most columns of Fig. 12. By increasing the regularization parameter, the method can be tuned to place less emphasis on capturing large amplitude, noisy variations and target less sensitive features in the data. In this case, for $\lambda_W \approx 1$ the convex coding basis vectors are in close agreement with the archetypes, while for $\lambda_W \approx 10$ they essentially coincide with the k-means centroids. In this sense, by appropriate choice of regularization it is possible to interpolate between a representation of the data in terms of points on the convex hull and mean features, depending on how much weight is placed on minimizing the reconstruction error. While here we consider only a few levels of regularization for illustrative purposes, in general the degree to which this is done, i.e., the choice of λ_W, can be guided by standard model selection methods, such as cross-validation.

In the absence of any regularization, the patterns obtained by a convex coding and by AA correspond to extreme departures from the mean state, but cannot necessarily be directly interpreted as individually representing particular physical extremes. As noted above, this arises due to the fact that such extremes are not necessarily associated with boundaries in state space but may instead be due to extended residence or persistence of a given (non-extreme) state. This difficulty in directly relating atmospheric extremes to a basis produced by AA or similar methods can be clearly demonstrated by considering the representation of a given event in terms of this basis. A dramatic example is provided by the 2010 summer heatwave in western Russia that saw an extended period of well above-average daily temperatures and poor air quality, and was associated with substantial excess mortality and economic losses (Barriopedro et al., 2011; Shaposhnikov et al., 2014). The upper-level circulation during July 2010 was characterized by persistent blocking over eastern Europe (Dole et al., 2011; Matsueda, 2011). The associated monthly mean pattern of height anomalies for that month (see, e.g., Figure 2 of Dole et al. (2011)) most closely resembles the pattern for cluster 3 obtained with k-means in Fig. 12. Consistent with this is the fact that most days during July 2010 are

assigned to this cluster, as shown in Fig. 14. Consequently, cluster 3 might be reasonably well interpreted as representing (a class of) European blocking events.

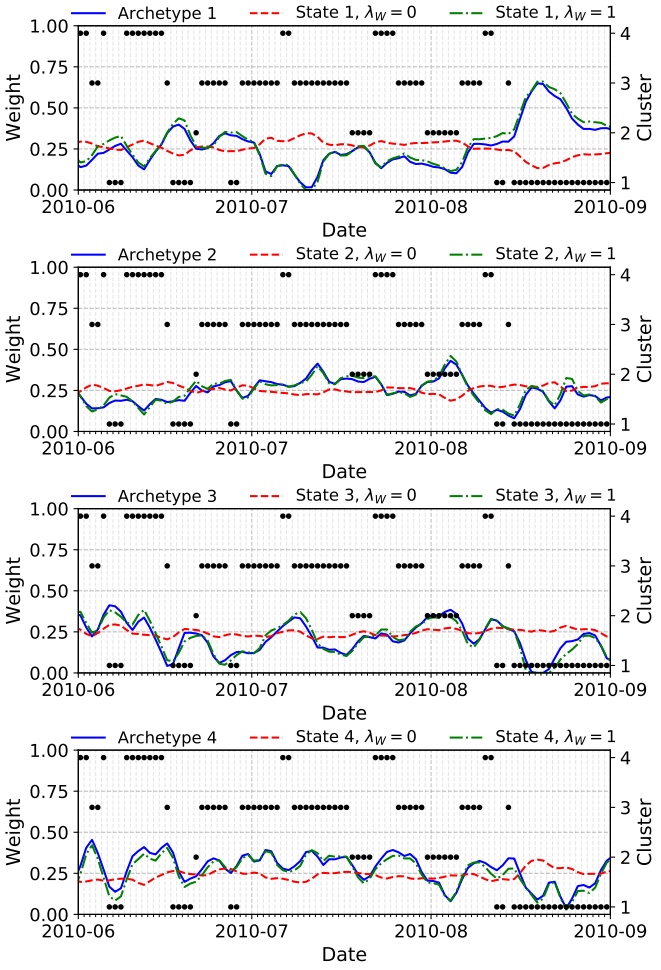

Figure 14. Time-series of basis weights (lines) associated with each archetype produced by AA and the states produced by convex coding with and without regularization, for $k = 4$ states during the 2010 boreal summer. The corresponding k-means cluster assignments for each day are shown as points. Note that the monthly mean $Z'_{g500\text{hPa}}$ for July 2010 associated with the heat wave event most closely resembles cluster 3 in the k-means clustering, to which most of the daily anomalies for that month are also assigned.

In contrast, despite the severity of this event, individual daily height anomalies during July 2010 are not unambiguously identified as extremes by AA or the less constrained convex coding. In Fig. 14 we show the time-series of weights associated with the basis vectors found by AA and by convex coding either without regularization, $\lambda_W = 0$, or with regularization parameter $\lambda_W = 1$, which for these data yields dictionary vectors very similar to those found by AA. For all three cases, roughly equal weights are assigned to each basis vector during July 2010. The complete anomaly is therefore represented

as a mixture of all of the dictionary vectors, rather than being assigned to a single characteristic type of extreme. Evidently, any single basis pattern extracted by these methods does not directly correspond to a particular indicator of extreme weather conditions; extreme events of practical concern may be spread across multiple basis vectors, making identifying such events in this representation more difficult. The hard clustering obtained using k-means is perhaps more easily interpreted in this case, as the resulting cluster affiliations show frequent occurrences of the European blocking cluster (i.e., cluster 3 in Fig. 12 and Fig. 14) during the peak heat-wave period, with a smaller number of days assigned to clusters 2 and 4, suggesting a clearer picture of residence in a single, persistent blocking state. On the other hand, the coarse representation of the actual anomalies by the single cluster 3 centroid is poor compared to the reconstruction provided by AA and convex coding, but could be improved by, for example, making use of a soft clustering algorithm instead. ~~As in the~~ 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, the ambiguous classification of events provided by the convex hull based methods is less of a problem when state space location alone (e.g., temperature anomalies) is in itself a relevant feature. When this is not the case, as here, methods that take advantage of state space density, either due to recurrence or persistence, may be more easily interpreted, or alternatively hybrid approaches could be used in order to partition the state space.

4 Conclusions

Representing a high-dimensional dataset in terms of a highly reduced basis or dictionary is an essential step in many climate analyses. Beyond the practical necessity of doing so, it is usually also desirable for the individual elements of the representation to be identifiable with physically relevant features for the sake of interpretation. A wide range of popular dimension reduction methods, including PCA, k-means clustering, AA, and convex coding, can be written down as particular forms of a basic matrix factorization problem. These methods differ in the details of the measure of cost that is optimized and the feasible solution regions, with the result that different methods yield representations suitable for targeting different features of the data. Of the methods considered here, PCA extracts directions in state space corresponding to maximal variability, k-means locates central points, and AA and similar convex codings identify points on or outside the convex hull of the observed data and so find a representation in terms of extreme points in state space. As different features may be relevant for different applications, it is important to consider these distinctions between these factorization methods and carefully choose a method that is effective for extracting the features of interest.

In some cases, the representations obtained using different dimension reduction methods are very similar, and one can identify more or less easily interpretable features using any given method. This is exemplified by our first case study of SST anomalies, in which the presence of the dominant ENSO mode ensures that PCA, k-means clustering, AA, and convex coding

all identify similar bases corresponding to well-known physical modes. In this case, the main distinction between the methods arises in the nature of the classification of the data (e.g., hard versus soft clustering) and the accuracy with which the original data can be reconstructed for a given level of compression.

As our second case study demonstrates, neglecting the important role played by temporal persistence in dynamically relevant features can lead to representations that are difficult to interpret and may not be as effective for studying persistent states. Clustering-based approaches, or more generally methods that attempt to approximate modes in the PDF rather than targeting the tails of the distribution, are likely to be a better choice in these circumstances. This can also be achieved by appropriate regularization so as to reduce sensitivity to transient features or outliers, which otherwise drive the definition of the basis in methods such as AA. In all of the methods that we have considered, a lack of independence in time is not explicitly modeled. Extensions to the simple methods that we have considered to account for non-independence are also possible, albeit usually at the cost of increased complexity. Singular spectrum analysis (see, e.g., Jolliffe (1986)) is a familiar example of one such extension for PCA. Similar generalizations can also be constructed for convex decompositions of the data. For instance, by virtue of the decomposability of the least squares cost function, it is possible to construct a joint convex discretization of instantaneous and lagged values of the variables of interest, which forms the basis of the scalable probabilistic approximation method (Gerber et al., 2020). In this approach, the decomposition may be parametrized in terms of a transition matrix relating the weights at different times, thus naturally incorporating a temporal constraint into the discretization. An associated regularization parameter allows the amount of temporal regularization to be appropriately tuned. Moreover, because the optimization problem remains separable in this case, the individual optimization steps can be parallelized and the method remains scalable. More sophisticated regularization strategies (Horenko, 2020) can further improve upon the performance of the method, even in the case of relatively small sample sizes and feature degeneracies.

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. An underlying probabilistic model is already suggested by the stochastic constraints that are imposed on the weights in AA and in convex coding, a feature that is already taken advantage of in the case of the scalable probabilistic approximation. Corresponding latent variable models can naturally be constructed for PCA (Roweis, 1998; Tipping and Bishop, 1999) and AA (Seth and Eugster, 2016). From the point of view of incorporating temporal dependence between samples, starting from a probabilistic model is fruitful as it is conceptually straightforward to incorporate a model for the dynamics of the latent weights $\boldsymbol{z}_t$ (Lawrence, 2005; Wang et al., 2006; Damianou et al., 2011). Taken together with a choice of conditional distribution for the observations, point estimation in the latent variable model is once again achieved by optimization of a suitable loss function. Regularization is in this context provided by suitable choice of prior distributions for the dictionary and weights. An additional advantage of starting from a generative model is the possibility of applying the full machinery of Bayesian inference rather than obtaining only point estimates (Mnih and Salakhutdinov, 2008; Salakhutdinov and Mnih, 2008; Virtanen et al., 2008; Shan and Banerjee, 2010; Gönen et al., 2013), although scaling such analyses remains a challenging issue. While this approach appears to be natural for constructing dimension reduction methods that may flexibly take into account more complicated dependence structures, it is by no means guaranteed to provide

improved low-dimensional representations of the data, and likely depends heavily on choices such as the assumed generative process for the weights. For example, in simple constructions with weights drawn according to a Gaussian process, in the absence of strong prior information the fitted bases are driven to be very similar to those found by ordinary PCA in order to maximize the likelihood of the observed data, while at the same time being substantially more expensive to fit. Thus, the development of temporally regularized methods derived from underlying stochastic models remains a topic for further investigation.

Appendix A: Numerical solution for convex coding dictionary and archetypes

As noted in Sect. 2, the optimization problem posed in either the convex coding case or by AA cannot be solved analytically. Moreover, the combined cost function is not convex in the full set of variables W and Z (or C and Z for AA). It is, though, convex in either W or Z separately, when the other is held fixed, and local stationary points may be straightforwardly found by alternating updates of the basis and weights. For instance, using the penalty function Eq. (12), we may update W with fixed Z via

$$W \leftarrow X^T Z \left[Z^T Z + \frac{4T\lambda_W}{dk(k-1)} \left(k I_{k\times k} - 1_{k\times k} \right) \right]^{-1}. \tag{A1}$$

For fixed W, Z may then be updated by projected gradient descent. An alternative that avoids having to perform direct projection onto the simplex (Mørup and Hansen, 2012) is to reparameterize the latent weights Z as

$$Z_{ti} = \frac{H_{ti}}{\sum_{i=1}^{k} H_{ti}}, \quad H_{ti} \geq 0, \tag{A2}$$

which automatically satisfies the stochastic constraint on $\boldsymbol{z}(t)$, and H_{ti} is required only to be non-negative, making the necessary projection trivial. The factors W and H may then be updated via, e.g., a sequence of projected gradient descent update steps of the form

$$\nabla_W F(W,Z;X) \leftarrow -\frac{1}{T}(X^T Z) + \frac{1}{T} W \left[Z^T Z + \lambda_W G_W \right], \tag{A3a}$$

$$W \leftarrow W - \eta_W \nabla_W F(W,Z;X), \tag{A3b}$$

$$\nabla_Z F(W,Z;X) \leftarrow -\frac{1}{T} XW + \frac{1}{T} Z W^T W, \tag{A3c}$$

$$H_{ti} \leftarrow \max \left\{ Z_{ti} - \eta_H \left[(\nabla_Z F(W,Z;X))_{ti} - \sum_{j=1}^{k} Z_{tj} (\nabla_Z F(W,Z;X))_{tj} \right], 0 \right\}, \tag{A3d}$$

$$Z_{ti} \leftarrow \frac{H_{ti}}{\sum_{i=1}^{k} H_{ti}}, \tag{A3e}$$

where, for simplicity, we have assumed that the derivative of the penalty term Φ_W may be written in the form

$$\frac{\partial \Phi_W}{\partial W} = \frac{1}{T} W G_W,$$

which is the case if, for instance, $\Phi_W \propto \mathrm{Tr}\left[W G_W W^T\right]$ for some symmetric matrix G_W. The step-size parameters η_W and η_H may either be fixed or determined by performing a line-search. This procedure may be iterated until the successive changes in the total cost fall below a given tolerance. As convergence to a global minimum is not guaranteed, in practice this procedure is repeated for multiple initial guesses for W and Z to try to improve the likelihood of locating the optimal solution. Inspecting the update equations Eq. (A3), the cost per iteration is seen to scale as $O(dTk) + O(k^2T) + O(k^2d) + O(kd) + O(kT)$. In particular, in the usual case of interest with $d, T \gg k$, the leading contribution to the cost is linear in each of d, T, and k, i.e., $O(dTk)$, making the method suitable for large datasets and comparable with k-means and similar decompositions (Mørup and Hansen, 2012; Gerber et al., 2020).

Code and data availability. The HadISST SST dataset used in this study is provided by the UK Met Office Hadley Centre and may be accessed at https://www.metoffice.gov.uk/hadobs/hadisst/. The JRA-55 geopotential height data used is made available through the JRA-55 project and may be accessed following the procedures described at https://jra.kishou.go.jp/JRA-55/index_en.html. All source code used to perform the analyses presented in the main text may be found at https://doi.org/10.5281/zenodo.3723948.

Author contributions. All of the authors designed the study. TO proposed the specific case studies, and DH implemented the code to perform the numerical comparisons and generated all plots and figures. All of the authors contributed to the direction of the study, discussion of results, and the writing and approval of the manuscript.

Competing interests. The authors declare that they have no conflicts of interest.

Acknowledgements. The authors would like to thank Illia Horenko for continual guidance and for his valuable inputs over the course of this project, and Vassili Kitsios and Didier Monselesan for encouragement and helpful discussions about the various methods. DH is supported by the Australian Commonwealth Scientific and Industrial Research Organisation (CSIRO) through a ResearchPlus postdoctoral fellowship. TO is supported by the CSIRO Decadal Climate Forecasting Project (https://research.csiro.au/dfp). The PCA and k-means results, and the visualizations using metric MDS, were obtained using the routines provided by the scikit-learn Python package (Pedregosa et al., 2011). Plots were generated using the Python package Matplotlib (Hunter, 2007).

References

Aloise, D., Deshpande, A., Hansen, P., and Popat, P.: NP-hardness of Euclidean sum-of-squares clustering, Machine Learning, pp. 245–248, https://doi.org/https://doi.org/10.1007/s10994-009-5103-0, 2009.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., and Perona, I.: An extensive comparative study of cluster validity indices, Pattern Recognition, 46, 243–256, 2013.

Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J.: Clustering with Bregman Divergences, Journal of Machine Learning Research, 6, 1705–1749, https://doi.org/10.1007/s10994-005-5825-6, 2005.

Barnston, A. G. and Livezey, R. E.: Classification, Seasonality and Persistence of Low-Frequency Atmospheric Circulation Patterns, Monthly Weather Review, 115, 1083–1126, https://doi.org/10.1175/1520-0493(1987)115<1083:csapol>2.0.co;2, 1987.

Barriopedro, D., Fischer, E. M., Luterbacher, J., Trigo, R. M., and García-Herrera, R.: The Hot Summer of 2010: Redrawing the Temperature Record Map of Europe, Science, 332, 220–224, https://doi.org/10.1126/science.1201224, https://science.sciencemag.org/content/332/6026/220, 2011.

Bezdek, J. C., Ehrlich, R., and Full, W.: FCM: The Fuzzy c-Means Clustering Algorithm, Computers and Geosciences, 10, 191–203, https://doi.org/10.1109/igarss.1988.569600, 1984.

Bregman, L.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming, USSR Computational Mathematics and Mathematical Physics, 7, 200 – 217, https://doi.org/https://doi.org/10.1016/0041-5553(67)90040-7, http://www.sciencedirect.com/science/article/pii/0041555367900407, 1967.

Bueh, C. and Nakamura, H.: Scandinavian pattern and its climatic impact, Quarterly Journal of the Royal Meteorological Society, 133, 2117–2131, https://doi.org/10.1002/qj.173, https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.173, 2007.

Cheng, X. and Wallace, J. M.: Cluster Analysis of the Northern Hemisphere Wintertime 500-hPa Height Field: Spatial Patterns, Journal of the Atmospheric Sciences, 50, 2674–2696, https://doi.org/10.1175/1520-0469(1993)050<2674:CAOTNH>2.0.CO;2, https://doi.org/10.1175/1520-0469(1993)050<2674:CAOTNH>2.0.CO;2, 1993.

Christiansen, B.: Atmospheric Circulation Regimes: Can Cluster Analysis Provide the Number?, Journal of Climate, 20, 2229–2250, https://doi.org/10.1175/JCLI4107.1, https://doi.org/10.1175/JCLI4107.1, 2007.

Cutler, A. and Breiman, L.: Archetypal Analysis, Technometrics, 36, 338–347, 1994.

Damianou, A. C., Titsias, M. K., and Lawrence, N. D.: Variational Gaussian Process Dynamical Systems, arXiv e-prints, 2011.

Ding, C. and He, X.: K-means clustering via principal component analysis, in: Proceedings of the twenty-first international conference on Machine learning, p. 29, ACM, 2004.

Dole, R., Hoerling, M., Perlwitz, J., Eischeid, J., Pegion, P., Zhang, T., Quan, X.-W., Xu, T., and Murray, D.: Was there a basis for anticipating the 2010 Russian heat wave?, Geophysical Research Letters, 38, https://doi.org/10.1029/2010GL046582, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2010GL046582, 2011.

Dole, R. M. and Gordon, N. D.: Persistent Anomalies of the Extratropical Northern Hemisphere Wintertime Circulation: Geographical Distribution and Regional Persistence Characteristics, Monthly Weather Review, 111, 1567–1586, https://doi.org/10.1175/1520-0493(1983)111<1567:PAOTEN>2.0.CO;2, https://doi.org/10.1175/1520-0493(1983)111<1567:PAOTEN>2.0.CO;2, 1983.

Dommenget, D. and Latif, M.: A cautionary note on the interpretation of EOFs, Journal of Climate, 15, 216–225, https://doi.org/10.1175/1520-0442(2002)015<0216:ACNOTI>2.0.CO;2, 2002.

Dunn, J. C.: A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, Journal of Cybernetics, 3, 32–57, https://doi.org/10.1080/01969727308546046, 1973.

Eckart, C. and Young, G.: The approximation of one matrix by another of lower rank, Psychometrika, 1, 211–218, https://doi.org/10.1007/BF02288367, https://doi.org/10.1007/BF02288367, 1936.

Efimov, V., Prusov, A., and Shokurov, M.: Patterns of interannual variability defined by a cluster analysis and their relation with ENSO, Quarterly Journal of the Royal Meteorological Society, 121, 1651–1679, 1995.

Fereday, D. R., Knight, J. R., Scaife, A. A., Folland, C. K., and Philipp, A.: Cluster Analysis of North Atlantic–European Circulation Types and Links with Tropical Pacific Sea Surface Temperatures, Journal of Climate, 21, 3687–3703, https://doi.org/10.1175/2007JCLI1875.1, https://doi.org/10.1175/2007JCLI1875.1, 2008.

Forgey, E.: Cluster analysis of multivariate data: Efficiency vs. interpretability of classification, Biometrics, 21, 768–769, 1965.

Gerber, S., Pospisil, L., Navandar, M., and Horenko, I.: Low-cost scalable discretization, prediction, and feature selection for complex systems, Science Advances, 6, https://doi.org/10.1126/sciadv.aaw0961, 2020.

Gönen, M., Khan, S., and Kaski, S.: Kernelized Bayesian matrix factorization, in: International Conference on Machine Learning, pp. 864–872, http://proceedings.mlr.press/v28/gonen13a.pdf, 2013.

Hannachi, A. and Legras, B.: Simulated annealing and weather regimes classification, Tellus A, 47, 955–973, https://doi.org/10.1034/j.1600-0870.1995.00203.x, https://onlinelibrary.wiley.com/doi/abs/10.1034/j.1600-0870.1995.00203.x, 1995.

Hannachi, A. and Trendafilov, N.: Archetypal analysis: Mining weather and climate extremes, Journal of Climate, 30, 6927–6944, https://doi.org/10.1175/JCLI-D-16-0798.1, 2017.

Hannachi, A., Jolliffe, I. T., and Stephenson, D. B.: Empirical orthogonal functions and related techniques in atmospheric science: A review, International Journal of Climatology, 27, 1119–1152, https://doi.org/10.1002/joc, 2007.

Harada, Y., Kamahori, H., Kobayashi, C., Endo, H., Kobayashi, S., Ota, Y., Onoda, H., Onogi, K., Miyaoka, K., and Takahashi, K.: The JRA-55 Reanalysis: Representation of Atmospheric Circulation and Climate Variability, Journal of the Meteorological Society of Japan. Ser. II, 94, 269–302, https://doi.org/10.2151/jmsj.2016-015, 2016.

Hartigan, J. A. and Wong, M. A.: Algorithm AS 136: A K-Means Clustering Algorithm, Journal of the Royal Statistical Society. Series C (Applied Statistics), 28, 100–108, http://www.jstor.org/stable/2346830, 1979.

Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer, 2005.

Horenko, I.: On a scalable entropic breaching of the overfitting barrier in machine learning, Neural Computation, in press, https://arxiv.org/abs/2002.03176, 2020.

Hunter, J. D.: Matplotlib: A 2D graphics environment, Computing in Science & Engineering, 9, 90–95, https://doi.org/10.1109/MCSE.2007.55, 2007.

Huth, R., Beck, C., Philipp, A., Demuzere, M., Ustrnul, Z., Cahynová, M., Kyselý, J., and Tveito, O. E.: Classifications of Atmospheric Circulation Patterns, Annals of the New York Academy of Sciences, 1146, 105–152, https://doi.org/10.1196/annals.1446.019, 2008.

Jenatton, R., Obozinski, G., and Bach, F.: Structured Sparse Principal Component Analysis, in: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, vol. 9, pp. 366–373, http://arxiv.org/abs/0909.1440, 2010.

Jolliffe, I.: Principal component analysis, Springer Verlag, New York, 1986.

Jolliffe, I. T., Trendafilov, N. T., and Uddin, M.: A Modified Principal Component Technique Based on the LASSO, Journal of Computational and Graphical Statistics, 12, 531–547, https://doi.org/10.1198/1061860032148, https://doi.org/10.1198/1061860032148, 2003.

Kaiser, E., Noack, B. R., Cordier, L., Spohn, A., Segond, M., Abel, M., Daviller, G., Östh, J., Krajnović, S., and Niven, R. K.: Cluster-based reduced-order modelling of a mixing layer, Journal of Fluid Mechanics, 754, 365–414, 2014.

Kaiser, H. F.: The varimax criterion for analytic rotation in factor analysis, Psychometrika, 23, 187–200, https://doi.org/10.1007/BF02289233, https://doi.org/10.1007/BF02289233, 1958.

Kidson, J. W.: THE UTILITY OF SURFACE AND UPPER AIR DATA IN SYNOPTIC CLIMATOLOGICAL SPECIFICATION OF SURFACE CLIMATIC VARIABLES, International Journal of Climatology, 17, 399–413, https://doi.org/10.1002/(SICI)1097-0088(19970330)17:4<399::AID-JOC108>3.0.CO;2-M, 1997.

Kidson, J. W.: An analysis of New Zealand synoptic types and their use in defining weather regimes, International Journal of Climatology, 20, 299–316, https://doi.org/10.1002/(SICI)1097-0088(20000315)20:3<299::AID-JOC474>3.0.CO;2-B, 2000.

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H., Miyaoka, K., and Takahashi, K.: The JRA-55 Reanalysis: General Specifications and Basic Characteristics, Journal of the Meteorological Society of Japan. Ser. II, 93, 5–48, https://doi.org/10.2151/jmsj.2015-001, 2015.

Lau, K.-M., Sheu, P.-J., and Kang, I.-S.: Multiscale Low-Frequency Circulation Modes in the Global Atmosphere, Journal of the Atmospheric Sciences, 51, 1169–1193, 1994.

Lawrence, N.: Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models, J. Mach. Learn. Res., 6, 1783–1816, http://dl.acm.org/citation.cfm?id=1046920.1194904, 2005.

Lee, D. D. and Seung, H. S.: Unsupervised Learning by Convex and Conic Coding, in: Advances in neural information processing systems, pp. 515–521, 1997.

Lee, H., Battle, A., Raina, R., and Ng, A. Y.: Efficient Sparse Coding Algorithms, in: Proceedings of the 19th International Conference on Neural Information Processing Systems, NIPS'06, pp. 801–808, MIT Press, Cambridge, MA, USA, http://dl.acm.org/citation.cfm?id=2976456.2976557, 2006.

Legras, B., Desponts, T., and Piguet, B.: Cluster analysis and weather regimes, in: Seminar on the Nature and Prediction of Extra Tropical Weather Systems. 7-11 September 1987, vol. II, pp. 123–150, ECMWF, ECMWF, Shinfield Park, Reading, https://www.ecmwf.int/node/10704, 1987.

Li, T. and Ding, C.: The Relationships Among Various Nonnegative Matrix Factorization Methods for Clustering, in: Sixth International Conference on Data Mining (ICDM'06), pp. 362–371, https://doi.org/10.1109/ICDM.2006.160, 2006.

Lloyd, S.: Least squares quantization in PCM, IEEE Transactions on Information Theory, 28, 129–137, https://doi.org/10.1109/TIT.1982.1056489, 1982.

Lorenz, E. N.: Empirical Orthogonal Functions and Statistical Weather Prediction, Tech. rep., Massachusetts Institute of Technology, Cambridge, 1956.

MacKay, D. J. C.: Information theory, inference and learning algorithms, Cambridge University Press, 2003.

MacQueen, J.: Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297, https://doi.org/10.1007/s11665-016-2173-6, 1967.

Mahajan, M., Nimbhorkar, P., and Varadarajan, K.: The planar k-means problem is NP-hard, Theoretical Computer Science, 442, 13 – 21, https://doi.org/https://doi.org/10.1016/j.tcs.2010.05.034, http://www.sciencedirect.com/science/article/pii/S0304397510003269, special Issue on the Workshop on Algorithms and Computation (WALCOM 2009), 2012.

Mairal, J., Bach, F., Ponce, J., and Sapiro, G.: Online Dictionary Learning for Sparse Coding, in: Proceedings of the 26th International Conference on Machine Learning, https://doi.org/10.1145/1553374.1553463, papers://d471b97a-e92c-44c2-8562-4efc271c8c1b/Paper/p46, 2009.

Matsueda, M.: Predictability of Euro-Russian blocking in summer of 2010, Geophysical Research Letters, 38, https://doi.org/10.1029/2010GL046557, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2010GL046557, 2011.

Michelangeli, P.-A., Vautard, R., and Legras, B.: Weather Regimes: Recurrence and Quasi Stationarity, Journal of the Atmospheric Sciences, 52, 1237–1256, https://doi.org/10.1175/1520-0469(1995)052<1237:WRRAQS>2.0.CO;2, https://doi.org/10.1175/1520-0469(1995)052<1237:WRRAQS>2.0.CO;2, 1995.

Mnih, A. and Salakhutdinov, R. R.: Probabilistic matrix factorization, in: Advances in neural information processing systems, pp. 1257–1264, http://papers.nips.cc/paper/3208-probabilistic-matrix-factorization.pdf, 2008.

Mo, K. and Ghil, M.: Cluster analysis of multiple planetary flow regimes, Journal of Geophysical Research: Atmospheres, 93, 10 927–10 952, https://doi.org/10.1029/JD093iD09p10927, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JD093iD09p10927, 1988.

Molteni, F., Tibaldi, S., and Palmer, T. N.: Regimes in the wintertime circulation over northern extratropics. I: Observational evidence, Quarterly Journal of the Royal Meteorological Society, 116, 31–67, https://doi.org/10.1002/qj.49711649103, 1990.

Monahan, A. H., Fyfe, J. C., Ambaum, M. H. P., Stephenson, D. B., and North, G. R.: Empirical Orthogonal Functions : The Medium is the Message, Journal of Climate, 22, 6501–6514, https://doi.org/10.1175/2009JCLI3062.1, 2009.

Mørup, M. and Hansen, L. K.: Archetypal analysis for machine learning and data mining, Neurocomputing, 80, 54–63, https://doi.org/10.1016/j.neucom.2011.06.033, 2012.

Neal, R., Fereday, D., Crocker, R., and Comer, R. E.: A flexible approach to defining weather patterns and their application in weather forecasting over Europe, Meteorological Applications, 23, 389–400, https://doi.org/10.1002/met.1563, 2016.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, 2011.

Pelly, J. L. and Hoskins, B. J.: A New Perspective on Blocking, Journal of the Atmospheric Sciences, 60, 743–755, https://doi.org/10.1175/1520-0469(2003)060<0743:ANPOB>2.0.CO;2, https://doi.org/10.1175/1520-0469(2003)060<0743:ANPOB>2.0.CO;2, 2003.

Pohl, B. and Fauchereau, N.: The Southern Annular Mode Seen through Weather Regimes, Journal of Climate, 25, 3336–3354, https://doi.org/10.1175/JCLI-D-11-00160.1, https://doi.org/10.1175/JCLI-D-11-00160.1, 2012.

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E. C., and Kaplan, A.: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, Journal of Geophysical Research: Atmospheres, 108, https://doi.org/10.1029/2002JD002670, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002JD002670, 2003.

Renwick, J. A.: Persistent Positive Anomalies in the Southern Hemisphere Circulation, Monthly Weather Review, 133, 977–988, https://doi.org/10.1175/MWR2900.1, https://doi.org/10.1175/MWR2900.1, 2005.

Richman, M. B.: Rotation of Principal Components, Journal of Climatology, 6, 293–335, https://doi.org/10.1177/1746847713485834, 1986.

Roweis, S. T.: EM algorithms for PCA and SPCA, in: Advances in neural information processing systems, pp. 626–632, http://papers.nips.cc/paper/1398-em-algorithms-for-pca-and-spca.pdf, 1998.

Ruspini, E. H.: A New Approach to Clustering, Information and Control, 15, 22–32, 1969.

Salakhutdinov, R. and Mnih, A.: Bayesian Probabilistic Matrix Factorization Using Markov Chain Monte Carlo, in: Proceedings of the 25th International Conference on Machine Learning, ICML '08, pp. 880–887, ACM, New York, NY, USA, https://doi.org/10.1145/1390156.1390267, http://doi.acm.org/10.1145/1390156.1390267, 2008.

Seth, S. and Eugster, M. J.: Probabilistic archetypal analysis, Machine Learning, 102, 85–113, https://doi.org/10.1007/s10994-015-5498-8, 2016.

Shan, H. and Banerjee, A.: Generalized probabilistic matrix factorizations for collaborative filtering, in: 2010 IEEE International Conference on Data Mining, pp. 1025–1030, IEEE, https://doi.org/10.1109/ICDM.2010.116, https://ieeexplore.ieee.org/abstract/document/5694079, 2010.

Shaposhnikov, D., Revich, B., Bellander, T., Bedada, G. B., Bottai, M., Kharkova, T., Kvasha, E., Lezina, E., Lind, T., Semutnikova, E., et al.: Mortality related to air pollution with the Moscow heat wave and wildfire of 2010, Epidemiology (Cambridge, Mass.), 25, 359, 2014.

Singh, A. P. and Gordon, G. J.: A Unified View of Matrix Factorization Models, in: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases - Part II, pp. 358–373, http://www.cs.cmu.edu/{~}ggordon/singh-gordon-unified-factorization-ecml.pdf, 2008.

Steinschneider, S. and Lall, U.: Daily Precipitation and Tropical Moisture Exports across the Eastern United States: An Application of Archetypal Analysis to Identify Spatiotemporal Structure, Journal of Climate, 28, 8585–8602, https://doi.org/10.1175/JCLI-D-15-0340.1, 2015.

Stone, E. and Cutler, A.: Introduction to archetypal analysis of spatio-temporal dynamics, Physica D: Nonlinear Phenomena, 96, 110 – 131, https://doi.org/https://doi.org/10.1016/0167-2789(96)00016-4, http://www.sciencedirect.com/science/article/pii/0167278996000164, measures of Spatio-Temporal Dynamics, 1996.

Stone, R. C.: Weather types at Brisbane, Queensland: An example of the use of principal components and cluster analysis, International Journal of Climatology, 9, 3–32, https://doi.org/10.1002/joc.3370090103, 1989.

Straus, D. M., Corti, S., and Molteni, F.: Circulation Regimes: Chaotic Variability versus SST-Forced Predictability, Journal of Climate, 20, 2251–2272, https://doi.org/10.1175/JCLI4070.1, https://doi.org/10.1175/JCLI4070.1, 2007.

Tibshirani, R., Walther, G., and Hastie, T.: Estimating the number of clusters in a data set via the gap statistic, Journal of the Royal Statistical Society. Series B: Statistical Methodology, 63, 411–423, https://doi.org/10.1111/1467-9868.00293, 2001.

Tipping, M. E. and Bishop, C. M.: Probabilistic Principal Component Analysis, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61, 611–622, https://doi.org/10.1111/1467-9868.00196, 1999.

Virtanen, T., Cemgil, A. T., and Godsill, S.: Bayesian extensions to non-negative matrix factorisation for audio signal modelling, ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pp. 1825–1828, https://doi.org/10.1109/ICASSP.2008.4517987, 2008.

Wang, C., Xie, S.-P., and Carton, J. A.: A Global Survey of Ocean–Atmosphere Interaction and Climate Variability, pp. 1–19, American Geophysical Union (AGU), https://doi.org/10.1029/147GM01, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/147GM01, 2004.

Wang, J., Hertzmann, A., and Fleet, D. J.: Gaussian process dynamical models, in: Advances in neural information processing systems, pp. 1441–1448, 2006.

Witten, D. M., Tibshirani, R., and Hastie, T.: A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, Biostatistics, 10, 515–534, https://doi.org/10.1093/biostatistics/kxp008, 2009.