Computational Geosciences manuscript No.
(will be inserted by the editor)

An Improved Methodology for Filling Missing Values in Spatio-Temporal Climate Dataset

Application to Tanganyika Lake Dataset

Antti Sorjamaa · Amaury Lendasse ·
Yves Cornet · Eric Deleersnijder

Received: date / Accepted: date

Abstract In this paper, an improved methodology for the determination of missing values in a spatio-temporal database is presented. This methodology performs denoising projection in order to accurately fill the missing values in the database. The improved methodology is called EOF Pruning and it is based on an original linear projection method called Empirical Orthogonal Functions (EOF). The experiments demonstrate the performance of the improved methodology and present a comparison with the original EOF and with a widely-used Optimal Interpolation method called Objective Analysis.

Keywords Missing Value Problem · Empirical Orthogonal Functions · EOF · Selection of Singular Values · Tanganyika Lake

Antti Sorjamaa
Helsinki University of technology
Adaptive Informatics Research Centre
Espoo, Finland
Tel.: +358-9-4515286
Fax: +358-9-4513277
E-mail: Antti.Sorjamaa@hut.fi

Amaury Lendasse
Helsinki University of technology
Adaptive Informatics Research Centre
Espoo, Finland

Yves Cornet
University of Liege
Unit of Geomatics
Liege, Belgium

Eric Deleersnijder
Universite catholique de Louvain
Louvain School of Engineering
Centre for Systems Engineering and Applied Mechanics (CESAME)
Louvain-la-Neuve, Belgium

1 Introduction

Meteorology and climate research are two rapidly growing fields with an increasing need for accurate and large measurement datasets. The African continent represents a clear example of the current challenges in these fields. The drought and humidity imbalance create extreme conditions for both the people on the continent and the very necessary research.

Lake Tanganyika is located in the African Rift in the center of the African continent. It is an important source of proteins for the people around it and the fish industry provides not only the food for the people, but also gives thousands of workers a job.

The importance to the people and the extraordinary size and shape of the lake make it really valuable for the climate research, but the size brings also difficulties. The size and the shape of the lake make it hard to adequately measure the bio-geo-hysical parameters, such as surface temperature. But due to the current political and economical situation in Africa, the satellite is the only valid option. The data measured by satellite includes a vast number of missing values, due to clouds, technical difficulties and even heavy smoke from forest fires. The missing values make *a posteriori* modelling a difficult problem and the filling procedure a mandatory preprocessing step [1] before climate modelling.

A great number of methods have been already developed for solving the problem by filling the missing values, for example, Kriging [2] and several other Optimal Interpolation methods, such as Objective Analysis.

One of the emerging approaches for filling the missing values is the Empirical Orthogonal Functions (EOF) methodology [3–5]. The EOF is a deterministic methodology, enabling a linear projection to a high-dimensional space. Moreover, the EOF models allow continuous interpolation of missing values even when a high percentage of the data is missing.

In this paper, an improvement to the standard EOF method is presented, called EOF Pruning. It enhances the accuracy of the EOF methodology and even speeds up the calculation process.

Sections 2 and 3 present the EOF methodology and the EOF pruning improvement, respectively. In Section 4 the validation and model selection procedure is briefly explained. Finally, the described methodologies are compared in Section 5 using the Tanganyika Lake temperature dataset.

2 Empirical Orthogonal Functions

Empirical Orthogonal Functions (EOF) [3] is a deterministic method allowing a linear, continuous projection to a high-dimensional space. The EOF has been used in climate research for finding the missing values as well as a denoising tool [4–8].

In this paper, the EOF is used as a denoising tool and for finding the missing values at the same time. The method presented here is based on the one presented in [4].

The EOF is calculated using the standard and well-known Singular Value Decomposition (SVD),

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^* = \sum_{k=1}^{K} \rho_k \mathbf{u}_k \mathbf{v}_k, \tag{1}$$

where $\mathbf{X}$ is a 2-dimensional data matrix, $\mathbf{U}$ and $\mathbf{V}$ are collections of singular vectors $\mathbf{u}$ and $\mathbf{v}$ in each dimension, respectively, $\mathbf{D}$ is a diagonal matrix with the singular values ρ in its diagonal and K is the smaller dimension of $\mathbf{X}$ (or the number of nonzero singular values if $\mathbf{X}$ is not full rank). The singular values and the respective vectors are sorted to decreasing order.

When the EOF is used to remove the noise from the data, not all singular values and vectors are used to reconstruct the data matrix. Instead, it is assumed that the vectors corresponding to larger singular values have larger signal-to-noise ratio than the ones corresponding to smaller values [3]. Therefore, it is logical to select the q largest singular values and the corresponding vectors and reconstruct the data matrix using only them.

When $q < K$, the reconstructed data matrix is obviously not the same than the original one. The larger q is selected, the more original data, which also includes more noise, is preserved. The optimal q is selected using validation methods; see for example [9].

The EOF (or the SVD method) cannot be directly used with databases including missing values. The missing values must be replaced by some initial values in order to use the EOF. This replacement can be for example the mean value of the whole data matrix $\mathbf{X}$, the row or column mean, linear regression or polynomial fitting row wise or column wise, depending on the structure of the data matrix.

After the initial value replacement the EOF process begins by performing the SVD and the selected q singular values and vectors are used to build the reconstruction. In order not to lose **any** information, only the missing values of $\mathbf{X}$ are replaced with the values from the reconstruction. After the replacement, the new data matrix is again broken down to singular values and vectors with the SVD and reconstructed again. The procedure is repeated until the convergence criterion is fulfilled. The procedure is summarized in Table 1.

3 EOF Pruning

In some cases, some of the biggest singular values contain so large noise levels that they disturb the selection process described in Table 1. For example, if the first n singular values are selected by the validation procedure, but not the $n+1$, it does not necessarily mean that all the rest from $n+2$ to K are as

Table 1 Summary of the EOF algorithm for finding the missing values.

1. Initial values are substituted into missing values of the original data matrix $\mathbf{X}$
2. For each q from 1 to K
    (a) q singular values and eigenvectors are calculated using the SVD
    (b) A number of values and vectors are used to make the reconstruction
    (c) The missing values from the original data are filled with the values from the reconstruction
    (d) If the convergence criterion is fulfilled, the validation error is calculated and saved and the next q value is taken under inspection. If not, then we continue from step a) with the same q value
3. The q with the smallest validation error is selected and used to reconstruct the final filling of the missing values in $\mathbf{X}$ starting from the originally initialized data of step 1

noisy as the $n+1$. Some of the smaller values can still hold some important information vital to the accurate estimation of the missing values.

Even the assumption of larger singular values holding more signal than noise is still valid; it does not mean that **all** smaller values are completely corrupted with noise. As described above, some smaller values can hold vital information even the amount of noise is increasing compared to the larger singular values.

If the purpose is to solely remove the noise from the dataset with the cost of accuracy, then the smaller values should not be used. But our goal here is to approximate the missing values as accurately as possible and the denoising is left as a secondary goal.

Therefore, instead of selecting a certain number of largest singular values and vectors to perform the reconstruction, we propose an alternative approach; selecting the values and vectors in a non-continuous fashion. The selection strategy is explained more deeply in Section 3.1.

The selection of singular values and vectors is done in each round of the EOF procedure. It means that when the initialization of the missing values is done and the singular values and vectors are calculated, the selection algorithm is used to select the most optimal values and vectors. Then, the initialized missing values are replaced by the reconstruction obtained using the selected set of singular values and vectors. In the next round, the new data matrix is again broken down to singular values and vectors and the selection is performed again.

The revised EOF Pruning algorithm is summarized in Table 2.

We have developed a toolbox[1] for the Matlab software to perform the EOF Pruning for the missing value imputation. The toolbox is free to use under GNU General Public License.

[1] http://www.cis.hut.fi/projects/tsp/index.php?page=research&subpage=downloads, A. Sorjamaa and A. Lendasse, 2008

Table 2 The EOF Pruning algorithm for finding the missing values.

1. Initial values are substituted into missing values of the original data matrix $\mathbf{X}$
2. Loop until convergence
    - (a) K singular values and eigenvectors are calculated using the SVD
    - (b) The selection process selects an optimal set of singular values and vectors from the K candidates. The selected set, q_r, is saved, where r represents the number of the current round.
    - (c) The values and vectors in the set q_r are used to make the reconstruction
    - (d) The missing values from the original data are filled with the values from the reconstruction
    - (e) The validation error is calculated and saved. If the convergence criterion is not fulfilled, we continue to the next round from step (a).
3. The selected singular values and vectors are used to reconstruct the final filling of the missing values in $\mathbf{X}$. The final filling uses as many rounds as the validation process. In each round the corresponding set q_r is used.

3.1 Forward Selection

In this greedy selection strategy, starting from the empty set S, the best available input is added to the set S one by one, until the size of the S is K. Here, one input consists of one singular value and the corresponding vectors together.

The validation error, see [10–12], of each selection step is saved and finally the set of values and vectors giving the smallest error is selected. The selection algorithm is summarized in Table 3.

Table 3 Forward Selection Strategy.

1. (Initialization)
   Set I to be the initial set of original K inputs, and S to be the empty set, which will contain the selected inputs. X^i denotes the singular value and vector currently under validation.
2. (Selection of the variables)
   Find:

   $$X^s = \arg\min_{X^i} \textsc{Validation}(\{S, X^i\}), \quad X^i \in I, \quad i = 1, ..., K$$

   Save $\textsc{Validation}(\{S, X^s\})$ and move X^s from I to S.
   Continue with the same way, until the size of S is K.
3. (Result)
   Compare the $\textsc{Validation}$ values for all sizes of sets S, the selection result is the set S that minimizes the $\textsc{Validation}$.

In Forward Selection, only $K(K+1)/2$ different combinations are evaluated. This is much less than the total possible number of combinations 2^K. Therefore, optimality is not guaranteed and the selected set may not be the global optimal one.

However, due to the possibly very large number of singular values and vectors, the complete search of all the possible combinations is not feasible. A good description of other selection strategies can be found in [13].

4 Optimization of the Method

For the used methodology we have to select the singular values and vectors q to be used for the EOF. In the original version of the EOF, the selection is done as a number of the largest values and vectors to be used, but for the EOF Pruning, the selection is done from K values and vectors in a non-continuous fashion, as described in Section 3.1.

All the selections are done using the same validation procedure with the same validation sets. Using the same validation sets and procedure for each method enables a reliable comparison of the results between different methods and the selected parameters are more secure and reliable.

To validate the parameters we use Monte-Carlo Cross-Validation (MCCV) method [9] with ten folds. Each fold consists of 2,5 percent of the known data of the full dataset and only one fold of validation data is removed at a time. The removal of the validation data is not completely random, but instead it is performed using artificial clouds placed randomly in the dataset. These artificial clouds are different groups of pixels and represent different shapes of real life clouds.

Finally, the optimal selection of the parameters is obtained by minimizing the Mean Square Error (MSE) over all validation sets and all values in each validation set. It is defined in Equation 2.

$$\mathrm{MSE} = \frac{1}{Z} \sum_{v=1}^{10} \sum_{i \in \mathrm{MC}_v} (\mathbf{x}_{\mathrm{MC}_v,i} - \hat{\mathbf{x}}_{\mathrm{MC}_v,i})^2, \tag{2}$$

where Z denotes the total number of samples in all validation sets, v denotes each Monte-Carlo validation set from 1 to 10, i denotes one measurement point in the validation set MC_v and $\mathbf{x}$ is the known measurement value and $\hat{\mathbf{x}}$ is the filled approximation.

5 Experiments

5.1 Tanganyika Lake Surface Temperature Dataset

For the comparison of the performances of the presented methods, we use a Tanganyika Lake Surface Temperature dataset[2]. The lake lies in the African Rift area and it is over 670 kilometres long with an average width of about 50 kilometres.

[2] Tanganyika Lake dataset, MODIS Data, a courtesy of Yves Cornet and the University of Liege, Belgium. The data comes from an RS dataset produced in the framework of the CKIMFISH project.

The measurements are obtained from the thermal infrared bands of the MODIS sensors of the satellite covering the lake with a spatial resolution of one kilometre.

The satellite has measured the lake a total of 666 times between years 2002 and 2006. The measuring frequency of the satellite is not constant during the five year period, instead it varies from one to 33 days. On average we have one image every 2,5 days.

The spatial resolution gives us more than 32 000 daily measurement points in one image. The amount of missing values in each image varies from 100 percent to four percent, meaning that some images have no measurement values and some have only four percent of the data missing. Finally, the whole dataset has over 63 percent of the data missing.

Before applying the methodology, each image is preprocessed, where the land is separated from the lake in each image and the missing values are defined using a mask provided with the measurements. Figure 1 shows four measurement examples with varying amount of missing values. The same example days are used in each figure later on.

Because of the huge size of the dataset, it is divided into slices. Each measurement image of the lake is cut to ten pieces in north-south direction. This is done in order to take into account the change in the dynamics of the long lake and to make the filling more local. Moreover, the percentage of missing values is on average greater in the northern part of the lake whereas the middle and the southern parts have more measurements present. Table 7 shows the specific percentages of missing values in each slice. Figure 2 shows the slicing.

As an example slice, the most southern part of the lake is shown in Figure 3 and used as an example slice in the following.

Because of the large number of missing data in the database, each day with more than 90 percent of the data missing from the slice is removed from the dataset before the learning phase. This is done for each slice individually. For the southernmost slice of the Tanganyika Lake, it leaves us 390 days with each day containing a total of 2947 measurement points with a total of 27 percent of the data missing on average.

However, this kind of removal prohibits us to fill every day of the dataset. In order to finally perform the complete filling of the dataset; all images must be taken into the calculation and finally filled. This is done after the initial comparison of the methods.

The percentage of missing values is still increased when the test and validation sets are removed. For the test set, 2,5 percent of the known data is removed from each slice separately. For the validation of the parameters for the methods, Monte Carlo Cross-Validation (MCCV) method with 10 folds is used. Each fold contains 2,5 percent of the known data, but only one of the folds is removed at a time. After removing the test set and the validation set, there is roughly 30 percent of the data missing from the southernmost slice when the validation is started. The overall statistics for the used datasets are collected to Table 4.

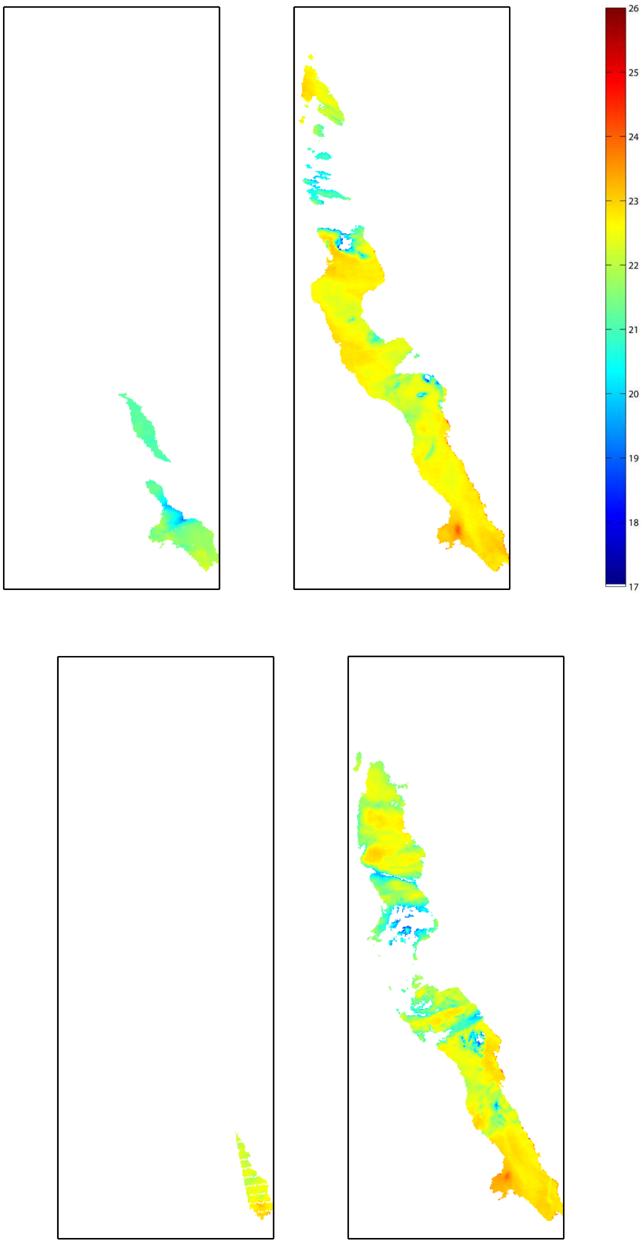

Fig. 1 Tanganyika Lake dataset, four example measurements from 2005. Days 182 and 183 are on the top from left to right and days 184 and 185 are on the bottom.

From Table 4, we can see that there are no big differences between any of the sets used in the experiments. The learning set has a slightly lower mean

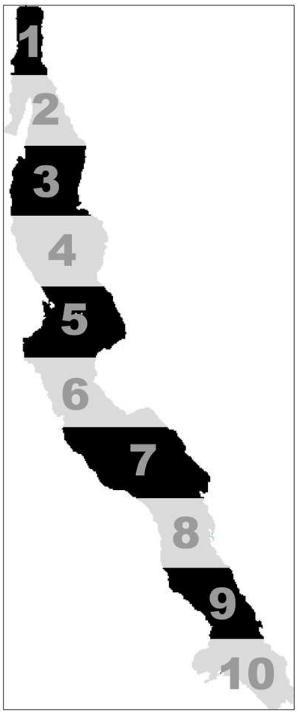

Fig. 2 The full shape of the lake. The slices are separated and numbered

Table 4 Overall statistics of the datasets covering the whole lake after the days with more than 90 percent of missing values have been removed.

	Learning	Validation	Test
Average	21,70	21,97	21,97
Standard Deviation	1,46	1,24	1,22
Number of measurements	7,36E+5	1,82E+4	1,87E+4

and a higher standard deviation due to the fact that it is roughly 40 times bigger than the other sets. Probably there are some outliers or heavily noise influenced measurements, which cause the slight difference with respect to validation and test sets.

5.2 Optimal Interpolation

For the comparisons with the EOF and EOF Pruning methodologies, a well-known Optimal Interpolation (OI) method, called Objective Analysis [14], is

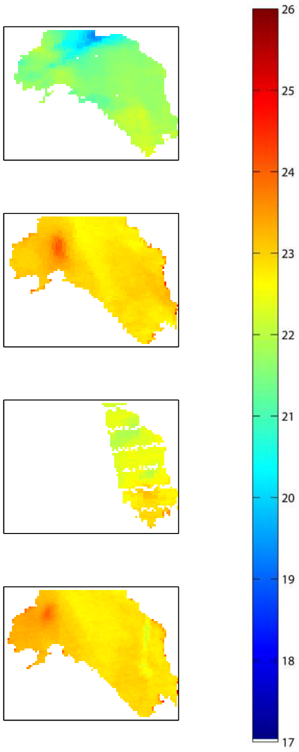

Fig. 3 Southern slice of the Tanganyika Lake dataset. The days are the same than in Figure 1, days from 182 to 185 from 2005

used. The method belongs to the large number of OI methods and it is applied using the OAX Toolbox[3].

The OI used here calculates the missing values using a weighted average of the nearest neighbors in a multidimensional space. The weights and the nearest neighbors are selected based on the covariance function and a pseudo-distance measure. The used covariance function is defined as

$$covariance(r) = \exp^{-r} * \left(1 + r + \frac{r^2}{3}\right) \tag{3}$$

and the pseudo-distance as

$$r = \sqrt{\sum_{i=1}^{n} \left(\frac{x_i - y_i}{a_i}\right)^2} \tag{4}$$

[3] More information of the specifics of the methodology and the toolbox can be found from http://www.mar.dfo-mpo.gc.ca/science/ocean/coastal_hydrodynamics/Oax/oax.html.

where x_i and y_i are the *ith* components of the x and y, respectively, and a_i is the local scale factor given for each coordinate.

The number of nearest neighbors must be determined by the user. In this paper, we used cross-validation methodology for the selection.

The OAX toolbox requires also the estimated noise in the measurements as an input. After extensive testing and several trials, we decided to set the estimation to zero, since the validation error of the OAX increased when the error estimation was increased.

5.3 Filling the Southern Slice

The validation results for the EOF method using the southern slice are shown in Figure 4.

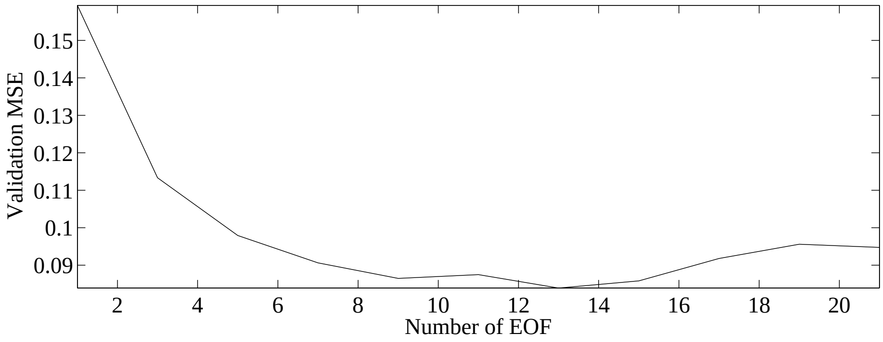

Fig. 4 Validation errors using the EOF method

From Figure 4 the optimal number of EOF is 13. It is a small number compared to the maximum of 390 EOF suggesting a strong noise influence in the data.

For the EOF Pruning method, the results are shown in Figure 5 and Table 5.

From Figure 5 we can see that three rounds are enough for the EOF Pruning. After the third round, the validation error starts to increase slowly.

Table 5 Selected EOFs using the EOF Pruning.

Round						
1	1 to 18	22				
2	1 to 45	47 to 51	56 to 59	62 to 64	74	81
3	1 to 50	52 to 54	56 to 100			

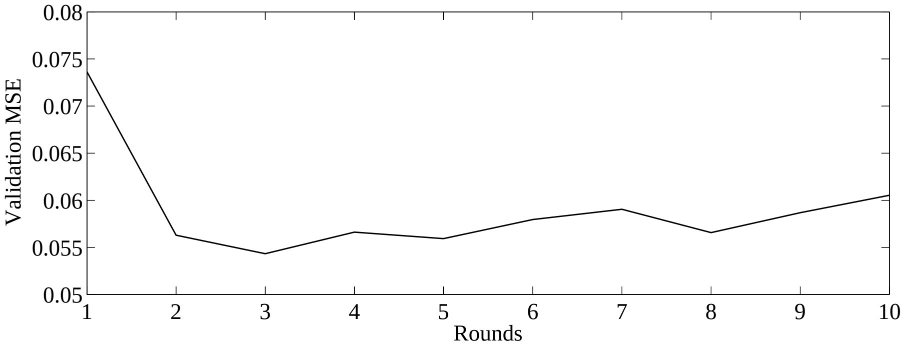

Fig. 5 Validation errors using the EOF Pruning method

From Table 5 we can see that the selected amount of singular values and vectors by the EOF Pruning varies. The first round, when the largest drop in the validation error is observed, the least amount of singular values and vectors are selected. The more rounds the EOF Pruning performs, the more values and vectors are used in the reconstruction process. This is not surprising, because each calculation round removes the noise from the data and therefore, more effective singular values and vectors can be found and used. The third round includes already almost all of the 100 EOF, which are used as a maximum in the Forward Selection. This suggests that almost all of the EOF are important and necessary for the projection and that they have been denoised.

For comparison, the OAX toolbox was used with the default settings. The validation of the number of neighbors used was performed using the same validation sets than with the EOF and EOF Pruning methods and the results of the validation can be seen in Figure 6.

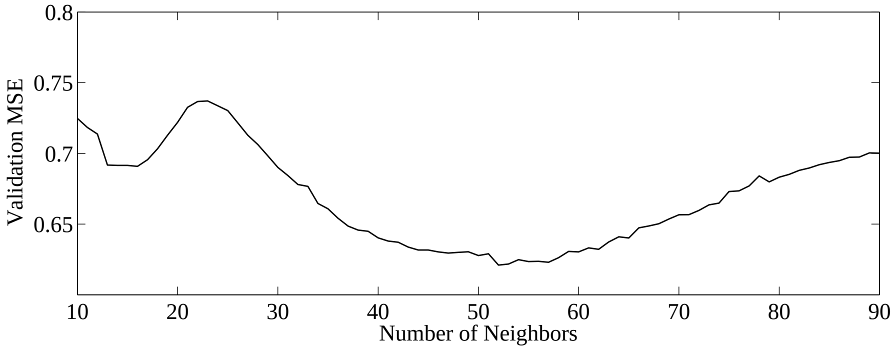

Fig. 6 10-Fold Monte Carlo Cross-Validation errors using the Objective Analysis and the OAX toolbox.

From Figure 6, we can see that there is a clear minimum in the validation error with respect to the number of neighbors. Based on this validation result, we selected the number of neighbors to be 52 for the OA.

Table 6 shows the summary of validation and test errors of all methods for the southern slice.

Table 6 Validation and Test Errors for the EOF, the EOF Pruning and the OA using the southern slice of the Tanganyika dataset.

	Validation MSE	nEOF	Test MSE
EOF	0,0839	13	0,0664
EOF Pruning	0,0543		0,0517
Objective Analysis	0,6210		0,6265

From the summary Table 6, it is quite clear that the Objective Analysis is not able to fill in the missing values accurately. As the OA relies on the distance between the neighbors in the approximation of the missing values, it might be that the highly varying temporal frequency of the measurements is disturbing the approximation. The distances between two days can vary from one to 33 days and, therefore, the neighborhood calculation is not reliable enough. The test error is on the same level than the validation error.

Furthermore, according to the table, the EOF Pruning outperforms the original EOF reducing the validation error roughly by one third and the test error by 23 percent.

Figure 7 shows the final filling of the southern part of the lake by the EOF Pruning.

5.4 Filling the whole Tanganyika Lake

In order to fill the whole dataset, we first fill all the slices. After that, the whole lake can be reconstructed from the filled slices. Here, we use all the images in the filling procedure regardless of the amount of missing values included.

The filling results of all slices are summarized in Table 7 and the final filling result of the full lake using the EOF Pruning is shown in Figure 8.

From Table 7 we can clearly see that the EOF Pruning method outperforms the original method in all slices. The selected number of EOF varies from slice to slice and the slice 1, which has the largest amount of data missing, has also the least amount of selected EOF.

The validation and test errors show no significant difference with respect to the amount of missing values. The errors are roughly the same with the slice 1 as with the slice 10, which are the slices with the most and with the least missing values, respectively.

The final filling in Figure 8 shows some evidence of the slicing, which emerge as slight discontinuities between slices. The smoothness is still better than

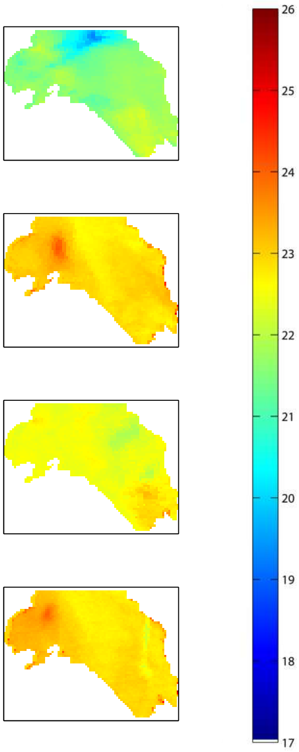

Fig. 7 Southern part of the Tanganyika Lake dataset. The pictures are the same examples filled by the EOF Pruning as in Figure 3

between some original data points and filled ones. Many of those discontinuities can already be found in the original dataset, as can be seen from the lower right picture in Figure 1.

The mean test errors including all the slices are summarized in Table 8. From this table, we can see that globally the EOF Pruning clearly outperforms the original method in comparison reducing the test error of the EOF by roughly 35 percent.

For the sake of comparison, the EOF Pruning is applied to the whole lake, without slicing. Otherwise, the methodology is the same than with the EOF Pruning with slicing. Both, the validation and the test sets are exactly the same.

Because of the length of the lake in north-south direction, the underlying dynamics are very different and, as the Table 8 shows, using the full lake is not achieving as good accuracy as the slicing. Slicing the lake into ten pieces roughly halves the test MSE, when using the EOF Pruning.

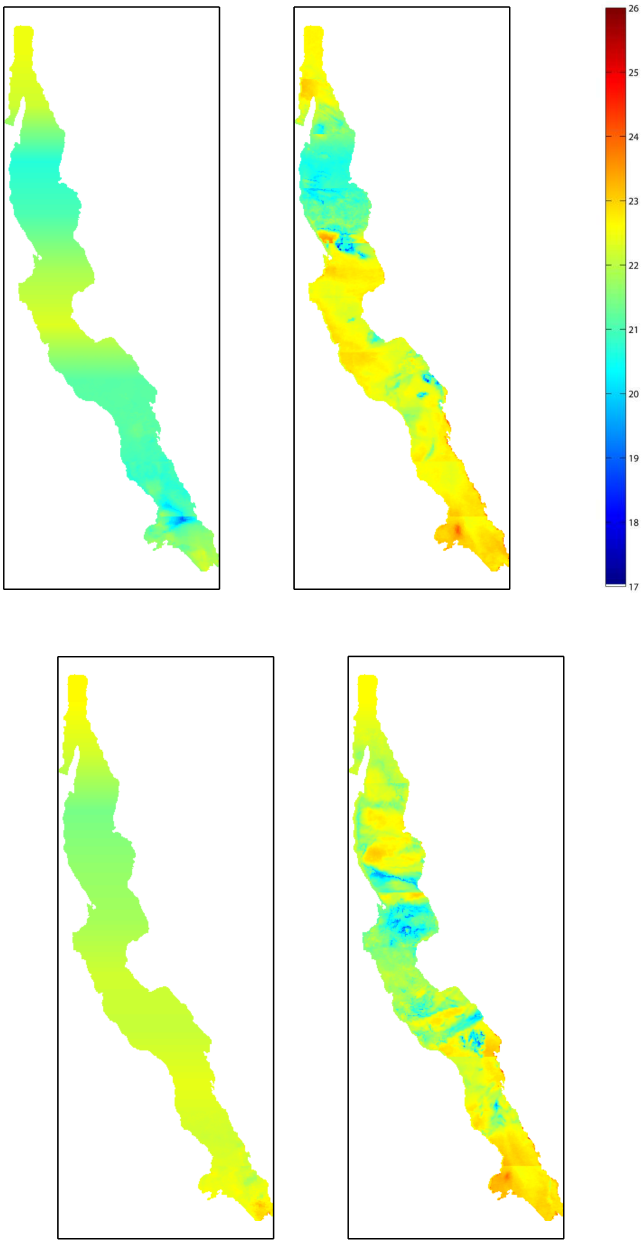

Fig. 8 Tanganyika Lake dataset. The pictures are the same examples filled by the EOF Pruning as in Figure 1

Table 9 shows an approximation of the calculation times of the used methods.

Table 7 Results of all slices with all methodologies. The percent number after each slice number is the amount of missing data in the slice before the processing starts.

	Validation MSE	Number of EOFs	Test MSE
Slice 1 (58,1%)			
EOF	0,0600	7	0,0635
EOF Pruning	0,0396		0,0470
Slice 2 (49,3 %)			
EOF	0,0745	13	0,0843
EOF Pruning	0,0523		0,0560
Slice 3 (38,7 %)			
EOF	0,1000	11	0,1040
EOF Pruning	0,0646		0,0615
Slice 4 (37,7 %)			
EOF	0,1100	15	0,1228
EOF Pruning	0,0753		0,0776
Slice 5 (39,7 %)			
EOF	0,0918	19	0,0858
EOF Pruning	0,0550		0,0635
Slice 6 (39,6 %)			
EOF	0,0977	15	0,0881
EOF Pruning	0,0689		0,0640
Slice 7 (33,6 %)			
EOF	0,0991	17	0,0865
EOF Pruning	0,0695		0,0570
Slice 8 (32,4 %)			
EOF	0,0881	19	0,0741
EOF Pruning	0,0613		0,0481
Slice 9 (29,3 %)			
EOF	0,0931	15	0,0821
EOF Pruning	0,0566		0,0469
Slice 10 (27,0 %)			
EOF	0,0839	13	0,0664
EOF Pruning	0,0543		0,0517
Objective Analysis	0,6476		0,6554

Table 8 Mean Test Errors for the whole Tanganyika Lake. Unsliced means that the EOF Pruning has been applied to the whole lake without slicing it to 10 slices.

	Test MSE
EOF	0,0858
EOF Pruning	0,0551
EOF Pruning unsliced	0,1120

The estimated calculation times show that not only the EOF Pruning achieve better accuracy, but it also does it roughly four times faster. Surprisingly, the unsliced EOF Pruning is only consuming twice the calculation time than the sliced one. However, when using the slicing of the lake, it is very easy to parallelize the computational load, which would speed up the filling substantially.

Table 9 Relative calculation times. The calculation times are given with respect to the calculation time of the EOF.

	Relative Calculation Time
EOF	1
EOF Pruning	0,23
EOF Pruning unsliced	0,41
Objective Analysis	0,05

6 Conclusion

We have presented an improved methodology for finding missing values in a spatio-temporal database. The improved EOF Pruning is compared with the original EOF using a spatio-temporal database with vast amount of missing values. The great size and the amount of missing values in the database make the problem difficult and eliminate the choices for methods that are able to fill it.

The experiments show that the improved version of the EOF is not only more accurate, but also decreases the calculation time needed to approximate the missing values. Therefore, the EOF Pruning is a very valuable and practical method for filling the missing values for spatio-temporal databases.

For further work, the EOF Pruning will be applied to other datasets in the field of climatology. The methodology will also be applied to datasets from other research fields, for example process industry and finance.

Acknowledgements The authors are indebted to Anouk de Brauwere and Christophe Leblanc for the useful comments and suggestions. Eric Deleersnijder is a research associate with the Belgian Fund for Scientific Research (FRS).

References

1. F.T. Tangang, B. Tang, A.H. Monahan, and W.W. Hsieh. Forecasting enso events: A neural network extended eof approach. *Journal of Climate*, 11:2941, January 1998.
2. Hans Wackernagel. *Multivariate Geostatistics - An Introduction with Applications*. Springer, Berlin, 1995.
3. R. Preisendorfer. *Principal Component Analysis in Meteorology and Oceanography*. Elsevier, 1988.
4. J. M. Beckers and M. Rixen. Eof calculations and data filling from incomplete oceanographic datasets. *Journal of atmospheric and oceanic technology*, 20(12):1839–1856, 2003.
5. J. Boyd, E. Kennelly, and P. Pistek. Estimation of eof expansion coefficients from incomplete data. *Deep Sea Research*, 41:1479–1488, 1994.
6. A. Alvera-Azcarate, A. Barth, M. Rixen, and J. M. Beckers. Reconstruction of incomplete oceanographic data sets using empirical orthogonal functions. application to the adriatic sea. *Ocean Modelling*, 9:325–346, 2005.
7. A. Alvera-Azcarate, A. Barth, J. M. Beckers, and R. H. Weisberg. Multivariate reconstruction of missing data in sea surface temperature, chlorophyll and wind satellite fields. *Journal of Geophysical Research*, (C03008), 2007.

8. J.-M. Beckers, A. Barth, and A. Alvera-Azcarate. Dineof reconstruction of clouded images including error maps. application to the sea surface temperature around corsican island. *Ocean Science*, 2(2):183–199, 2006.
9. Amaury Lendasse, V. Wertz, and Michel Verleysen. Model selection with cross-validations and bootstraps - application to time series prediction with rbfn models. In *LNCS*, number 2714, pages 573–580, Berlin, 2003. ICANN/ICONIP (2003), Springer-Verlag.
10. R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, volume 2, 1995.
11. B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. Chapman et Hall, 1993.
12. B. Efron and R. J. Tibshirani. Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92:548–560, 1997.
13. Antti Sorjamaa, Jin Hao, Nima Reyhani, Yongnan Ji, and Amaury Lendasse. Methodology for long-term prediction of time series. *Neurocomputing*, 70(16-18):2861–2869, October 2007.
14. L.S. Gandin. Objective analysis of meteorological fields. *Israel Program for Scientific Translations, Jerusalem*, page 242, 1969.