Archetypal Analysis of Geophysical Data illustrated by Sea Surface Temperature

Amanda S. Black, Didier P. Monselesan, James S. Risbey, Bernadette M. Sloyan, Christopher C. Chapman, Abdelwaheb Hannachi, Doug Richardson, Dougal T. Squire, Carly R. Tozer, Nikolay Trendafilov

a International Laboratory for High-Resolution Earth System Prediction, Texas A&M University, College Station, TX, USA
b CSIRO Oceans & Atmosphere, Hobart, Tasmania, Australia

c The Centre for Southern Hemisphere Oceans Research, Hobart, Tasmania, Australia
d Department of Meteorology, Stockholm University, Stockholm, Sweden
e Department of Human and Social Sciences, University of Naples ‘L’Orientale’, Italy

Amanda Black and Didier Monselesan are co-first authors

Corresponding author: Didier P. Monselesan, didier.monselesan@csiro.au

1

Early Online Release: This preliminary version has been accepted for publication in Artificial Intelligence for the Earth Systems, may be fully cited, and has been assigned DOI 10.1175/AIES-D-21-0007.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

© 2022 American Meteorological Society
ABSTRACT: The ability to find and recognize patterns in high-dimensional geophysical data is fundamental to climate science and critical for meaningful interpretation of weather and climate processes. Archetypal analysis (AA) is one technique that has recently gained traction in the geophysical science community for its ability to find patterns based on extreme conditions. While traditional empirical orthogonal function (EOF) analysis can reveal patterns based on data covariance, AA seeks patterns from the points located at the edges of the data distribution. The utility of any objective pattern method depends on the properties of the data to which it is applied and the choices made in implementing the method. Given the relative novelty of the application of AA in geophysics it is important to develop experience in applying the method. We provide an assessment of the method, implementation, sensitivity, and interpretation of AA with respect to geophysical data. As an example for demonstration, we apply AA to a 39-year sea surface temperature (SST) reanalysis data set. We show that the decisions made to implement AA can significantly affect the interpretation of results, but also, in the case of SST, that the analysis is exceptionally robust under both spatial and temporal coarse-graining.
1. Significance statement

Archetypal analysis (AA), when applied to geophysical fields, is a technique designed to find typical configurations or modes in underlying data. This technique is relatively new to the geophysical science community and has been shown to be beneficial to the interpretation of extreme modes of the climate system. The identification of extreme modes of variability and their expression in day-to-day weather or state of the climate at longer time scales may help in elucidating the interplay between major teleconnection drivers and their evolution in a changing climate. The purpose of this work is to bring together a comprehensive report of the AA methodology using a SST reanalysis for demonstration. It is shown that the AA results are significantly affected by each implementation decision, but also can be resilient to spatio-temporal averaging. Any application of AA should provide a clear documentation of the choices made in applying the method.

2. Introduction

Pattern recognition plays an important role in quantitative climatology helping in diagnosing and understanding climate processes. Archetypal analysis (AA) is one technique that is gaining traction in the geophysical science community for its ability to find patterns based on extreme modes of data. Given the relatively new discovery of the utility of this analysis to geophysical problems, resources and references are scattered for the researcher who wishes to implement this technique. The goal of this paper is to present the AA method along with a detailed description of the decisions made in its implementation, and the effect each decision may have on the final output. We also provide a discussion on the interpretation of AA with respect to geophysical data.

Empirical orthogonal function (EOF) decomposition or factorisation has become a hallmark of statistical analysis and data reduction (Hotelling 1933; Jolliffe 1986) since its application in the mid-50s by Lorenz (1956) to weather and climate studies. Known also as principal component analysis (PCA), EOF analysis constructs patterns in the spatial dimension that maximize variance. The constructed EOFs are not directly interpretable in terms of the original data and therefore any attempt to attribute a particular dynamical mechanism to any one EOF pattern is discouraged when analysing geophysical data (Monahan et al. 2009). By comparison, AA seeks patterns from the extreme points of a convex hull, or envelope, surrounding the data in state space. It follows that the constructed archetypes may be interpreted in terms of the original data, as shown in the derivation.
section of this paper and by previous studies (Mørup and Hansen 2012; Bauckhage 2014). Other pattern recognition types discussed here are non-negative matrix factorization (Cichocki et al. 2009; Gillis 2020; Mairal 2014, 2017), clustering, and optimisation on manifold (Boumal et al. 2014; Hannachi and Trendafilov 2017; Hannachi 2021; Trendafilov and Gallo 2021).

Like any data mining tool, many decisions can be made in archetypal analysis to tweak or optimise certain parameters based on the needs of the user. The output of archetypal analysis will depend strongly on the decisions made along the way. A 39-year sea surface temperature (SST) reanalysis dataset is used here as paradigm for the method and to illustrate some of these choices.

We demonstrate the utility of this analysis method and the benefits that arise particularly when analysing climate and weather datasets. We show that the decisions made to implement AA can greatly affect the interpretation of results and should therefore be considered carefully and documented thoroughly in all work involving AA. The structure of this paper is as follows. Sections 1 and 2, provide an historical perspective and the rationale for the work. Next, we introduce the data sets in Section 3. In Section 4, we contrast both PCA and AA methods, and introduce the minimization algorithm for reduced space archetypal analysis (RSAA). Section 5 describes some decisions required on the input data, on the archetypal analysis, and their impacts to the final result. We then examine potential generalisation and extension of the AA method in Section 6. Section 7 illustrates teleconnections derived from extreme conditions resolved by AA applied to SST. Lastly, Section 8 provide a summary and conclusion statement. Appendix A touches on the available computing packages.

3. Data

We apply AA to the Optimum Interpolation Sea Surface Temperature (OISST) v2.1 high resolution dataset (Reynolds et al. 2007) provided SST on a 0.25° global grid, of which a subset from 1982-2020 is re-interpolated to a 4° by 4° resolution when computational efficiency is required. Some illustrations of the technique consider daily anomalies, but here we focus on monthly anomalies. SST anomalies (SSTA) are defined the standard way. Daily and monthly SST anomalies represent here a departure from daily and monthly climatological values, both defined as time mean for each day-of-year and month-of-year across all years considered in the interval 1982-2020, respectively. Only complete years are considered. The archetypes spatial and temporal imprints,
and their linkages to extra-tropical atmospheric circulation are revealed by compositing JRA-55 reanalysis atmospheric fields (Kobayashi et al. 2015) at the surface and on isobaric levels with corresponding level of temporal aggregation.

The SST dataset is used principally for illustration purpose of the AA method as the weather and climate research community is familiar with its variability patterns across multiple spatio-temporal scales and its teleconnections have been extensively studied and could be readily compared to the AA results presented hereafter.

4. Mathematical derivation

AA belongs to an ever increasing class of data analysis methods called matrix factorisation (Cichocki et al. 2009; Elad 2010; Elden 2019; Gan et al. 2020; Gillis 2020; Hannachi 2021), where factorisation allows one to represent the original data as a combination of factors or components that are provably easier to interpret. Another advantage of factorisation is that the dimensionality of dataset can also be reduced, and so its complexity. Due diligence when these methods are employed is to thoroughly test their domain of applicability. Hereafter, we will focus on PCA, one of the oldest and most widespread techniques in statistical data analysis (Hotelling 1933; Jolliffe 1986), and AA, a lesser known one, but emerging when applied to geophysical problems (Steinschneider and Lall 2015; Hannachi and Trendafilov 2017; Richardson et al. 2021; Risbey et al. 2021).

The goal of factorisation, when it is suitable for the problem at hand, is to extract a reduced representation of the underlying processes generating some imprint in the data. In particular, let’s assume that the underlying phenomenon is El Niño Southern Oscillation (ENSO), leading to SST anomalies (SSTA), the data set we wish to investigate. Without loss of generality, we assume SSTA to be represented by a two-dimensional data matrix \( X = X_{sxt} \), where the dimensions \( s \) and \( t \) represent the space and time coordinates, respectively. Here, the space coordinate \( s \) maps into a 2-dimensional longitude and latitude coordinates domain. SSTA are typically centred with respect to the time dimension \( t \),

\[
\sum_{j=1}^{t} X(i, j) = 0, \forall i,
\]  
(1)
for all geographical locations if the climatological mean is considered over the entire time domain.
Throughout the paper, the subscripts, $s \times t$, are used to remind the reader of matrix dimensions, whereas round brackets, $(i, j)$, refer to single matrix elements.

Mathematically, $X$ can be decomposed or factored by a standard algorithm called singular value decomposition (SVD), available to most data analysis software in use today,

$$X_{s \times t} \approx X'_{s \times t} = U_{s \times r} \Lambda_{r \times r} V_{r \times t}^T = \sum_{i=1}^{r} \lambda_i PC_i(t) EOF_i(s)$$

where $r \leq \min(s, t)$. When the rank $r$ of $X$ equals $\min(s, t)$, $X = X'$ to machine precision. The factors, the matrices $U$ and $V$, or component-wise the vectors $EOF_i$ and $PC_i$, identify data-driven spatial patterns and associated time-series, in contrast to other decomposition methods such as Fourier or spherical harmonics for example, where basis functions are a priori chosen. In geophysics, the spatial patterns are referred to as empirical orthogonal functions (EOFs) and the time-series, principal components (PCs). An underlying constraint in the SVD algorithm is that both EOFs and PCs are orthonormal,

$$\sum_{i=1}^{s} \sum_{j=1}^{s} U(i, t) U(j, t) = \delta_{ij}, \forall t,$$

and, correspondingly,

$$\sum_{i=1}^{t} \sum_{j=1}^{t} V(s, i) V(s, j) = \delta_{ij}, \forall s.$$

The matrix $\Lambda = \lambda_i \delta_{ij}$, with $i, j = 1, \ldots, r \leq \min(s, t)$, is diagonal and positive definite, $\Lambda \geq 0$. The $\lambda_i \geq 0$ are where the true physical scales of the data reside. An interesting property of SVD is that the factors,

$$\frac{\lambda_i^2}{\text{Trace}(XX^T)},$$

correspond to the fraction of the variance explained by individual product or mode, $PC_i(t) EOF_i(s)$, and are related to the eigenvalues, $\lambda_i^2$, of the covariance of $X$ (Jolliffe 1986; Jolliffe and Cadima 2016; Hannachi 2021). The positiveness of $\Lambda$ has the added advantage that modes can be ranked as function of explained variance. This property can be further exploited in two ways, 1) to isolate modes of variability based on the explained variance of $X$ and 2) to reduce the dimensionality.
of $X$ for computational convenience and interpretability. PCA modes display also the important property of ‘nestedness’ whereby the variance explained by the lower rank approximation $X^r$ is also contained in $X^{r+1}$, $\forall r \leq \min(s,t) - 1$. Note however that PCA modes of variability do not necessarily map neatly into physical modes of variability (North 1984; Mo and Ghil 1987; Hasselmann 1988; Monahan et al. 2009). As in Richardson et al. (2021) and Risbey et al. (2021), we will apply PCA as a dimension reduction tool.

AA has been applied to the analysis of weather and climate data only recently. For a comprehensive description of various implementations of the method, the reader is referred to the following publications, (Steinschneider and Lall 2015; Hannachi and Trendafilov 2017; Richardson et al. 2021; Risbey et al. 2021; Han et al. 2021). Only a brief derivation is presented hereafter. Similar to PCA in Eq. 2, AA corresponds also to a matrix factorisation method, whereby the original data matrix $X = X_{s\times t}$ is approximated by the product of two factors, $XC$ and $S$, according to the optimization problem

$$\arg \min_{C,S} \|X - XCS\|^2_F,$$

where $\| \cdot \|^2_F$ stands for the Frobenius norm and the solutions, $C$ and $S$, are to be sought in sets of left-stochastic$^1$ matrices $C = C_{t \times p}$ and $S = S_{p \times t}$ for a prescribed order $p$. The factor $XCS_{s \times p} = X_{s \times t}C_{t \times p}S_{p \times t}$, $p$ convex combinations of $t$ data points, are called archetypes. Furthermore, each data point in $X$ can also be approximated by a convex combination of $p$ archetypes according to $XCS = XCS_{s \times p}S_{p \times t} = X_{s \times t}C_{t \times p}S_{p \times t}$. Although not immediately apparent in the formulation of the AA problem, the archetypes can be shown to lie on the convex hull$^2$ enclosing the data; a welcome characteristic if one is interested in identifying extremes or outliers of data sets (Cutler and Breiman 1994; Bauckhage and Thurau 2009; Bauckhage 2014).

We consider hereafter a modification of AA, called reduced space AA (RSAA) introduced by Richardson et al. (2021), Risbey et al. (2021) and Han et al. (2021) to substantially reduce the computational burden incurred by large datasets. RSAA is performed on a reduced form of the data matrix, $X^r$, derived by applying a PCA driven dimension reduction step to $X$, where the factorisation in Eq. (2) is truncated to a given order $r$ informed by a predefined percentage of the total variance explained to be kept in the analysis. This consideration is mainly driven by

---

$^1$A left-stochastic matrix is a non-negative matrix with each column summing to 1.

$^2$In geometry, the convex hull of a set $S$ of points sampled from $r$-dimensional Euclidean space, is the smallest convex $r$-polytope enclosing the entire data set and which vertices are points of $S$. 

7

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-21-0007.1.
computational resources available and by an investigation of what processes, if any, account for the residual variance excluded. RSAA seeks to minimize the following cost function

\[
\|X' - X'C'S'\|^2_F = \|U\Lambda V^T - U\Lambda V^T C' S'\|^2_F \\
= \|U(\Lambda V^T - \Lambda V^T C' S')\|^2_F \\
= \|\Lambda V^T - \Lambda V^T C' S'\|^2_F ,
\]

for the left-stochastic matrices \(C' = C'_{r \times p}\) and \(S' = S'_{p \times t}\), the superscript \(r\) indicating here the solution for the reduced problem in contrast to Eq. 6. Element-wise, the stochasticity constraints are such that both \(C', S' \geq 0\) with \(\sum_{i=1}^{r} C'(i, p) = 1, \forall p, \text{ and } \sum_{i=1}^{p} S'(i, t) = 1, \forall t\). The invariance of the Frobenius norm, \(\|\cdot\|^2_F\), under unitary transformation has been employed in Eq. (7) to eliminate the spatial dependency, \(U\), from the minimisation. The parameter \(p\), the archetype cardinality\(^3\) or order, is set a priori. The reduced space archetypes are defined as the products \(\Lambda V^T C'\), and all records in \(\Lambda V^T\), the matrix built on the eigenvalues \(\lambda_i\) and \(PC_i\), can be approximated by \(\Lambda V^T C' S'\).

To recover the archetypes of the original (albeit reduced) data set \(X'\), the reduced space archetypes are simply left-multiplied by the matrix \(U\), built on the \(EOF_i\), \(U\Lambda V^T C'\). Similarly, all records in \(X'\) are approximated by \(U\Lambda V^T C' S'\).

PCA and AA are data driven factorisation methods belonging to unsupervised clustering techniques (Mørup and Hansen 2012). Only the PCA truncation order \(r\) and archetype cardinality \(p\) (akin to a number of clusters in clustering methods) allowed in the decomposition are predefined. Although not explicitly indicated in our notation, as not to render it too cumbersome, we weight spatially the SST data matrix \(X_{s \times t}\) by \(\sqrt{\cos(\theta)}\), \(\theta\) being the latitude of the grid cell \(s\), prior to applying SVD. It is also important to realize that, in both PCA and AA cases, the factorisation does not take into account serial correlations existing in the data. Indeed, any permutation of time records changes neither the \(EOF_i\) nor the archetypes \(XC\); and vice versa, any permutation of the space records \(s\) changes neither the \(PC_i\) nor the stochastic matrix \(S\), given that permutations are also unitary transformations leaving the Frobenius norm invariant. Further post-processing steps are therefore required to tease out any spatiotemporal relationship.

The constraints in the minimisation procedure in both cases critically affect the factorisation. For PCA and a predefined truncation of level \(r = \min(s, t)\), the factorisation is lossless. In other

\(^3\)The cardinality of a set corresponds to the number of elements in the set.
words, the original data set $X$ is identical to its factorisation expression within machine precision. For AA, the positiveness and stochasticity (convexity) constraints on both factors $C$ and $S$, lead to sparser and lossy representations as a function of the predefined cardinality selected $p$ and retained dimension $r$. AA representations will never explain the total variance of the original data set, even when $r = \min(s, t)$, unless the number of archetypes trivially equals the number of observations, i.e. $p = t$ (Bauckhage 2014). In the following, the predefined truncation level $r \leq \min(s, t)$ corresponds to the number of retained dimensions or $PC_i$ with $i = 1, \ldots, r$, ranked by fraction of explained variance, used in RSAA.

The appeal of AA over PCA, has been succinctly summarised by Bauckhage (2014):

*Archetypes are convex combinations of data points and data points are approximated in terms of convex combinations of archetypes.*

The convexity characteristic of AA is crucial and leads to a probabilistic interpretation of both archetypes, $XC$, and data point representations, $XCS$, the matrices $C$ and $S$ being stochastic. However, Bauckhage (2014) summary description eludes the fact that archetypes are ‘closer’ to individual data points and possibly more representative than EOF patterns. EOF patterns may never be observed, as the covariance of $X$ is the mean (expectation) of the product of anomalies derived from the climatology computed over the entire data set.

Finally, to reveal potential teleconnections based on AA, the composites $A_F$ for $p$ archetypes from any geophysical field, $F = F_{s \times t}$, can be constructed directly from the stochastic matrices, $C = C_{t \times p}$ or $S = S_{p \times t}$, as the products $A_F = FC$ or $F \tilde{S}^T$, both $C$ and $\tilde{S}_{p \times t} = S_{p \times t}/(\sum_{j=1}^{t} S(p, j))$ depending solely on time. We note that the field $F$ to composite needs not to be interpolated on the grid of the original data set $X$, but time records and aggregation levels should overlap and be matched respectively for consistency (i.e. daily to daily, monthly to monthly averaged records).

5. Data-driven characteristics impacting AA

a. Data structure and distribution

Since the AA inception by Cutler and Breiman (1994), it is recognised that the location of archetypes are intimately linked to the underlying distribution of data under investigation and driven by outliers or extremes thereof. As with any clustering technique, the behaviour of the
method on selected data sets should be therefore well-understood prior to its implementation. As illustrated by elliptically-shaped density level sets⁴ of points drawn from a normal distribution $N(\mu, \Sigma)$ with mean $\mu = [1, 1]$ and covariance $\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$, Cutler and Breiman (1994) show AA solutions for 4 archetypes are preferentially located along the level set ellipse’s major $2a$ and minor $2b$ axes with $a/b = 3^5$. When applied to 100 samples of 1000 points each, the non-isotropic character of the underlying distribution comes to the fore for eccentricities $e = \sqrt{1 - \frac{b^2}{a^2}} > 0$. One can easily convince oneself that, when $a = b$ ($e = 0$), the distribution is spherical and no preferential directions away from the mean $\mu$ are privileged by the method. The archetype locations in data space will strongly depend on outliers in each of the samples, as the underlying distribution is invariant under rotations centred on the mean, $\mu$, and outliers are equally likely in any direction.

When 100 samples of 1000 points each are drawn from an isotropic distribution in 3 dimensions, for example a multivariate normal distribution, with mean $\mu = [0, 0, 0]$ and trivial covariance $\Sigma = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, the eight archetypes resulting from AA applied to each individual samples in Figure 1a tend to be uniformly distributed on a spherical shell loosely bounded by the convex hulls of all samples. As expected, no preferential direction can be detected by the method, but the ensemble of archetypes over all samples are mainly located on the outer shell, the furthest away from the center. In contrast, if the samples are drawn from a multivariate uniform distribution centered on the origin, the distribution support being a 3-dimensional cube with edges of unit length, the eight archetypes are preferentially located close to the cube vertices and therefore correspond to extreme points of the distribution. In Figure 1b, the archetypes are labelled as a function of their distances to a given vertex and the method detects preferential directions in the data space, along the eight vertices of the cube.

In general, geophysical observables do not have a spherically- or elliptically-shaped distribution. As a tractable example, Figure 1c displays the results of AA for eight archetypes applied to 100 samples of 1000 daily SSTA records drawn from the 14250 records of sea surface temperatures over the 1982 to 2020 period. When the first 3 scaled PCs, $\lambda_{1-3}PC_{1-3}$, are retained for AA, the reduced data set seems at first glance elliptically distributed given that the eigenvalues in the singular value decomposition of SSTA are such that $\lambda_1 > \lambda_2 > \lambda_3$. However, a closer inspection of the 3-dimensional cloud points reveals that, at times, excursions occur away from the broad

---

⁴https://en.wikipedia.org/wiki/Level_set
⁵The square root $\sqrt{\frac{\lambda_1}{\lambda_2}}$ of the ratio of the covariance matrix $\Sigma$ eigenvalues, $\lambda_1 = 1.8$ and $\lambda_2 = 0.2$. 

ellipsoidally-shaped distribution and outliers are readily identified by the AA method and are consistently arranged in small spatially coherent clusters across all 100 samples.

To represent AA results for dimension larger than 3, a 2-dimensional representation of the stochastic matrix $S = S_{p\times t}$ and corresponding archetypes $X_{s\times t}C_{t\times p}$ can be constructed given that the points $S(\cdot, t)$ belong to the simplex

$$\Delta_{p-1} = \left\{ (s_1, \ldots, s_p) \in \mathbb{R}^p : \sum_{i=1}^{p} s_i = 1, \ \forall s_i \geq 0 \right\}.$$ 

Points in $\Delta_{p-1}$ can be projected on the plane perpendicular to the diagonal of the non-negative orthant of the $p$-dimensional hyper-cube centred at the origin of $\mathbb{R}^p$. The simplex projection footprint corresponds to the edges and interior points of regular $p$-polygons (Seth and Eugster 2016; Hannachi and Trendafilov 2017) where vertices correspond to the pure archetypes $XC(s, \cdot)$ if all but one element of $S(\cdot, t)$ at time $t$ equal zero and the sole non-zero element being necessarily equal to 1. Any other point either on the edges or in the interior of the $p$-polygons corresponds to a convex mixture of archetypes. Figure 2 displays simplex projections of the stochastic matrix $S_{p\times t}$ for detrended monthly SSTA over 1982-2020 for 4 (2a) and 8 (2b) archetypes respectively. The vertex colour indicates how close the data record representation lies to a single archetype. Mixed

---

*The non-negative orthant is the generalization of the first quadrant in 2 dimensions to $n$ dimensions.
colours away from vertices towards the 4- and 8-polygon centers indicate records represented by mixtures of archetypes. The simplex representation may help to visualize the clustering of ‘extreme’ records in the data sets.

**Fig. 2:** Simplex representations of detrended monthly SST anomalies over the 1982-2020 for archetypes cardinalities of 4 in a) and 8 in b).

**b. Dimensionality reduction**

The spatio-temporal dimensions of the SST and JRA-55 data sets considered throughout this work are commensurate to the resolutions of other major global reanalysis efforts and ocean-atmosphere general circulation model (OAGCM) simulations. For OISST v2.1, the daily (monthly) data matrix $X = X_{s \times t}$ consists for $s = 691119$ spatial points times $t = 14245$ (468) daily (monthly) records, in total about $O(10^{10})$ data points for daily records and an order of magnitude less for monthly records, $O(10^9)$. Similarly, JRA-55 atmospheric reanalysis fields used hereafter to investigate teleconnections correspond to 41760 spatial points $\times 14245$ day records, also a $O(10^9)$ order of magnitude for single pressure level data on a $1.25^\circ \times 1.25^\circ$ latitude-longitude grid. The large and inflationary increase in spatial and temporal resolution over the last four decades has led to exponential growth in size of available data sets and is mainly due to both the advent of space-born observing and super-computing platforms used to collect and assimilate observations into
circulation models. Due to its high dimensionality, the voluminous amount of data is particularly hard to manipulate when one is interested in the study of large scale phenomena and their global teleconnection imprints. A direct application of AA to the original data matrix $X$, an example of a Frobenius distance clustering problem, incurs large computational burden and is likely to be ill-posed. AA optimization times are driven by the NP-hard character of the minimization algorithms (Aloise et al. 2009; Bauckhage and Manshaei 2014).

To alleviate the dimensionality issue, we may first consider a reduction of the domain size or spatio-temporal averaging. This is suitable for certain problems; for example, when AA is used to find extreme patterns involving meso-scale ocean processes and both local and remote linkages to atmospheric circulation, or when variability at smaller scales and higher frequencies can safely be ignored. For example, applying $4^\circ \times 4^\circ$ spatial averaging to the original $0.25^\circ \times 0.25^\circ$ OISST v2.1 data set decreases the spatial dimension by factor $16 \times 16 = 256$, two orders of magnitude. Additional aggregation of the time dimension from daily to monthly records further reduces the total number of data points by a factor of $14245/468 \times 256 \approx 7792$, without substantially changing SST AA results for large scale phenomena such as ENSO, for example.

Another class of dimensionality reduction rests on general matrix factorisation procedures (Mørup and Hansen 2012; Nguyen and Holmes 2019), where the original data matrix $X$ can be decomposed into the sum or product of factors with the view to reducing its complexity and dimensionality, diminishing the computational burden of clustering algorithms such as AA. Throughout this work, we focus on one dimension reduction method, PCA, as formulated in the Mathematical Derivation section. PCA is an unsupervised, linear dimension reduction method with the following advantages: 1) the dimension reduction level can be informed by the fractional amount of explained variance desired to be maintained in the reduced data passed to AA and 2) the global data structure, its spatial covariance that is, is preserved in contrast to other more sophisticated methods (Izenman 2008; Nguyen and Holmes 2019). It is important to note however that linear reduction methods do not necessarily ‘remove’ the inherent non-linear characteristics of the reduced data set.

To illustrate the equivalence between AA and RSAA methods as in Eq. (7), we apply AA and RSAA to spatially averaged OISST v2.1 monthly anomalies on $4^\circ \times 4^\circ$ grid. The AA and RSAA data matrix dimensions are $X = X_{2643 \times 468}$ and $\Lambda V^T = X_{RSAA,468 \times 468}$, where all 468 scaled PCs,

---

*Qualify non-deterministic polynomial acceptable problems in reference to the computing time needed to find their ‘near-optimal’ solutions.
Fig. 3: AA and RSAA results for the stochastic matrices, $S_{AA}$, $C_{AA}$, $S_{RSAA}$ and $C_{RSAA}$, performed for 8 archetypes on detrended monthly SST anomalies over 1982-2020 period, 468 monthly records. The top row shows the resulting $S$ (left) and $C$ (right) for AA on the full data set, to be compared to the middle row showing $S$ and $C$ when all (468) principle components are retained in RSAA. The bottom row shows the absolute difference in $S$ and $C$ between AA and RSAA.

$AV^T$, in the singular vector decomposition $UAV^T$ of $X$, have been retained. No information loss has incurred given that 100% of the variance is explained by 468 modes in this case, leading to a dimension reduction factor of $2643/468 \approx 6$. Figure 3 compares the 8 archetypes AA and RSAA stochastic matrix components, $S_{AA}$, $C_{AA}$, $S_{RSAA}$ and $C_{RSAA}$ (top and middle rows), and their absolute difference $|S_{AA} - S_{RSAA}|$ and $|C_{AA} - C_{RSAA}|$ (bottom row). Differences in the resulting $S$ and $C$ matrices are of a small order of magnitude $O(10^{-3})$. The data points approximated by the convex combinations of archetypes, $XC_{AA}S_{AA}$ and $UX_{RSAA}C_{RSAA}S_{RSAA}$, are similar but not identical, the AA optimization procedure being a NP-hard problem for both. We note that the absolute differences $|S_{AA} - S_{RSAA}|$ and $|C_{AA} - C_{RSAA}|$ remain of the order of $O(10^{-3})$ for archetype cardinality ranging from 2 to at least 20 (not shown).
A difficulty encountered by most dimensionality reduction methods is to justify the level of truncation or reduction to apply to the original data set. Usually for PCA, researchers rely on the first ‘significant’ step change between consecutive values of ranked fraction of variance explained (Eq. 5) when displayed on a scree plot. However, the optimal truncation level is conditioned by the data itself. The SSTA data matrix spectral characteristics show a rather smooth and incremental decrease between consecutive eigenvalues without an obvious step change and, therefore, the truncation levels applied throughout this work are mainly informed by the percentage of variance one wishes to retain and driven by computational considerations: typically, 90 to 100% of the variance corresponding to reduced dimensions of the order of $O(10^3)$ or less. The residual variance made out of the excluded modes of variability can be displayed to gain insight on where it lays and what it represents. Similarly, a spectral analysis of the associated PCs informs on the time scales excluded from the reduced data set. Hereafter, the variance explained is given as the ranked fraction of variance of the full or reduced data set and ranges from 0 to 1.

The attentive reader may realize that the difficulty mentioned for the dimension reduction problem also applies to the number of archetypes to be chosen in AA. How to decide how many clusters are needed in order to detect existing or new climate and weather regimes? Christiansen (2007) cautions against the application of clustering techniques to atmospheric regimes without a comprehensive analysis of the robustness and reproducibility of the results. In Figure 4, the variance explained and sum of squares errors for increasing level of 1) PCA truncation (blue curves) and 2) archetype cardinality (red curves) are shown to gradually change for both SSTA and detrended SSTA. Here, the fraction variance explained by the AA factorisation or in short the AA explained variance is equal to the difference between data matrix total variance, $\Sigma^2 = \|X\|^2_F = \text{Trace}(XX^T)$, and the AA residual sum of squares, $\|X - XCS\|^2_F$, divided by $\|X\|^2_F$,

$$\Sigma^2_{AA} = 1 - \frac{\|X - XCS\|^2_F}{\|X\|^2_F} = 1 - \frac{\|X - XCS\|^2_F}{\Sigma^2}. \quad (8)$$

No obvious spectral gaps can be readily identified and the spectral characteristics of the RSAA factorisation, as for PCA, are of little help.

In investigating the impact of dimension reduction on RSAA factorisation, one does not fail to notice that both stochastic matrices, solutions of Eq. (7) $C$ and $S$, display levels of sparsity.
Fig. 4: Explained variances (top) and sum of squares errors (bottom) for increasing number of principal components or archetypes for PCA (blue) and AA (red) for both detrended (continuous) and full (dashed) monthly SST anomalies. The AA explained variance (Eq. 8) reported for both the full and detrended cases corresponds to RSAA results for cardinality from 3 to 20 when all 468 PCs are retained.

depending on both the truncation order and the archetype cardinality. First, one notes that the level of sparsity for both C and S increases as a function of the number of archetypes. This behaviour is understandable when one considers the extreme case where the number of archetypes equals the
number of observation records \( t \). Trivially, both \( C \) and \( S \) stochastic matrices are the sparsest with entries taking the values of 1 or 0 when \( X_{sx} = XCS \), i.e. \( I = CS \) exactly, with \( C \) and \( S \) equal to the identity matrix \( I = I_{sx} \).

The sparsity dependence on the retained dimension order is less obvious and an objective measure of sparsity is required. Following Hurley and Rickard (2009), who systematically compared a range of metrics, we adopt here a novel application of an evaluation metric called the Gini coefficient or index (Gini 1921; Hurley and Rickard 2009; Zonoobi et al. 2011; Abrol and Sharma 2020) and is defined as follows. Given a set of \( N \) observations, \( O = (o_1, \ldots, o_N) \), which values can be ordered from the smallest to largest such that \( o_{(1)} \leq \ldots \leq o_{(N)} \) with \( (1), \ldots, (N) \) now indexing the sorted elements of \( O \), the Gini coefficient \( \Gamma(O) \) is defined as

\[
\Gamma(O) = 1 - 2 \frac{1}{N} \sum_{i=1}^{N} \frac{o_{(i)}}{\|O\|_1} \left( \frac{N - i + 1/2}{N} \right),
\]

where \( \|O\|_1 \) corresponds to the \( l^1 \)-norm of \( O \), \( \|O\|_1 = \sum_{i=1}^{N} |o_i| \).

The Gini coefficient has been used previously in economics as a measure of wealth inequality, where Gini = 1 indicates maximum inequality and Gini = 0 indicates perfectly distributed wealth across \( o_i \) values. In the application of AA, Gini = 1 for the probability matrix \( C \) indicates that the resulting archetype patterns are each expressed by a single data record, whereas a lower Gini number would indicate archetype patterns expressed by more than one record. In general, we would like the sparse matrix, \( C \), to be fairly distributive so as not to be overwhelmed by one particular record for each archetype, but still representative of extremes in the data distribution. Therefore, there lies some ‘sweet spot’ which may be indicated by the Gini coefficient for maintaining a distributed \( C \) matrix while simultaneously staying true to the extreme mode nature of AA.

Figure 5 illustrates the effect of increasing retained dimensions on how well dispersed the mixture weights are, as well as a comparison of the variance explained by the reduced problem versus the explained variances compared to the total variance of the original data. The top panel plots the Gini coefficient as a function of retained dimension (the number of scaled PCs) for seven selected archetype numbers (shown by different colored lines). For \( C \), the Gini coefficients monotonically decrease from values close to 1, total inequality, to values ranging from 0.98 to 0.91 for the full

---

\(^8\)Abrol and Sharma (2020) recently make use of the Gini sparsity measure to develop computationally efficient greedy AA (GAA) algorithm, where it is implemented in the AA optimization procedure to update the sparse stochastic matrix \( C \).
problem (100% of the total original variance with 468 PCs retained) over the range of archetypal numbers from 3 to 20 considered here. Clearly, the sparsity of $C$ has only been marginally affected by the number of dimensions retained. The ‘wealth redistribution’ or sparsity lessening in $C$ occurs more quickly as function of retained dimensions with fewer archetypes.

The impact on $S$ is however remarkable in Figure 5 bottom panel. The Gini coefficient values are shown as solid lines, overlaid by the AA explained variances for the reduced problem (dashed) and the explained AA variances compared to the total variance (dotted). For RSAA applied to monthly SSTA, we observe that the Gini coefficients increase approximately monotonically as a function of archetype numbers and asymptote as function of the number of retained dimensions, the earlier the lower number of archetypes selected and without any substantial gain in the total variance explained by the AA factorisation. For any given AA cardinality, the ‘sweet spot’ for the number of retained dimensions used for the analysis 1) corresponds approximately to the number of the fastest change in the AA explained variances for the reduced problem on the dashed curves and 2) where the total variance explained on the dotted lines remains unchanged in Figure 5.

When the dimensionality of the original data becomes computationally prohibitive even for PCA, ‘approximate’ methods of dimension reduction can be deployed. For geophysical data sets, Seitola et al. (2014) and Hannachi (2021) have recently illustrated that the issue of dimensionality could be addressed through random projections (RP), where the PCA decomposition factors, $U$, $Λ$ and $V$ in Eq. (2) can be approximated at low computational cost. We refer the reader to the appendix in Seitola et al. (2014) for details on how to generate RP approximations $Λ_{RP}V_{RP}^T$ as proxies for $ΛV^T$ in RSAA.

c. Trend analysis

Geophysical data often display non-stationary or trending behaviour, of which a notable example is the observed warming of SST caused by globally rising temperatures due to increasing greenhouse gas concentrations (IPCC 2013, 2019). It is important to ascertain the impact of this trend on natural climate modes. For example, there is clear observational evidence that significant changes in the nature of key ENSO indicators happen posterior to 1980, when at least 3 major El Niño episodes occurred in the 39-years period between 1982 and 2020 compared to the previous 39 years (Capotondi and Sardeshmukh 2015, 2017; Capotondi et al. 2020). Throughout this work,
Fig. 5: Gini coefficients as a function of retained dimensions (scaled principle components) for matrices $C$ (top) and $S$ (bottom) of in terms of archetype cardinality (shown as different colors). Gini coefficients (solid), AA explained variances for the reduced problem (dashed) and AA explained variances (Eq. 8) as fractions of the total variance (dotted).

we focus only on SSTA over the satellite era (Reynolds et al. 2007). Given the relatively short record of near-global SST coverage available since the advent of satellite observing platforms, the power of any statistical analysis to investigate the interplay between SSTA variability and a
‘warming’ or slowly changing mean state is limited. Therefore, we will not attempt hereafter to explain this interaction. However, if we are interested in detecting the different ‘flavours’ of ENSO, removing a linear trend to SSTA prior to AA implementation is a legitimate step to prevent the global warming signature from potentially ‘washing out’ the ENSO global extreme imprints if one assumes that internal variability of the climate system can be neatly separated from anthropogenic forcing effects. Conversely, if we are interested in the change of ENSO extreme impacts under climate change, we would want to retain that trend. It remains part of the due diligence in the application of AA to properly formulate the questions that this method aims to address.

To illustrate the impact of trend removal, we compare the resulting archetype patterns, $XC$ and $XST$, and stochastic weights, $C$ and $S$, for full and linearly detrended SSTA for $n_{AA} = 4$, 6, and 8 numbers of archetypes in Figures 6b, 7b and 8b, respectively. When no SSTA trend is removed, the overall warming is captured in all cases by two archetypes dominated by global cooling and warming patterns: rows 1 and 3 in Figure 6a for $n_{AA} = 4$, rows 1 and 2 in Figure 7a for $n_{AA} = 6$, and rows 2 and 4 in Figure 8a for $n_{AA} = 8$. The associated weights, $C$ and $S$, also reflect this non-stationary behaviour with stochastic weights $S$ approximately decreasing from 1982 to 2010 for the cooling pattern and increasing from 2011 to 2020 for the warming pattern. The remaining archetypes resemble patterns associated with the different ‘flavours’ of ENSO (for a recent review see Capotondi et al. (2020)). When removing the linear trend prior to applying AA, the global warming and cooling patterns disappear and, for equal number of archetypes, are ‘replaced’ by additional patterns also bearing resemblance to typical ENSO conditions studied in the literature (Ashok et al. 2007; Cai and Cowan 2009). The effect of the trend removal is especially clear in $C$ and $S$, both taking non-zero values spread over the entire time period under consideration.

6. Archetypal analysis computation and extensions

a. Initialization and convergence

Esposito (2021) has recently reviewed initialisation methods for non-negative matrix factorization (NMF). NMF shares algorithmic similarities with AA and so do initialization strategies employed to solve the optimization problem sketched in Section 4, Eq. (7). Here, we combine random based and clustering based or data driven initialisation procedures. Being simpler to implement, random based procedures are used as a benchmark for more sophisticated ones, but require a

---

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-21-0007.1.
Fig. 6: AA spatial pattern and time series results using full a) and detrended b) monthly SST anomalies over 1982–2020 for a selected archetype number of 4. The two left columns on each subplot show archetypes constructed by $XC$ and $X\hat{S}'$, with $\hat{S}_{p\times i} = S_{p\times i}/\sum_{i=1}^{I} S(p,i)$, respectively. The two right columns show $C$ and $\hat{S}$ matrix time series, respectively, with Multivariate ENSO Index (MEI) time series (grey) included on both.

thorough investigation of their robustness and reproducibility. For AA, a suite of strategies have been adopted in the literature (Baukhage and Thurau 2009; Thurau et al. 2009, 2011; Eugster and Leisch 2011; Mørup and Hansen 2012; Baukhage and Manshæi 2014; Suleman 2017a,b; Mair
et al. 2017; Mair and Brefeld 2019). Our benchmark in all cases, the first trial of typically a number of randomly sampled initialisations per optimization, is always the data driven ‘FurthestSum’ procedure advocated by Mørup and Hansen (2012). The algorithm ‘FurthestSum’ initialises the AA procedure with a number of observation points equal to the desired AA cardinality through the matrix $C$, such that 1) the points are located on the convex hull of the dataset and 2) are the furthest away from each other, by taking into account the distance, in this instance the Frobenius norm, between all data points and the initial candidates.

However, Suleman (2017a) criticises ‘FurthestSum’ as ‘ill-conceived’ and potentially leading to archetype redundancy after convergence for increasing archetype cardinality. To protect against this eventuality, we implement the initialisation procedure prescribed by Mair and Brefeld (2019) based on ‘coreset’ construction for AA. Algorithm 2 in Mair and Brefeld (2019), where initial
seed archetypes are randomly drawn from a distribution constructed from the square Euclidean distance of each data point \(X(\cdot,t)\) from the time mean of \(X_{\text{t}}\) which sits at the origin in our case given that \(X\) corresponds to anomalies. The AA ‘coreset’ strategy privileges initialisation points the furthest away from the mean, but are not necessarily located on the data convex hull in contrast to ‘FurthestSum’.

For all results reported in the paper, we implement the MatLab code of Mørup and Hansen (2012) ‘PCHA’, suitably modified to accommodate both ‘FurthestSum’ and AA ‘coreset’ initialisations, see Table 2 for references. The AA procedure runs through an outer loop consisting of 1000 initialisation trials, 999 random ‘coreset’ and one ‘FurthestSum’ trials, where for each individual trial, the iterative non-linear least square algorithm in ‘PCHA’, the inner loop, is allowed to converge with relative sum of square error (SSE) stopping criterion of \(10^{-8}\). We report the solution that
minimizes the relative SSE across all 1000 trials. We note that ‘FurthestSum’, throughout our many experiments, never corresponds to the optimum. All computations are performed in double precision.

b. Archetype cardinality

We observe no clear ‘knee point’ in the evolution of neither the AA explained variance nor the sum of square errors between the full SSTA data and the AA representations as a function of archetype cardinality in Figure 4. Therefore a balance has to be struck between AA cardinality and representation of extremes conditions in the original data set. To avoid the pitfalls of archetype redundancy mentioned by Suleman (2017a) for example, several initialisation procedures, dimension reduction truncation and aggregation levels have to be tested for a number of archetype orders and the results compared (Baukhage and Thurau 2009; Suleman 2017b).

It is interesting to note, somewhat unexpectedly, that the global SSTA archetypes ‘nest’ in contrast to the assumption of Risbey et al. (2021) for AA applied to geopotential anomalies at 500 hPa. A pattern correlation distance is applied to identify archetype correspondence for different cardinalities. The correspondence is corroborated by visual inspection directly in Figure 9. Overall, pattern correlations across near-identical archetypes are typically larger than 0.8. Table 1 shows the AA correspondence across several archetype orders from $n_{AA} = 4$ to 10, for both the full and detrended cases, where each row corresponds to near-identical archetypes independent of the order $n_{AA}$, at least when $n_{AA}$ is small ≤ 10. Such a correspondence could not be readily established when comparing the full with $n_{AA} = p$ and the detrended with $n_{AA} = p − 2$ sets of archetypes. For the non-detrended SSTA, 2 archetypes account for the cooling and warming patterns. As a linear trend has been removed, the cooling and warming patterns found in the full case have to be absent in the detrended case. For example, Figure 6b detrended archetypes for a cardinality of 4 ($n_{AA} = p − 2$ with $p = 6$) can only be compared to Figure 7a archetypes 3 to 6 for a cardinality of 6 ($n_{AA} = p$) given that archetypes 1 and 2 correspond to global cooling and warming patterns. The mismatch between Figure 6b detrended archetypes for a cardinality of 4 and the remaining archetypes 3 to 6 in Figure 7a for the full problem possibly indicates that a clean separation between a slow-changing mean state and natural modes of variability is elusive, at least as far as the distribution of extremes is concerned.
Fig. 9: a) Stacked bar-plots of $S$-matrix probabilities for detrended monthly SST anomalies over 1982-2020 and b) corresponding matched archetypes using pattern correlation for cardinalities ranging from 2 to 8. Each row in a) corresponds to AA results ranging from 2 (top row) to 8 (bottom row), with MEI time series (black) included on all. The bar color codes correspond to matched archetypes referenced to AA results for a cardinality of 8, whereas the labels $A_i$ in each row indicate the archetype ranks based on the time mean of the AA stochastic matrix $S_{n_{AA}}$ with $n_{AA} = 2, \ldots, 8$ and $i = 1, \ldots, n_{AA}$ in decreasing order of $\bar{S}_{n_{AA}}(i)$. The 7 columns by 8 rows AA patterns in b) correspond to matched archetypes (rows) referenced to the AA results for cardinality of 8 (last column) across cardinalities ranging from 2 (first column) to 8 (last column).
The nesting properties of AA, when applied to SSTAs, could be utilized to increase computational speed when stepping through the archetype orders by ‘recycling’ the results from the previous order or perturbations thereof as initial seeds for the next order. Throughout this paper, we assume at the outset that archetypes do not nest, even approximately, and we reinitialise randomly our trials chosen independently of the previous order results.

As companion to Table 1, Figure 9 illustrates the patterns ‘nestedness’ and the impact of cardinality on the affiliation probabilities expressed by the matrix $S = S_{pxt}$ for detrended SSTAs and cardinalities ranging from $p = 2$ to 8. The stacked bar-plots of $S$-matrix affiliation probabilities as function of time, Figure 9a, take advantage of the stochastic constraint along the cardinality dimension $p$, where each row corresponds to AA results for cardinalities ranging from 2 (top row) to 8 (bottom row). The bar color codes correspond to matched archetypes referenced to AA results for a cardinality of 8 and are given in the legend of the last row. For each time record, the length of each coloured bar corresponds to the probability of a given archetype to be expressed. The labels $A_i$ in each row indicate the archetype ranks based on the time mean of the AA stochastic matrix $S_{nAA}$, $\bar{S}_{nAA}(i) = (\sum_{j=1}^t S(i,j))/t$ with $n_{AA} = 2, \ldots, 8$ and $i = 1, \ldots, n_{AA}$ in decreasing order of $\bar{S}_{nAA}(i)$. As a guide, the 7 columns by 8 rows AA patterns in Figure 9b correspond to matched archetypes (rows) referenced to the AA results for cardinality of 8 (last column) across all tested cardinalities ranging from 2 (first column) to 8 (last column).

Figure 9a illustrates the hierarchical nature of the AA power of discrimination for extreme SSTAs conditions as a function of increasing cardinality. For example, it shows that no spurious blending of archetypes occurs when ‘it matters’, that is, when extreme conditions occur. This can be seen when the archetype corresponding to the 3 major Niño intervals between 1982 and 2020 matches across 2-8 cardinalities as the pink coloured intervals in 1982-1983, 1997-1998 and 2015-2016 indicate. For a cardinality of 2 in Figure 9a first row, all records, being convex combinations of archetypes, have to be expressed as ‘blended’ patterns by construction unless they correspond to strongly expressed Niño and Niña intervals (respectively pink and light green) as depicted in Figure 9b fifth and third rows. Individual data records not corresponding to extremes will lead to a low discrimination score $\Delta_p(t)$ introduced in Eq. 10 or a low Gini coefficient $\Gamma_p(t)$ in Eq. 11 for $S$ introduced in Section 5c. This occurs when the probabilities of being expressed are approximately the same ($\approx 1/p$) for all archetypes for a given cardinality $p$. The extra patterns introduced
with trend

| $AA_4$ | $AA_5$ | $AA_6$ | $AA_7$ | $AA_8$ | $AA_10$ | linear trend removed
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>8</td>
<td>9</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 1: Archetypes nesting for SSTA with trend (left hand-side) and SSTA with linear trend removed (right hand-side). The numbers in each row label the archetype rank based on the time mean of the AA stochastic matrix $S_{nAA}$, $\bar{S}_{nAA}(i) = (\sum_{j=1}^{t} S(i,j))/t$ with $n_{AA} = 4, \ldots, 10$ and $i = 1, \ldots, n_{AA}$ in decreasing order of $\bar{S}_{nAA}(i)$. Each row corresponds to near-identical archetypes found across $AA_{nAA}$ for $n_{AA} = 4, \ldots, 10$. Circled archetype numbers on the left hand-side label the trending patterns.

by increasing the cardinality could be viewed as transition patterns or potentially new ‘extreme’ conditions in SSTA identified by AA.

c. Serial correlation and causality

As mentioned in Section 4, a direct AA factorisation of SSTA, $X_{s\times t} \approx X_{s\times p} S_{p\times t} = X_{s\times t} C_{t\times p} S_{p\times t}$ with $s \times t$ spatiotemporal dimensions and $p$ archetype cardinality, does not explicitly capture serial correlation or temporal patterns existing in the data. Serial correlation or causal relationship between records need to be extracted from the matrix $S_{p\times t}$ which explains how individual records are expressed in terms of archetypes $XC_{s\times p}$ and how the AA representation evolves with time.

Risbey et al. (2021) introduced a simple metric, a discrimination score, to isolate persistent atmospheric flow patterns based on the matrix $S_{p\times t}$. The discrimination score $\Delta_p(t)$ takes advantage of the stochastic nature of $S$, $\sum_{i=1}^{p} S(i,t) = 1, \forall t$, such that,

$$\Delta_p(t) = 1 - \left(1 - \frac{1}{p-1}\right)\left(\frac{1}{\max_{i=1,\ldots,p} S(i,t)} - 1\right) = 1 - \left(1 - \frac{1}{p-1}\right)\left(\frac{1}{S(i_{\text{max}},t)} - 1\right),$$

where $\max_{i=1,\ldots,p} S(i,t)$ and $i_{\text{max}} = i_{\text{max}}(t)$ correspond to the maximum values of $S(\cdot,t)$ and archetypes labeled by $i_{\text{max}}$ contributing with highest probability to the snapshot $X(\cdot,t)$ for each
time record \( t \). Combined with a persistence criterion, the discrimination score \( \Delta_p(t) \) and archetype affiliation \( \alpha_{\text{max}}(t) \) with \( \alpha_{\text{max}}(t) \in \{1, \ldots, p\} \) can be used to study the evolution of ‘extreme’ conditions or their AA representation over the period of interest. A discrimination score equal to 1 means that the archetype corresponding to \( \alpha_{\text{max}}(t) \) is expressed with a probability weight of 1 in record \( t \), whereas a zero discrimination score means that all archetypes are equally expressed in record \( t \), all with the same probability of \( 1/p \).

Unsurprisingly, the Gini coefficient\(^9\) (Eq. 9), \( \Gamma_p(t) \), based on the sum of \( S = S_{p \times t} \) row vector components weighted by their normalised ascending rank index for each record \( t \),

\[
\Gamma_p(t) = 1 - 2 \sum_{i=1}^{p} \frac{S(\pi_i^p(i), t)}{\|S(\cdot, t)\|} \left( \frac{p - i + 1/2}{p} \right),
\]

highly correlates with the discrimination score \( \Delta_p(t) \) (Eq.10) with correlation values \( \geq 0.95 \) for archetype cardinality \( p \) ranging from 3 to 20. Here, the permutation \( \pi_i^p \) ranks in increasing order the values of \( S(\cdot, t), \forall t \), noting that \( S \) is left-stochastic, i.e., \( \sum_{j=1}^{p} S(j, t) = 1, \forall t \).

To complement the previous approach based on an univariate discrimination score, \( \Delta_p(t) \) or \( \Gamma_p(t) \), based on the ‘winning’ archetype for time record \( t \), the whole archetype ranking in \( S_{p \times t} \) can be exploited. The AA representation can be viewed as a form of discretisation or categorisation of the dynamics where, for a given archetype cardinality \( p \), the time evolution can be represented by a series of permutation \( \pi_i^p \),

\[
\pi_i^p = \begin{pmatrix}
1 & \cdots & p \\
\pi_i^p(1) & \cdots & \pi_i^p(p)
\end{pmatrix}.
\]

Transitions from one record \( S(\cdot, t) \) to the next \( S(\cdot, t + 1) \) can be now analysed from the changes from \( \pi_i^p \) to \( \pi_i^{p+1} \), where \( \pi_i^p \) corresponds to the permutations needed to rank in decreasing order the probabilities recorded in \( S_{p \times t} \). The time mean of the AA stochastic matrix \( S \), \( \overline{S}(i) = (\sum_{j=1}^{p} S(i, j))/t \) with \( i = 1, \ldots, p \), ranked in decreasing order of \( \overline{S}(i) \) corresponds here to the identity or the reference permutation and is originally used to numerically label the set of archetypes by Mørup and Hansen

(2012). A suitable distance function or metric over the set of permutations of \( p \) elements could be chosen (Fligner and Verducci 1986) not only to track the changes in \( \pi^p_t \), but also to compare clustering methods as in Meilă (2007) for example. We intend to pursue this line of enquiry in further works with a focus on onset, persistence and decay of archetypal states.

As an alternative to post-processing results obtained from the direct application of AA, one could imagine AA to be applied to data sets where serial correlation has been explicitly modelled. Horenko (2009, 2010b,c,a), O’Kane et al. (2013), Risbey et al. (2015), Franzke et al. (2015), O’Kane et al. (2016), Yu et al. (2016), O’Kane et al. (2017), Gerber et al. (2020) and Quinn et al. (2021) extend matrix factorisation techniques to time series predictions, where lags of the data set under investigation are included in vector auto-regressive or dynamic linear models. These methods have to be combined with an appropriate level of regularisation as the number of free parameters typically increases quadratically with the spatial, latent or retained\(^{10}\) dimensions \( s \) multiplied by the number of lags \( L, O(Ls^2) \). Additional levels of regularisation are often further imposed on solutions to handle ill-conditioning and over-fitting plaguing problems of these types for high-dimensional data sets.

A simpler approach would be to apply time-embedding (Takens 1981) to construct an augmented data matrix \( \Xi = \Xi_{(m+1)s \times t} = [X(\cdot, m+1 : t); \ldots; X(\cdot, 1 : t-m)] \) by stacking lagged versions of the original data matrix \( X_{s \times t} \) where the embedding dimensions \( m \) is the maximum number of selected lags. Dimension reduction techniques could be applied to the augmented data matrix \( \Xi_{(m+1)s \times t} \) to further reduce the spatial dimensions as in RSAA implemented in Section 4.

d. Multivariate RSAA

The time-embedding construction followed by RSAA described previously is a special case of a more general method where a combined EOF analysis is employed (O’Kane et al. 2017; Hannachi 2021) followed by RSAA on the augmented data matrix \( \Xi_{(s_1 + \ldots + s_m) \times t} = [X^1; \ldots; X^m] \). \( \Xi \) is constructed by stacking \( m \) geophysical fields \( X^i = X^i_{s_i \times t} \) with \( i = 1, \ldots, m \) defined over an identical temporal domain. Each individual field \( X^m \) needs to be suitably scaled as not to favour any particular field in the PCA dimension reduction step as \( X^m \) may have wildly different scales and physical units. A conventional AA is then performed on the retained scaled principal components \( \Lambda V^T_{r \times t} \) which now capture some of the variance common to all \( X^m \). One notes that \( \Xi_{(s_1 + \ldots + s_m) \times t} = U_{(s_1 + \ldots + s_m) \times r} \Lambda_{r \times r} V^T_{r \times t} \)

\(^{10}\)If a dimension reduction step has been implemented for example.
for \( r \leq \min(s_1 + \ldots + s_m, t) \) retained dimensions and the resulting archetype field patterns can be simply recovered as \( U(s_1 + \ldots + s_m) \times r \Lambda_{r \times r} V^T_{r \times t} C_{t \times p} \) for a cardinality equal to \( p \).

7. Application of archetypal analysis

In section 5, we illustrate on Figure 1c how AA identifies extreme SST conditions in the reduced space spanned by the first 3 scaled PCs of SST anomaly fields and we describe the impact of detrending the dataset to separate extremes in the interannual variability from those driven by anthropogenic forcing. Hereafter, we present an application of AA to characterise the El Niño-Southern Oscillation (ENSO), building on the example of Hannachi and Trendafilov (2017). For a choice of four archetypes applied to global SST anomaly fields, the resulting archetypes display patterns indicative of the four ENSO types: the classical eastern Pacific-type El Niño and La Niña (Rasmusson and Carpenter 1982) and the central Pacific-type (coined ‘Modoki’ by Ashok et al. (2007)) El Niño and La Niña (e.g. Fu et al. 1986; Wang 1995; Trenberth and Smith 2006; Kao and Yu 2009; Cai et al. 2009). These patterns are shown in column 1 in Figure 10.

The time series of the archetype coefficients, \( C \) or \( S \), can be used as indices of the ENSO types, and to form composites of other flow fields corresponding to each archetype. Lagged compositing could be used to investigate the state of the ocean or atmosphere in a certain period leading up to, or caused by extreme conditions, though composite results reported in this paper correspond to lag zero. There are a multitude of compositing techniques one might employ using either of the \( C \) or \( S \) matrix time series. We have already discussed two methods for building (time) mean composites of any given field \( F \) based on AA in Section 4, where the stochastic matrices \( C \) and \( S \) are employed to derive composites by simple matrix multiplications where all individual entries of \( C \) and \( S \) are taken into account. However, the discrimination score \( \Delta_p(t) \) (Eq. 10) or the Gini coefficient \( \Gamma_p(t) \) (Eq. 11) can also be used to exclude time records from the stochastic matrices, \( C \) and \( S \), upon which the composites of \( F \) are formed, if these entries are below a given probability threshold informed by the distribution of probabilities across time records of any given archetype. Furthermore, as applied by Risbey et al. (2021) to isolate long lasting atmospheric blocking events, a persistence criterion combined to a discrimination score or Gini coefficient based ‘thresholding’ can be applied to only retain time records to composite upon when a given archetype has 1) a high probability of being expressed and 2) persists over several time records.
In the following examples, we formed composites based on the $S$ matrix as explained in Section 4. Columns 2 and 3 in Figure 10 provide two sets of atmospheric diagnostic composites to elucidate teleconnection patterns associated with the classical (archetypes 3 and 4) and ‘Modoki’ (archetypes 1 and 2) ENSO patterns.

Column 2 in Figure 10 illustrates the anomalous 300 hPa zonal wind component, 500 hPa geopotential height anomalies, and thermal wind anomaly vectors corresponding to the spatial pattern of SST anomaly for each ENSO archetype to the left. Several key features match with previously identified ENSO behavior. Conventional understanding of a Northern Hemisphere wintertime La Niña episode includes a strong westerly Pacific jet stream that splits around a well-developed North Pacific high pressure system (Alexander et al. 2002; Newman et al. 2016; Christensen et al. 2017; Capotondi et al. 2020). Both classical and ‘Modoki’ La Niña archetypes (rows 2 and 3) follow this behavior, though the classical La Niña pattern has a more coherent strengthening of the subtropical jet over Asia into the North Pacific. This strengthening of the jet occurs in association or response to the enhanced thermal wind induced by the warm and cold SST anomalies across the North Pacific Ocean in this pattern. In the Southern Hemisphere the main response for La Niña is in the polar jetstream. For La Niña ‘Modoki’ (row 2) there is an almost circumglobal change from warm SST anomalies to cold SST anomalies around 50°S latitude, which results in westerly thermal wind anomalies and an enhanced polar jetstream. The stronger polar jetstream is associated with lower geopotential height poleward of the jet, indicating enhanced storminess at high latitudes. At lower latitudes in the South Pacific, the subtropical jetstream is weakened for La Niña ‘Modoki’ by the strong easterly thermal wind anomaly.

The ‘Modoki’ El Niño in the top row has a strong anticyclone at 500 hPa in geopotential height over the Gulf of Alaska with evidence of an in situ response to thermal wind anomalies, consistent with findings by Kao and Yu (2009) that indicated ‘Modoki’ ENSO tends to favor in situ development forced primarily by the atmosphere. In the Southern Hemisphere, both El Niño types (rows 1 and 4) feature SST gradients that enhance the thermal wind in the vicinity of the subtropical jet in the Pacific, though this is much stronger for classical El Niño (row 4). The geopotential height anomalies for both El Niño types feature a ridge and trough about South America reminiscent of the Pacific South America pattern.

Accepted for publication in Artificial Intelligence for the Earth Systems.
FIG. 10: AA composite results using detrended monthly SST anomalies over 1982-2020 for a selected archetype number of 4. The first column includes the resulting spatial patterns of SST anomalies constructed with the $S$ matrix values, followed in column 2, by 300 hPa zonal wind component (shading) with superimposed 500 hPa geopotential height anomalies (green contours) and thermal wind anomalies components (vectors) and, in column 3, by monthly averaged daily maximum surface wind speed anomalies (shading) with superimposed velocity potential difference anomalies between 150 and 850 hPa levels (green contours) and wave activity flux anomalies components at 200 hPa (vectors).
A different set of atmospheric diagnostic composites for the same archetype patterns is provided in column 3 of Figure 10. Here, monthly averaged maximum daily wind speed anomalies at the surface are shaded and superimposed with contours of velocity potential difference anomalies between 150 and 850 hPa as in Adames and Wallace (2014), as well as vectors of anomalous wave activity flux (WAF) at 200 hPa (Takaya and Nakamura 2001). In line with recent studies (Liang et al. 2021; Chen et al. 2015), both flavours of El Niño are associated with a lessening of the easterly trade winds. Conversely, both flavours of La Niña correspond to a strengthening of the easterly trade winds. The well-developed north Pacific block in the ‘Modoki’ La Niña (row 2) is supported by strong WAF activity from the tropics into the North Pacific.

The composite behavior of WAF in the Southern Hemisphere may prompt questions about the influence of ENSO on the South Pacific Convergence Zone (SPCZ), which has been described as a ‘graveyard for fronts’ (Trenberth 1976) and more recently associated with Rossby wave breaking (Matthews 2011). In each of the ENSO archetype composites there is a flux of wave activity in the Southern Hemisphere polar waveguide, which tends to move equatorwards across the Australian continent and into the SPCZ region, consistent with the analysis of Matthews (2011). However, there are variations in this picture from case to case as indicated by the set of longitudes where the flux out of the waveguide moves equatorward. For El Niño ‘Modoki’ (row 1) and classical La Niña (row 3) the strong equatorward flux is in the Indian Ocean and Australian continent region. For classical El Niño (row 4) and La Niña (row 2) ‘Modoki’ the equatorward flux is over the Australian continent and there is convergence of wave activity flux in the South Pacific region. In these two cases the impact on the SPCZ seems stronger as indicated by the more coherent northwest-southeast oriented anomaly in maximum surface wind for these composites.

The strongest anomaly composites are found in the tropical to sub-tropical bands for classical El Niño (row 4) and ‘Modoki’ La Niña (row 2) in both surface and atmospheric fields aloft. Monthly averages of maximum daily surface wind speed, velocity potential difference, $\Delta_{150-850}$, and thermal wind anomaly composites show a clear correspondence to ENSO phases. For classical El Niños there is a slackening of the surface trade winds in the Central Pacific, enhanced/reduced convection activity in the Eastern/Western Pacific, and symmetrically diverging thermal wind anomalies from the equatorial region. The opposite conditions are observed for ‘Modoki’ La Niñas with a reinforcement of the trade winds in the Western to Central Pacific, a reduced/enhanced
convection activity in the Eastern/Western Pacific and symmetrically convergent northeasterly and southeasterly thermal wind anomalies toward the equator.

One notable feature of the composite teleconnection patterns associated with the four ENSO types is that they are not particularly symmetric. That is, the teleconnections for classical versus ‘Modoki’ forms are different for the same ENSO type, just as the El Niño and La Niña forms are different. In some cases the teleconnection for the ‘Modoki’ form more closely resembles that for the opposite ENSO type than it does the classical equivalent. This is understandable in that the teleconnections form in response to the global SST patterns for each type, which can be very different outside the tropics. More local gradients in SST in the archetype patterns can drive thermal wind responses that modify the jets and dynamical response in a region.

8. Summary and conclusions

This paper has demonstrated the utility of the AA method and the benefits that arise particularly when analyzing geophysical data. A derivation of RSAA is first provided as the foundation for working with large datasets that require an initial dimensionality reduction step to increase computational efficiency. Using a prototype dataset of monthly SST anomalies between 1982–2020, we have shown how outliers around a broadly ellipsoid-shaped distribution may be readily identifiable as archetypes of the data. These spatial archetype patterns resemble anomalies of SST associated with ENSO. If trends are of interest, the non-detrended data yields archetypes that show gradual warming from an initially cold pattern to a warmer one. Detrending the data prior to AA may remove a global warming pattern and instead reveal different flavours of ENSO, like the central Pacific (aka ‘Modoki’) versus classical eastern Pacific ENSO. As the number of archetypes increase from 4 to 8, the spatial patterns increase in diversity while still retaining familiar ENSO patterns. The Gini coefficient is introduced as a tool to inform on the number of principal components to be retained in the analysis based on the total variance explained by RSAA and conditioned on archetype cardinality. The Gini coefficient can also be used as a univariate discrimination score to identify extreme conditions and their persistence. Lastly, a useful application of AA is presented to show that compositing around the AA matrix factors time series reveals familiar atmospheric teleconnection patterns associated with extreme SST anomaly patterns.
We show that the decisions made to implement AA can greatly affect the interpretation of results. There is \textit{a priori} no guarantee that solutions exist for the minimization problem for any given task, or that the solutions found will be meaningful. Results of AA should therefore be considered carefully through the lens of each individual decision made. Chosen methods should be documented thoroughly in all work involving AA to encourage reproducibility and understanding.

\textit{Acknowledgments.} This research was supported by the Multiyear Climate Project at the Commonwealth Scientific and Industrial Research Organisation, Oceans and Atmosphere. Bernadette Sloyan and Christopher Chapman were also funded by The Centre for Southern Hemisphere Oceans Research, Hobart, Tasmania, Australia. Abdelwaheb Hannachi and Nikolay Trendafilov received no external support for this work. We would like to thank Richard Matear at the Commonwealth Scientific and Industrial Research Organisation, Oceans and Atmosphere, for his constant support and encouragement. We are also grateful to the reviewers whose insightful comments led to substantial improvement of the paper.

\textit{Data availability statement.} Sea surface temperature (SST) data are from the Optimum Interpolation Sea Surface Temperature (OISST) v2.1 high resolution dataset provided by the NOAA/OAR/ESRL Physical Sciences Laboratory. This data is available at https://psl.noaa.gov/. The atmospheric reanalysis data used to relate extreme events to large-scale climate modes comes from the Japanese 55-year Reanalysis (JRA-55) project carried out by the Japan Meteorological Agency (JMA). JRA-55 data is available at https://jra.kishou.go.jp.

\section*{A. Available AA packages}

A number of open-source AA packages are available on-line. They have been implemented for most major computing languages in use today such as MatLab, Python and R. The reader is referred to Table 2 for a non-exhaustive selection of package URLs. Throughout the paper, we have used exclusively the AA package developed by Mørup and Hansen (2012), after trialing several implementations listed in Table 2. We have found the pure MatLab script ‘PCHA’ extremely robust and easy to modify for our purpose. ‘PCHA’ computation speed compares to the SPAMS v2.6 MatLab version of Mairal (2014, 2017) resulting in near-identical solutions in the stochastic matrices $C$ and $S$ with sum of square differences of the order $O(10^{-4})$ for monthly SSTA and an archetype cardinality of 6. We note however that python ‘pymf’-class packages may not be
as robust for high-dimensional data set due to an internal issue with the optimization routine implemented therein. This issue was also reported by Chen et al. (2014). Finally, we have also tested the elegant implementation AA of Hannachi and Trendafilov (2017) and Trendafilov and Gallo (2021) using the optimization on manifold package ‘Manopt’ developed by Boumal et al. (2014) and have found it computationally efficient and twice as fast as ‘PCHA’ for the ‘Trust Region’ optimization solver. All AA optimizations are performed on Intel(R) Core(TM) i7-10875H CPU @ 2.30GHz laptop for small problems, on a dual Intel(R) Xeon(R) Gold 6132 CPU @ 2.60GHz desktop for intermediate problems and a dual Intel(R) Xeon(R) CPU E5-2697A v4 @ 2.60GHz blade server for large problems.

<table>
<thead>
<tr>
<th>URL</th>
<th>Language</th>
<th>Main reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="https://CRAN.R-project.org/package=archetypes">https://CRAN.R-project.org/package=archetypes</a></td>
<td>R</td>
<td>Eugster and Leisch (2011)</td>
</tr>
<tr>
<td><a href="https://CRAN.R-project.org/package=Anthropometry">https://CRAN.R-project.org/package=Anthropometry</a></td>
<td>R</td>
<td>Vinué (2017)</td>
</tr>
<tr>
<td><a href="http://www.mortenmorup.dk">http://www.mortenmorup.dk</a></td>
<td>MatLab</td>
<td>Mørup and Hansen (2012)</td>
</tr>
<tr>
<td><a href="https://github.com/AlonLabWIS/ParTI.git">https://github.com/AlonLabWIS/ParTI.git</a></td>
<td>MatLab</td>
<td>Boumal et al. (2014)</td>
</tr>
<tr>
<td><a href="http://github.com/NicolasBoumal/manopt">http://github.com/NicolasBoumal/manopt</a></td>
<td>MatLab</td>
<td>Gillis (2020)</td>
</tr>
<tr>
<td><a href="https://github.com/jhausser/ParTI.git">https://github.com/jhausser/ParTI.git</a></td>
<td>MatLab</td>
<td>Mørup and Hansen (2012)</td>
</tr>
<tr>
<td><a href="https://github.com/AlonLabWIS/ParTI.git">https://github.com/AlonLabWIS/ParTI.git</a></td>
<td>MatLab</td>
<td>Mørup and Hansen (2012)</td>
</tr>
<tr>
<td><a href="https://github.com/smair/archetypalanalysis-coreset">https://github.com/smair/archetypalanalysis-coreset</a></td>
<td>Python</td>
<td>Mair et al. (2017); Mair and Brefeld (2019)</td>
</tr>
<tr>
<td><a href="https://pypi.org/project/PyMF">https://pypi.org/project/PyMF</a></td>
<td>Python</td>
<td>Thurau et al. (2009, 2011)</td>
</tr>
<tr>
<td><a href="https://github.com/ChrisSchinnerl/pymf3.git">https://github.com/ChrisSchinnerl/pymf3.git</a></td>
<td>Python</td>
<td>Keller et al. (2021)</td>
</tr>
<tr>
<td><a href="https://doi.org/10.25919/5d39588889f7ff">https://doi.org/10.25919/5d39588889f7ff</a></td>
<td>Python</td>
<td></td>
</tr>
</tbody>
</table>

TABLE 2: A non-exhaustive list of archetypal analysis package URLs with corresponding computing language types and main references.

References


Lorenz, E., 1956: Empirical orthogonal functions and statistical weather prediction.


43

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-21-0007.1.


Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-21-0007.1.


