

# An improved methodology for filling missing values in spatiotemporal climate data set

## Application to Tanganyika Lake data set

Antti Sorjamaa · Amaury Lendasse ·  
Yves Cornet · Eric Deleersnijder

Received: 20 April 2008 / Accepted: 22 January 2009 / Published online: 12 February 2009  
© Springer Science + Business Media B.V. 2009

**Abstract** In this paper, an improved methodology for the determination of missing values in a spatiotemporal database is presented. This methodology performs denoising projection in order to accurately fill the missing values in the database. The improved methodology is called empirical orthogonal functions (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

### 1 Introduction

Meteorology and climate research are two rapidly growing fields with an increasing need for accurate and

large measurement data sets. 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 food for the people, but it also gives thousands of workers jobs.

The importance to the people and the extraordinary size and shape of the lake make it really valuable for climate research, but the size also brings difficulties. The size and the shape of the lake make it hard to adequately measure the biogeophysical parameters, such as surface temperature. However, due to the current political and economical situation in Africa, the satellite is the only valid option. The data measured by satellite include 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 already been developed for solving the problem by filling the missing values, for example, Kriging [2] and several other optimal interpolation (OI) 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.

---

A. Sorjamaa (✉) · A. Lendasse  
Adaptive Informatics Research Centre,  
Helsinki University of Technology, Espoo, Finland  
e-mail: Antti.Sorjamaa@hut.fi

Y. Cornet  
Unit of Geomatics, University of Liege, Liege, Belgium

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

**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
- 

**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
- 

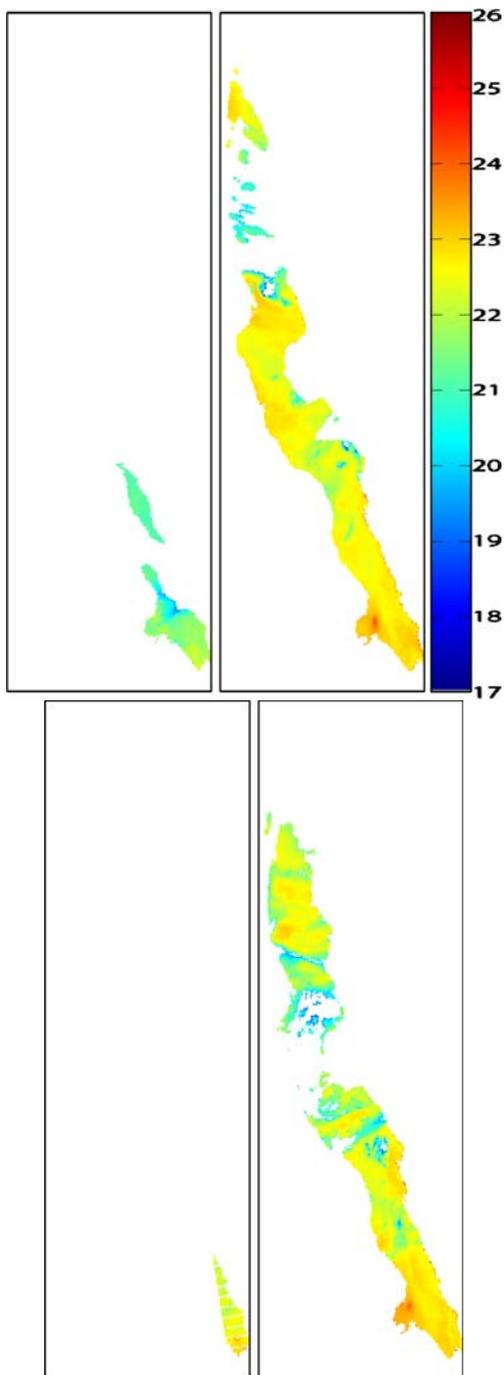
**Table 3** Forward selection strategy

- 
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:  

$$X^s = \arg \min_{X^i} \text{VALIDATION}(\{S, X^i\}), \quad X^i \in I, \quad i = 1, \dots, K$$
 Save  $\text{VALIDATION}(\{S, X^s\})$  and move  $X^s$  from  $I$  to  $S$   
 Continue the same way, until the size of  $S$  is  $K$
  3. (Result)  
Compare the  $\text{VALIDATION}$  values for all sizes of sets  $S$ , the selection result is the set  $S$  that minimizes the  $\text{VALIDATION}$
-

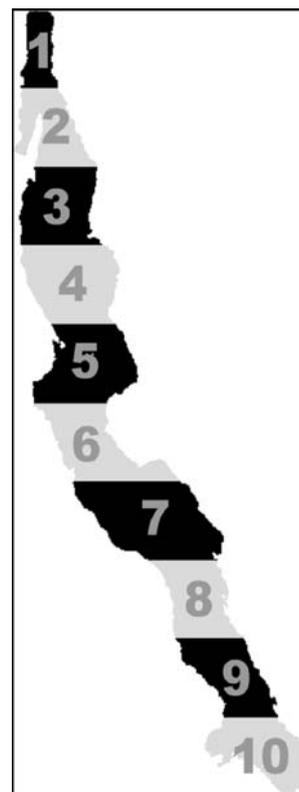
In this paper, an improvement on the standard EOF method is presented, called EOF pruning. It enhances the accuracy of the EOF methodology and even speeds up the calculation process.

Sections 2 and 3 present the EOF methodology and the EOF pruning improvement, respectively. In



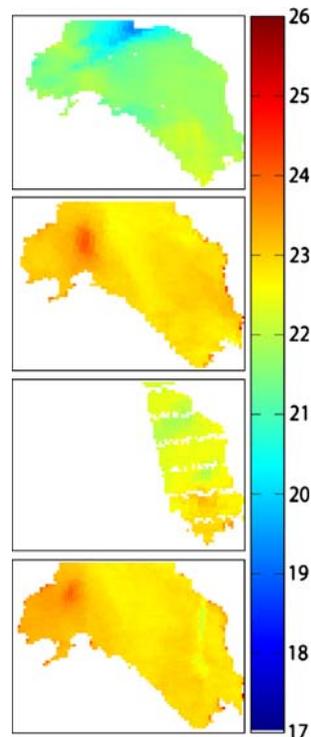
**Fig. 1** Tanganyika Lake data set, four example measurements from 2005. Days 182 and 183 are on the *top* from left to right and days 184 and 185 are on the *bottom*

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



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 data set.

**Fig. 3** Southern slice of the Tanganyika Lake data set. The days are the same as in Fig. 1, days 182 to 185 from 2005



**Table 4** Overall statistics of the data sets covering the whole lake after the days with more than 90% 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

## 2 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),

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

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  $\rho$  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 a 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 as the original one. The larger the  $q$  that 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.

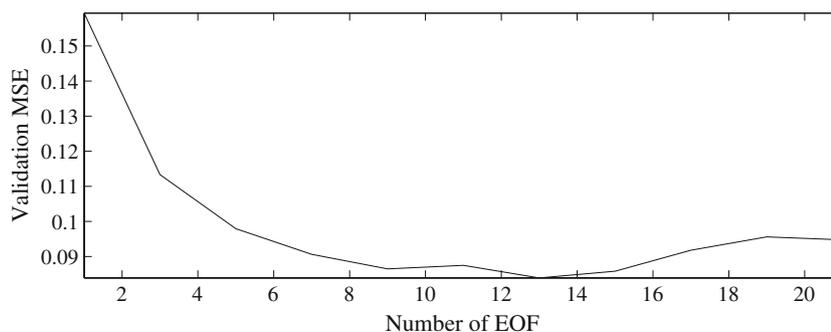
After the initial value replacement, the EOF process begins by performing the SVD and the selected  $q$  singular values, and vectors are used to build the reconstruction. In order not to lose *any* information, only the missing values of  $\mathbf{X}$  are replaced with the values from the reconstruction. 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

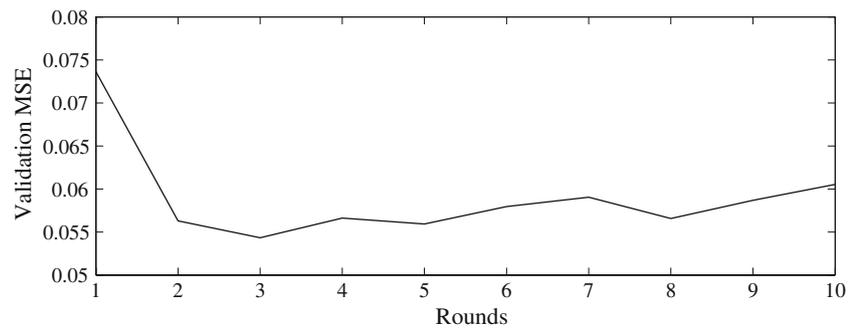
In some cases, some of the biggest singular values contain noise levels that are so large that they disturb the selection process described in Table 1. For example, if the first  $n$  singular values are selected by the validation procedure, but not the  $n + 1$ , it does not necessarily mean that all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

**Fig. 4** Validation errors using the EOF method



**Fig. 5** Validation errors using the EOF pruning method



set with the cost of accuracy, then the smaller values should not be used. However, 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 noncontinuous 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. This 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<sup>1</sup> 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.

### 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.

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 noncontinuous 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.

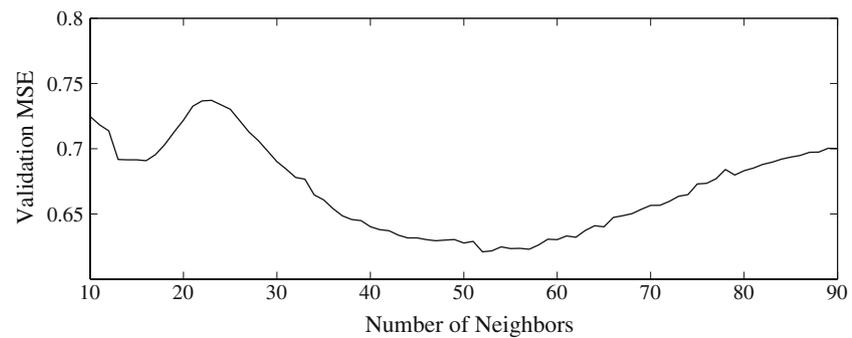
To validate the parameters, we use the Monte-Carlo cross-validation method [9] with ten folds. Each fold consists of 2.5% of the known data of the full data set, and only one fold of validation data is removed at a time. The removal of the validation data is not completely random, but instead, it is performed using artificial clouds placed randomly in the data set. These

**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			

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

**Fig. 6** Ten-fold Monte Carlo cross-validation errors using the objective analysis and the OAX toolbox



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 Eq. 2.

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

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  $\text{MC}_v$  and  $\mathbf{x}$  is the known measurement value, and  $\hat{\mathbf{x}}$  is the filled approximation.

## 5 Experiments

### 5.1 Tanganyika Lake surface temperature data set

For the comparison of the performances of the presented methods, we use a Tanganyika Lake surface temperature data set.<sup>2</sup> The lake lies in the African Rift area, and it is over 670 km long with an average width of about 50 km. The measurements are obtained from the thermal infrared bands of the MODIS sensors of the satellite covering the lake with a spatial resolution of 1 km.

The satellite has measured the lake a total of 666 times between the years 2002 and 2006. The measuring frequency of the satellite was not constant during the 5-year period; instead, it varied from 1 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% to 4%, meaning that some images have no measurement values and some have only 4% of the data missing. Finally, the whole data set has over 63% of the data missing.

Before applying the methodology, each image is pre-processed, 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 amounts of missing values. The same example days are used in each figure later on.

Because of the huge size of the data set, it is divided into slices. Each measurement image of the lake is cut into ten pieces in the 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 Fig. 3 and is used as an example slice in the following.

Because of the large amount of missing data in the database, each day with more than 90% of the data missing from the slice is removed from the data set before the learning phase. This is done for each slice individually. For the southernmost slice of the

**Table 6** Validation and test errors for the EOF, the EOF pruning, and the OA using the southern slice of the Tanganyika data set

	Validation MSE	nEOF	Test MSE
EOF	0.0839	13	0.0664
EOF pruning	0.0543		0.0517
Objective analysis	0.6210		0.6265

<sup>2</sup>Tanganyika Lake data set, MODIS Data, a courtesy of Yves Cornet and the University of Liege, Belgium. The data come from an RS data set produced in the framework of the CKIMFISH project.

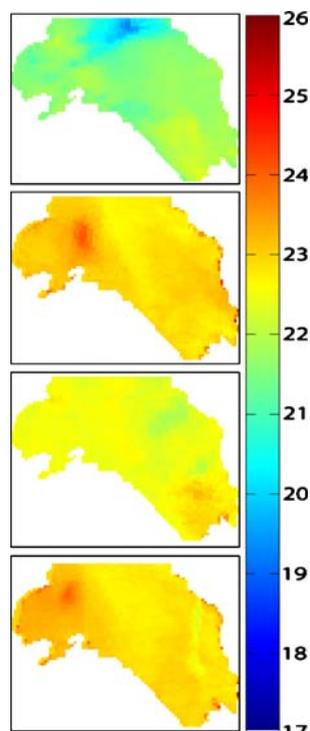
Tanganyika Lake, it leaves us 390 days, with each day containing a total of 2,947 measurement points, with a total of 27% of the data missing on average.

However, this kind of removal prohibits us from filling every day of the data set. In order to finally perform the complete filling of the data set, 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% of the known data is removed from each slice separately. For the validation of the parameters for the methods, Monte-Carlo cross-validation method with ten folds is used. Each fold contains 2.5% 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% of the data missing from the southernmost slice when the validation is started. The overall statistics for the used data sets are collected in Table 4.

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 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.

**Fig. 7** Southern part of the Tanganyika Lake data set. The pictures are the same examples filled by the EOF pruning as in Fig. 3



**Table 7** Results of all slices with all methodologies

	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

The percent number after each slice number is the amount of missing data in the slice before the processing starts.

### 5.2 Optimal interpolation

For the comparisons with the EOF and EOF pruning methodologies, a well-known OI method, called objective analysis [14], is used. The method belongs to the large number of OI methods, and it is applied using the OAX toolbox.<sup>3</sup>

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

<sup>3</sup>More information of the specifics of the methodology and the toolbox can be found in [http://www.mar.dfo-mpo.gc.ca/science/ocean/coastal\\_hydrodynamics/Oax/oax.html](http://www.mar.dfo-mpo.gc.ca/science/ocean/coastal_hydrodynamics/Oax/oax.html).

and a pseudodistance measure. The used covariance function is defined as

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

and the pseudodistance as

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

where  $x_i$  and  $y_i$  are the  $i$ th 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 also requires 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

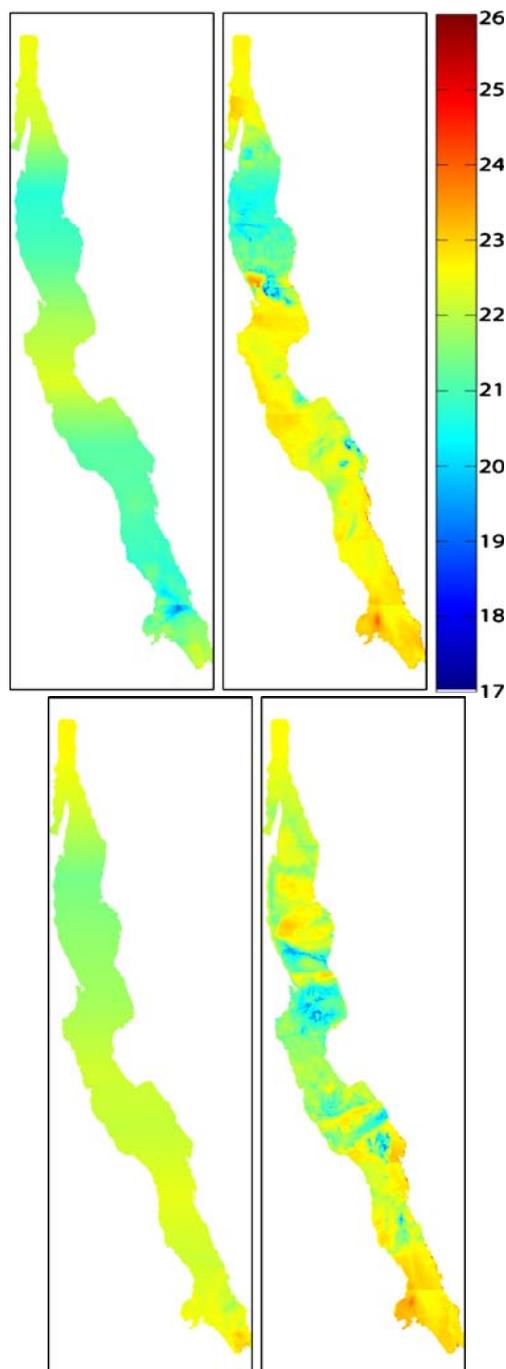
The validation results for the EOF method using the southern slice are shown in Fig. 4. From Fig. 4, the optimal number of EOFs is 13. This is a small number compared to the maximum of 390 EOFs suggesting a strong noise influence in the data.

For the EOF pruning method, the results are shown in Fig. 5 and Table 5. From Fig. 5, we can see that three rounds are enough for the EOF pruning. After the third round, the validation error starts to increase slowly.

From Table 5, we can see that the selected amount of singular values and vectors by the EOF pruning varies. In the first round, when the largest drop in the validation error is observed, the least amount of singular values and vectors is 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 EOFs, which are used as a maximum in the forward selection. This suggests that almost all of the EOFs 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 as with the EOF and EOF pruning methods, and the results of the validation can be seen in Fig. 6.

From Fig. 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.



**Fig. 8** Tanganyika Lake data set. The pictures are the same examples filled by the EOF pruning as in Fig. 1

From the summary in 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 2 days can vary from 1 to 33 days, and, therefore, the neighborhood calculation is not reliable enough. The test error is on the same level as 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%. 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 data set, 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 Fig. 8.

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

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

The final filling in Fig. 8 shows some evidence of the slicing, which emerge as slight discontinuities between slices. The smoothness is still better than between some original data points and filled ones. Many of those discontinuities can already be found in the original data set, as can be seen from the lower right picture in Fig. 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

**Table 8** Mean test errors for the whole Tanganyika Lake

	Test MSE
EOF	0.0858
EOF pruning	0.0551
EOF pruning unsliced	0.1120

Unsliced means that the EOF pruning has been applied to the whole lake without slicing it to ten slices.

**Table 9** Relative calculation times

	Relative calculation time
EOF	1
EOF pruning	0.23
EOF pruning unsliced	0.41
Objective analysis	0.05

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

original method in comparison, reducing the test error of the EOF by roughly 35%.

For the sake of comparison, the EOF pruning is applied to the whole lake, without slicing. Otherwise, the methodology is the same as 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 the north–south direction, the underlying dynamics are very different and, as Table 8 shows, using the full lake does not achieve as good accuracy as the slicing. Slicing the lake into ten pieces roughly halves the test MSE when using the EOF pruning. Table 9 shows an approximation of the calculation times of the used methods.

The estimated calculation times show that not only does the EOF pruning achieve better accuracy, but it also does it roughly four times faster. Surprisingly, the unsliced EOF pruning only consumes twice the calculation time as 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.

## 6 Conclusion

We have presented an improved methodology for finding missing values in a spatiotemporal database. The improved EOF pruning is compared with the original EOF using a spatiotemporal database with a 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 spatiotemporal databases.

For further work, the EOF pruning will be applied to other data sets in the field of climatology. The methodology will also be applied to data sets from

other research fields, for example, process industry and finance.

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

## References

1. Tangang, F.T., Tang, B., Monahan, A.H., Hsieh, W.W.: Forecasting enso events: a neural network—extended eof approach. *J. Climate* **11**, 29–41 (1998)
2. Wackernagel, H.: *Multivariate Geostatistics—An Introduction with Applications*. Springer, Berlin (1995)
3. Preisendorfer, R.: *Principal Component Analysis in Meteorology and Oceanography*. Elsevier, Amsterdam (1988)
4. Beckers, J.M., Rixen, M.: Eof calculations and data filling from incomplete oceanographic datasets. *J. Atmos. Ocean. Technol.* **20**(12), 1839–1856 (2003)
5. Boyd, J., Kennelly, E., Pistek, P.: Estimation of eof expansion coefficients from incomplete data. *Deep Sea Res.* **41**, 1479–1488 (1994)
6. Alvera-Azcarate, A., Barth, A., Rixen, M., Beckers, J.M.: Reconstruction of incomplete oceanographic data sets using empirical orthogonal functions. Application to the adriatic sea. *Ocean Model.* **9**, 325–346 (2005)
7. Alvera-Azcarate, A., Barth, A., Beckers, J.M., Weisberg, R.H.: Multivariate reconstruction of missing data in sea surface temperature, chlorophyll and wind satellite fields. *J. Geophys. Res.* **112**, C03008 (2007)
8. Beckers, J.-M., Barth, A., Alvera-Azcarate, A.: Dineof reconstruction of clouded images including error maps. Application to the sea surface temperature around Corsican Island. *Ocean Sci.* **2**(2), 183–199 (2006)
9. Lendasse, A., Wertz, V., Verleysen, M.: Model selection with cross-validations and bootstraps—application to time series prediction with rbf models. In: LNCS, no. 2714, pp. 573–580. Springer, Berlin (2003)
10. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, vol. 2, Montreal, 20–25 August 1995
11. Efron, B., Tibshirani, R.J.: *An Introduction to the Bootstrap*. Chapman and Hall, London (1993)
12. Efron, B., Tibshirani, R.J.: Improvements on cross-validation: the .632+ bootstrap method. *J. Am. Stat. Assoc.* **92**, 548–560 (1997)
13. Sorjamaa, A., Hao, J., Reyhani, N., Ji, Y., Lendasse, A.: Methodology for long-term prediction of time series. *Neurocomputing* **70**(16–18), 2861–2869 (2007)
14. Gandin, L.S.: *Objective Analysis of Meteorological Fields*, p. 242. Israel Program for Scientific Translations, Jerusalem (1969)