An extended approach for spatiotemporal gapfilling: dealing with large and systematic gaps in geoscientific datasets

Author(s):
von Buttlar, Jannis; Zscheischler, Jakob (iD); Mahecha, Miguel D.

Nonlinear Processes
in Geophysics

Open Access

An extended approach for spatiotemporal gapfilling: dealing with large and systematic gaps in geoscientific datasets

J. v. Buttlar[1], **J. Zscheischler**[1,2,3], **and M. D. Mahecha**[1]

[1]Max Planck Institute for Biogeochemistry, P.O. Box 100164, 07701 Jena, Germany
[2]Max Planck Institute for Intelligent Systems, P.O. Box 2169, 72012 Tübingen, Germany
[3]ETH Zürich, Rämistraße 101, 8092 Zürich, Switzerland

Correspondence to: J. v. Buttlar (jbuttlar@bgc-jena.mpg.de)

Abstract. Spatiotemporal observations in Earth System sciences are often affected by numerous and/or systematically distributed gaps. This data fragmentation is inherited from instrument failures, sparse measurement protocols, or unfavourable conditions (e.g. clouds or vegetation thickness in case of remote-sensing data). Missing values are problematic as they may cause analytic biases and often inhibit advanced statistical analyses. Hence, gapfilling is an undesired but necessary task in Earth System sciences. State-of-the-art gapfilling algorithms based on Singular Spectrum Analysis (SSA) exploit the information contained in periodic temporal patterns to fill gaps in the observations. Here we propose an extension of this method in order to additionally consider the spatial processes and patterns underlying most geoscientific datasets. The latter has been made possible by including a recently developed 2-D-SSA approach. Using both artificial and real-world test data, we show that simultaneously exploiting spatial and temporal patterns improves the gapfilling substantially. We outperform conventional approaches particularly for large and systematically recurring gaps. The new method is reasonably fast and can be applied with a minimum of a priori assumptions regarding the structure of the data and the distribution of gaps. The algorithm is available as a ready-to-use open source software package.

1 Introduction

The global monitoring of the atmosphere, the land surface, and oceans via in situ measurements and remote sensing has opened unprecedented opportunities for studying various aspects of the functioning of the Earth System (Overpeck et al.,

2011; Reichstein et al., 2013). However, several analyses in Earth System sciences depend on gap-free data: In the simplest case, the estimation of unbiased annual sums and budgets requires reliable gapfilling techniques if the fragmentation does not happen at random (Falge et al., 2001). Also many advanced analyses, for instance exploratory statistical or machine learning approaches (Mjolsness and DeCoste, 2001), as well as as spectral (time series) analysis (Ghil et al., 2002), generally need gap-free and evenly sampled data. Process-oriented modeling approaches depend on continuous observations as drivers for predicting system properties. Likewise, model benchmarking is limited by missing data (Luo et al., 2012).

In reality this need for continuous data is often not fulfilled. Instrumental failures or unfavourable measurement conditions (e.g. cloud cover, aerosols, or complex surface properties in the context of remote sensing) cause gaps in both in-situ or remote-sensing data. Examples of such datasets are most remotely sensed land surface properties including Soil Moisture (SM; Liu et al., 2011), Leaf Area Index (LAI), Normalised Difference Vegetation Index (NDVI; Huete et al., 2002), or Land Surface Temperature (LST; Justice et al., 1998). Hence, filling missing data points by empirical estimates is a generally undesired but often crucial step to tap the full information in a data rich world.

Several methods have been proposed that exploit multivariate empirical relationships between the variable of interest and other variables available at gap positions (Moffat et al., 2007). A more import argument for not including ancillary observations is that the independence amongst data sources should be maintained. Otherwise, any subsequent multivariate analysis investigating relationships between the

variable and the ancillary data will be spurious. The following brief literature overview focuses explicitly on univariate methods.

A widely used method for gapfilling is expectation maximisation (EM; cf. Dempster et al., 1977; Schneider, 2001, for specific extensions to climate data) where the mean and covariance of a dataset are iteratively estimated and used to predict missing values. A classical parametric set of univariate gapfilling methods is based on optimal interpolation (OI, cf. Reynolds and Smith, 1994; Smith et al., 1996; Kaplan et al., 1997) which uses interpolations form "optimal" periods to replace gaps. However, these methods require a priori assumptions about the covariance structure of the data and the structure of the gaps. Beckers and Rixen (2003) present a method where missing values are estimated via Empirical Orthogonal Functions (EOFs) and the initially filled dataset is iteratively used to update the EOF estimation. Their approach can capture spatial patterns but largely ignores temporal correlations. An alternative approach is to interpolate missing data in all available dimensions. Examples are presented by Garcia (2010) and Wang et al. (2012) .

Another alternative is offered by the 1-D temporal gapfilling approaches relying on Singular Spectrum Analysis (SSA; Broomhead and King, 1986; Vautard and Ghil, 1989), where the first concept was presented by Schoellhamer (2001). Golyandina and Osipov (2007) modify the classical SSA algorithm and estimate the SSA components based on nonmissing data only. The values of the reconstructions are then imputed to the missing values. Kondrashov and Ghil (2006) generalise the EOF-based iterative procedure of Beckers and Rixen (2003) and propose a method that fills gaps using either univariate or multi-channel SSA (M-SSA). For univariate time series, Kondrashov and Ghil (2006) exploit the periodic and non-periodic temporal structures (for instance the annual cycle and trend) of a given dataset to fill the gaps. In this setting, spatial information can be partly used with the help of M-SSA based on all grid locations. Note that there are also other methods exploiting the periodic structure of the signal, like the HANTS algorithm using Fourier decomposition (Roerink et al., 2000) and the approach of Hocke and Kaempfer (2009) based on the Lomb Scargle periodogram, which operate in the 1-D temporal domain.

All state-of-the-art SSA-based methods are strongly biased by periodic and/or long continuous gaps. This is particular problematic for remote-sensing products that are often affected by seasonal occurrences of unfavorable conditions (examples are winter snow cover or seasonal cloud distributions). For example, Musial et al. (2011) compared the Kondrashov and Ghil (2006) method to splines and the approach by Hocke and Kaempfer (2009) and found generally good gap predictions with SSA – with the exception of artifacts in the presence of periodic winter gaps. The cause for these artifacts is the initial filling of gaps with a mean value as input for the SSA runs (see Sect. 2.3 for details). Periodic gaps then tend to produce spurious periodic patterns that persist throughout the iteration process. Gaps longer than the period of the oscillation used by SSA can similarly not be filled as they directly influence and bias the shape of the reconstructions of these oscillations.

In this paper, we propose to extend the iterative Kondrashov and Ghil (2006) approach in order to explicitly capture and exploit the spatial information from geo-datasets. The spatial patterns can be either used as unbiased first guess (see Sect. 2.3) for one-dimensional temporal SSA under difficult conditions (e.g. periodic or long gaps) or as full alternative for temporal SSA. The extraction of spatial patterns is facilitated by a recently developed 2-D-SSA variant (Golyandina and Usevich, 2009). The 2-D-SSA method was developed to decompose spatial data, e.g. orographic maps, into a set of overlaid spatial patterns of different detail. Emphasising the spatial auto-correlation structure is of paramount importance in remote sensing data, where regional anisotropic features need to be maintained. The integration of the 2-D-SSA method into the gapfilling scheme by Kondrashov and Ghil (2006) is also important to allow the processing of high resolution spatiotemporal data as they are currently made available to the scientific community.

2　Methods

2.1　Singular Spectrum Analysis (SSA)

In general terms, the SSA algorithm decomposes a signal into a set of superimposed (i.e. additive) independent sub-signals. In the context of Earth observation time series typical sub-signals are diurnal and annual cycles. Also less regular processes for instance ENSO patterns or long-term trends can be extracted. Likewise, short-term stochastic variability may play a role. Analogously, in the 2-D case, the overlaid patterns represent spatial patterns at different scales (Golyandina and Usevich, 2009). SSA has some advantages compared to other more popular spectral methods (Ghil et al., 2002) such as Fourier decomposition: in particular, SSA can extract phase-amplitude modulated oscillations from relatively short and noisy signals (Golyandina and Zhigljavsky, 2013).

In the following, we will briefly discuss the different steps of SSA and the relevant parametric choices. An in-depth mathematical description is given in the appendix. The process of SSA basically consists of four subsequent steps. First, the time series or spatial field is *embedded*, i.e. a moving window is shifted along the time series (or spatial matrix) and the vectors (or blocks) inside this window are arranged to form a trajectory matrix. Second, this matrix is subject to Singular Value Decomposition (SVD) which yields a set of *eigentriples* (i.e. individual SSA components, see the Appendix A for details). These eigentriples represent all individual and statistically independent (i.e. orthogonal) sub-signals. Third, the eigentriples have to be *grouped* as some sub-signals are represented by a set of complementary

Table 1. Overview of the parametric choices used for the test runs and their respective argument names in the GNU-R code.

parameter	R argument name	value
window length (1-D/2-D)	M	$45/20 \times 20$
iterations (outer/inner)	max.steps, amnt.iters	10/10
eigentriples extracted	n. comp	20
amount of artificial gaps	amnt.artgaps, gaps.cv	10 %

eigentriples. Examples are oscillatory patterns that are represented by two (for sinusoidal signals) or more eigentriples (for non-sinusoidal patterns including harmonics). Fourth, in a *reconstruction* step each individual group of eigentriples is transformed back into the temporal (or spatial) dimension of the original series (or matrix).

2.2 Parametric choices

The analyst has to make two parametric choices for SSA (Ghil et al., 2002): the *window length* (or *embedding dimension*) L (or L_x and L_y in the 2-D case) and the grouping of the eigentriples. Due to the symmetry of the SVD, $L \leq N/2$ (with $N =$ length of time series; Golyandina and Zhigljavsky, 2013). Increasing L generally improves the separability between different independent signals which is especially important for short time series as in our case. For extracting oscillatory signals of period P, L should be an integer multiple of P, i.e. $L = n \cdot P$ (Golyandina and Zhigljavsky, 2013). As the annual cycle of our remote-sensing test datasets (c.f. Sect. 2.4) with a sampling interval of 16 days has a period $P \approx 23$ we set $L = 45$ for all 1-D SSA runs. Comparable recommendations for 2-D SSA have not (yet) been developed. Preliminary tests yielded better results for small L_x and L_y, so we chose a value of 20×20 for our experiments. The user should bear in mind that this window size influences the shape of the SSA reconstructions (Golyandina and Zhigljavsky, 2013). Additionally its optimal size yielding the best gap filling results depends on the size or amount and structure of the gaps present. An optimum value can be obtained by cross validation with artificial gaps as done by (Kondrashov and Ghil, 2006). In our test framework, however, this gap amount and structure was varied systematically. To ensure comparability between the different test cases, we used the fixed value of 20×20 consistently throughout our experiments.

The grouping is usually done manually via a visual inspection of the shape of the eigentriples and the spectrum of their variance (i.e. the "Scree diagram"; Ghil et al., 2002; Golyandina and Zhigljavsky, 2013). This is not possible with the high amount of independent SSA runs in our case. Hence, we use an automated method provided by the *Rssa* GNU-R package (Golyandina and Korobeynikov, 2013; Golyandina and Zhigljavsky, 2013). Basically it identifies the groups via a complete-linkage hierarchical clustering algorithm based
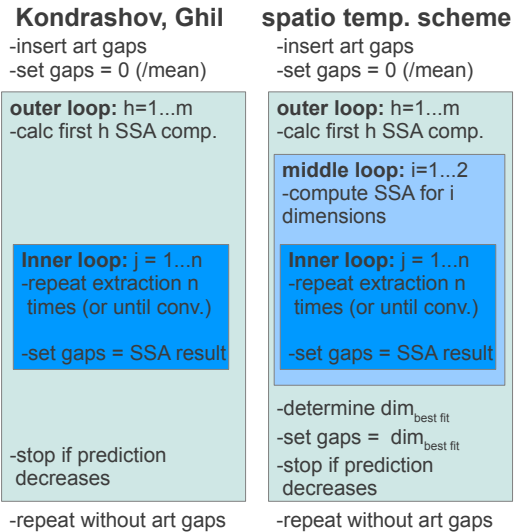
Fig. 1. Comparison between the concepts of the Kondrashov and Ghil (2006) gap filling scheme and the proposed spatiotemporal scheme.

on the so called w-correlation matrix. This w-correlation matrix contains the weighted correlations between the individual reconstructed SSA components.

Table 1 provides an overview of the parametric choices and the corresponding parameters in the GNU-R function in the package *spectral.methods*.

2.3 Spatiotemporal scheme

With the proposed spatiotemporal gapfilling scheme we follow the conceptual idea of Kondrashov and Ghil (2006, see Fig. 1). Their algorithm identifies independent (temporal) sub-components or regular patterns and uses them sequentially to interpolate missing values. As a first step, all gaps are filled with the mean of the series (as SSA itself can not handle missing values). Subsequently, the spectral SSA component with the highest variance is computed and its values are inserted to the gap positions. The process is iterated in an *inner loop* to minimise the effect of the previously inserted mean values. In the following *outer loop* iteration steps, additional spectral components of lower variance are computed and processed in the same manner. Cross-validation can be used to identify the optimal number of outer (and inner) loop steps. This is done by the insertion of additional artificial gaps and comparing predictions at these locations to the original data.

We generalise this idea of independently extracting specific spectral components by treating spatial components in the same manner. At each outer loop step n we compute both, the n 1-D SSA components and the n 2-D SSA components with the highest variance (referred to as *middle loop* in
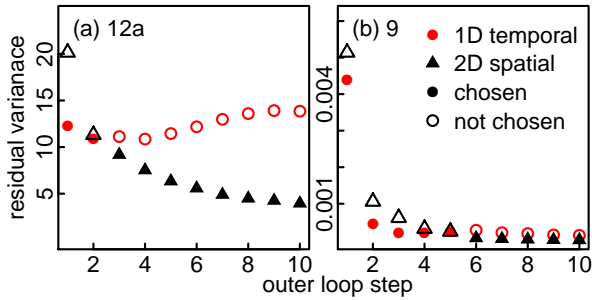
Fig. 2. Demonstration of the step-wise cross-validation scheme. The algorithm compares the variance of the residua at the artificial gap positions and chooses the dimension setting with the lowest residual variance. Shown are, as examples, **(a)** scenario 12a and **(b)** scenario 9 with 55 % and, respectively, 20 % gaps.

Fig. 1). For both settings we compute the residual variance:

$$\text{Var}_{\text{resid}} = \frac{1}{n} \sum_{i=1}^{n} (R_i - \bar{R})^2 \qquad (1)$$

with the residuum $R_i = P_i - O_i$, O_i being the prediction (i.e. gapfilled value), and P_i being the observation (i.e. original data) at additional artificial gap locations. The dimensional setting with the lower $\text{Var}_{\text{resid}}$ is expected to yield the better predictions, and its results are used as the *first guess* (i.e. inserted into the gap positions) in the next outer loop step (see Fig. 2). Here again, both SSA methods are used to extract the first $n+1$ components and so forth. The Kondrashov and Ghil (2006) inner iteration loop is performed alike for each outer loop step and dimension independently. After a maximum of 10 outer loop steps, $\text{Var}_{\text{resid}}$ is also used to determine that step with the overall best prediction. This is not necessarily the last step in cases of decreasing prediction capabilities (e.g. due to overfitting). Finally, the process is repeated without artificial gaps, and the outer loop path with the lowest $\text{Var}_{\text{resid}}$ from the cross-validation is followed with all available data.

The principal setting of the algorithm allows for the use of any possible number and combination of 1-D or 2-D cuts through a 3-D datacube. This would, for example, also allow for a two-dimensional time × latitude SSA computation. In this paper, however, we only considered 1-D temporal and 2-D spatial (i.e. latitude × longitude) SSA. In cases where the cross-validation chooses a dimension with which it is not possible to fill all gaps (i.e. 2-D SSA is chosen, but whole time slices $\mathbf{X}_t$ are missing), the algorithm fills these gaps with the respective other, non-chosen, dimension. Additional features include the possibility to run the cross-validation only on parts of the data (in our test runs 20 %) to increase speed, the padding of the input time series to reduce edge effects, and the possibility to supply an ocean mask which will not be filled. To reduce computational cost, the algorithm uses an optimised SSA routine that only computes the major SSA eigentriples by truncating the computationally
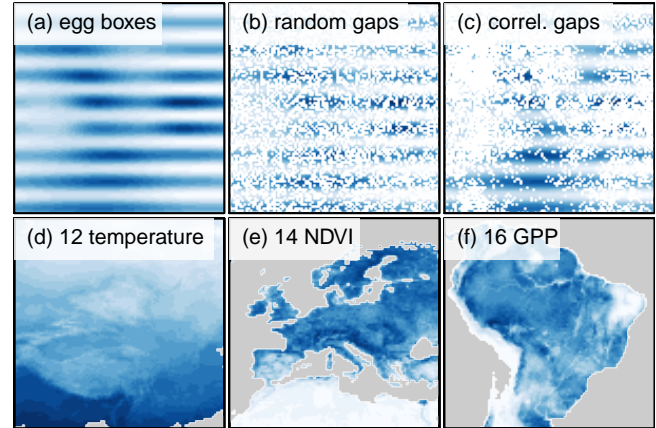
Fig. 3. Datasets and gap scenarios used for testing the gapfilling method. Artificial data (egg boxes, panels **a–c**) and real-world datasets (**d–e**). Panel **(b)** and **(c)** show 50 % of gaps for random gaps (**b**, scenario 1 and 2) and spatiotemporally correlated gaps (**c**, scenario 3 and 4). Scenario 3 gap positions were also used for the real-world test datasets. The ocean masks shown in (**d**)–(**f**) were also used for each individual real-world dataset and the egg boxes sets containing oceans (scenario 2a-c and 4a-c). Shadings of blue denote the actual value with light blue for the minimum and dark blue for the maximum of the respective variable range. See also Fig. 4 for a visualisation of the temporal structure of the data.

costly SVD step (Golyandina and Korobeynikov, 2013). For all our test runs, 20 individual eigentriples were computed. The whole algorithm has been programmed in GNU-R (R Development Core Team, 2013) and is available accompanying this paper as the package *spectral.methods* on R-Forge (https://r-forge.r-project.org/projects/jbtools/).

2.4 Test datasets

We used four different types of test datasets to evaluate the gap filling performance of the spatiotemporal scheme (cf. Fig. 3).

The first dataset was artificial and constructed by a superposition of multiple sine waves (cf. Fig. 3, top panels). The period of this sine in the y direction (i.e. latitude) of the datacube was ≈ 16, and 50 in the x direction (i.e. longitude). This pattern was multiplied with a combination of harmonic sines that mimic a yearly cycle along the third (i.e. time) dimension of the datacube (Fig. 4, top panel). Due to their similar structure along the spatial dimensions these datasets are here referred to as "egg boxes". To study the effects of not-to-fill ocean gaps, real-world coastlines were imposed on the egg boxes datasets (the real-world datasets already contain differently shaped oceans).

The second suite of datasets consisted of a selection of different real-world datasets. We used cutouts of air temperature (ERA 40 reanalysis, Weedon et al., 2011), remotely sensed NDVI (Huete et al., 2002), and empirically upscaled gross

Table 2. Overview of datasets and gap scenarios used for testing the gapfilling method.

	code	data/gap type		code	data/gap type
egg boxes	1/2	random gaps	real-world	12	temperature
	3/4	correlated gaps		14	NDVI
	5/6	missing series		16	GPP
	7/8	missing t.steps			
	9/10	mix of 5–8	cutout	a	Asia
nr.	odd	no ocean		b	Europe
	even	ocean		c	Sth.America

Fig. 4. Visualisation of the temporal patterns in the test datasets. All time series are plotted, and the colour code shows the densities of points per pixel (yellow = 1, blue = 1500).

primary productivity (GPP; Beer et al., 2010). All datasets had a 0.5° spatial resolution and cover different geographical regions.

To assess the method's prediction capabilities in different climatic regimes, cutouts from the global data covering Central Asia (mainly China and Northern India), Europe, and South America were used. A global run of the algorithm would not have been computationally feasible due to the huge amount of datasets tested (cf. Sect. 4.3 for a discussion). All test datasets had a size of $100 \times 100 \times 100$ (longitude $\times$ latitude $\times$ time) and contained about 4 yr of data with a 16 day temporal resolution. See Fig. 4 for a visualisation of the temporal patterns in these datasets.

We inserted artificial gaps of different structure and amount into these originally (nearly) gap-free datasets to compare the prediction performance of the different methods (cf. Table 2). We used five different gap scenarios: randomly distributed gaps (scenario 1/2), gaps that are spatially and temporally correlated or clustered (scenario 3/4), a scenario with time series completely missing (scenario 5/6), a scenario with time steps missing completely (scenario 7/8), and a mix of these gap structures (scenario 9/10). Even scenario numbers denote data with ocean, odd numbers ocean-free scenarios. For each of these scenarios different datasets were created with a gap ratio (i.e. the percentage of missing values) varying between 5 % and up to 70 %. With 4 different datasets per scenario (one without ocean and three with different land masks) and 5 gap scenarios we obtained 435 test datasets in total. We used the full set of gap scenarios only for the egg boxes data and restricted the real-world datasets to spatially and temporally correlated gaps (i.e. identical gap locations to the scenarios 3 and 4).
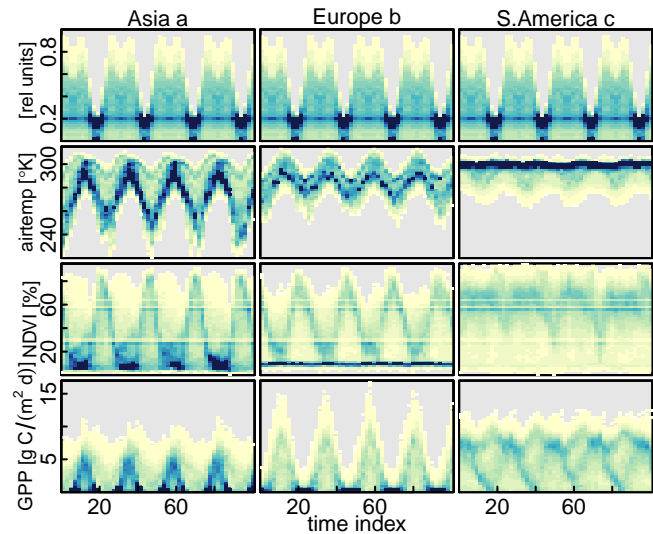
2.5 Performance measures

We quantified the prediction performance for each gapfilling run via the modelling efficiency MEF (Janssen and Heuberger, 1995):

$$\text{MEF} = 1 - \frac{\sum_{i=1}^{n}(P(i) - O(i))^2}{\sum_{i=1}^{n}(O(i) - \bar{O})^2}. \tag{2}$$

$\bar{O}$ is the empirical mean over the original values of all gaps. $O(i)$ and $P(i)$ are the original (i.e. **o**bserved) and filled (i.e. **p**redicted) values of gap i, respectively. A MEF value of 1 would be perfect agreement, and a value of zero would indicate a prediction comparable to simply inserting the mean of all (not-filled) values into gap positions. As 1-D temporal SSA was not able to fill all gaps (see Sect. 3.2 for details), we calculated MEF values for different subsets of the data:

– $\text{MEF}_{\text{1-Dfill}}$: for all data points that could be filled with 1-D SSA (allowing a direct comparison between 1-D and 2-D spatiotemporal SSA)

– MEF_{tot}: for all data points that could be filled with the respective method (meaning all for spatiotemporal SSA and a value identical to $\text{MEF}_{\text{1-Dfill}}$ for 1-D SSA)

– $\text{MEF}_{\text{Mfill}}$: for all gap positions with remaining gaps filled with the mean of the gappy dataset. In case a totally gap free dataset is required, using the mean is the best available guess to fill the gaps not filled by temporal SSA. This measure, hence, compares the predictions in such a usage scenario.

3   Results

The goal of our experiments with artificial (egg boxes) and real-world test datasets was to compare the gapfilling scheme using spatiotemporal SSA with state-of-the-art gapfilling scheme using only temporal 1-D SSA. Additionally we wanted to explicitly identify situations (i.e. gap scenarios and gap ratios) under which one of the two methods outperforms the other. Scenarios 5–10 were mainly developed and filled to test the algorithm's treatment of empty series and time steps. They showed results comparable to those described below and will not be discussed here in detail.

3.1   Choice of the temporal vs. the spatial scheme

In general, the 2-D spatial SSA filling scheme was chosen due to the Var_{resid} criterium (cf. Sect. 2.3) as a first guess in around 25 % of all cases (i.e. including all individual outer loop steps) for each filling process (cf. Fig. 5). This percentage was remarkably higher for the first outer loop step. It dropped sharply for the subsequent steps but rose monotonously again up to 40 % at the last (i.e. 10th step) (cf. Fig. 5a). Overall, the usage of the spatial 2-D scheme increased with increasing gap ratio. The type of gap scenario (i.e. random vs. spatially and temporally correlated gaps) had no strong influence on the choice of dimension for filling.

The choice of the SSA dimension used for filling, however, changed drastically between the different data types and also between the geographic locations of the cutouts (cf. Fig. 5c). For scenario 12a and b (air temperature; Asia and Europe), nearly all steps were filled with spatial 2-D SSA. The same was true, to a lesser extent (up to 50 %), for scenario 14c (NDVI, South America). For most other real-world scenarios, 1-D SSA provided better results and was chosen as the filling method in the majority of cases. For filling the egg boxes, 2-D spatial SSA was used in roughly 45 % of the cases in the ocean-less scenarios 1 and 3 and for the China cutout which contained very few ocean and a nearly continuous land mass (ocean a). For all other egg boxes, 1-D SSA was used in nearly all cases.

3.2   Prediction performance

Overall, the modelling efficiency for all filled data points compared to the original data (MEF_{tot}, cf. Sect. 2.5) was relatively high and well above 0.9 for most tests up to gap ratios of 60 % (cf. Figs. 6 and 7). MEF_{tot} of the 1-D temporal and the 2-D spatiotemporal scheme did not differ for most test datasets for gap ratios below 40–50 % (for correlated gaps and most real-world tests) or even 60 % (scenario 1/2: random gaps, egg-boxes). For most gap ratios higher than this value, spatiotemporal SSA yielded better predictions (i.e. higher MEF_{tot}) than 1-D temporal SSA. Remarkably, MEF_{tot} for spatiotemporal SSA was higher for all gap ratios from 5 to 75 % for scenarios 12a, 12b, and 14c. On the
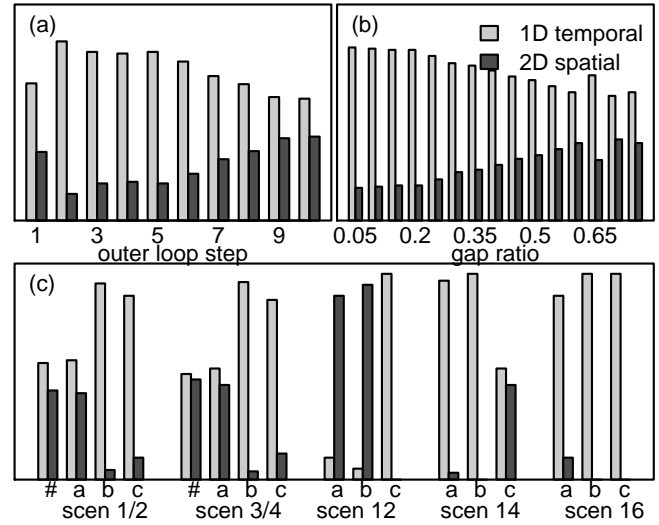
Fig. 5. Dimensions chosen (i.e. temporal 1-D vs. spatial 2-D) for all test datasets and all steps plotted for **(a)** the dimension choosing outer loop steps, **(b)** the ratio of missing values, and **(c)** the different gap and data type scenarios. # denotes the gap free egg-boxes of scenarios 1 and 3.

contrary, temporal 1-D SSA yielded better results and higher MEF_{tot} for the high gap ratio regimes for scenarios 12c, 14b, and 16 b and c.

Temporal 1-D SSA could not fill all gap values as it can not be used to reliably extrapolate the signal into continuous gaps at the beginning and end of a time series. The amount of these margin gaps increased with increasing gap ratios and was higher with correlated gaps (scenario 3). This resulted in up to 60 % of the gaps not being filled by temporal 1-D SSA for a gap ratio of 70 % (see Fig. 8). Comparing MEF_{tot} for 1-D and 2-D SSA, hence, yields a bias as it refers to different amounts of data points. The prediction performance at data points both methods were able to fill ($MEF_{1-Dfill}$, cf. Sect. 2.5) was similar for 1-D and 2-D SSA for gap ratios of up to 50 % but higher for 2-D spatiotemporal SSA above this value for nearly all scenarios. For MEF_{Mfill} (cf. Sect. 2.5), this aspect was even more pronounced.

Scenario 14c (NDVI; South America) showed an exceptionally different behaviour with a much lower MEF_{tot} ($\approx$ 0.8) even for very low gap ratios for both methods. This stayed relatively constant with an increasing amount of gaps for the spatiotemporal scheme and droped down nearly linearly to 0.4 for 1-D temporal SSA.

3.3   Step-wise development

An in-depth investigation of the prediction capabilities (quantified by Var_{resid}, cf. Sect. 2.3) of the different SSA methods for each outer loop step revealed a different behaviour for each scenario and test dataset (cf. Fig. 9). The
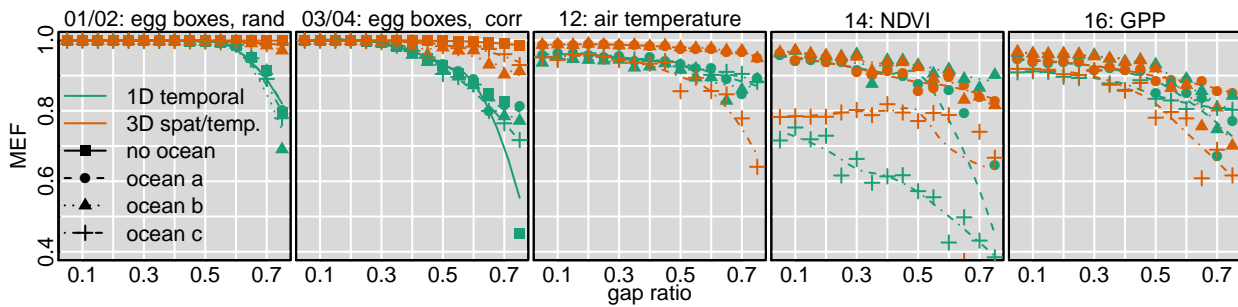
Fig. 6. Prediction performance of the 2-D spatiotemporal gapfilling scheme in comparison to the state-of-the-art (1-D temporal) SSA method. Solid or dashed lines were obtained using a loess smoother for each scenario and dataset.
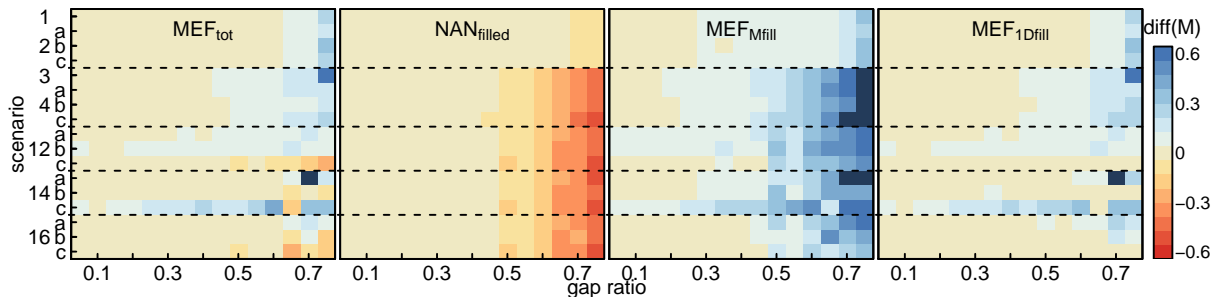
Fig. 7. Prediction performance of the spatiotemporal gapfilling scheme in comparison to state-of-the-art (1-D temporal) SSA methods for the different scenarios (y axis) as a function of the gap ratio (x axis). Shown is the difference (1-D temporal – 2-D spatiotemporal) for the different MEF measures (cf. Sect. 2.5). Red colours of increasing intensity indicate higher values for the temporal, and blue colours indicate higher values for the spatiotemporal scheme.
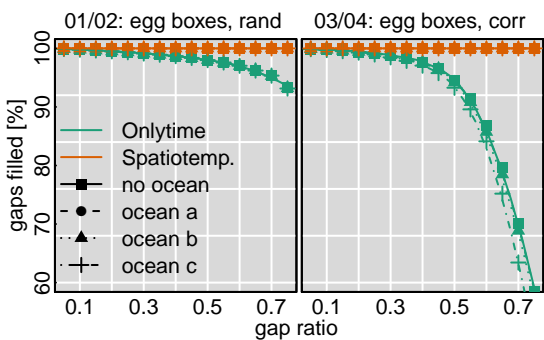
Fig. 8. Percentage of non-filled gaps for the different test scenarios. Solid or dashed lines were obtained using a loess smoother for each scenario and dataset. Comparable for scenario 12–16 would be identical to the "correlated" gaps figure as they have identical gap locations.

Var_{resid} of 1-D temporal SSA dropped quickly to a minimum after only 3 steps and remained constant thereafter for all egg boxes scenarios (only 1 and 2b are shown). For most real-world datasets this optimum (i.e. minimum) was reached a few steps later, and the prediction quality decreased with subsequent outer loop steps in most cases. For identical scenar-

ios Var_{resid} was generally higher with higher gap ratios but showed the same general pattern as a function of gap ratio.

The residual variance for the 2-D spatial SSA calculations decreased slower and more gradually down to its optimum for ocean-free egg boxes datasets (scenario 1 is shown as an example) and most real-world data. For most egg boxes data containing ocean (e.g. 2b), Var_{resid} increased with additional outer loop steps. For some scenarios (12a, 12b, and 14c) spatial SSA resulted in lower Var_{resid} during the final steps than 1-D temporal SSA. For some scenarios (all ocean egg boxes, 14b, 16b, and 16c) 1-D temporal SSA yielded better predictions for these steps. In many other cases (1, 2c, 3, 4c, 12c, 14c, and 16a), however, the results from both methods were similar at the final outer loop steps.

4 Discussion

In general, the presented results show that the spatial 2-D SSA can extract additional information and provides better results than temporal 1-D SSA in the cases where the analyst is confronted with highly fragmented datasets. However, there are several conceptual and methodological aspects that could be further optimised and extended for better results or performance.
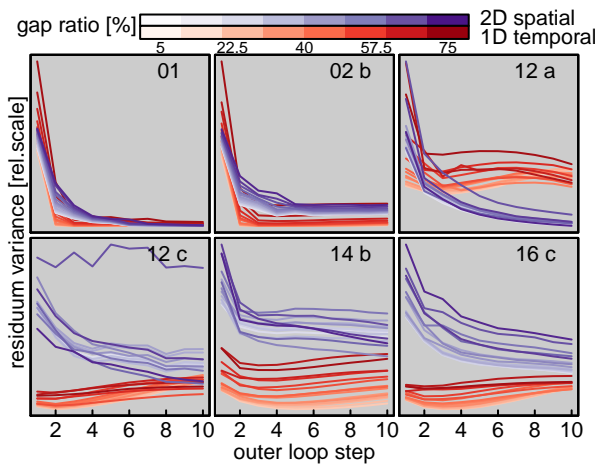
Fig. 9. Comparison of the residual variance of the spatial 2-D vs. the temporal 1-D SSA reconstructions for selected test datasets as a function of the outer loop step. Test datasets not visualised here show similar patterns as the example datasets plotted.

4.1   General performance and behavior

It is not surprising that the spatiotemporal scheme performs at least equally well as state-of-the-art temporal SSA in most cases. The algorithm is designed to use temporal 1-D SSA as one of two dimensional choices and will use it for gapfilling if spatial SSA performs worse. In the theoretical case where 1-D SSA is chosen in each outer loop step, the only difference between 1-D SSA and the spatiotemporal scheme is that for the latter the same optimum outer loop step is chosen for all time series whereas 1-D SSA allows for a different choice for each individual series. However, this algorithmic difference is expected be of marginal importance as demonstrated by the very small performance differences in such cases in our experiments.

As temporal 1-D SSA was chosen in the majority of cases, the interesting results are those cases where using the spatiotemporal scheme actually yields better results than temporal SSA. In the general picture this was the case mostly for gap scenarios with gap ratios above 50 %. In addition, the spatiotemporal algorithm uses the inferior dimension combination to fill remaining gaps not filled by the superior dimension. In our experiments this meant that for all test datasets filled with 1-D SSA, 2-D SSA was still used to fill gaps at margin locations or totally empty time series, yielding completely filled datasets. These advantages build a strong argument to use both the spatial and the temporal information available in geoscientific datasets for gapfilling, in particular at large gap ratios.

The choice between 1-D and 2-D SSA differed strongly between the type of data to be filled and also the geographical location of the cutouts. However, it is difficult to identify a general pattern. The remarkably bad performance of 1-D SSA (and hence the big difference to 2-D SSA) for scenario

14c can be explained by the NDVI characteristics of the tropical rainforest. Due to the tropical climate most of the signals do not show a strong seasonal (i.e. temporal) signal that can be used to fill the gaps (c.f. Fig. 4). This leads to a low signal/noise ratio (SNR) and a poor prediction capacity. Apparently, the SNR in the spatial dimension is lower and the spatial SSA is is still able to capture certain spatial patterns.

Compared to NDVI and GPP, whose spatial patterns reflect the patchiness of the vegetation cover, the spatial gradients in air temperature are much smoother. 2-D SSA is able to capture these smooth gradients and is chosen in most cases for filling scenario 12a and b. Interestingly, this is the opposite for the tropical cutout 12c. A reason for this may be the relatively weak spatial differences over the rain forests compared to the steep and locally confined gradients in the Andean mountains (cf. Fig. 3). Spatial SSA seems to have difficulties capturing such small-scale patterns. For most vegetation related real-world test cases (except 14c), the temporal patterns seem to be more pronounced compared to the rather noisy spatial patterns, so that 1-D SSA is mostly used to fill the gaps.

The conceptual idea of the iteration scheme of the spatiotemporal gapfilling algorithm is to separate temporal and spatial patterns of different scales and to use these patterns step wise for gap predictions. If such clearly separable patterns exist in the data, they would be visible, for example, in drops and steps (i.e. "elbows") in the residual variance between different outer loop steps. Such an elbow is clearly visible for the temporal domain, where the dominant annual cycle (and not many more patterns) are captured by SSA. For spatial 2-D SSA, however, the decrease of $\mathrm{Var_{resid}}$ is much slower and more gradual. This most probably is caused by the specific structure of spatial patterns in the (bio)geoscientific data which can not be separated into several distinct components in such a straightforward manner. Additionally, for any SSA reconstruction, several of the individual eigentriples have to be grouped to yield one sub-signal. For the annual cycle, for example, two eigentriples have to be summed. For spatial SSA, without clear separations between these ungrouped eigentriples (i.e. via marked differences in their variance), such a grouping scheme is much less developed and more difficult. A strict implementation of the conceptual idea would hence require a more in depth methodological and theoretical development of the grouping and separation of 2-D SSA eigentriples with respect to the special structure of geoscientific data.

4.2   Advantages compared to M-SSA

The study by Kondrashov and Ghil (2006) shows that multichannel SSA can likewise be used to ensure a lateral (i.e. spatial) transfer of correlation structures. In particular, this approach allows for a "flow of information" from other grid cells to more fragmented time series. There are, however, some conceptual concerns that need to be discussed. The

M-SSA approach is likely to fail when individual time series contain very little or no actual values. The extended approach presented in our study, however, even allows filling time series that are fully missing. In addition, M-SSA only incorporates bivariate correlations between time series whereas our approach can exploit true 2-D structures inherent in the data.

Both methods, however, have their strengths and weaknesses and their application depends strongly on the temporal and spatial patterns present in the data to be filled. While M-SSA is faster, the spatiotemporal scheme will produce better results for high gap ratios or situations where complete time series are missing or time periodic gaps occur.

4.3 Limitations and open issues

Gapfilling in particular and other interpolation schemes in general have the danger to feign the existence of knowledge or information that can factually not be extracted from the data. For example, many remote sensing datasets of vegetation indices, show continuous gaps for tropical evergreen forest areas due to prevalent cloud cover (cf. e.g. Musial et al., 2011). The spatial scheme interpolating into these areas can use no factual knowledge and, hence, "invents" data here which would also influence gap predictions in adjacent areas. The test data used here did not show such scenarios, but care has to be taken when filling such datasets. One possible strategy would be to (pre-)fill such locations with educated guesses (i.e. close to zero values for vegetation greenness parameters like GPP or NDVI during winter) or data from other sources.

One open issue to be solved in the future is the influence of ocean boundaries in these terrestrial datasets. Even though the ambiguous results for scenario 14 and 16 show that other factor also play a role, the frequent selections of spatial SSA for egg boxes scenarios without or with very small ocean coverage compared to the other cutouts show that spatial SSA is strongly influenced by the existence of oceans. As 2-D SSA also needs gap-free data, the ocean locations are simply treated as gaps, iteratively filled during the inner loops, and only set to empty or missing-values at the end of the process. As a consequence, they act as very big, continuous spatial gaps. One solution would be to fill only a set of (mainly) ocean-free cutouts of a global dataset. Another approach for future developments could be to "pad" the coastlines by simple repetitions of the last available terrestrial value.

One limitation and disadvantage compared to the Kondrashov and Ghil (2006) method are the computational demands of the spatiotemporal scheme. Due to the iterative nature of the Kondrashov and Ghil (2006) scheme itself, the rather costly SSA procedure is repeated several times (at maximum of 10 outer × 10 inner = 100 iterations in our case) even in the 1-D temporal case. The application of the truncated Golyandina and Korobeynikov (2013) SSA algorithm alleviates though not removes this constraint.

In our tests, one 2-D SSA run of a 100×100 grid and one 1-D run of a time series of length 100 consumed roughly comparable amounts of CPU time. Hence, the speed limiting step in each case is the full run of 100×100 1-D temporal SSA runs per step compared to only 100 runs of 2-D SSA. As this is repeated during each outer loop step, our method consumes at a maximum roughly 10 times the CPU time as the traditional 1-D SSA. Due to these constraints and especially due to the huge amount of repetitions in our experimental setup, we constrained our method testing analysis to a rather small (compared to current high resolution remote sensing products) grid of 100×100 pixels.

In the application case, however, we do not expect these demands to play a crucial role. First, several options are implemented in the algorithm to reduce the amount a iterations necessary (e.g. the possibility to run the cross validation only with a subset of the data). Second, our algorithm is highly parallelized to fully utilise the capacities of modern multi-core or cluster machines. Run parallelized on 4 CPUs, for example, our high gap test runs needed ≈ 12 h to complete. Filling larger datacubes for real world test cases would scale roughly linearly with the amount of grid cells. Last and most important, for gapfilling only one single run per dataset is necessary so a computation time of, for example, a few days would pose a rather negligible constraint. One upper limit, however, is the memory needed to load the full datacube, a problem especially pronounced with GNU-R. For the present time, this constraint restricts the filling of, for example, 0.5×0.5 degree global datasets to high performance computing environments with high memory capabilities.

4.4 Future directions

The conceptual framework presented here which uses only SSA to separate temporal and spatial patterns of different scales can be easily extended to other methods. One example would be Empirical Mode Decomposition (EMD; Huang et al., 1998) which can be applied both in a temporal 1-D and a spatial 2-D setting (Nunes et al., 2003) or simply multidimensional smoothing splines (Garcia, 2010). For EMD the grouping of eigentriples which is necessary for SSA is relatively straightforward. In addition it has been shown to yield equally good results as SSA (Wu et al., 2010). This makes it a promising candidate for future tests and extensions. It is also possible to apply a mixture of different methods during the same gapfilling run, for example, by using the results from one method as first guesses for the other method.

For this paper, we tested only the most obvious combination of dimensional settings, i.e the choice between 1-D temporal and 2-D spatial SSA. Theoretically, however, the three dimensions of the datacube can be used in 6 different ways of combining 1-D or 2-D cutouts (6 combinations = 3 different single dimension settings + 3 different 2 dimension settings). Using SSA to decompose a longitude × time 2-D matrix, for example, might produce improved results as it partly over-

comes several of the difficulties encountered in the current setup. Such a matrix would include the clear periodic pattern of the annual cycle as opposed to the rather patchy and non-periodic spatial patterns. In addition, a cut along one latitude band would group together a set of data from potentially similar vegetation located in similar climatic zones. Such a cutout would, hence, overcome the limitation of 1-D SSA with completely (or nearly) missing time series and still exploit the strength of SSA to detect periodic patterns.

This algorithm was specifically developed to facilitate univariate gapfilling especially for the case where no additional data is available or when multivariate gapfilling would bias a subsequent exploratory analysis. 2-D SSA, however, can also be used in a multivariate framework (hence the alternative name M-SSA) via decomposing sets of time series of different variables. This may be particularly helpful for situations with continuously missing data for large adjacent regions.

Finally, the use of spectral separation methods working in three dimensions would greatly simplify and speed up the iteration scheme. It would remove the necessity to run different combinations of fewer dimensions after each outer loop step and to pick the combination yielding the best results. It would also allow for a more consistent use of the information from other dimensions into the filling of one particular dimension. In this algorithm such information is only transported via its use as a first guess, and its influence may be reduced significantly during the many inner loop iterations. Hence, the incorporation of a 3-D method in the algorithm or a 3-D extension of an existing 1-D and 2-D method may yield promising results and further improvements. In the SSA case, however, such a method has not yet been developed.

5 Conclusions

We presented a gapfilling framework based on SSA to simultaneously extract spatial and temporal patterns in geoscientific datasets. The algorithm iteratively determines which dimension yields the better results and uses its gapfilling results as a first guess for subsequent steps. The results show that even though state-of-the-art 1-D SSA is used in the majority of cases, spatial SSA can improve the results especially with high gap ratios. In addition it yields totally gap-free data. Whether 1-D, M-SSA, or spatiotemporal SSA provide better predictions depends on the amount and type of spatial and temporal patterns in the data and on the amount and structure of the gaps. The merit of the new method proposed here is that it simultaneously applies the different methodological and dimensional settings. Our gapfilling framework combines the advantages of temporal and spatial SSA. This integration is highly flexible and frees the user from a priori assumptions and the restriction of the analysis to one particular dimensional choice. In the future, our conceptual framework can be extended to integrate other temporal and spatial methods.

Appendix A

Detailed SSA description

The following description of 2-D SSA follows the work and notation of Golyandina and Usevich (2009). A good overview and a discussion of the different ways to perform SSA is given in Ghil et al. (2002).

Suppose we want to decompose a 2-D array of data which is a sum of unknown components $\mathbf{F} = \mathbf{F}^{(1)} + \ldots + \mathbf{F}^{(m)}$. The task of 2-D-SSA is to produce a decomposition $\mathbf{F} = \tilde{\mathbf{F}}^{(1)} + \cdots + \tilde{\mathbf{F}}^{(m)}$, where the terms approximate the initial components.

A1 Embedding

Let

$$\mathbf{F} = \begin{pmatrix} f(1,1) & f(1,2) & \cdots & f(1,N_y) \\ f(2,1) & f(2,2) & \cdots & f(2,N_y) \\ \vdots & \vdots & \ddots & \vdots \\ f(N_x,1) & f(N_x,2) & \cdots & f(N_x,N_y) \end{pmatrix}. \tag{A1}$$

The algorithm is based on the SVD of a Hankel-block-Hankel (HbH) matrix constructed from the 2-D array. The dimensions are defined by the window sizes (L_x, L_y), which are restricted by $1 < L_x \leq N_x$, $1 < L_x \leq N_x$ and $1 < L_x L_y \leq N_x N_y$. Let $K_x = N_x - L_x + 1$ and $K_x = N_x - L_x + 1$.

A1.1 Embedding

First, we arrange the input 2-D array into a Hankel-block-Hankel matrix of size $L_x L_y \times K_x K_y$:

$$\mathbf{W} = \begin{pmatrix} \mathbf{H}_1 & \mathbf{H}_2 & \mathbf{H}_3 & \cdots & \mathbf{H}_{K_y} \\ \mathbf{H}_2 & \mathbf{H}_3 & \mathbf{H}_4 & \cdots & \mathbf{H}_{K_y} \\ \mathbf{H}_3 & \mathbf{H}_4 & \cdot^{\cdot^{\cdot}} & \cdot^{\cdot^{\cdot}} & \vdots \\ \vdots & \vdots & \cdot^{\cdot^{\cdot}} & \cdot^{\cdot^{\cdot}} & \vdots \\ \mathbf{H}_{L_y} & \mathbf{H}_{L_y+1} & \cdots & \cdots & \mathbf{H}_{N_y} \end{pmatrix}. \tag{A2}$$

where

$$\mathbf{H}_j = \begin{pmatrix} f(1,j) & f(2,j) & \cdots & f(K_x,j) \\ f(2,j) & f(3,j) & \cdots & f(K_x+1,j) \\ \vdots & \vdots & \ddots & \vdots \\ f(L_x,j) & f(L_x,j) & \cdots & f(N_x,j) \end{pmatrix}. \tag{A3}$$

Obviously, there is a one-to-one correspondence between 2-D arrays of size $N_x \times N_y$ and HbH matrices (Eq. A2).

A2 SVD

We apply SVD onto the HbH (Eq. A2):

$$\mathbf{W} = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^T. \tag{A4}$$

Here λ_i ($1 \leq i \leq d$) are the non-zero eigenvalues of the matrix WW^T arranged in decreasing order. $\{U_1, \ldots, U_d\}$ is a system of orthonormal eigenvectors of WW^T of length $L_x L_y$; $\{V_1, \ldots, V_d\}$ is an orthonormal system of vectors in $R^{K_x K_y}$. The factors V_i's can be expressed as follows: $V_i = \mathbf{W}^T U_i / \sqrt{\lambda_i}$. The triple $(\sqrt{\lambda_i}, U_i, V_i)$ is said to be the ith eigentriple.

A3   Grouping

Depending on their structure, different sub-signals relate to single (often the case for trend components), pairs (sinusoidal sub-signals) or even large groups of these eigentriples (non-sinusoidal signals with several harmonics). To obtain the original sub-signals corresponding to the eigentriples one has to group (see also Sect. 2.2) the latter accordingly and project (see below) these groups independently. One chooses m disjoint subsets of indices I_k (groups of eigentriples),

$$I_1 \cup I_2 \cup \cdots \cup I_m = \{1, \ldots, d\}. \tag{A5}$$

Then, one obtains the decomposition of the HbH matrix

$$\mathbf{W} = \sum_{k=1}^{m} \mathbf{W}_{I_k}, \quad \text{where} \quad \mathbf{W}_{I_k} = \sum_{i \in I_k} \sqrt{\lambda_i} U_i V_i^T. \tag{A6}$$

This step controls the resulting decomposition of the 2-D array and thus is the critical step in the algorithm.

A4   Projection

In order to obtain a decomposition of the initial 2-D array, projection is necessary. First, matrices $\mathbf{W}_{I_k}$ are reduced to Hankel-block-Hankel matrices $\tilde{\mathbf{W}}_{I_k}$. Then 2-D arrays $\tilde{F}_{I_k}$ are obtained from $\tilde{\mathbf{W}}_{I_k}$ by the above-mentioned one-to-one correspondence.

The matrices $\tilde{\mathbf{W}}_{I_k}$ are obtained by a two-step *hankelization*. That means that first one averages over the secondary diagonals within the blocks of $\mathbf{W}_{I_k}$ (*within-block* hankelization) and then the blocks of the whole resulting matrix are averaged between themselves (*between-block* hankelization). The result of the algorithm is then

$$F = \sum_{k=1}^{m} \tilde{F}_{I_k}. \tag{A7}$$

A5   Special case – 1-D-SSA

It turns out that one can consider 1-D-SSA as a special case of 2-D-SSA (Golyandina and Usevich, 2009). It occurs when the input array has only one dimension (e.g. a time series). In this case one only needs one parameter L, also called *window length*. The algorithm is exactly the same (see e.g. Ghil et al. (2002) for a description).

References

Beckers, J. and Rixen, M.: EOF Calculations and Data Filling from Incomplete Oceanographic Datasets, J. Atmos. Ocean. Technol., 20, 1839–1856, doi:10.1175/1520-0426(2003)020<1839:ECADFF>2.0.CO;2, 2003.

Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., Rodenbeck, C., Arain, M. A., Baldocchi, D., Bonan, G. B., Bondeau, A., Cescatti, A., Lasslop, G., Lindroth, A., Lomas, M., Luyssaert, S., Margolis, H., Oleson, K. W., Roupsard, O., Veenendaal, E., Viovy, N., Williams, C., Woodward, F. I., and Papale, D.: Terrestrial Gross Carbon Dioxide Uptake: global Distribution and Covariation with Climate, Science, 329, 834–838, doi:10.1126/science.1184984, 2010.

Broomhead, D. S. and King, G. P.: Extracting Qualitative Dynamics From Experimental data, Physica D, 20, 217–236, doi:10.1016/0167-2789(86)90031-X, 1986.

Dempster, A., Laird, N., and Rubin, D.: Maximum Likelihood from Incomplete Data Via the EM Algorithm, J. Royal Stat. Soc. Ser B., 39, 1–38, WOS:A1977DM46400001, 1977.

Falge, E., Baldocchi, D., Olson, R., Anthoni, P., Aubinet, M., Bernhofer, C., Burba, G., Ceulemans, R., Clement, R., Dolman, H., Granier, A., Gross, P., Grunwald, T., Hollinger, D., Jensen, N. O., Katul, G., Keronen, P., Kowalski, A., Lai, C. T., Law, B. E., Meyers, T., Moncrieff, H., Moors, E., Munger, J. W., Pilegaard, K., Rannik, U., Rebmann, C., Suyker, A., Tenhunen, J., Tu, K., Verma, S., Vesala, T., Wilson, K., and Wofsy, S.: Gap filling strategies for defensible annual sums of net ecosystem exchange, Agr. Forest Meteorol., 107, 43–69, doi:10.1016/S0168-1923(00)00225-2, 2001.

Garcia, D.: Robust smoothing of gridded data in one and higher dimensions with missing values, Comput. Stat. Data Anal., 54, 1167–1178, doi:10.1016/j.csda.2009.09.020, 2010.

Ghil, M., Allen, M. R., Dettinger, M. D., Ide, K., Kondrashov, D., Mann, M. E., Robertson, A. W., Saunders, A., Tian, Y., Varadi, F., and Yiou, P.: Advanced spectral methods for climatic time series, Rev. Geophys., 40, 1003, doi:10.1029/2000RG000092, 2002.

Golyandina, N. and Korobeynikov, A.: Basic Singular Spectrum Analysis and forecasting with R, Comput. Stat. Data Anal., doi:10.1016/j.csda.2013.04.009, 2013.

Golyandina, N. and Osipov, E.: The "Caterpillar"-SSA method for analysis of time series with missing values, J. Stat. Plann. Infer., 137, 2642–2653, doi:10.1016/j.jspi.2006.05.014, 2007.

Golyandina, N. and Usevich, K.: Matrix Methods: Theory, Algorithms, Applications, chap. 2-D-extensions of singular spectrum analysis: algorithm and elements of theory, 450–474, World Scientific Publishing, 2009.

Golyandina, N. and Zhigljavsky, A.: Singular spectrum analysis for time series, Springer, available at: http://www.springer.com/statistics/statistical+theory+and+methods/book/978-3-642-34912-6 (last access: 19 September 2013), 2013.

Hocke, K. and Kämpfer, N.: Gap filling and noise reduction of unevenly sampled data by means of the Lomb-Scargle periodogram, Atmos. Chem. Phys., 9, 4197–4206, doi:10.5194/acp-9-4197-2009, 2009.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N., Tung, C. C., and Liu, H. H.: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis, Proc. Royal Soc. London A, 454, 903–995, doi:10.1098/rspa.1998.0193, 1998.

Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., and Ferreira, L. G.: Overview of the radiometric and biophysical performance of the MODIS vegetation indices, Remote Sens. Environ., 83, 195–213, doi:10.1016/S0034-4257(02)00096-2, 2002.

Janssen, P. H. M. and Heuberger, P. S. C.: Calibration of Process-oriented Models, Ecol. Modell., 83, 55–66, doi:10.1016/0304-3800(95)00084-9, 1995.

Justice, C. O., Vermote, E., Townshend, J. R. G., Defries, R., Roy, D. P., Hall, D. K., Salomonson, V. V., Privette, J. L., Riggs, G., Strahler, A., Lucht, W., Myneni, R. B., Knyazikhin, Y., Running, S. W., Nemani, R. R., Wan, Z. M., Huete, A. R., van Leeuwen, W., Wolfe, R. E., Giglio, L., Muller, J. P., Lewis, P., and Barnsley, M. J.: The Moderate Resolution Imaging Spectroradiometer (MODIS): Land remote sensing for global change research, IEEE Trans. Geosci. Remote Sens., 36, 1228–1249, doi:10.1109/36.701075, 1998.

Kaplan, A., Kushni, Y., Cane, M., and Blumenthal, M.: Reduced space optimal analysis for historical datasets: 136 years of Atlantic sea surface temperatures, J. Geophys. Res., 102, 27–27, doi:10.1029/97JC01734, 1997.

Kondrashov, D. and Ghil, M.: Spatio-temporal filling of missing points in geophysical datasets, Nonlin. Processes Geophys., 13, 151–159, doi:10.5194/npg-13-151-2006, 2006.

Liu, Y. Y., Parinussa, R. M., Dorigo, W. A., De Jeu, R. A. M., Wagner, W., van Dijk, A. I. J. M., McCabe, M. F., and Evans, J. P.: Developing an improved soil moisture dataset by blending passive and active microwave satellite-based retrievals, Hydrol. Earth Syst. Sci., 15, 425–436, doi:10.5194/hess-15-425-2011, 2011.

Luo, Y. Q., Randerson, J. T., Abramowitz, G., Bacour, C., Blyth, E., Carvalhais, N., Ciais, P., Dalmonech, D., Fisher, J. B., Fisher, R., Friedlingstein, P., Hibbard, K., Hoffman, F., Huntzinger, D., Jones, C. D., Koven, C., Lawrence, D., Li, D. J., Mahecha, M., Niu, S. L., Norby, R., Piao, S. L., Qi, X., Peylin, P., Prentice, I. C., Riley, W., Reichstein, M., Schwalm, C., Wang, Y. P., Xia, J. Y., Zaehle, S., and Zhou, X. H.: A framework for benchmarking land models, Biogeosciences, 9, 3857–3874, doi:10.5194/bg-9-3857-2012, 2012.

Mjolsness, E. and DeCoste, D.: Machine Learning for Science: State of the Art and Future Prospects, Science, 293, 2051–2055, doi:10.1126/science.293.5537.2051, 2001.

Moffat, A. M., Papale, D., Reichstein, M., Hollinger, D. Y., Richardson, A. D., Barr, A. G., Beckstein, C., Braswell, B. H., Churkina, G., Desai, A. R., Falge, E., Gove, J. H., Heimann, M., Hui, D. F., Jarvis, A. J., Kattge, J., Noormets, A., and Stauch, V. J.: Comprehensive comparison of gap-filling techniques for Eddy Covariance net carbon fluxes, Agr. Forest Meteorol., 147, 209–232, doi:10.1016/j.agrformet.2007.08.011, 2007.

Musial, J. P., Verstraete, M. M., and Gobron, N.: Technical Note: Comparing the effectiveness of recent algorithms to fill and smooth incomplete and noisy time series, Atmos. Chem. Phy., 11, 7905–7923, doi:10.5194/acp-11-7905-2011, 2011.

Nunes, J., Bouaoune, Y., Delechelle, E., Niang, O., and Bunel, P.: Image analysis by bidimensional empirical mode decomposition, Image Vision Comput., 21, 1019–1026, doi:10.1016/S0262-8856(03)00094-5, 2003.

Overpeck, J. T., Meehl, G. A., Bony, S., and Easterling, D. R.: Climate data challenges in the 21 st century, Science (Washington), 331, 700–702, doi:10.1126/science.1197869, 2011.

R Development Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, available at: http://www.R-project.org, 2013.

Reichstein, M., Bahn, M.and Ciais, P., Frank, D., Mahecha, M. D., Seneviratne, S. I., Zscheischler, J., Beer, C., Buchmann, N., Frank, D. C., Papale, D., Rammig, A., Smith, P., Thonicke, K., van der Velde, M., Vicca, S., Walz, A., and Wattenbach, M.: Climate extremes and the carbon cycle, Nature, 500, 287–295, doi:10.1038/nature12350, 2013.

Reynolds, R. W. and Smith, T. M.: Improved Global Sea Surface Temperature Analyses Using Optimum Interpolation, J. Climate, 7, 929–948, doi:10.1175/1520-0442(1994)007<0929:IGSSTA>2.0.CO;2, 1994.

Roerink, G. J., Menenti, M., and Verhoef, W.: Reconstructing cloudfree NDVI composites using Fourier analysis of time series, Int. J. Remote Sens., 21, 1911–1917, doi:10.1080/014311600209814, 2000.

Schneider, T.: Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values, J. Climate, 14, 853–871, doi:10.1175/1520-0442(2001)014<0853:AOICDE>2.0.CO;2, 2001.

Schoellhamer, D. H.: Singular Spectrum Analysis for time series with missing data, Geophys. Res. Lett., 28, 3187–3190, doi:10.1029/2000GL012698, 2001.

Smith, T. M., Reynolds, R. W., Livezey, R. E., and Stokes, D. C.: Reconstruction of Historical Sea Surface Temperatures Using Empirical Orthogonal Functions, J. Climate, 9, 1403–1420, doi:10.1175/1520-0442(1996)009<1403:ROHSST>2.0.CO;2, 1996.

Vautard, R. and Ghil, M.: Singular Spectrum Analysis in nonlinear dynamics, with applications to paleoclimatic time series, Physica D: Nonlinear Phenomena, 35, 395–424, doi:10.1016/0167-2789(89)90077-8, 1989.

Wang, G., Garcia, D., Liu, Y., de Jeu, R., and Johannes Dolman, A.: A three-dimensional gap filling method for large geophysical datasets: Application to global satellite soil moisture observations, Environ. Modell. Softw., 30, 139–142, doi:10.1016/j.envsoft.2011.10.015, 2012.

Weedon, G. P., Gomes, S., Viterbo, P., Shuttleworth, W. J., Blyth, E., Österle, H., Adam, J. C., Bellouin, N., Boucher, O., and Best, M.: Creation of the WATCH Forcing Data and Its Use to Assess Global and Regional Reference Crop Evaporation over Land during the Twentieth Century, J. Hydrometeorol., 12, 823–848, doi:10.1175/2011JHM1369.1, 2011.

Wu, Z., Schneider, E. K., Kirtman, B. P., Sarachik, E. S., Huang, N. E., and Tucker, C. J.: The modulated annual cycle: an alternative reference frame for climate anomalies, Clim. Dynam., 31, 823–841, doi:10.1007/s00382-008-0437-z, 2010.