water

MDPI

Article

TIEOF: Algorithm for Recovery of Missing Multidimensional Satellite Data on Water Bodies Based on Higher-Order Tensor Decompositions

Leonid Kulikov [1,2], Natalia Inkova [2], Daria Cherniuk [1,2], Anton Teslyuk [2,3] and Zorigto Namsaraev [3,*]

[1] Skolkovo Institute of Science and Technology, 121205 Moscow, Russia; kulikov.la@phystech.edu (L.K.); daria.cherniuk@skoltech.ru (D.C.)
[2] Moscow Institute of Physics and Technology, 141701 Moscow, Russia; inkova.ns@phystech.edu (N.I.); anthony.teslyuk@grid.kiae.ru (A.T.)
[3] NRC "Kurchatov Institute", 123182 Moscow, Russia
* Correspondence: zorigto@gmail.com

Abstract: Satellite research methods are frequently used in observations of water bodies. One of the most important problems in satellite observations is the presence of missing data due to internal malfunction of satellite sensors and poor atmospheric conditions. We proceeded on the assumption that the use of data recovery methods based on spatial relationships in data can increase the recovery accuracy. In this paper, we present a method for missing data reconstruction from remote sensors. We refer our method to as Tensor Interpolating Empirical Orthogonal Functions (TIEOF). The method relies on the two-dimensional nature of sensor images and organizes the data into three-dimensional tensors. We use high-order tensor decomposition to interpolate missing data on chlorophyll a concentration in lake Baikal (Russia, Siberia). Using MODIS and SeaWiFS satellite data of lake Baikal we show that the observed improvement of TIEOF was 69% on average compared to the current state-of-the-art DINEOF algorithm measured in various preprocessing data scenarios including thresholding and different interpolating schemes.

Keywords: satellite observations of water bodies; missing data reconstruction; higher-order tensor decomposition; chlorophyll; lake baikal

Citation: Kulikov, L.; Inkova, N.; Cherniuk, D.; Teslyuk, A.; Namsaraev, Z. TIEOF: Algorithm for Recovery of Missing Multidimensional Satellite Data on Water Bodies Based on Higher-Order Tensor Decompositions. *Water* **2021**, *13*, 2578. https://doi.org/10.3390/w13182578

Academic Editors: Monica Pinardi, Mariano Bresciani, Viktor Tóth and Igor Ogashawara

Received: 14 July 2021
Accepted: 9 September 2021
Published: 18 September 2021

1. Introduction

Satellite research methods are actively involved in observations of water bodies. They are used for a wide variety of tasks, including study of lake morphodynamic characteristics (water level, surface, volume) [1], surface temperature [2], chlorophyll content, dynamics of phytoplankton and primary production [3], protection of endangered species [4], etc. However, satellite methods for observing water bodies have their own drawbacks. One of the most important problems in satellite observations is the presence of missing data due to internal malfunction of satellite sensors and poor atmospheric conditions.

Missing value reconstruction is a problem known for decades that emerges in a vast range of research areas [5,6]. When analyzing measurements that contain partially missing information, a researcher has to adjust data analysis methods to take into account the missing data. A common approach for dealing with missing data is data imputation, where missing values are filled with new data reconstructed from observations. To make the reconstruction, one needs to define a model which will capture particular patterns in observed data and use the model to fill the gaps in measurements. Various approaches exist to model the missing data, including a rich set of statistical and probabilistic methods [7], factor analysis [8], deep learning networks [9], autoencoders [10], and many more.

One of the most widely used methods to model remote sensing data with missing information is the Data Interpolating Empirical Orthogonal Functions (DINEOF) algorithm [11].

It relies on the empirical orthogonal functions (EOF) method which approximates the data using low dimensional linear decomposition. The DINEOF algorithm is closely related to principal component analysis method (PCA) [12] which is one of the most common methods for low-dimensional data representation. A number of modifications for PCA were proposed to account for missing data. A detailed review of them was presented in [13]. PCA with missing data allows various formulations including: least square problem [14], alternating algorithm [15], probabilistic model (PPCA) [16], and variational Bayesian method (VBPCA) [17].

When applied to spatio-temporal data, EOF- and PCA-based methods represent the data using two-dimensional matrix where one dimension represents timeline and the other represents a flattened n-dimensional measurement. When dealing with remote sensing data like satellite images, three-dimensional tensors preserve a two-dimensional measurement without unraveling it into a single vector. One dimension is used for the timeline and the other two for longitude and latitude coordinates of satellite images. In this case, the information about 2D spatial data structure does not get lost, and one can take advantage of it in further analysis. Although there is no single straightforward generalization of singular matrix decomposition underlying the EOF and PCA methods to the case of multidimensional tensors, a number of methods have been developed for the latter to find low-rank representations of multidimensional data [18] which include canonical polyadic decomposition (CPD) [19], Tucker decomposition [19,20], and its generalization as a higher-order singular value decomposition (HOSVD) [20,21].

In this paper we show that the use of the tensors for data representation and high-order tensor decomposition can significantly improve the quality of missing information recovery compared to the DINEOF algorithm. We refer to our algorithm as the Tensor Interpolating Empirical Orthogonal Functions (TIEOF).

We tested the developed algorithm using satellite data on the concentration of chlorophyll a in Lake Baikal obtained within the period from 1998 to 2020. Lake Baikal is a UNESCO world heritage site situated in southern Siberia (Russia) (coordinates 53.5 N, 108 E). It is the world's largest freshwater lake by volume, containing 20% of the world's fresh surface water, and it is also the world's seventh-largest lake by surface area. Lake Baikal is the oldest (25 million years) and deepest (1700 m) lake in the world [22]. Since 2011, the blooms of filamentous green algae have regularly been observed in the nearshore zone of the lake. Besides, since 2016 severe cyanobacterial blooms have been observed in the coastal and pelagic zones of the lake [23]. Given the large size of the lake, satellite data provide valuable information for studying the long-term trends in lake eutrophication, but the amount of missing satellite data makes this challenging. We hope that our algorithm will make it possible to assess the long-term trends in the development of lake ecosystems more accurately, which in turn is necessary to take effective measures to preserve them.

2. Materials and Methods

2.1. Multi-Way Tensor Decompositions

Tensor decompositions is an area that has been actively explored over the last decades [24]. Having emerged in the 1920s tensor decompositions were later successfully applied in a variety of fields: signal processing [18,25], image processing [26,27], biomedical data analysis [28,29], bioinformatics [30], geophysics [31], and topic modelling [32,33] have added tensor decomposition use cases. The first work, introducing CP-decomposition, was published back in 1927 [19]. Then it was reinvented in 1970 in two independent research papers: CANDECOMP [34] and PARAFAC [35] in psychometrics and linguistics domains. After that, a significant number of papers have been published, both introducing new methods and building upon the old ones. A great overview of methods and their applications for data analysis is given by Tamara Kolda in [36].

In this work we employed three popular decomposition methods: Higher Order Singular Value Decomposition (HOSVD), Higher Order Orthogonal Iteration (HOOI), and PARAllel FACtors (PARAFAC).

HOSVD [20,21] is the simplest generalization of renowned Singular Value Decomposition (SVD) on tensors of order higher than 2. The main idea here is to find Empirical Orthogonal Functions (EOFs) and a kernel tensor of the same order as the original data which contains singular values of different tensor unfoldings. When combining all that information with mode products we get a reconstruction according to Figure 1.

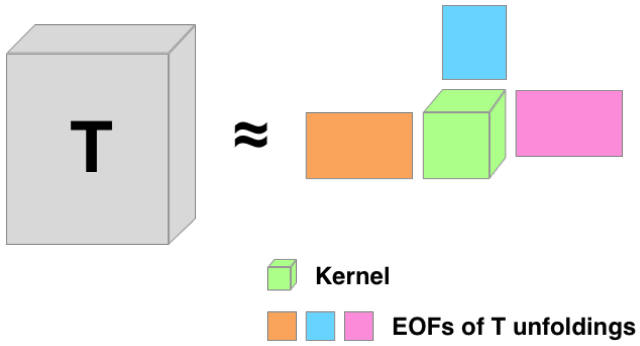

Figure 1. HOSVD.

PARAFAC [35] is a decomposition of a multi-way tensor on the sum of tensors of rank 1 (i.e., sum of outer products of vectors) like in Figure 2. To calculate those vectors the Alternating Least Squares (ALS) algorithm [37] can be used. It is simple yet an effective algorithm which iteratively fine-tunes part of the vectors while keeping others fixed [38]. Such a sum of outer products can actually be formulated as a mode product with diagonal kernel similar to HOSVD.

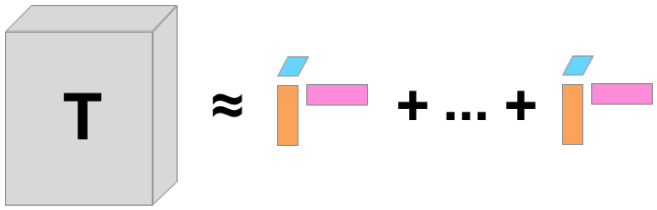

Figure 2. PARAFAC.

The HOOI [39] algorithm simply combines ideas from both decompositions: HOSVD and PARAFAC. Firstly, we use HOSVD to initialize a kernel and EOFs of unfoldings and then apply the ALS algorithm to further improve the approximation.

2.2. Interpolation on a Static Rectangular Grid

Every satellite image contains a set of points with chlorophyll a measurements made along the path of the satellite at a certain date and time. Coordinates of these points are not constant in time and change with every image. As the first step of data preprocessing, we interpolated the measurements on a rectangular grid, where each node corresponds to a certain pair of coordinates and steps along each axis are constant. The grid is universal for all images in the timeline. We have used two methods of interpolation: regression with k-nearest neighbours and regression with neighbours within a fixed radius [40]. In both methods the weight of a neighbouring point was calculated as the inverse of its proximity. The rationale under using two interpolation methods is that the satellite data tends to be sparse and sometimes the k-nearest neighbours method should employ far away points to make a prediction. The comparison of two methods is given in the Results section.

2.3. Data Thresholding

The next preprocessing step is satellite signal thresholding, clipping signal with high chlorophyll a concentration. In situ studies [41,42] showed that chlorophyll a concentration determined using satellites can be significantly overestimated due to presence of

coloured organic matter. In particular, a large discrepancy was observed in Lake Baikal at concentrations of chlorophyll a higher than 2 mg/m^3. To reduce the effect of satellite chlorophyll a overestimation we have replaced satellite measurements exceeding 2 mg/m^3 to the threshold value of 2 mg/m^3. The value of the threshold was motivated by the in situ HPLC determination of chlorophyll a concentration in lake Baikal water and comparison with satellite observations performed by Heim [41]. In this way we introduce some mistrust to the raw satellite data, but at the same time reduce the noise coming from satellite chlorophyll a estimation algorithm OC3M [43]. In the Results section we show that such a threshold increases the accuracy of every algorithm in the validation subset.

2.4. Validation Dataset

The performance of all the models used in reconstruction was measured by calculating the error on a separate validation subset (5% of the available data) with known chlorophyll a measurements. As chlorophyll a concentration tends to be zero in the majority of points, sampling measurements uniformly would result in a validation subset that mostly consists of zeros. Such set is not representative for reconstruction quality evaluation, since zero is an initial guess for missing values (Section 2.5). Instead, we used a stratified sampling approach: data points were grouped into several bins according to their chlorophyll a values with the K-Means clustering method [44]. The number of bins were chosen to be 10. This number is highly dependent on the dataset, thus we have tried several options: 3, 5, and 10 on MODIS Aqua and found 10 to be the most appropriate as it ensures that each bin contains measurements with close enough magnitude. Then an equal number of points from each bin were sampled and added to the validation set. Such procedure provides a closer to uniform distribution of chlorophyll a values within the validation dataset. While reconstructing data, these points are marked as missing. This allows us to compare reconstructed chlorophyll a values to the reserved satellite measurements.

2.5. Missing Data Reconstruction

We have devised a novel approach to satellite data reconstruction—TIEOF algorithm (Figure 3). The core idea is similar to DINEOF: to reconstruct missing values with the help of low-rank data representation. However, DINEOF relies on matrices for data representation and uses SVD decomposition to find the best low-rank approximation, while TIEOF works with three dimensional tensors and utilizes tensor decomposition methods. In our work we have used three different multi-way tensor decompositions: Truncated Higher Order Singular Value Decomposition (HOSVD) [20,21], Higher Order Orthogonal Iteration (HOOI) [39], and Parallel Factors (PARAFAC) [38].

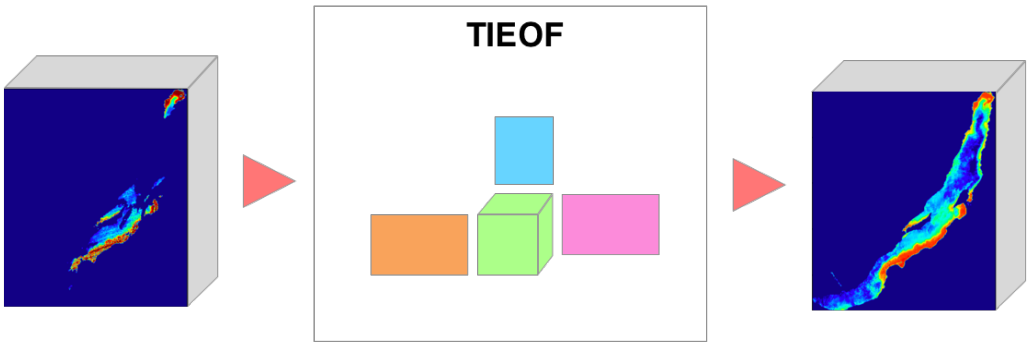

Figure 3. Simplified visual representation of TIEOF. The algorithm takes a batch of satellite images with missing values as an input, performs a series of tensor decompositions, and outputs reconstructed data.

The TIEOF pseudo-code is presented in Algorithm 1. The input is an array of satellite images—a three dimensional tensor *Data* with dimensions (lat, lon, t) which correspond to latitude, longitude, and time respectively. At the initial step we calculate an average

along spatial dimensions (latitude, longitude) as well as an average along time dimension. Then subtract these averages from initial data to get it centered along spatial and time axes. All averages are calculated ignoring missing values. In PCA and EOF methods this step is usually performed implicitly while calculating correlation matrix. However, when using decompositions, we have to perform this computation explicitly [45].

Algorithm 1 TIEOF

Data—input set of images (lat, lon, t)
Mask—boolean mask of known pixels
r—number of components in reconstruction
eps—a threshold on r-rank approximation error
es—boolean variable defining whether to use early stopping or not
procedure TIEOF(*Data*, *Mask*, r, $eps = 1e-3$, $es = False$)
    $Data, averages \leftarrow Center(Data)$ ▷ Substract averages on (lat, lon) and (t) dimensions
    $Data[!mask] = 0$ ▷ Initial guess for unknown values
    $Error \leftarrow \infty$
    while $Error > eps$ and $\nabla Error > eps$ if es else $True$ **do**
        K_r, F_r = [HOSVD | HOOI | PARAFAC]($Data, r$)
        $Data_r = Reconstruct(K_r, F_r)$
        $Data_r[mask] = Data[mask]$
        $Error \leftarrow NRMSE(Data[!mask], Data_r[!mask])$
        $Data \leftarrow Data_r$
    $Data \leftarrow !Center(Data, averages)$ ▷ Add memorized averages back
    return *Data*

In the next step we fill missing data with zero values. Then an iterative algorithm starts until convergence: tensor decomposition of the current representation of *Data* tensor is calculated using one of three algorithms (HOSVD, HOOI, PARAFAC). Tensor decomposition is used to estimate a reconstructed tensor $Data_r$. Reconstructed tensor is used to fill unknown values of the data. Reconstruction error is calculated as a normalized root mean squared error (NRMSE) difference of reconstructed values for missing data between successive iterations defined by Equation (1) ($y^{'}$—reconstructed missing values of the current iteration and y—reconstructed missing values from previous iteration).

$$NRMSE = \frac{\sqrt{E[(y' - y)^2]}}{\sigma[y]} \quad (1)$$

A large number of iterations with a low amount of training data can lead to severe overfitting. To avoid overfitting we have included an early stopping technique that is typically used in deep learning routines [46]. As the stopping criteria, we first used a threshold on NRMS error between *Data* and its r-rank approximation (Algorithm 1) on a subset of unknown points. Later, we introduced an early stopping technique [46], by putting a threshold on NRMS error gradients (absolute difference between r-rank reconstruction errors in successive iterations). Effects of early stopping are presented in the Results section. When the stopping criteria is satisfied, we cease iterations, fill missing values in the initial data with corresponding points in the last reconstructed tensor and apply reverse centering with the previously memorized averages.

2.6. The Lake Baikal Dataset

To test the TIEOF algorithm, we used three publicly available datasets distributed by NASA's Ocean Biology Processing Group. These three datasets contain products obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) located aboard the Terra (EOS AM) and Aqua (EOS PM) satellites and from the Sea-Viewing Wide Field-of-View Sensor (SeaWiFS) located aboard the OrbView-2 (SeaStar) satellite. MODIS Terra

provides data for the period from 24 February 2000, to the present day. MODIS Aqua—from 4 July 2002, to the present day. The SeaWiFS data is available for the period from 4 September 1997, to 11 December 2010. Each dataset provides data for each day in the corresponding period. For our experiments we selected data on chlorophyll a concentration (Level 2, 1 km resolution) for 93 days (1 June to 1 September) each year during 2003–2020 for MODIS Aqua, 2001–2020 for MODIS Terra, and 1998–2010 for SeaWiFS.

During preprocessing stage described in Section 2.2, we have interpolated satellite images onto a static spatial grid with fixed coordinates. Coordinates of the static grid were chosen to cover the whole region of interest (Baikal) with a close to 1 km density. Lake Baikal is located approximately between 51.5 degrees and 56 degrees latitude and between 104 degrees and 110 degrees longitude. This roughly corresponds to a rectangular area of 482 by 406 km. Taking into account a sample of 93 days for each year we sum up to the dataset tensors of shape (482, 406, 93). To extract only coordinates within the Lake Baikal borders in the downloaded pieces, we used the shapefile with Baikal coastline [47]. It allowed us to mask out the points outside the Lake borders.

The amount of missing information for the Baikal region differs between satellites. In MODIS Aqua and MODIS Terra the average amount of missing measurements throughout the year is around 75%, whereas the SeaWiFS shows the average of 95% missing values. The average data missing ratios for each satellite and year are depicted in Figure 4.

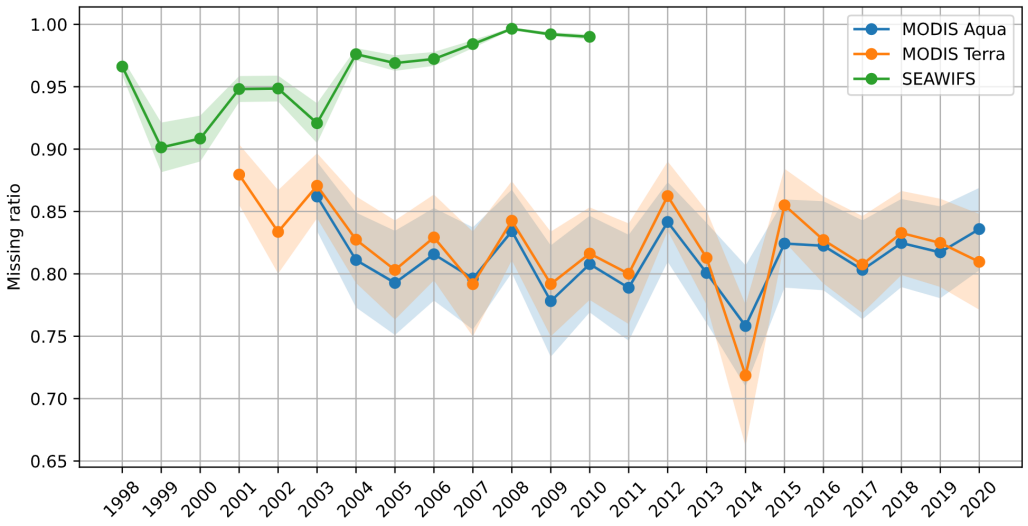

Figure 4. The average ratio of missing data for each year and satellite with a 95% confidence interval obtained by bootstrapping.

3. Results

3.1. Interpolation on a Rectangular Grid

To compare two interpolation methods of aligning satellite data on a constant rectangular grid (Section 2.2) , we have conducted the following experiment: we aligned MODIS Aqua data using the K-neighbors method with $K = 3$ and regression with neighbors within a 5 km radius, reconstructed missing data with all discussed methods, and evaluated NRMS error on the validation set. $K = 3$ was chosen due to high sparsity of data. A 5 km radius was chosen heuristically, but has proven to bring an improvement to the algorithm. Experiments with larger set of possible values were not conducted due to high computational costs.

Results are presented in Figure 5. With the exception of the time period from 2008 to 2015, interpolation with neighbours within a 5 km radius has resulted in a significant drop of NRMS error: by 9.5% for DINEOF; by 11.2% for TIEOF HOSVD; by 11.5% for TIEOF HOOI; by 11.3% for TIEOF PARAFAC.

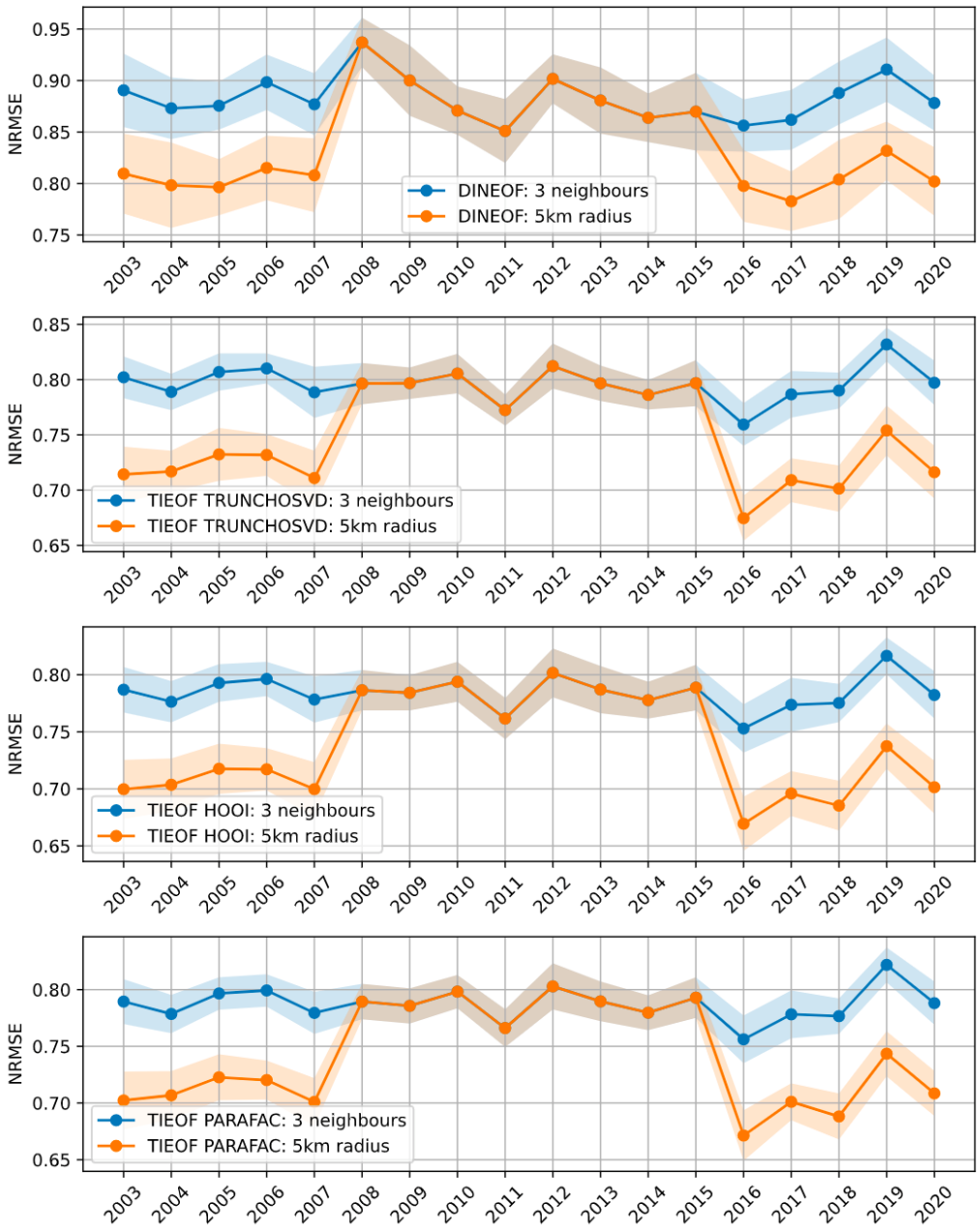

Figure 5. NRMS error on validation set with K-neighbours (3 neighbours) and neighbours within a radius (5 km radius). Shaded area defines a 95% confidence interval obtained by bootstrapping.

3.2. Effect of Early Stopping

To check the effect of the early stopping technique, we reconstructed the tensors from MODIS Aqua dataset using all of the algorithms and measured NRMS error on a validation subset. In this experiment we used the 5 km radius neighbours regression for interpolating on a static rectangular grid. The stopping criteria value for both reconstruction error and reconstruction error gradients between iterations was chosen to be 0.001.

The results are depicted in Figure 6. In the case of DINEOF, the early stopping leads to a significantly lower NRMSE (by 8.4% on average). In the case of TIEOF, the early stopping brings a slight increase in the NRMSE (by 0.3% for HOOI, by 0.4% for PARAFAC, by 1.8% for HOSVD), but all differences are within the confidence interval. Thus we conclude that TIEOF has the better resistance to the overfitting phenomenon than DINEOF, because full convergence mode is very close to early stopping mode (max difference is only 1.8%).

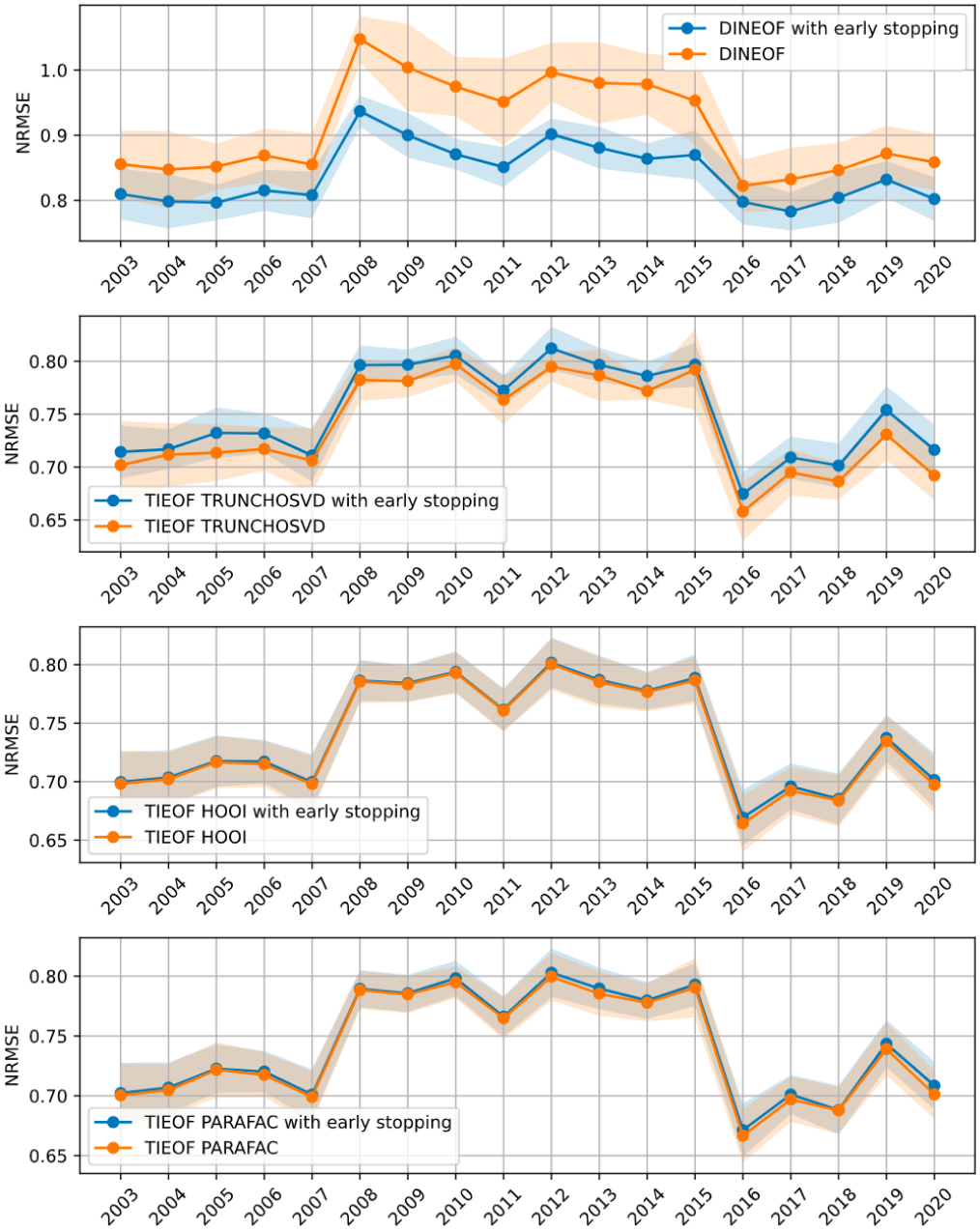

Figure 6. NRMS error on validation set with and without using early stopping. Shaded area defines a 95% confidence interval obtained by bootstrapping.

All further experiments were conducted with early stopping mode to slightly alleviate the computational burden.

3.3. Data Thresholding Reduces Reconstruction Errors

As described in Section 2.3, the satellite signal of chlorophyll a beyond 2 mg/m^3 tends to be greatly overestimated when compared to in situ values. To check the effect of data thresholding, we reconstructed the same dataset with and without threshholding step and compared NRMS errors on validation subset.

For this experiment we used the MODIS Aqua dataset, neighbours regression within 5 km radius to interpolate on a static rectangular grid, 2 mg/m^3 for threshholding step, and early-stopping for model convergence strategy. Reconstruction of initially thresholded tensor led to significantly lower NRMSE on the validation subset which is demonstrated in Figure 7. The average factor of NRMSE decrease between all years is presented in Table 1.

Table 1. Factor of improvement (NRMSE decrease) when using thresholding.

Method	Improvement Factor
DINEOF	1.24
TIEOF: HOSVD	1.56
TIEOF: PARAFAC	1.56
TIEOF: HOOI	1.57

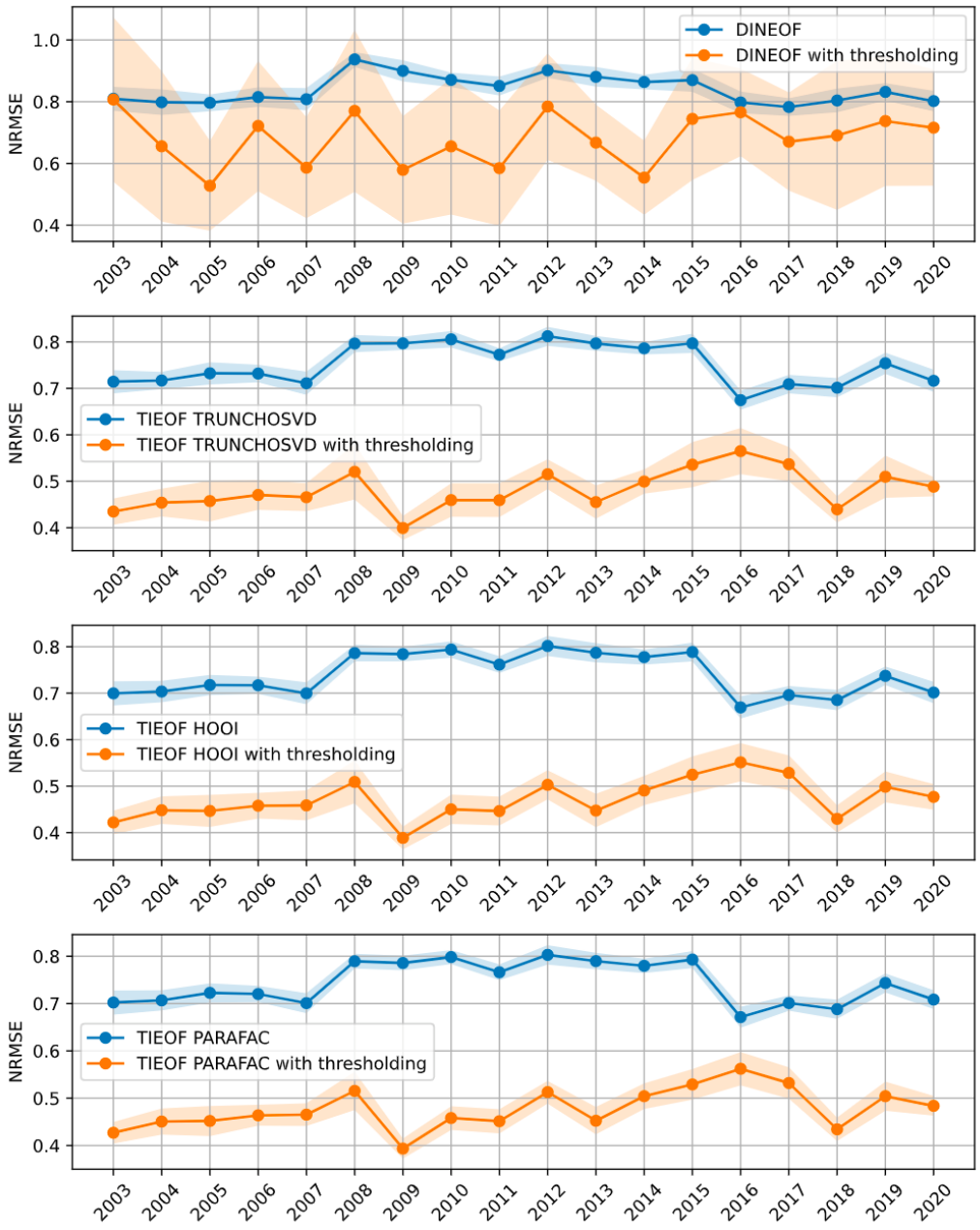

Figure 7. Effect of thresholding data signal above 2 mg/m^3. Shaded area defines a 95% confidence interval obtained by bootstrapping.

3.4. Comparison of the Tensor Decomposition Methods

To perform a fair comparison of DINEOF and TIEOF algorithms, we reconstructed data from all the available datasets and compared NRMS errors on validation subsets. All TIEOF variations performed better than the DINEOF on all the available datasets (Figure 8).

The difference between different TIEOF variations is not as significant (errors for different TIEOF variations are within each other's 95% confidence intervals) as compared to DINEOF. This shows us that TIEOF is a robust choice of tensor decomposition and its main benefit is not in the specific tensor decomposition but in the usage of original multidimensional feature space. We depicted the mean errors between all years for all datasets and methods in Table 2.

We observe that the improvement of TIEOF over DINEOF varies between 9 and 309% (min and max improvement among all years and TIEOF variations). High improvement is noticed when the data is poor, i.e., the error (NRMSE) is much larger (>25%) than a median error calculated in all years and low improvement is observed when the error is close to the median. Therefore, TIEOF provides a stable (69% on average) improvement compared to DINEOF, and the difference between methods is most evident on poor data where the DINEOF NRMS error increases.

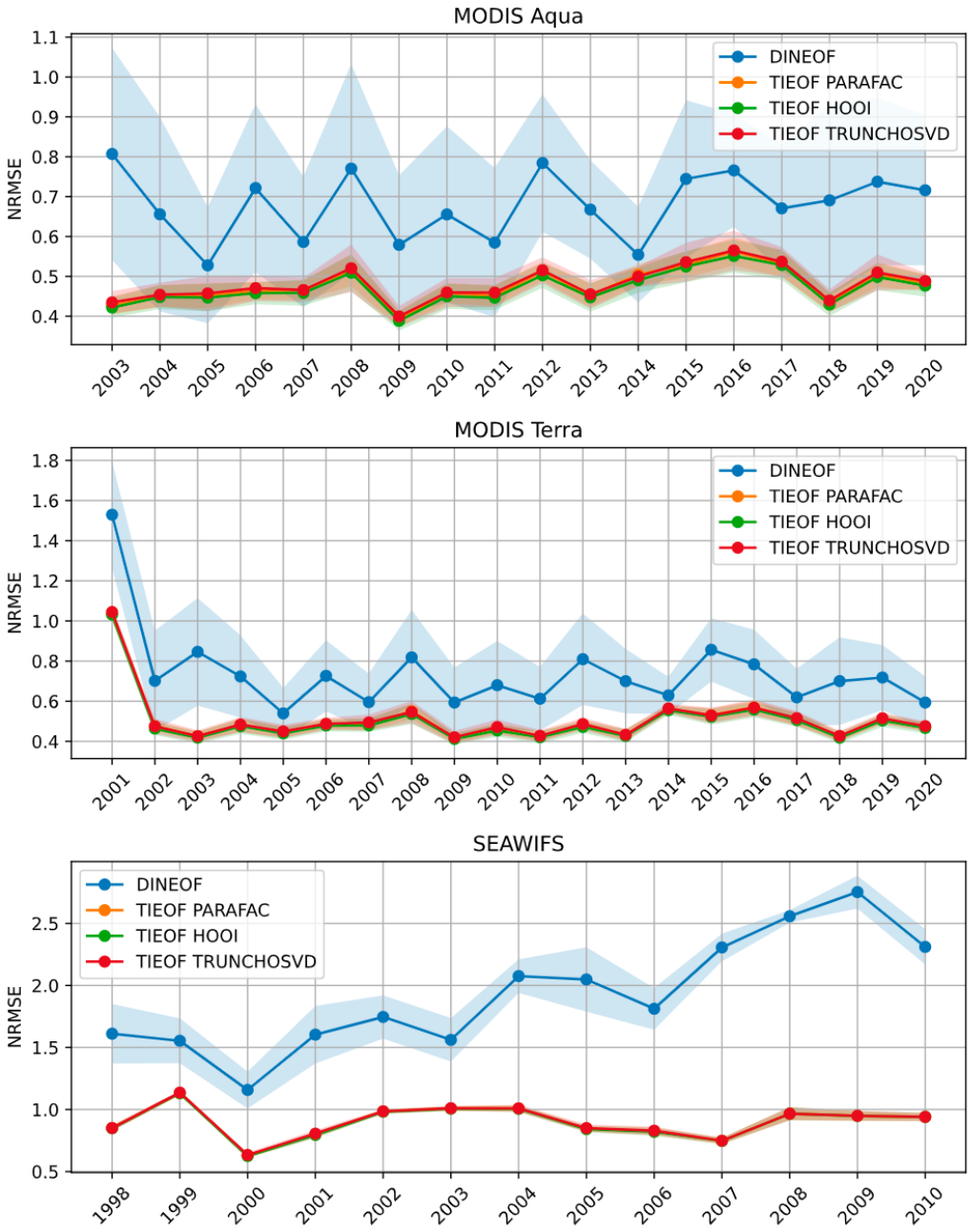

Figure 8. Performance comparison. Shaded area defines a 95% confidence interval obtained by bootstrapping.

Table 2. Mean errors (NRMSE).

Data	Method	Mean NRMSE
AQUA	DINEOF	0.68
	TIEOF: HOSVD	0.48
	TIEOF: PARAFAC	0.48
	TIEOF: HOOI	0.47
TERRA	DINEOF	0.74
	TIEOF: HOSVD	0.51
	TIEOF: PARAFAC	0.51
	TIEOF: HOOI	0.50
SEAWIFS	DINEOF	1.93
	TIEOF: HOSVD	0.90
	TIEOF: PARAFAC	0.90
	TIEOF: HOOI	0.90

4. Discussion

In Section 3.3 we show that both DINEOF and TIEOF methods demonstrate better results when using a signal thresholding step. We believe that cropping the data from satellite instruments reduced the error and brought it closer to the ground truth chlorophyll a levels. When compare Heim's in situ data [41] to the MODIS and SeaWIFS chlorophyll a estimations, we observe that discrepancy between them increases when concentration goes beyond 2 mg/m^3. For satellite data we get less error when we use a threshold value instead of the calculated chlorophyll a level value. As a result, we obtain more accurate data with less noise, which allows missing data recovery algorithms to find patterns and recover missing information more accurately. The need for this step and the choice of the threshold value of 2 mg/m^3 may be specific for the Baikal lake. For other lakes with different trophic ranges, different heights above sea level, and other parameters which affect the quality of chlorophyll a estimation, a separate analysis with in situ data is needed.

Our results demonstrate that tensor decompositions give consistently better results than the DINEOF matrix-based algorithm. This is not only proven quantitatively (Section 3.4) but also qualitatively. Figure 9 depicts reconstruction for two days. When the input is not completely empty both algorithms (DINEOF and TIEOF HOOI) are able to restore the missing data. Though the restorations differ, they both generated plausible results as chlorophyll a concentration increases in the shallower parts of lake Baikal near the coastline [23]. In the second row we show the reconstruction of the input that is fully empty. Probably, the weather was cloudy that day. This occludes the reception of a spectrometer on board the satellite. Still, data can be reconstructed using measurements from other days. We conclude that the reconstruction of TIEOF HOOI is better compared to DINEOF as the latter did not catch the difference between the coastline of the lake and its depths and assigned even smaller values of chlorophyll a near the shore. On the contrary, TIEOF HOOI provided the results that agree with the lake depth growth. The biggest difference is presented in Figure 10. It is worth mentioning that the reconstruction can have values above the threshold we applied earlier as no constraints were applied during the iterations, thus the bias (thresholding) that was employed before the restoration can be corrected by the algorithm.

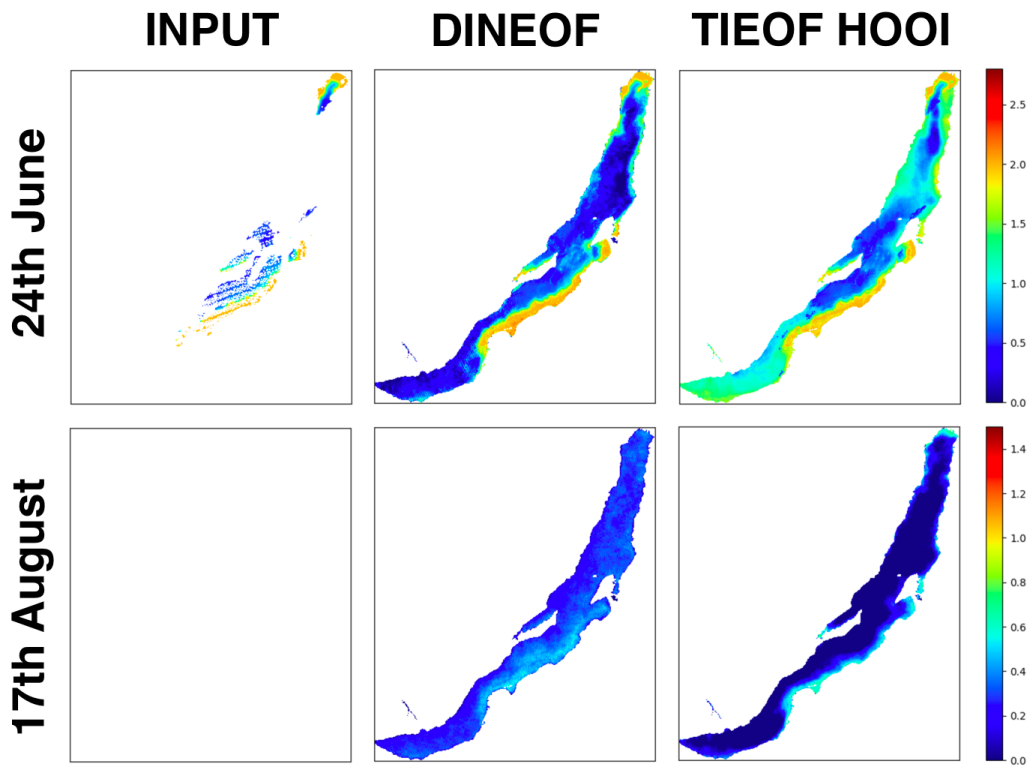

Figure 9. Example of data restoration for two days on MODIS Aqua summer 2012 dataset. In the second row input is completely unknown, but the data can still be reconstructed using measurements from other days. The validation NRMS error for TIEOF HOOI is 26% lower comparing to DINEOF.

While decomposing the data matrix with the DINEOF algorithm, we operate in two spaces: a space of vectors of length t (denoted as U in classical SVD notation) and a space of vectors of length lat times lon (denoted as V in classical SVD notation). Thus we either calculate the pixel-by-pixel correlation of satellite images between pairs of days or a day-by-day correlation between pairs of coordinates. Far away points are usually not correlated with each other, and close points correlate immensely. The DINEOF algorithm does not take into the consideration the proximity of elements in each pair. So the majority of point pairs provide no informative signal for the model. In the TIEOF algorithm, we decompose the data tensor onto spaces of vectors of lengths: *lat*, *lon*, and t. Thus, we compute correlations not between points within the whole spatial grid but between columns and rows of data which correspond to lines along the same latitude and longtitude on the map. This way, we take into account more features in images of different scope, not only correlations between single coordinates but also correlations between latitude and longtitude lines. This idea is close to the renowned convolutional neural networks [48]. The convolution operation (or more precisely cross-correlation) uses typically small patches to calculate features on the next layer. Thus the new features are constructed from only some local part of the information from the previous layer of a neural network. For high-level tasks (e.g., image classification) the features that are based on local regions surely are not enough. A stacking of convolutional layers leads to a gradual increase of a receptive field which is important for the abstractness of features on deep levels. Based on these abstract features it is intuitively easier for a simple function (e.g., linear function) to classify the image. Yet for low-level tasks (e.g., super-resolution, image completion) even small shallow convolutional neural networks bring surprisingly good results [49]. Therefore, we make a conclusion that searching correlations between lines rather than points within the whole spatial grid is of crucial importance. Tensor decompositions employ exactly that treatment of data.

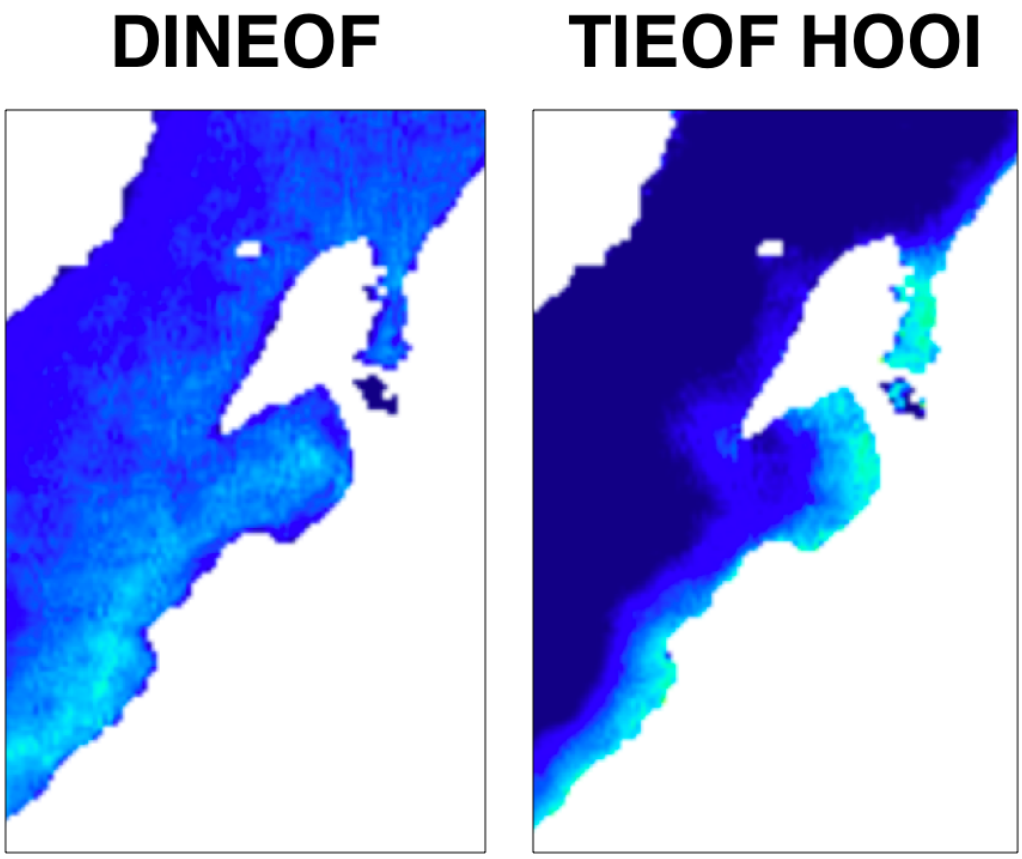

Figure 10. Close up area of the 2nd row in Figure 9.

5. Conclusions

In this work we proposed the new set of methods to reconstruct sparse data from satellite images and have proven their effectiveness on chrolophyll a measurements of Lake Baikal. In comparison with the renowned DINEOF, our methods have lower error (69% improvement on average). The significance of the error difference was proven with a 95% confidence interval.

We have also investigated additional preprocessing techniques: data thresholding, early stopping, and different interpolation methods to align satellite data to a static rectangular grid. In a series of experiments we have shown the following: data thresholding is an important part of the process, since it improves the quality significantly; early stopping decreases the reconstruction error for DINEOF but is unnecessary for the TIEOF since our algorithm has proven to be resistant to overfitting; regression with neighbours within a fixed radius appears to be a better approach than *K*-neighbours regression, but this is highly dependent on the dataset and can be proven wrong for some other satellite, measurement, or geographical area.

The TIEOF algorithm can be used to achieve a wide range of goals related to the restoration of missing satellite data. The algorithm can be used to analyze short-term and long-term trends, identify correlations between various geophysical variables, as well as for environmental modelling and solving applied problems of ecosystem management [50].

Author Contributions: Conceptualization, L.K., A.T. and Z.N.; methodology, L.K., D.C. and A.T.; software, L.K., D.C. and N.I.; writing—original draft preparation, L.K., N.I., A.T. and Z.N.; writing—review and editing, D.C., L.K., N.I., A.T. and Z.N.; visualization, L.K.; supervision, A.T. and Z.N.; project administration, Z.N.; funding acquisition, Z.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science and Higher Education of the Russian Federation, grant #075-15-2019-1659.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The source code has been made available on github (https://github.com/theleokul/tieof, accessed on 14 September 2021).

Acknowledgments: The work has been carried out using computing resources provided by NRC Kurchatov institute project "Development of modular platform for scientific data processing and mining" (Project No. 1571). Thanks to the developers of Algorithm: The Logic Game (https://apps.apple.com/us/app/algorithm-the-logic-game/id1475410194, accessed on 14 September 2021) for advising us on the topic of algorithms.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ALS	Alternating Least Squares
DINEOF	Data Interpolating Empirical Orthogonal Functions
EOF	Empirical Orthogonal Functions
HOOI	Higher Order Orthogonal Iteration
HOSVD	Truncated Higher Order Singular Value Decomposition
MODIS	Moderate Resolution Imaging Spectroradiometer
NRMSE	Normalized Root Mean Squared Error
PARAFAC	Parallel Factors
PCA	Principal component analysis
SeaWiFS	Sea-Viewing Wide Field-of-View Sensor
SVD	Singular Value Decomposition
TIEOF	Tensor Interpolating Empirical Orthogonal Functions

References

1. Crétaux, J.F.; Abarca-del Río, R.; Berge-Nguyen, M.; Arsen, A.; Drolon, V.; Clos, G.; Maisongrande, P. Lake volume monitoring from space. *Surv. Geophys.* **2016**, *37*, 269–305. [CrossRef]
2. Ganzedo, U.; Alvera-Azcarate, A.; Esnaola, G.; Ezcurra, A.; Saenz, J. Reconstruction of sea surface temperature by means of DINEOF: A case study during the fishing season in the Bay of Biscay. *Int. J. Remote Sens.* **2011**, *32*, 933–950. [CrossRef]
3. Bergamino, N.; Horion, S.; Stenuite, S.; Cornet, Y.; Loiselle, S.; Plisnier, P.D.; Descy, J.P. Spatio-temporal dynamics of phytoplankton and primary production in Lake Tanganyika using a MODIS based bio-optical time series. *Remote Sens. Environ.* **2010**, *114*, 772–780. [CrossRef]
4. Breece, M.W.; Oliver, M.J.; Fox, D.A.; Hale, E.A.; Haulsee, D.E.; Shatley, M.; Bograd, S.J.; Hazen, E.L.; Welch, H. A satellite-based mobile warning system to reduce interactions with an endangered species. *Ecol. Appl.* **2021**, *31*, e02358. [CrossRef] [PubMed]
5. Pigott, T.D. A Review of Methods for Missing Data. *Educ. Res. Eval.* **2001**, *7*, 353–383. [CrossRef]
6. Ibrahim, J.G.; Molenberghs, G. Missing data methods in longitudinal studies: A review. *Test* **2009**, *18*, 1–43. [CrossRef]
7. Little, R.J.; Rubin, D.B. *Statistical Analysis with Missing Data*; John Wiley & Sons: Vienna, Austria, 2019; Volume 793.
8. Kamakura, W.A.; Wedel, M. Factor analysis and missing data. *J. Mark. Res.* **2000**, *37*, 490–498. [CrossRef]
9. Zhang, Q.; Yuan, Q.; Zeng, C.; Li, X.; Wei, Y. Missing data reconstruction in remote sensing image with a unified spatial–temporal–spectral deep convolutional neural network. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 4274–4288. [CrossRef]
10. Jaques, N.; Taylor, S.; Sano, A.; Picard, R. Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction. In Proceedings of the 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), San Antonio, TX, USA, 23–26 October 2017; pp. 202–208.
11. Beckers, J.M.; Rixen, M. EOF calculations and data filling from incomplete oceanographic datasets. *J. Atmos. Ocean. Technol.* **2003**, *20*, 1839–1856. [CrossRef]
12. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1901**, *2*, 559–572. [CrossRef]
13. Ilin, A.; Raiko, T. Practical approaches to principal component analysis in the presence of missing values. *J. Mach. Learn. Res.* **2010**, *11*, 1957–2000.
14. Wiberg, T. Umea, Computation of Principal Components when Data are Missing. *Proc. Second Symp. Comput. Stat.* **1976**, 229–236.
15. Grung, B.; Manne, R. Missing values in principal component analysis. *Chemom. Intell. Lab. Syst.* **1998**, *42*, 125–139. [CrossRef]
16. Tipping, M.E.; Bishop, C.M. Probabilistic principal component analysis. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **1999**, *61*, 611–622. [CrossRef]

17. Bishop, C.M. Pattern recognition. *Mach. Learn.* **2006**, *128*, 580–583.
18. Sidiropoulos, N.D.; De Lathauwer, L.; Fu, X.; Huang, K.; Papalexakis, E.E.; Faloutsos, C. Tensor decomposition for signal processing and machine learning. *IEEE Trans. Signal Process.* **2017**, *65*, 3551–3582. [CrossRef]
19. Hitchcock, F.L. The expression of a tensor or a polyadic as a sum of products. *J. Math. Phys.* **1927**, *6*, 164–189. [CrossRef]
20. Tucker, L.R. Some mathematical notes on three-mode factor analysis. *Psychometrika* **1966**, *31*, 279–311. [CrossRef]
21. De Lathauwer, L.; De Moor, B.; Vandewalle, J. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.* **2000**, *21*, 1253–1278. [CrossRef]
22. UNESCO World Heritage Centre: Lake Baikal. Available online: https://whc.unesco.org/en/list/754 (accessed on 1 September 2021).
23. Namsaraev, Z.; Melnikova, A.; Ivanov, V.; Komova, A.; Teslyuk, A. Cyanobacterial bloom in the world largest freshwater lake Baikal. In *IOP Conference Series: Earth and Environmental Science*; IOP Publishing: Bristol, UK, 2018; Volume 121, p. 032039.
24. Cichocki, A.; Mandic, D.; De Lathauwer, L.; Zhou, G.; Zhao, Q.; Caiafa, C.; Phan, H.A. Tensor decompositions for signal processing applications: From two-way to multiway component analysis. *IEEE Signal Process. Mag.* **2015**, *32*, 145–163. [CrossRef]
25. Sidiropoulos, N.D.; Bro, R.; Giannakis, G.B. Parallel factor analysis in sensor array processing. *IEEE Trans. Signal Process.* **2000**, *48*, 2377–2388. [CrossRef]
26. Vasilescu, M.A.O.; Terzopoulos, D. Multilinear analysis of image ensembles: Tensorfaces. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2002; pp. 447–460.
27. Wang, Y.; Peng, J.; Zhao, Q.; Leung, Y.; Zhao, X.L.; Meng, D. Hyperspectral image restoration via total variation regularized low-rank tensor decomposition. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *11*, 1227–1243. [CrossRef]
28. Cong, F.; Lin, Q.H.; Kuang, L.D.; Gong, X.F.; Astikainen, P.; Ristaniemi, T. Tensor decomposition of EEG signals: A brief review. *J. Neurosci. Methods* **2015**, *248*, 59–69. [CrossRef]
29. Hamdi, S.M.; Wu, Y.; Boubrahimi, S.F.; Angryk, R.; Krishnamurthy, L.C.; Morris, R. Tensor decomposition for neurodevelopmental disorder prediction. In *International Conference on Brain Informatics*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 339–348.
30. Hore, V.; Viñuela, A.; Buil, A.; Knight, J.; McCarthy, M.I.; Small, K.; Marchini, J. Tensor decomposition for multiple-tissue gene expression experiments. *Nat. Genet.* **2016**, *48*, 1094–1100. [CrossRef] [PubMed]
31. McNeice, G.W.; Jones, A.G. Multisite, multifrequency tensor decomposition of magnetotelluric data. *Geophysics* **2001**, *66*, 158–173. [CrossRef]
32. Franz, T.; Schultz, A.; Sizov, S.; Staab, S. Triplerank: Ranking semantic web data by tensor decomposition. In *International Semantic Web Conference*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 213–228.
33. Anandkumar, A.; Ge, R.; Hsu, D.; Kakade, S.M.; Telgarsky, M. Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.* **2014**, *15*, 2773–2832.
34. Carroll, J.D.; Chang, J.J. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika* **1970**, *35*, 283–319. [CrossRef]
35. Harshman, R. *Foundations of the PARAFAC Procedure: Models and Conditions for an "Explanatory" Multi-Modal Factor Analysis*; UCLA Working Papers in Phonetics; University of California: Los Angeles, CA, USA, 1970.
36. Kolda, T.G.; Bader, B.W. Tensor Decompositions and Applications. *SIAM Rev.* **2009**, *51*, 455–500. [CrossRef]
37. Kroonenberg, P.M.; De Leeuw, J. Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika* **1980**, *45*, 69–97. [CrossRef]
38. Rabanser, S.; Shchur, O.; Günnemann, S. Introduction to Tensor Decompositions and their Applications in Machine Learning. *arXiv* **2017**, arXiv:1711.10781.
39. Sheehan, B.; Saad, Y. Higher Order Orthogonal Iteration of Tensors (HOOI) and Its Relation to PCA and GLRAM. In *Proceedings of the 2007 SIAM International Conference on Data Mining (SDM)*; 2007. Available online: https://epubs.siam.org/doi/10.1137/1.9781611972771.32 (accessed on 14 September 2021)
40. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Müller, A.; Nothman, J.; Louppe, G.; et al. Scikit-learn: Machine Learning in Python. *arXiv* **2018**, arXiv:1201.0490.
41. Heim, B. Qualitative and Quantitative Analyses of Lake Baikal's Surface-Waters Using Ocean Colour Satellite Data (SeaWiFS). Ph.D. Thesis, Universität Potsdam, Potsdam, Germany, 2005.
42. Abbas, M.M.; Melesse, A.M.; Scinto, L.J.; Rehage, J.S. Satellite Estimation of Chlorophyll-a Using Moderate Resolution Imaging Spectroradiometer (MODIS) Sensor in Shallow Coastal Water Bodies: Validation and Improvement. *Water* **2019**, *11*, 1621. [CrossRef]
43. Hooker, S.B.; Firestone, E.R.; OReilly, J.E.; Maritorena, S.; OBrien, M.C.; Siegel, D.A.; Toole, D.; Mueller, J.L.; Mitchell, B.G.; Kahru, M.; et al. *Postlaunch Calibration and Validation Analyses*; SeaWiFS Postlaunch Technical Report Series; Goddard Space Flight Center: Greenbelt, MD, USA, 2000; Volume 11.
44. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A K-Means Clustering Algorithm. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **1979**, *28*, 100–108. [CrossRef]
45. Wall, M.; Rechtsteiner, A.; Rocha, L. Singular Value Decomposition and Principal Component Analysis. In *A Practical Approach to Microarray Data Analysis*; Springer: Boston, MA, USA, 2002; Volume 5. [CrossRef]

46. Prechelt, L. Early stopping-but when? In *Neural Networks: Tricks of the Trade*; Springer: Berlin/Heidelberg, Germany, 1998; pp. 55–69.
47. Claus, S.; De Hauwere, N.; Vanhoorne, B.; Deckers, P.; Souza Dias, F.; Hernandez, F.; Mees, J. Marine regions: Towards a global standard for georeferenced marine names and boundaries. *Mar. Geod.* **2014**, *37*, 99–125. [CrossRef]
48. Albawi, S.; Mohammed, T.A.; Al-Zawi, S. Understanding of a convolutional neural network. In Proceedings of the 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, 21–23 August 2017; pp. 1–6. [CrossRef]
49. Hui, Z.; Gao, X.; Yang, Y.; Wang, X. Lightweight Image Super-Resolution with Information Multi-distillation Network. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019. [CrossRef]
50. Hilborn, A.; Costa, M. Applications of DINEOF to satellite-derived chlorophyll-a from a productive coastal region. *Remote Sens.* **2018**, *10*, 1449. [CrossRef]