

Supplement: Data Processing, Algorithm Details, Additional Results from the Simulation Study and the Application

Mengyi Gong^{*1}, Claire Miller², Marian Scott², Ruth O'Donnell², Stefan Simis³,
Steve Groom³, Andrew Tyler⁴, Peter Hunter⁴ and Evangelos Spyarakos⁴

¹*Department of Mathematics and Statistics, Lancaster University*

²*School of Mathematics and Statistics, University of Glasgow*

³*Plymouth Marine Laboratory*

⁴*School of Biological and Environmental Science, University of Stirling*

1 MERIS Chlorophyll-*a* data processing

Time series of chlorophyll-*a* (Chl) were derived from Level 1B (third reprocessing) satellite imagery of the Medium Resolution Imaging Spectrometer (MERIS) onboard the European Space Agency ENVISAT satellite (2002-2012), using the Calimnos (v1.04) processing chain operated at PML. This is a highly automated series of processing stages, optimised to handle many small processing jobs (one per lake per satellite overpass) in parallel on a high-performance computing grid. Several processing stages call on external software such as modules in the Sentinel Application Platform (SNAP, v6.0) toolbox or in the case of Calimnos v1.04 using MERIS data, the BEAM Earth Observation Toolbox and Development Platform (v5). Individual processing steps and software included: (1) geolocation correction of the MERIS satellite passes using AMORGOS (v3.0) which performs precise orbit determination, instrument pointing and ortho-rectification (Bicheron et al. 2011); (2) spatial subsetting of satellite passes around the lake location using the subset operator included in BEAM with a 10% areal buffer around the pre-defined maximum-water extent of the lake; (3) MERIS Radiometric Correction (v5.0.3) including radiometric equalisation (Bouvet and Ramoino, 2010) included in BEAM; (4) Pixel identification as land, cloud, water, and ice/snow with Idepix-Water v2.2.10 included in BEAM; (5) Atmospheric correction using POLYMER v3.5 (Steinmetz et al. 2011), parameterized to use the Park and Ruddick (2005) bidirectional reflectance distribution function, and operating only on pixels identified as water by Idepix; (6) Optical Water Type (OWT) fuzzy classification analogous to Moore et al. (2001) but using standardized water-leaving reflectance. Class memberships were calculated against the 13 inland water OWTs defined in Spyarakos et al. (2018); (7) Chl concentration retrieval based on Gons et al. (2005) for OWTs 1, 4, 5 and 6, Gilerson et al. (2010) for OWTs 2, 8, 11 and 12, Mishra et al. (2014) for OWT 7, and the OC2 algorithm (oceancolor.gsfc.nasa.gov/cms/atbd/chlor_a) for OWTs 3, 9,

^{*}Correspondence email: m.gong1@lancaster.ac.uk

10 and 13. Specific tuning of each algorithm against in situ data corresponding to the respective OWTs, contained in the LIMNADES database, was carried out as part of the GloboLakes project (Neil et al. pers. comm.). The EO-derived Chl concentration was then calculated taking into account the algorithm results corresponding to the three most similar OWTs of each pixel, using the class membership scores to calculate a weighted average.

The 10-year time-series of Lake Taruo followed a seasonal pattern of OWT membership (left panel of Figure 1), with fractional OWT-coverage dominated by classes 13 (36.2%) and 3 (41.3%), described as very clear and clear waters by Spyarakos et al. (2018), respectively. It is noted that for these waters, Chl retrieval is met with high uncertainty because the Blue/Green band ratio algorithms adopted from ocean colour remote sensing are sensitive to the presence of coloured dissolved organic matter, which can only partly be compensated through OWT-specific algorithm tuning.

The OWT-based approach to Chl retrieval is designed to prevent, among other considerations, assignment of concentration values to pixels that do not show sufficient similarity to any of the known optical types (Type None in the left panel of Figure 1). These pixels represented 20.8% of observations with a sharp peak in February-March, most likely corresponding to seasonal ice and cloud cover. Ideally, the pixel identification step (step 4 above) distinguishes water, land, ice/snow and cloud pixels using a neural network algorithm, but the lack of thermal channels on MERIS and occurrence of mixed-type pixels can cause misidentification as water. Cumulative pixel identification results over the observed area (including a margin of land around the maximum water extent of the lake), indicate that relatively few observations in February, March and July were treated as water due to high cloud cover, whereas the period from August to January had the largest observable water area (right panel of Figure 1).

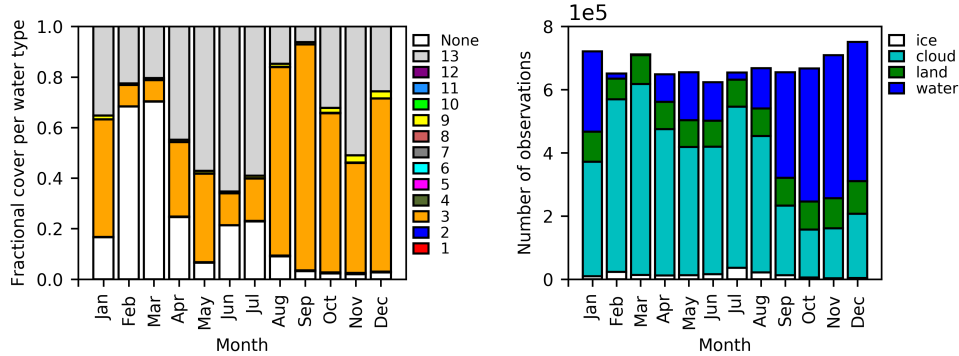


Figure 1: (Left) Optical water typology of Lake Taruo, aggregated by month in the period 2002-2012. (Right) Cumulative pixel identification results, aggregated by month in the period 2002-2012.

2 Additional information on the 2-cycle AECM algorithm

This section provides additional derivation and computation details of the 2-cycle AECM algorithm for the SS-FPCA model. Following the expressions of the target functions of the 2 cycles in section 2.3 of the main document, the computation forms (the expression used in the actual computation and programming) of $\mathcal{Q}^{[1]}(\Psi^{[1]}; \Psi^{(it-1)})$ can be derived

to be

$$\begin{aligned} & \sum_{t=1}^T \left[\log(|G_t^{(it)}|) + \text{tr} \left\{ \left(G_t^{(it)} \right)^{-1} \left(\Phi_t \mathbf{B}_{t|T} \Phi_t^\top + (\mathbf{Z}_t - \Phi_t \beta_{t|T}) (\mathbf{Z}_t - \Phi_t \beta_{t|T})^\top \right) \right\} \right] \\ & + \log(|\mathbf{B}_0|) + \text{tr} \left\{ \mathbf{B}_0^{-1} \left(\mathbf{B}_{0|T} + (\beta_{0|T} - \beta) (\beta_{0|T} - \beta)^\top \right) \right\} \\ & + T \log(|\mathbf{H}|) + \text{tr} \left\{ \mathbf{H}^{-1} \left(\mathbf{V}_{11} - \mathbf{V}_{10} \mathbf{M}^\top - \mathbf{M} \mathbf{V}_{10}^\top + \mathbf{M} \mathbf{V}_{00} \mathbf{M}^\top \right) \right\} + \text{constant} \end{aligned}$$

and that of $\mathcal{Q}^{[2]}(\Psi^{[2]}; \Psi^{(it, it-1)})$ can be derived as

$$\begin{aligned} & \sum_{t=1}^T \left[n_t \log(\sigma^2) + \frac{1}{\sigma^2} \text{tr} \left\{ \Phi_t \mathbf{B}_{t|T} \Phi_t^\top + (\mathbf{Z}_t - \Phi_t \beta_{t|T}) (\mathbf{Z}_t - \Phi_t \beta_{t|T})^\top \right\} \right. \\ & + \frac{1}{\sigma^2} \text{tr} \left\{ \Phi_t \Theta \widehat{\alpha_t \alpha_t^\top} \Theta^\top \Phi_t^\top \right\} - \frac{2}{\sigma^2} \text{tr} \left\{ \Phi_t \Theta \hat{\alpha}_t \mathbf{Z}_t^\top - \Phi_t \Theta \widehat{\alpha_t \beta_t^\top} \Phi_t^\top \right\} \left. \right] \\ & + T \log(|\mathbf{H}^{(it)}|) + \text{tr} \left\{ \left(\mathbf{H}^{(it)} \right)^{-1} \left(\mathbf{V}_{11} - \mathbf{V}_{10} \left(\mathbf{M}^{(it)} \right)^\top - \mathbf{M}^{(it)} \mathbf{V}_{10}^\top + \mathbf{M}^{(it)} \mathbf{V}_{00} \left(\mathbf{M}^{(it)} \right)^\top \right) \right\} \\ & + \log(|\mathbf{B}_0|) + \text{tr} \left\{ \mathbf{B}_0^{-1} \left(\mathbf{B}_{0|T} + (\beta_{0|T} - \beta) (\beta_{0|T} - \beta)^\top \right) \right\} \\ & + T \log(|\Lambda|) + \sum_{t=1}^T \text{tr} \left\{ \Lambda^{-1} \widehat{\alpha_t \alpha_t^\top} \right\} + \text{constant}, \end{aligned}$$

where $\mathbf{G}_t^{(it)}$, $\mathbf{V}_{11}$, $\mathbf{V}_{00}$ and $\mathbf{V}_{10}$ are defined as in section 2.3 of the main document. The smoothed states $\{\beta_{t|T}\}_{t=1}^T$ in cycle 1 are obtained by Kalman filter/smoothing with $\mathbf{H}^{(it-1)}$, $\mathbf{M}^{(it-1)}$ and $\mathbf{G}_t^{(it)}$. The smoothed states $\{\beta_{t|T}\}_{t=1}^T$ in cycle 2 are obtained with $\mathbf{H}^{(it)}$ and $\mathbf{M}^{(it)}$ from the previous cycle. The estimation of $\hat{\alpha}_t$ and $\widehat{\alpha_t \alpha_t^\top}$ are based on the conditional distribution of $\alpha_t | \mathbf{Z}_{1:T}, \Psi^{(it, it-1)}$, under the independence assumption of α_t and β_t , given the data.

As β_t in the SS-FPCA is the coefficient of the dynamic component, the derivation of $\hat{\alpha}_t$ and $\widehat{\alpha_t \alpha_t^\top}$ turns out to be slightly different from that in the FPC model of James, Hastie & Sugar (2000). This paper approached the problem in the following way.

1. The covariance matrix of $\mathbf{Z}_t$ is $\Phi \text{Cov}[\beta_t] \Phi^\top + \Phi_t \Theta \Lambda \Theta^\top \Phi_t^\top + \sigma^2 \mathbf{I}$.
2. As $\beta_{t|T}$ is obtained knowing only the variance, not the value, of α_t , giving information of $\mathbf{Z}_{1:T}, \Psi^{(it, it-1)}$ is sufficient to determine the values of $\beta_{t|T}$ and $\mathbf{B}_{t|T}$. Meanwhile, α_t is assumed to be independent from β_t and is only related to $\mathbf{Z}_t$, the observations at time point t . Hence, the following dependence can be deduced,

$$\left[\alpha_t \mid \mathbf{Z}_{1:T}, \Psi^{(it, it-1)} \right] = \left[\alpha_t \mid \mathbf{Z}_{1:T}, \beta_{t|T}, \mathbf{B}_{t|T}, \Psi^{(it, it-1)} \right] = \left[\alpha_t \mid \mathbf{Z}_t, \beta_{t|T}, \mathbf{B}_{t|T}, \Psi^{(it, it-1)} \right].$$

3. It is relatively easy to figure out the conditional distribution of $\alpha_t | \mathbf{Z}_t$, given the multivariate normal distribution of $(\alpha_t, \mathbf{Z}_t)^\top$, which is

$$\mathcal{N} \left(\begin{pmatrix} \mathbf{0} \\ \Phi_t \mathbf{E}[\beta_t] \end{pmatrix}, \begin{pmatrix} \Lambda & \Lambda \Theta^\top \Phi_t^\top \\ \Phi_t \Theta \Lambda & \Phi_t \text{Cov}[\beta_t] \Phi_t^\top + \Phi_t \Theta \Lambda \Theta^\top \Phi_t^\top + \sigma^2 \mathbf{I} \end{pmatrix} \right),$$

but the task remains to find $\mathbf{E}[\boldsymbol{\beta}_t]$ and $\mathbf{Cov}[\boldsymbol{\beta}_t]$. This can be done using the property of the iterative expectation, which gives $\mathbf{E}[\boldsymbol{\beta}_t] = \mathbf{E}[\boldsymbol{\beta}_{t|T}]$ and $\mathbf{Cov}[\boldsymbol{\beta}_t] = \mathbf{E}[\mathbf{B}_{t|T}]$. That is, the smoothed state $\boldsymbol{\beta}_{t|T}$ and its covariance matrix $\mathbf{B}_{t|T}$ could be used as the estimates of $\mathbf{E}[\boldsymbol{\beta}_t]$ and $\mathbf{Cov}[\boldsymbol{\beta}_t]$.

Putting all the information in 1-3 together gives the following distributional results,

$$\begin{aligned}\mathbf{E}[\boldsymbol{\alpha}_t | \mathbf{Z}_{1:T}, \Psi^{(it, it-1)}] &= \left(\boldsymbol{\Phi}_t \boldsymbol{\Theta}^{(it-1)} \boldsymbol{\Lambda}^{(it-1)} \right)^\top \left(\boldsymbol{\Sigma}^{(it)} \right)^{-1} (\mathbf{Z}_t - \boldsymbol{\Phi}_t \boldsymbol{\beta}_{t|T}), \\ \mathbf{Cov}[\boldsymbol{\alpha}_t | \mathbf{Z}_{1:T}, \Psi^{(it, it-1)}] &= \boldsymbol{\Lambda}^{(it-1)} - \left(\boldsymbol{\Phi}_t \boldsymbol{\Theta}^{(it-1)} \boldsymbol{\Lambda}^{(it-1)} \right)^\top \left(\boldsymbol{\Sigma}^{(it)} \right)^{-1} \boldsymbol{\Phi}_t \boldsymbol{\Theta}^{(it-1)} \boldsymbol{\Lambda}^{(it-1)},\end{aligned}$$

where $\boldsymbol{\Sigma}^{(it)} = \boldsymbol{\Phi}_t \mathbf{B}_{t|T} \boldsymbol{\Phi}_t^\top + \boldsymbol{\Phi}_t \boldsymbol{\Theta}^{(it-1)} \boldsymbol{\Lambda}^{(it-1)} \boldsymbol{\Theta}^{(it-1)\top} \boldsymbol{\Phi}_t^\top + \sigma^{2(it-1)} \mathbf{I}$. Using the above conditional expectation and variance, both $\hat{\boldsymbol{\alpha}}_t$ and $\widehat{\boldsymbol{\alpha}_t \boldsymbol{\alpha}_t^\top}$ can be obtained.

When the VAR coefficient matrix $\mathbf{M}$ is parameterized to be a diagonal matrix, with diagonal elements $m_1, \dots, m_K$, the MLEs of the parameters are derived using the general approach in Xu & Wikle (2007), as

$$\begin{pmatrix} m_1 \\ \vdots \\ m_K \end{pmatrix} = \mathbf{L}^{-1} \mathbf{b},$$

where

$$\begin{aligned}\mathbf{L} = (l_{ij}) &= \left(\text{tr} \left\{ \mathbf{H}^{-1} \frac{\partial \mathbf{M}}{\partial m_j} \mathbf{V}_{00} \frac{\partial \mathbf{M}^\top}{\partial m_i} \right\} \right) = (\mathbf{H}^{-1}(i, j) \mathbf{V}_{00}(j, i)) \\ \mathbf{b} = (b_i) &= \left(\text{tr} \left\{ \mathbf{H}^{-1} \frac{\partial \mathbf{M}}{\partial m_i} \mathbf{V}_{10}^\top \right\} \right) = \left(\sum_{j=1}^K \mathbf{H}^{-1}(j, i) \mathbf{V}_{10}^\top(i, j) \right),\end{aligned}$$

with $\mathbf{H}^{-1}(i, j)$, $\mathbf{V}_{00}^\top(i, j)$ and $\mathbf{V}_{10}^\top(i, j)$ denoting the (i, j) -th element in the corresponding matrices.

3 Additional information on the simulation study

This section provides additional figures and tables supporting the simulation study. Figure 2 shows the simulated data from one repetition, under the scenario of weak spatial signal $\kappa = 1.25$ and small noise $\sigma^2 = 0.01$ with missing patterns. The boxplots in Figure 3 show the distribution of the estimated eigenvalues λ_1 and λ_2 (subject to the signal adjusting factor κ) from three scenarios, S4, S8 and S12, with the red dots indicating the true values. These box plots show an underestimation of λ_1 and λ_2 , especially for λ_2 . The increase in the strength of the spatial signal and the more precise separate initialization method did not seem to improve the estimation, which can be seen from the middle and right panels.

Table 2 and Figure 5 provide some details of the simulation study in chapter 3 of the PhD thesis of Gong (2017). Images with 30% and 50% data missing at random or with a spatial pattern were simulated. The mean integrated squared prediction errors (MISE) and bias of the reconstruction of the dominant spatial pattern (i.e. the leading PC) from

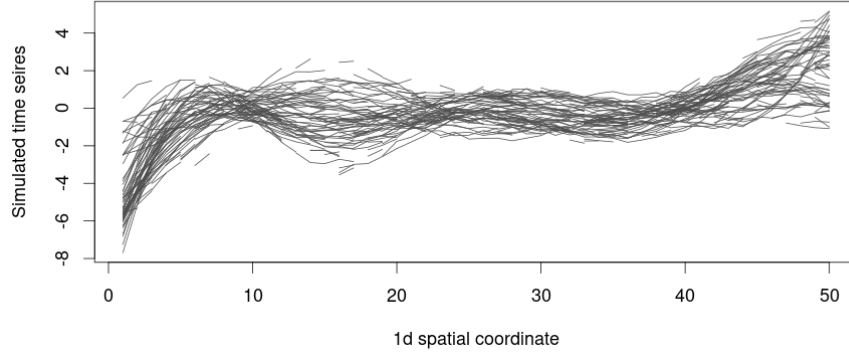


Figure 2: The time series of curves from one simulated replicate. These curves are generated using weak spatial signal $\kappa = 1.25$, small noise $\sigma^2 = 0.01$ with missing patterns. Each curve consists of observations at one time point.

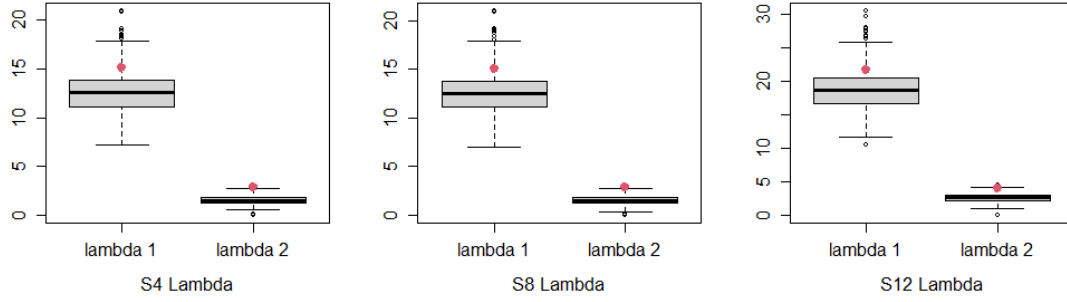


Figure 3: The boxplots of the estimated eigenvalues λ_1 and λ_2 from 500 replicates, with the red dots representing the true values of the two eigenvalues. The three panels represent scenario S4, S8 and S12 respectively.

Table 1: The means and the 95% confidence intervals of the MLEs of σ^2 and the means of RSS (numbers in *italics*) based on the 500 replicates from 12 simulation scenarios.

		small $\sigma^2 = 0.01$		large $\sigma^2 = 0.25$	
		complete	sparse	complete	sparse
weak + standard	MLE	0.0690	0.0708	0.3114	0.3136
	CI	(0.0589, 0.0782)	(0.0600, 0.0808)	(0.2956, 0.3281)	(0.2933, 0.3320)
	RSS	<i>0.0132</i>	<i>0.0139</i>	<i>0.2365</i>	<i>0.2323</i>
weak + separate	MLE	0.0689	0.0709	0.3117	0.3140
	CI	(0.0591, 0.0782)	(0.0606, 0.0806)	(0.2965, 0.3281)	(0.2963, 0.3330)
	RSS	<i>0.0132</i>	<i>0.0139</i>	<i>0.2366</i>	<i>0.2323</i>
strong + standard	MLE	0.0816	0.0838	0.3250	0.3276
	CI	(0.0700, 0.0928)	(0.0716, 0.0957)	(0.3061, 0.3429)	(0.3063, 0.3488)
	RSS	<i>0.0136</i>	<i>0.0145</i>	<i>0.2367</i>	<i>0.2322</i>

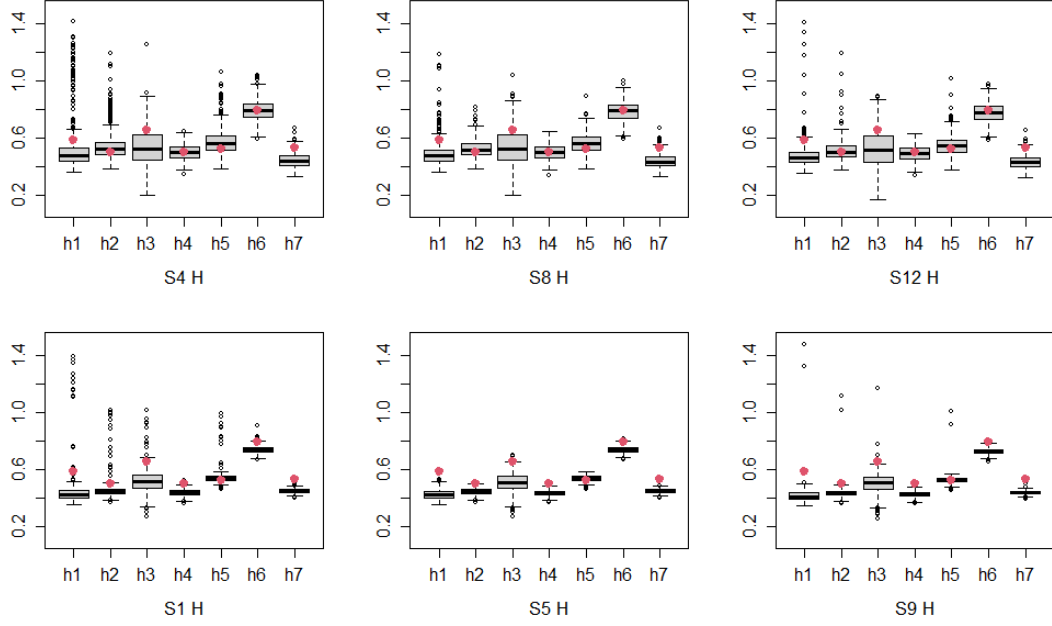


Figure 4: The boxplots of the estimated $h_1, \dots, h_7$ from scenario S4, S8 and S12 (top) and S1, S5 and S9 (bottom), with the red dots representing the true values.

200 repetitions were recorded. In Table 2 and Figure 5, ‘spatial I’ to ‘spatial IV’ refers to the different spatial variation scenarios of the simulated images (not for the missing data). It can be seen that the big increase in MISE and bias occurs mainly in the missing 50% data with spatial pattern (P50) scenario. There is a small increase in MISE for the missing 30% data with spatial pattern (P30) scenario (Table 2), but not much in bias as compared the missing 30% and 50% data at random (R30 and R50) scenarios (Figure 5).

Table 2: The percentage increments in the mean of MISE under different missing scenarios from the complete data scenario.

	spatial I $d = 1.5, \sigma_{ng}^2 = 0.01$	spatial II $d = 1.5, \sigma_{ng}^2 = 0.1$	spatial III $d = 1, \sigma_{ng}^2 = 0.01$	spatial IV $d = 1, \sigma_{ng}^2 = 0.1$
	No pattern			
30%	0.17%	0.26%	0.19%	0.57%
50%	0.79%	0.89%	1.84%	2.10%
	Pattern			
30%	1.07%	0.91%	7.58%	3.50%
50%	18.02%	7.52%	30.07%	12.87%

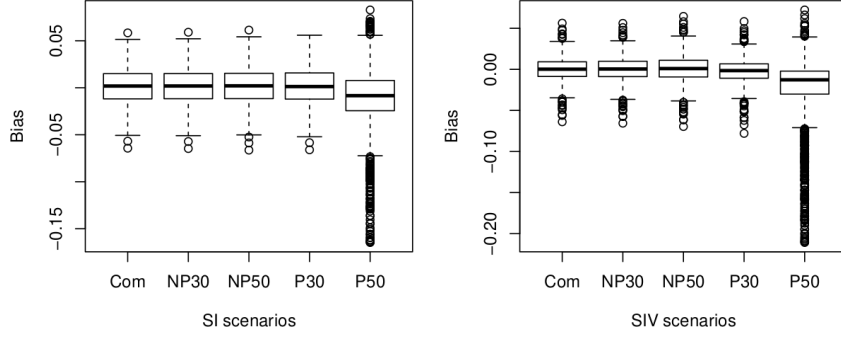


Figure 5: Box plots of the bias of the imputation from spatial scenarios I (left) and spatial scenarios IV (right). In each panel, from left to right are complete data, missing 30% without pattern, missing 50% without pattern, missing 30% with pattern and missing 50% with pattern.

4 Additional information on application to Taruo Lake Chl-*a* data

The following paragraphs provide additional results from the SS-FPCA model applied to the Taruo Lake data. An example of the estimated system dynamic component $\Phi\beta_{t|T}$ from February to July 2004 is given in Figure 6. This shows how the dynamic component, categorized by a VAR(1) process, evolves over time.

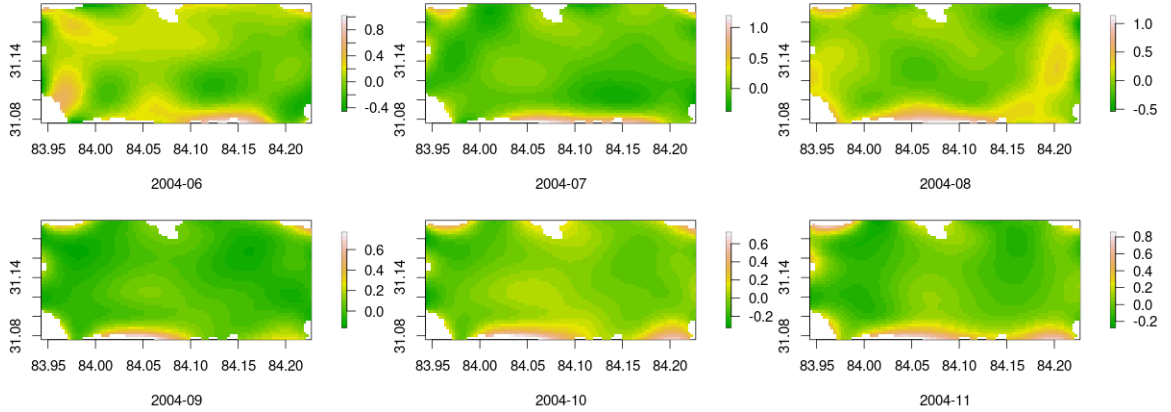


Figure 6: The system dynamic component $\Phi\beta_{t|T}$ from June to November 2004.

Figure 7 presents the residual sum of squares of the interpolation over time and space. The RSS over time (left panel) appeared to be related to data availability. Specifically, the values varies from close to 0 for well observed time points to over 0.7 for time points with very sparse observations. The RSS over space (right panel) are of the similar scale (with a medium of 0.05) for the majority of the pixels, apart from a few pixels around an island in the lake, where the RSS tend to be quite large.

Figures 8 and 9 present the standard errors of the parameters produced from the 200 bootstrap samples. Figure 8 presents the 95% confidence intervals of the estimated

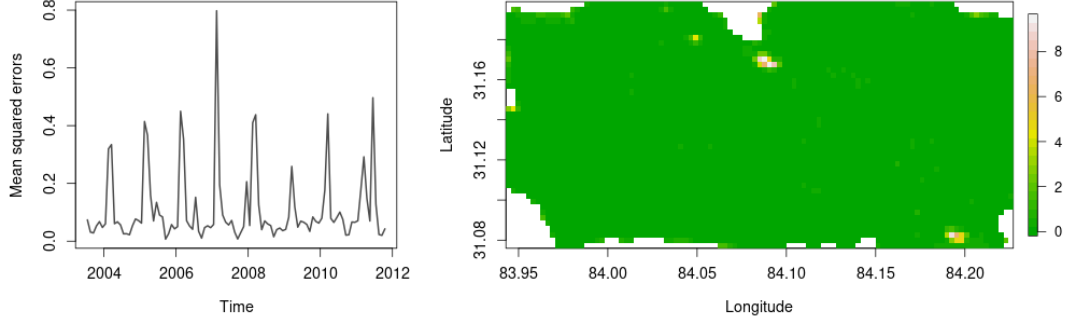


Figure 7: (Left panel) The RSS of the fitted values from the SS-FPCA model over time. (Right panel) The RSS of the fitted values over space.

eigenvalues $\lambda_1, \dots, \lambda_6$ in the left panel and the 95% confidence intervals of the estimated VAR(1) coefficients $m_1, \dots, m_{48}$ in the right panel. The red dots represent the estimated values from the original data; the gray lines represent the confidence intervals. Figure 9 presents the 95% confidence intervals of the estimated functional PC coefficients $\theta_1, \dots, \theta_6$. The red solid curves represent the estimated values from the original data; the gray dashed curves represent the confidence intervals. In particular, for the PC coefficients, the bootstrap estimations are aligned first before the calculation. Note that the coefficients of the PCs sometimes differ by a sign, but they are essentially the same, only with different orientations.

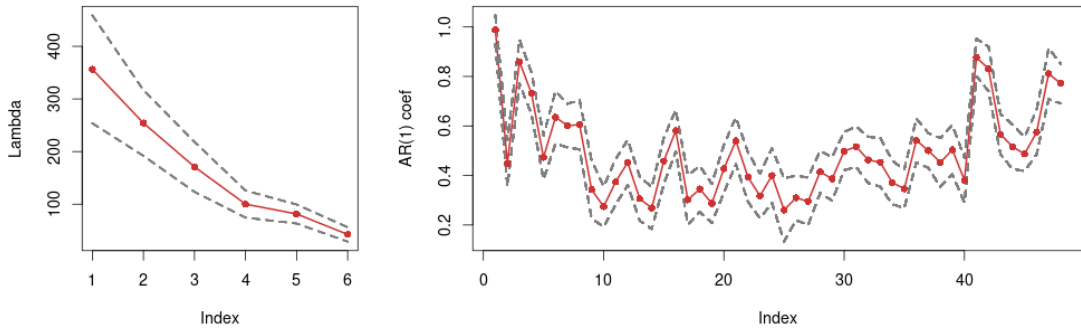


Figure 8: The 95% confidence intervals of the estimated eigenvalues $\lambda_1, \dots, \lambda_6$ (left) and the 95% confidence intervals of the estimated VAR(1) coefficients $m_1, \dots, m_{48}$ (right) from the bootstrap samples. The red dots represent the estimated values from the original data; the gray dashed curves represent the confidence intervals.

Finally, to complete the investigation, the FPC model by James et al. (2000) was applied to the Taro Lake Chl data. The same K-L expansion order, $P = 6$, was used in the model, giving 6 functional principal components. These are the functional PCs estimated without taking into account the temporal correlation in the data. Figure 10 presents the first 3 functional PCs from the FPC model (top panels) and the SS-FPCA (bottom panels). They account for over 80% (FPC model) and over 77% (SS-FPCA) of

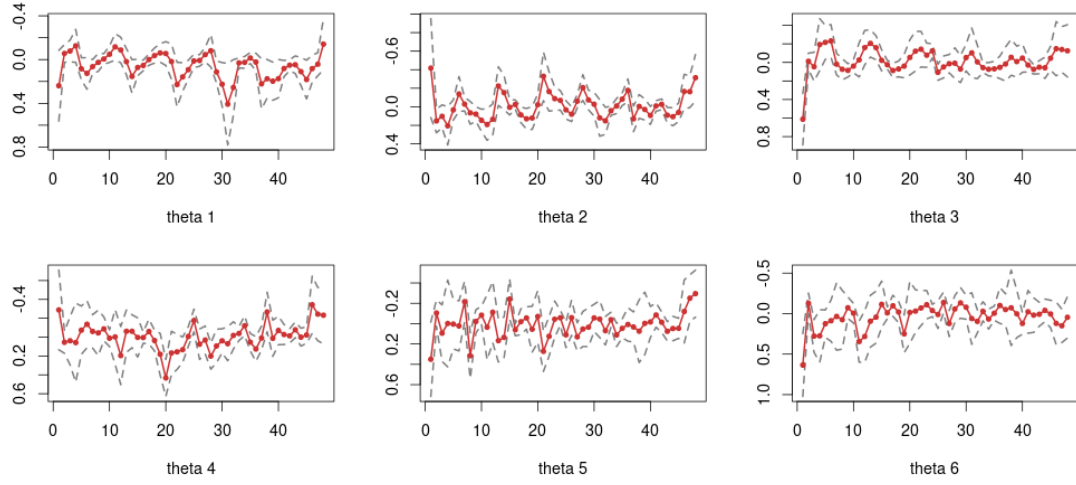


Figure 9: The 95% confidence intervals of the estimated coefficient vectors $\theta_1, \dots, \theta_6$ of the eigenfunctions from the bootstrap samples. The red curves represent the estimated values from the original data; the gray dashed curves represent the confidence intervals.

the variations in the data explained by the functional PCA. It appears that the first two eigenfunctions show similar patterns, although with different orders. The interpolations from the SS-FPCA model achieved smaller mean squared errors than those from the FPC model (0.0889 and 0.1054 respectively). They also captured more details in the data, which can be seen in Figure 11.

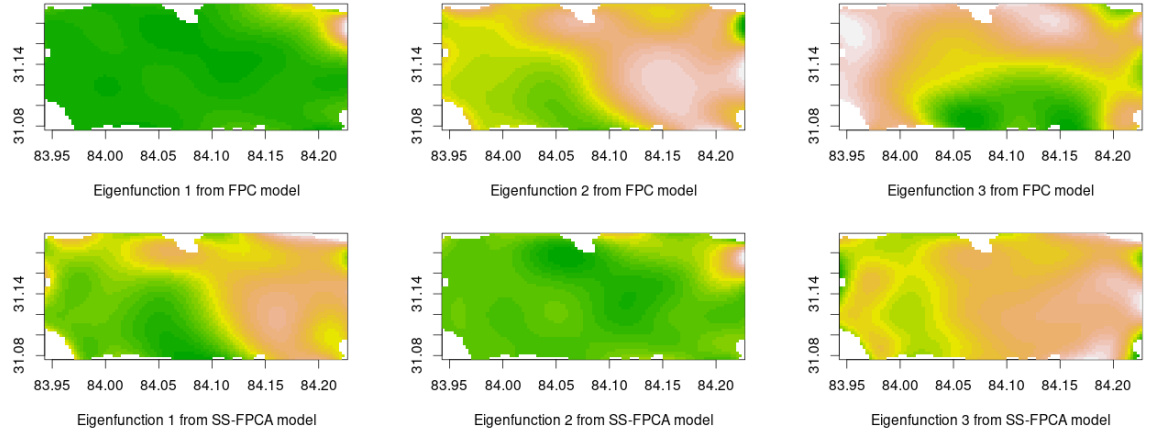


Figure 10: The leading 3 eigenfunctions (or functional PCs) estimated from the FPC model (top panels) and the SS-FPCA (bottom panels) applied to the Taruo Lake Chl data.

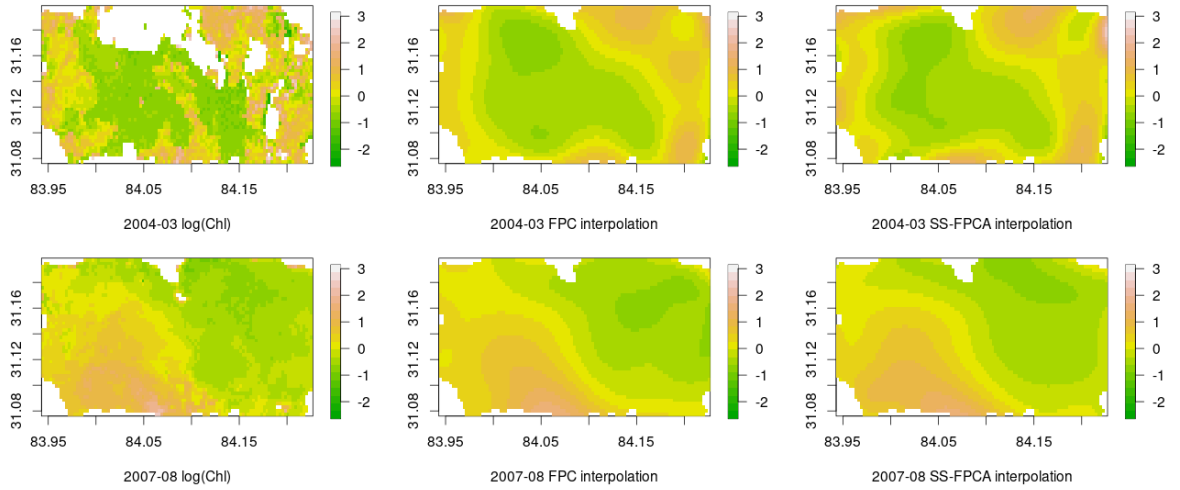


Figure 11: The log(Chl) data (left), the interpolations from the FPC model (middle) and the interpolations from the SS-FPCA (right), for March 2004 and August 2007.