Dataset Reduction Techniques to Speed Up SVD Analyses on Big Geo DataSets

Laurens Bogaardt [1] , **Romulo Goncalves** [1] , **Raul Zurita-Milla** [2],* and **Emma Izquierdo-Verdiguier** [3]

1   Netherlands eScience Center; l.bogaardt@esciencecenter.nl, r.goncalves@esciencecenter.nl
2   Faculty ITC, University of Twente; r.zurita-milla@utwente.nl
3   IVFL, University of Natural Resources and Life Sciences (BOKU); emma.izquierdo@boku.ac.at
*   Correspondence: r.zurita-milla@utwente.nl

1   **Abstract:** The Singular Value Decomposition (SVD) is a linear algebra procedure with multiple
2   applications in the geosciences. For instance in dimensionality reduction and as support operator
3   for various analytical tasks applicable to spatio-temporal data. Performing SVD analyses on
4   large datasets, however, can be computationally costly, time consuming and sometimes practically
5   infeasible. Yet, techniques exist to arrive at the same output, or at a close approximation, which
6   require far less effort. This article examines several such techniques in relation to the inherent scale of
7   the structure within the data. Datasets with large structures contain geographical and time based
8   fields with large autocorrelation, e.g. in a spatial field of temperature over a country the temperature
9   values vary slowly. Hence, they do not need a high resolution to describe such fields and their
10 analysis can benefit from alternative SVD techniques based on rank decomposition, coarsening or
11 matrix factorization approaches. We use both simulated Gaussian Random Fields with various levels
12 of autocorrelation and real-world geospatial datasets to illustrate our study while examining the
13 accuracy of various SVD techniques. As the main result, this article provides researchers with a
14 decision tree indicating which technique to use when and predicting the resulting level of accuracy
15 based on the dataset's structure scale.

16 **Keywords:** Singular value decomposition, autocorrelation, rank deficiency, data reduction,
17 coarsening, approximate SVD, Gaussian Random Fields

1. Introduction

19   The *Singular Value Decomposition* (SVD) is a linear algebra procedure used to factorize a matrix
20 A as a product of three matrices (**U**,**V**, and **S**), where the matrices **U** and **V** are orthonormal and **S**
21 is a diagonal matrix that contains the so-called singular values [**?**]. This matrix decomposition has
22 found multiple applications in both engineering and scientific disciplines [1–3]. In the geosciences,
23 SVD helps to identify structure when analysing a spatial field for a single time period. In many
24 real-world applications, however, spatial fields include multiple time periods or contain multiple
25 thematic attributes (e.g. spectral bands in the case of images). In these cases, SVD helps to eliminate
26 some of the redundancy between time stamps or attributes as well as between neighbouring locations
27 (pixels in images and grid cells in rasters). The SVD can thus be seen as an efficient dimensionality
28 reduction technique that also facilitates further analysis. For instance, SVD helps to obtain accurate
29 products of spatio-temporal fields (matrices), and to reduce the deficiencies of the acquired data (noise
30 or non-linear nature) [4]. Moreover, several analytical approaches commonly used in the geosciences
31 are based on the SVD. For example, the Partial Least Square method commonly used in classification

[5] and regression [6] tasks. The *Maximum Covariance Analysis* (MCA) and *Canonical Correlation Analysis* (CCA) methods, which are often used to analyse multi-temporal datasets, are also based on the SVD [7]. A combination of CCA with *Minimum Noise Fraction* (MNF) has proven successful at filtering out noise from images [8]. Last but not least, the SVD has been used to compare ground data and measurements done by Earth observation satellites [9,10]. This is because SVD offers an effective method to find frequent and simultaneous patterns between two spatio-temporal fields [11,12].

SVD-based analysis are not only applicable to spatio-temporal fields and Earth observation images. Many geo-datasets can be represented as a matrix. For instance, temperatures measured at m locations for n time periods can be arranged in a matrix with m rows and n columns. From a linear algebra perspective, this matrix is nothing more than a linear combination of basis vectors that indicate direction, each with a coefficient that indicates magnitude. As an extension of vectors, matrices have two bases, the row- and the column bases, which can be changed via a rotation. A clever basis to rotate into is one where the product of the first row- and column basis vectors explains as much of the variance in the dataset as possible. Subsequent pairs of basis vectors, known as *modes*, explain as much of the remaining variance as possible while being orthogonal to all previous modes. Such basis vectors are called *Principle Components* (PC's) or *Empirical Orthogonal Functions* (EOF's) and they are commonly found via an SVD of the data matrix. Again these vectors can be derived using SVD and are their use is widespread in the geosciences [13,14].

Besides noise, real-world datasets also contain redundancy. For example, datasets often contain values that vary slowly and neighbouring values are not entirely independent of one another, neither in space nor in time [11]. In these cases, there is a high level of autocorrelation and the data matrix is said to contain large scale structures. Such an autocorrelation means redundancy, which in turn means that the data can be reduced in size without losing much relevant information. Often, techniques exist to arrive at the same output, or at a close approximation, which require far less effort. This article examines several SVD procedures which exploit autocorrelation and rank decomposition to analyse data in an efficient manner.

In this study we first use simulated spatio-temporal fields with different levels of autocorrelation to study the relationship between autocorrelation and the accuracy of different SVD implementations. Then, the lessons learned with simulated data are verified using real-world datasets. Finally, we discuss which type of problems could be tackled with which SVD implementation and predict the error incurred by using approximations based on the level of autocorrelation of the input data. To the best of our knowledge, this is the first review that examines various alternative SVD implementations for large datasets from the point of view of autocorrelation and data reduction [15–17].

In fact, our review is not only relevant to datasets with high autocorrelation but also to big geo-datasets. Indeed, performing a standard SVD on a big matrix is computationally costly and sometimes even prohibitive. Yet the various SVD implementations presented and discussed in this article help to overcome this computational limitation.

The paper is outlined as follows. Section **??** reviews matrix decomposition and the standard formulation of SVD. Section 3 explains the different SVD implementations and presents the the simulated and real-world data sets used in their study as well as our experimental results. Finally, Section **??** summarizes our main findings and recommendations.

2. Materials and Methods

2.1. Matrix Decomposition

As briefly mentioned in the introduction, a matrix can be seen as a combination of basis vectors with their coefficients. If there exists a rotation for which some coefficients become zero, the matrix needs fewer basis vectors to describe it than are available. In a sense, it is underdetermined; its internal dimension is smaller than what would be expected from its m by n size. This is the concept of matrix *rank*; if the rows and columns both span a subspace of dimension r, a matrix has rank r. A matrix is

said to have full rank if $r = \min(m, n)$, the maximum number of linearly independent basis vectors. It is rank deficient if $r < \min(m, n)$. A rank decomposition or factorization is the splitting of a matrix into a product where each factor has full rank. For example, an *m* by *n* matrix of rank *r* can be decomposed into an *m* by *r* matrix multiplied by an *r* by *n* one.

For rank deficient matrices, some of the singular values are zero. These zero singular values and their associated basis vectors can be truncated without affecting the data but reducing storage and computational requirements. Whereas the entire, original dataset requires $m \times n$ units of storage and computing the product with a vector requires $m \times n$ flops, the rank decomposed version requires $m \times r + r \times n$ units of storage and an equal number of flops for the vector multiplication [18]. If *r* is small, this can be a substantial improvement. Thus, it is possible to use the rank decomposed version to reduce storage and have faster computation. In many real-world applications, the mathematical rank *r* of a dataset is usually not relevant because of the inherent noise level of data gathered by machines with finite precision [18]. Hence, some singular values may be small enough to be considered noise and be rounded off to zero. In these cases, we can approximate a dataset by another matrix of rank *l*, with $l < r$, without losing much relevant information.

Following the Eckart-Young theorem, the best approximation is one described in the same bases as the original dataset, taking a subset of the *l* largest singular values and truncating the remainder [19]. Setting a threshold ϵ, the dataset is approximate rank deficient if some singular values fall below ϵ. Then, it has an ϵ-rank of *l* and the spectral norm of the difference with its approximation is at most ϵ [18]. The idea is to reduce the data to the point where the error due to reduction is around the noise level.

2.2. Simulated Spatio-Temporal Fields

This section expands on a method to simulate fields which resemble those often encountered in real-world applications. Additionally, we discuss how to measure the level of autocorrelation of a field. As simulated spatio-temporal fields, real-valued *Gaussian Random Fields* (GRF's) are particularly useful because their structure scale can be captured in a single parameter. For such rotational invariant fields, the spectrum follows the power law described by $P(\vec{k}) \sim |\vec{k}|^{-\alpha}$ where $\vec{k}$ is the wavevector and α the parameter which controls the level of autocorrelation. Figure 1 shows fields with various α's.

(a) $\alpha = 1$	**(b)** $\alpha = 3$	**(c)** $\alpha = 5$
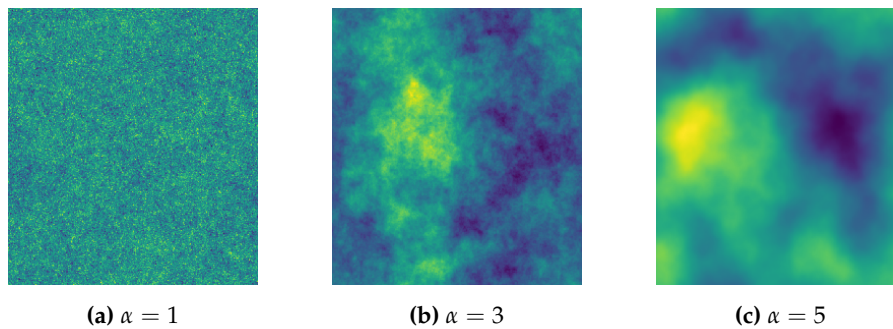
Figure 1. Gaussian Random Fields for various α's

Just as there is spatial autocorrelation, there is temporal autocorrelation, when the values of the field over the entire time period do not change drastically. In principle, there can be different levels of autocorrelation over time and over space. However, for simplicity, in our analyses we will use the same level of autocorrelation in all directions, determined by parameter α.

2.3. SVD

The SVD is used factorize a matrix while finding its eigenvectors and eigenvalues. A sort review of SVD from a mathematical perspective is as follows:
Let **A** be a matrix ($\mathbb{R}^{n \times m}$, where if $n = m$, **A** is a rectangular matrix and if $n = m$, **A** is a square

115   matrix), the SVD is the factorization of the matrix: $\mathbf{A} = \mathbf{USV}^\top$ where $\mathbf{U} \in \mathbb{R}^{n \times n}$ is the eigenvector
116   matrix associated with non-zero eigenvalues of $\mathbf{AA}^\top$, $\mathbf{V} \in \mathbb{R}^{m \times m}$ is the eigenvector matrix associated
117   with non-zero eigenvalues of $\mathbf{A}^\top \mathbf{A}$, and $\mathbf{S}$ is a diagonal matrix that contains the Singular Values of $\mathbf{A}$.
118   The Singular Values are the square root eigenvalues of $\mathbf{AA}^\top$. Matrix $\mathbf{A}$ can be a singular matrix or a
119   product between two matrix and we can express also the SVD in vectorial from: $\mathbf{A} = \sum_{i=1}^{d_f} \mathbf{s}_i \mathbf{u}_i \mathbf{v}_i^\top$.

2.4. SVD Techniques

121   This section reviews several SVD implementations that can be used to analyse large datasets
122   efficiently by exploiting autocorrelation and rank deficiency. Additional methods exist, though the
123   ones covered here have three main benefits: coarsening is an easy-to-implement method, randomised
124   dimensionality reduction provides the best possible approximation for any level of reduction and the
125   QR decomposition provides an exact and efficient result when interested in the SVD of the product of
126   two matrices.
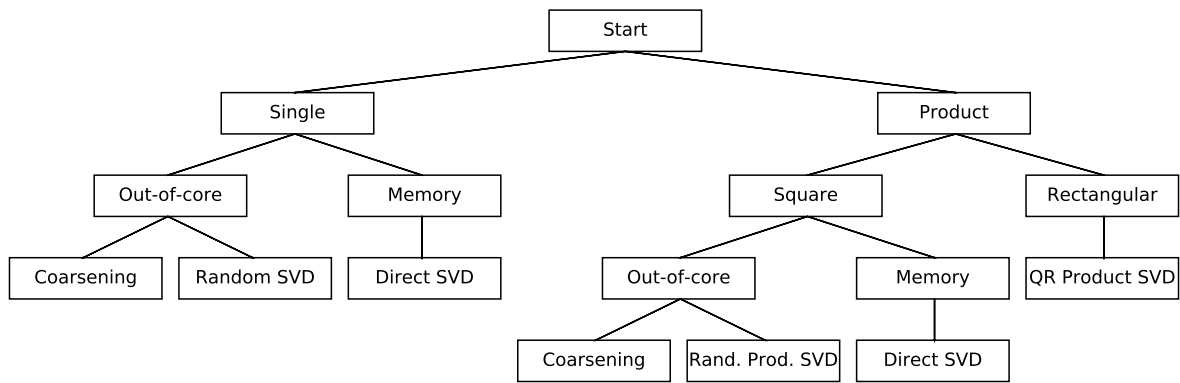
Figure 2. Decision tree describing the possible SVD techniques

127   To help researchers identify situations where different SVD approaches are beneficial, we have
128   constructed the decision tree in figure 2. The first question to be answered is whether the SVD will be
129   applied to a single matrix or to the product of two matrices. For single fields, the data may be small
130   enough to fit in the memory of a computer. Then, a regular SVD is the best option. If the dataset is
131   too large, two alternatives exists which provide an approximate answer: coarsening and randomised
132   dimensionality reduction. These will be discussed in section 3.2.
133   When the SVD is performed on the product of two matrices, the best course of action depends on
134   whether the matrices are square or rectangular. The rank of a matrix is at most the size of the smallest
135   side, which, for rectangular matrices, can be small. How to exploit this fact is described in section 3.3.
136   Square matrices small enough to fit in memory can be analysed directly. Variations of coarsening and
137   dimensionality reduction can assist analyses of larger datasets, discussed in section 3.4.

3. Results

3.1. Measures of Autocorrelation

140   In the geosciences, there are additional measures of spatial autocorrelation [11,12]. One frequently
141   used is Moran's I [20–22]. Figure 3 shows the relationship between Moran's I, using a uniform kernel
142   with a bandwidth equal to 10, and the α of our simulated GRF's.
143   One can also devise an autocorrelation measure from the singular values of a dataset. Each
144   singular value indicates the amount of variance explained by its associated mode. For fields with
145   autocorrelation, the sorted list of singular values decays quickly. A power law can be fitted to this list,
146   with an exponent which we call β.

147   A high α implies a high Moran's *I* and a high β, which, in turn, implies some singular values are
148   close to zero. Therefore, spatial fields with high levels of autocorrelation are described by matrices
149   which are approximate rank deficient. This allows for data reduction without losing much information.
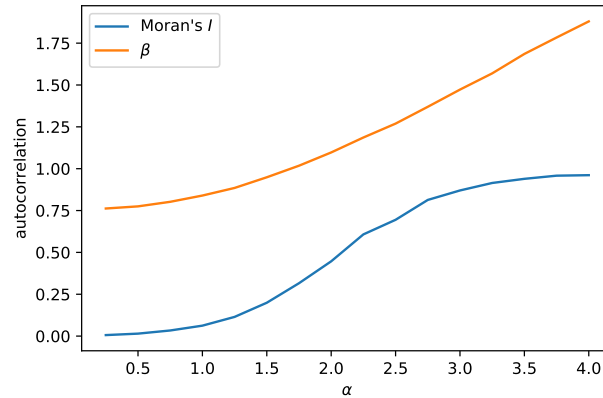

Figure 3. Measures of autocorrelation as a function of α

3.2. SVD of a Single Matrix

150

151   3.2.1. Approximate SVD of a Single Matrix via Coarsening

152   When a spatial field has large scale structure, the values of neighbouring cells do not change
153   drastically. Perhaps these cells can be aggregated together to produce a smaller dataset which still
154   faithfully describes the original field. Here, we exploit the data redundancy described in section **??**
155   and coarsen two dimensional GRF's.

156   Figure 4 shows the error in a coarsening process for matrices of various α's and for different
157   coarsening window sizes. The error is determined as the norm of the difference between the original
158   matrix and the coarsened version, divided by the norm of the original [23]. Other measures of
159   similarity, such as the correlation between the two datasets, are likely to show the same effects. Note
160   that coarsening a two dimensional field at level 5 reduces the dataset by 25 times. For fields with high
161   autocorrelation, e.g. $\alpha = 3$, the coarsened version differs by less than 10% from the original.
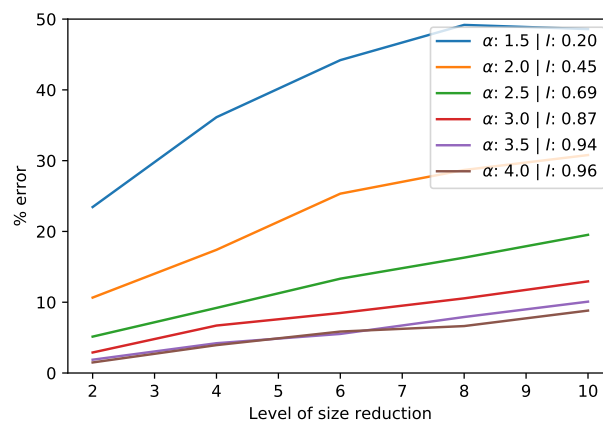
Figure 4. Error in the SVD of a coarsened spatial field for various α's

162   3.2.2. Approximate SVD of a Single Matrix via Dimensionality Reduction

163   An alternative method of reducing the size of a matrix is via randomised dimensionality reduction.
164   During dimensionality reduction, the number of basis-vectors of a matrix is truncated to the *l* most
165   important ones, similar to finding an ϵ-rank approximation. Remember from section **??** that, if a

166 dataset is noisy, low rank approximations can contain most of the relevant information. Furthermore.
167 matrices with low rank require less storage space and fewer flops for vector multiplications.

168   A random algorithm exists which performs an approximate rank decomposition of a large matrix
169 efficiently. The mathematics behind this algorithm is reviewed extensively elsewhere, here we examine
170 its performance on datasets with autocorrelation [24,25]. The main idea is to sample from the original
171 values, taking as many values as needed to achieve the required accuracy. One of its benefits is that
172 this algorithm requires only a small number of passes over the data, which, for large matrices stored
173 out-of-core, reduces reading time. Additionally, the incurred error in the approximation can be made
174 arbitrarily small by adjusting l and ϵ, giving the researcher full control over the balance between
175 computation cost and accuracy.

176   Figure 5 depicts the calculation in this process, which first reduces the input matrix to a smaller
177 square matrix of l by l. It also provides two projection matrices which rotate the rows and columns of
178 this smaller matrix back as close as possible to the bases of the original input. Subsequently, the SVD is
179 applied to the small l by l matrix, which results in a fast and efficient approximation of the matrix's
180 decomposition with an error of the order of the size of the largest truncated singular value [18,24].

Figure 5. Visualising the calculation of an approximate SVD via dimensionality reduction

181   The calculations in the *Jupyter Notebook*, which is provided in the suplementary material, show
182 that the errors induced by the randomised SVD procedure are very small, even for high levels of
183 reduction [23]. This is especially true for fields with autocorrelation. Looking at figure 3, this is not
184 surprising; high autocorrelation means a high β which implies the singular values decay quickly and
185 that the last modes contribute little information to the dataset.

186   This randomised technique achieves much lower errors than coarsening. The coarsening
187 procedure has several advantages though. For one, it is intuitive and the results are easy to interpret.
188 It is also trivial to implement. Additionally, different coarsening levels can be applied to different
189 directions. This is especially advantages when directions have different levels of autocorrelation or are
190 recorded at different resolutions. Finally, the predictions of figure 4 can help researchers determine at
191 what resolution to gather their data in the first place. In domains were satellite data is used, datasets
192 are often not very detailed because the imaging resolution is low. Unlike local analyses of developed
193 countries, where high resolution data is becoming more accessible, for continental or global analyses,
194 coarse spatial resolution data may simply be the only option.

3.2.3. Case Study of an SVD of a Single Matrix

196   Coarsening and dimensionality reduction are particularly useful for spatial fields with high
197 levels of autocorrelation. As an example of this, we examine humidity and cloud cover data from the
198 ERA5 datasets for a single time period measured on 1 April 2012. ERA5 is an atmospheric reanalysis
199 of the global climate using high spatial resolution forecasts, produced by combining models with
200 observations [26]. It contains estimates of atmospheric parameters such as air temperature, pressure
201 and wind at different altitudes.

202   Clearly, the ERA5 humidity and cloud cover fields are not GRF's. Nonetheless, we can get an idea
203 of the accuracy of an SVD after data reduction if we estimate the levels of autocorrelation. This can
204 verify whether simulated GRF's are reasonable representations of real-world datasets. The humidity
205 field has a Moran's $I \approx 0.98$, while the cloud cover data shows less structure with a Moran's $I \approx 0.82$.
206 The estimations for α were unreliable as the power law did not fit properly, though figure 3 can help
207 us translate the measures and suggests the fields are equivalent to GRF's with an $\alpha \sim 3.5$ and $\alpha \sim 2.5$,
208 respectively.

209　　The coarsening predictions of figure 4 indicate the first field is expected to incur errors around
210　a few percent for size reductions between 2 and 8, while for the second field we should see errors
211　between 5% and 15%. The calculations in the *Jupyter Notebook* in the supplementary material show
212　that this prediction is fairly accurate, perhaps slightly pessimistic. The humidity field, which has the
213　largest autocorrelation, only incurred an error of 1% when halved in size. For the cloud cover data, this
214　values was closer to 4%. Likewise, we can apply the randomised dimensionality reduction technique
215　to the fields. As expected, this resulted in very small errors, below 1%. If a researcher cares most about
216　accuracy, dimensionality reduction is the best option.

3.3. Product SVD of Rectangular Matrices

218　　In real-world applications, researchers often wants to find the relation between two fields.
219　Analyses such as the MCA and CCA, discussed in section 1, expose patterns in the data which
220　occur frequently and simultaneously [11,12]. In some domains, the term SVD is used synonymously
221　with MCA. If the input datasets have the various spatial gridpoints as rows and the sample of recorded
222　values over time as columns, multiplying them gives their cross-covariance matrix. An SVD of this
223　cross-covariance matrix provides the row- and the column vectors which covary maximally [27].
224　Multiplying the two spatio-temporal fields, however, can result in a rather large matrix which may not
225　fit in *RAM*-memory and would be difficult to analyse. This section and section 3.4 describe possible
226　solutions.

### 3.3.1. Exact Product SVD of Rectangular Matrices via QR Decomposition

228　　For highly rectangular datasets, when there are many spatial gridpoints but few temporal samples,
229　the resulting cross-covariance matrix is obviously rank deficient. Performing a rank decomposition on
230　each dataset before multiplying them allows the SVD to be calculated in an efficient manner [17,28].
231　Figure 6 depicts the calculation in this technique. Using the QR decomposition, the input datasets are
232　first transformed into two small, square matrices together with rectangular orthonormal basis vectors.
233　The SVD is then performed on the product of the square matrices, giving a mathematically identical
234　result to the full SVD while never forming an unnecessarily large, intermediate matrix.

Figure 6. Visualising the calculation of the exact SVD of a product via QR decomposition

### 3.3.2. Case Study of a Product SVD of Rectangular Matrices

236　　Let's apply the QR decomposition technique to phenological datasets. Phenology is the science
237　that studies recurring biological events such as leafing and blooming as well as their causes and
238　variations in space and time. Spatio-temporal fields of remotely sensed images can be used to
239　derive various phenological metrics. One important family of such metrics are the *Extended Spring*
240　*Indices* (SI-x), which are based on a suite of models that transform daily temperatures into consistent
241　phenological metrics [29]. In particular, we take versions of the Leaf and the Bloom index, depicted in
242　Figure 7, which were recently generated for the US by adapting the SI-x models to a cloud computing
243　environment [30].
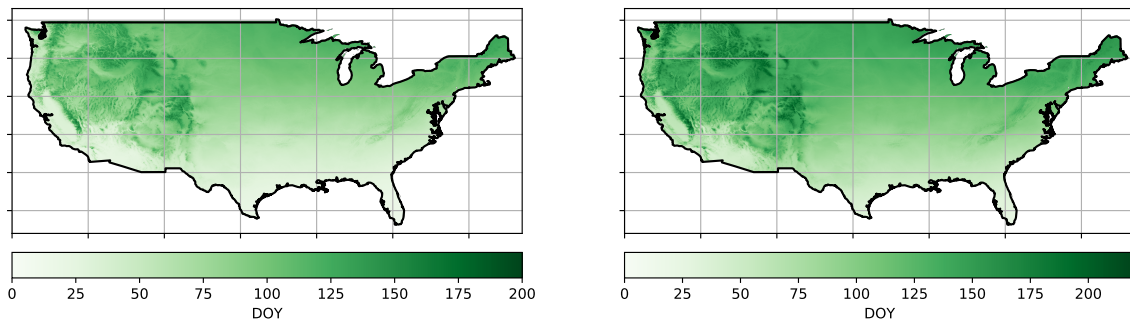
Figure 7. Average of phenology products: Leaf [l] and Bloom [r] from 1989 to 2014

Both datasets span from 1989 to 2014 and have a 1km^2 spatial resolution, meaning there are far fewer time periods than spatial gridpoints, giving highly rectangular matrices. In fact, each dataset contains 30 million rows and 26 columns and is about 3.1 GB large. Their cross-covariance would be 30 million rows by 30 million columns and about 3.6 PB in size. Luckily, this product matrix need not be created; the SVD is performed on a 26 rows by 26 columns matrix, stored in memory.

For the SVD via QR decomposition, the level of autocorrelation is irrelevant because this technique provides a mathematically exact result. Indeed, the *Jupyter Notebook* in the supplementary material, as well as work being prepared for publication, shows that this technique provides the full SVD of the cross-covariance matrix for these datasets in a matter of seconds, without ever exceeding the *RAM*-memory [23,31]. The first mode for both Leaf and Bloom data can be mapped back into the original geographical shape, as depicted in Figure 8.
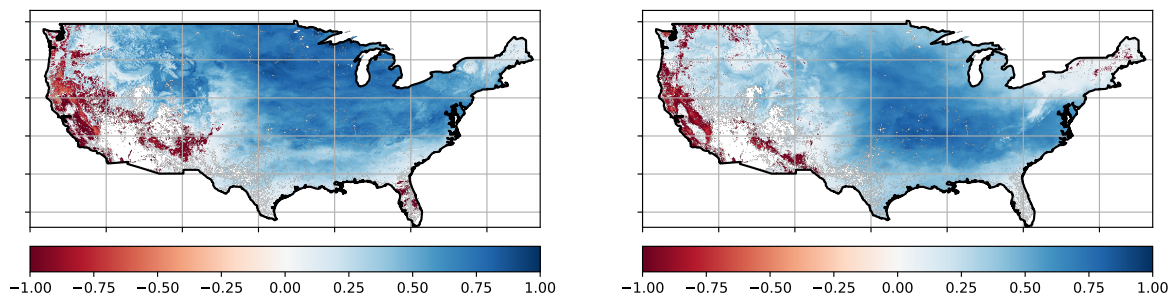

Figure 8. Phenology products: Leaf [l] and Bloom [r] data projected onto the first principal component

3.4. Product SVD of Square Matrices

3.4.1. Approximate Product SVD of Square Matrices via Coarsening

In section 3.2.1, we saw coarsening can reduce the size of a single spatial field. We can also coarsen two spatio-temporal fields before analysing their cross-covariance matrix. Figure 9 shows the percentage error for various simulated spatio-temporal fields. Note that only the spatial directions are coarsened in our calculation. This is because the time direction gets consumed in the matrix product of the MCA or CCA and coarsening it will not decrease the size of the cross-covariance matrix nor speed up the SVD. While coarsening two spatial directions means each field is reduced by the square of the coarsening level, the cross-covariance matrix is reduced by this level to the power 4. As a result, the typical error in this product is larger than for the single field, though the speed up is also substantial. Clearly, the level of autocorrelation plays an important part, with larger α's leading to less error.
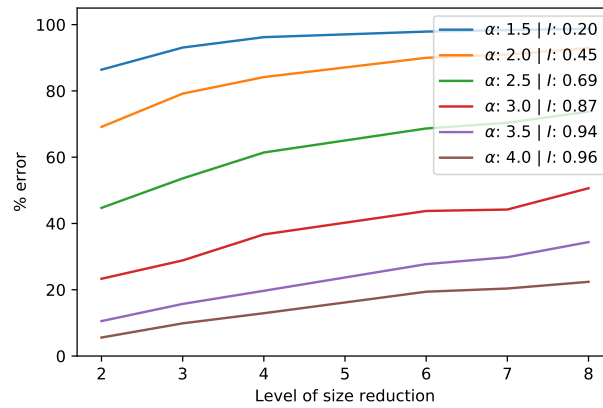
Figure 9. Error in the SVD of the product of two coarsened fields for various α's

3.4.2. Approximate Product SVD of Square Matrices via Dimensionality Reduction

The randomised dimensionality reduction process may also speed up the SVD analysis of two spatio-temporal fields. The reduction can be applied to each of the fields before they are multiplied into the cross-covariance matrix. Similar to the QR decomposition of section 3.3.1, it has the advantage that the SVD is performed on a small l by l matrix. This calculation is visualised in figure 10. In the first step, the input datasets are decomposed using the algorithm discussed in section 3.2.2 and reviewed extensively elsewhere [24,25]. Again, the main idea behind the algorithm is to sample from the original values, taking as many values as needed to achieve the required accuracy. Subsequently, the inner matrices are multiplied and the SVD is performed on the resulting, small matrix.

Figure 10. Visualising the calculation of an approximate SVD for the product of two fields using dimensionality reduction

After generating various GRF's, the reduced cross-covariance matrix is compared with the original. Figure 11 shows that the results are not satisfactory for fields with a small α, but high levels of autocorrelation allow for substantial savings in computation time without incurring much error. Note that, when performing an MCA or CCA on a spatio-temporal field, the spatial directions are flattened and some of the spatial autocorrelation is lost. This partially explains why the error is substantial for low α.

The reduction of the number of dimensions of each input dataset before an MCA or CCA is actually advised by some researchers, as a method to filter out noise [32]. Especially when the number of temporal samples is small, outliers and random fluctuations could affect the result [27]. This is because any statistical analysis will choose its regression-coefficients so as to optimize the fit. It may occur that two noise-vectors in the two fields coincidentally covary and show up as dominant modes. Prefiltering can alleviate this risk.
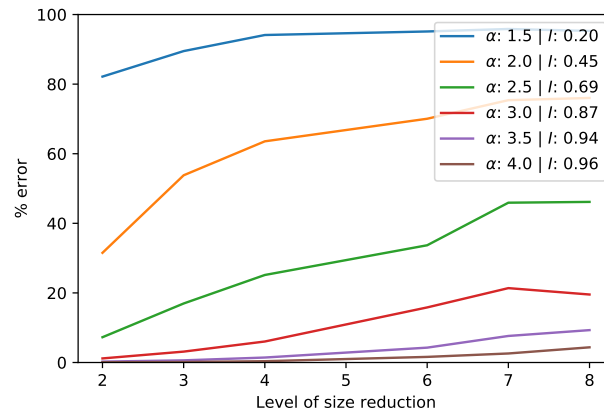
Figure 11. Error in the SVD of the product of two reduced fields for various α's

3.4.3. Case Study of a Product SVD of Square Matrices

The JRA55 data is an atmosphere reanalysis product which includes quantities such as humidity, pressure and temperature [33]. Recently, these quantities were used to determine the total meridional energy transport and latent heat, measures important to understand the global climate [34]. As an example of our reduction techniques for matrix products, we are using a Mercator projection of the energy transport and latent heat, recorded monthly from 1979 to 2015. Although the spatially flattened matrices are not completely square, their high resolution in the time direction make them substantially less rectangular than the phenology data from section 3.3.2. Therefore, this serves as a good use case for the coarsening and dimensionality reduction techniques for a square product SVD.

The energy field has a Moran's $I \approx 0.93$, while the latent heat field has a Moran's $I \approx 0.86$. The estimations for α were unreliable, though figure 3 can help us translate the measures and suggests the fields are equivalent to GRF's with an $\alpha \sim 3.0$ and $\alpha \sim 2.5$, respectively. The analyses of section 3.4.1 and of section 3.4.2 make the prediction that, for such α levels, our data reduction techniques result in errors which are quite high. In contrast, we found that coarsening the fields before applying an SVD on their product merely resulted in an error of 8% when the data was halved and an error around 16% when the data was reduced by a factor of 4. For the dimensionality reduction technique, as well, the predictions overstated the observed error, which were around a few percent, even for high levels of size reduction.

4. Discussion

4.1. Further Work and Caveats

Much of the analysis here relies on knowledge of the level of autocorrelation, which may be difficult to determine for large datasets. The *Jupyter Notebook* accompanying this article includes an algorithm which estimates Moran's I based on a sample of gridpoints. This speeds up the calculation substantially compared with the full calculation and gives very similar results. A spin-off from the current article could be to further develop this approximate measure of Moran's I. Additional areas of future research include relaxing the assumption that the autocorrelation in the time direction is similar to that in the spatial directions. In fact, it is more realistic to allow for different levels of autocorrelation in all directions and to have a version of Moran's I which can estimate these values.

Furthermore, a warning about autocorrelation and standardisation. In MCA's, the time series of each spatial gridpoint is centred about its mean and in CCA's, each gridpoint is standardised. These operations destroy much of the spatial autocorrelation, as it can affect neighbouring cells differently. When researcher choose the coarsening technique, this should occur before any additional data processing steps.

Unlike the coarsening procedure, the dimensionality reduction is not applied on each spatial field for each time period, but rather on the entire spatially flattened time series. Therefore, the level of spatial autocorrelation may not be as important as the level of temporal autocorrelation. Further work can examine how to apply the reduction to the spatial part of the spatio-temporal fields, before it is flattened. Alternative solutions, which retain the spatial structure, may include 3D tensor operations such as *Higher-Order Singular Value Decomposition* (HOSVD) [35].

4.2. Conclusion

Randomised dimensionality reduction works best for datasets which are too large for internal memory. It requires only a small number of passes over the data, which decreases storage reading time. It also allows the researcher to balance computation cost with accuracy, by tuning the algorithm's parameters. Performing analyses at a coarse level can be beneficial when data collection is difficult and provides an intuitive and easy-to-implement alternative. These techniques require at least some autocorrelation in the fields, which results in approximate rank deficient datasets. In general, rank decompositions can speed up calculations by splitting datasets into smaller matrices of full rank. For the product of rectangular matrices, this can even give a mathematically exact result.

Author Contributions: Conceptualization, R.Z.M.; Methodology, L.B.; Software, L.B.; Validation, R.G., R.Z.M. and E.I.V.; Formal Analysis, L.B.; Investigation, L.B.; Resources, R.G., R.Z.M. and E.I.V.; Data Curation, R.G., R.Z.M. and E.I.V.; Writing—Original Draft Preparation, L.B.; Writing—Review & Editing, R.G., R.Z.M. and E.I.V.; Visualization, L.B.; Supervision, R.G. and R.Z.M.; Project Administration, R.G.; Funding Acquisition, R.Z.M.

Abbreviations

The following abbreviations are used in this manuscript:

SVD	Singular value decomposition
PLS	Partial least squares
MCA	Maximum covariance analysis
CCA	Canonical correlation analysis
MNF	Minimum noise fraction
PC	Principle component
EOF	Empirical orthogonal function
GRF	Gaussian random field
SI-x	Extended spring indices
ERA5	European fifth generation reanalysis
JRA55	Japanese 55-year reanalysis
HOSVD	Higher-order singular value decomposition

1. Rajwade, A.; Rangarajan, A.; Banerjee, A. Image Denoising Using the Higher Order Singular Value Decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2013**, *35*, 849–862.

2. Khoshbin, F.; Bonakdari, H.; Ashraf Talesh, S.H.; Ebtehaj, I.; Zaji, A.H.; Azimi, H. Adaptive neuro-fuzzy inference system multi-objective optimization using the genetic algorithm/singular value decomposition method for modelling the discharge coefficient in rectangular sharp-crested side weirs. *Engineering Optimization* **2016**, *48*, 933–948.

3. Meuwissen, T.H.; Indahl, U.G.; Ødegård, J. Variable selection models for genomic selection using whole-genome sequence data and singular value decomposition. *Genetics Selection Evolution* **2017**, *49*, 94.

4. Izquierdo-Verdiguier, E.; Laparra, V.; Marí, J.M.; Chova, L.G.; Camps-Valls, G. Advanced Feature Extraction for Earth Observation Data Processing. In *Comprehensive remote sensing, volume 2: data processing and analysis methodology*; Elsevier, 2017.

5. Izquierdo-Verdiguier, E.; Gómez-Chova, L.; Bruzzone, L.; Camps-Valls, G. Semisupervised kernel feature extraction for remote sensing image analysis. *IEEE trans. on geoscience and remote sensing* **2014**, *52*, 5567–5578.

6. Hansen, P.; Schjoerring, J. Reflectance measurement of canopy biomass and nitrogen status in wheat crops using normalized difference vegetation indices and partial least squares regression. *Remote sensing of environment* **2003**, *86*, 542–553.

7. Munoz-Mari, J.; Gomez-Chova, L.; Amoros, J.; Izquierdo, E.; Camps-Valls, G. Multiset Kernel CCA for multitemporal image classification. Analysis of Multi-temporal Remote Sensing Images, MultiTemp 2013: 7th International Workshop on the. IEEE, 2013, pp. 1–4.

8. Nielsen, A.A. The regularized iteratively reweighted MAD method for change detection in multi-and hyperspectral data. *IEEE Transactions on Image processing* **2007**, *16*, 463–478.

9. Li, J.; Carlson, B.E.; Lacis, A.A. Application of spectral analysis techniques in the intercomparison of aerosol data. Part II: Using maximum covariance analysis to effectively compare spatiotemporal variability of satellite and AERONET measured aerosol optical depth. *J. of Geophysical Research: Atmospheres*, *119*, 153–166.

10. Li, J.; Carlson, B.E.; Lacis, A.A. Application of spectral analysis techniques to the intercomparison of aerosol data. Part IV: Synthesized analysis of multisensor satellite and ground-based AOD measurements using combined maximum covariance analysis. *Atmospheric Measurement Techniques* **2014**, *7*, 2531–2549.

11. Eshel, G. *Spatiotemporal data analysis*; Princeton University Press, 2011.

12. von Storch, H.; Zwiers, F.W. *Statistical analysis in climate research*; Cambridge University Press, 1999.

13. Demirel, H.; Ozcinar, C.; Anbarjafari, G. Satellite Image Contrast Enhancement Using Discrete Wavelet Transform and Singular Value Decomposition. *IEEE Geoscience and Remote Sensing Letters* **2010**, *7*, 333–337.

14. Hannachi, A.; Jolliffe, I.T.; Stephenson, D.B. Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology* **2007**, *27*, 1119–1152.

15. Golub, G.H.; Reinsch, C. Singular value decomposition and least squares solutions. *Numerische Mathematik* **1970**, *14*, 403–420.

16. Björck, Å.; Golub, G.H. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation* **1973**, *27*, 579–594.

17. Chan, T.F. An improved algorithm for computing the svd. *ACM Trans. Math. Softw.* **1982**, pp. 72–83.

18. Martinsson, P.G. Randomized methods for matrix computations and analysis of high dimensional data. *ArXiv* **2016**.

19. Eckart, C.; Young, G. The approximation of one matrix by another of lower rank. *Psychometrika* **1936**, pp. 211–218.

20. Moran, P.A.P. Notes on continuous stochastic phenomena. *Biometrika* **1950**, *37*, 17–23.

21. Hubert, L.J.; Golledge, R.G.; Costanzo, C.M. Generalized procedures for evaluating spatial autocorrelation. *Geographical Analysis* **1981**, *13*, 224–233.

22. Rey, S. PySAL. http://pysal.readthedocs.io, 2009–2013.

23. Bogaardt, L. Dataset reduction techniques to speed up svd analyses. https://github.com/phenology/, 2018.

24. Halko, N.; Martinsson, P.G.; Tropp, J.A. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review* **2011**, *53*, 217–288.

25. Li, H.; Kluger, Y.; Tygert, M. Randomized algorithms for distributed computation of principal component analysis and singular value decomposition. *CoRR* **2016**, *abs/1612.08709*.

26. Dee, D.P.; Uppala, S.M.; Simmons, A.J.; Berrisford, P.; Poli, P.; Kobayashi, S.; Andrae, U.; Balmaseda, M.A.; Balsamo, G.; Bauer, P.; Bechtold, P.; Beljaars, A.C.M.; van de Berg, L.; Bidlot, J.; Bormann, N.; Delsol, C.; Dragani, R.; Fuentes, M.; Geer, A.J.; Haimberger, L.; Healy, S.B.; Hersbach, H.; Holm, E.V.; Isaksen, L.; Kallberg, P.; Köhler, M.; Matricardi, M.; McNally, A.P.; Monge-Sanz, B.M.; Morcrette, J.J.; Park, B.K.; Peubey, C.; de Rosnay, P.; Tavolato, C.; Thepaut, J.N.; Vitart, F. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society* **2011**, *137*, 553–597.

27. Bretherton, C.S.; Smith, C.; Wallace, J.M. An intercomparison of methods for finding coupled patterns in climate data. *Journal of Climate* **1992**, *5*, 541–560.

28. Tygert, M. Suggested during personal communication, 2017.

29. Schwartz, M.D.; Ault, T.R.; Betancourt, J.L. Spring onset variations and trends in the continental united states: past and regional assessment using temperature-based indices. *International Journal of Climatology* **2013**, pp. 2917–2922.

30. Izquierdo-Verdiguier, E.; Zurita-Milla, R.; Ault, T.R.; Schwartz, M.D. Using cloud computing to study trends and patterns in the extended spring indices. *Third International Conference on Phenology* **2015**, p. 51.

31. Zurita-Milla, R.; Bogaardt, L.; Izquierdo-Verdiguier, E.; Gonçalves, R. Analyzing the cross-correlation between the extended spring indices and the AVHRR start of season phenometric. *EGU General Assembly: Geophysical Research Abstracts* **2018**.

32. Barnett, T.P.; Preisendorfer, R. Origins and levels of monthly and seasonal forecast skill for us surface air temperatures determined by canonical correlation analysis. *Monthly Weather Review* **1987**, *115*, 1825–1850.

33. Kobayashi, S.; Ota, Y.; Harada, Y.; Ebita, A.; Moriya, M.; Onoda, H.; Onogi, K.; Kamahori, H.; Kobayashi, C.; Endo, H.; Miyaoka, K.; Takahashi, K. The JRA-55 reanalysis: general specifications and basic characteristics. *Journal of the Meteorological Society of Japan* **2015**, *93*, 5–48.

34. Liu, Y.; Attema, J.; Moat, B.; Hazeleger, W. Synthesis and evaluation of historical meridional heat transport from midlatitudes towards the arctic. *Climate Dynamics* **2018**, *Submitted*.

35. Tucker, L.R. The extension of factor analysis to three-dimensional matrices. In *Contributions to mathematical psychology*; Gulliksen, H.; Frederiksen, N., Eds.; Holt, Rinehart and Winston, 1964; pp. 110–127.