Empirical Orthogonal Functions for ecology

Baptiste Alglave, Benjamin Dufée, Said Obakrim, James T Thorson

Journal: **Ecology**

Submission Type: **Statistical Report**

Title: Empirical Orthogonal Functions for Ecology

Baptiste Alglave[1], Benjamin Dufée[2], Said Obakrim[3], and James T. Thorson[4]

Keywords: spatio-temporal data, dimension reduction, satellite data, NDVI, species distribution model, multivariate analysis, canonical correlation analysis.

1 Lab-STICC, Université Bretagne Sud, 8 rue Michel de Montaigne, 56000, Vannes, France. *Contact author*: baptiste.alglave@univ-ubs.fr
2 Laboratory of Mathematics in Bretagne Atlantique, Rue de Saint-Maudé, 56100 Vannes, France.
3 Institute of Earth Surface Dynamics, University of Lausanne, Lausanne, Switzerland.
4 Resource Ecology and Fisheries Management Division, Alaska Fisheries Science Center, National Marine Fisheries Service, NOAA, Seattle, WA 98112, USA. *Corresponding author*: james.thorson@noaa.gov

Abstract: Spatio-temporal data are ubiquitous in ecology, involving large spatial domains, long time periods, and multiple variables. To take advantage of these data for ecological analysis, it is crucial to rely on both efficient and ecologically interpretable statistical methods. Empirical Orthogonal Functions (EOFs) are a widely used method for reducing the dimensionality of spatio-temporal data. While EOFs are common in meteorology and climate studies, their use in ecology is still emerging.

This paper introduces the basics of EOFs and presents new methods to specifically address key ecological questions. Since EOFs lack explicit ecological constraints, their results may be difficult to interpret. We propose a method to rotate EOFs using a temporal ecological variable, improving interpretability by better representing ecological patterns. We also explore the use of EOFs in multivariate analysis to identify shared spatial and temporal patterns in community and ecosystem dynamics.

We illustrate these methods on two case studies including NDVI satellite data over France and spatio-temporal predictions from fish species distribution models.

Our goal is to provide a clear and concise roadmap for applying EOFs to ecological problems. We emphasize practical guidance through code snippets and two worked examples, making the approach accessible to a broad audience of ecologists.

1 Introduction

Spatial and temporal data are ubiquitous in ecological research. Survey data, citizen-science data, satellite data or biogeochemical model outputs provide huge information for unraveling ecological processes at various temporal and spatial scales (Hampton et al., 2013; Grémillet, Chevallier, and Guinet, 2022). Given the extensive spatial and temporal coverage of all these datasets, dimension reduction (or ordination) methods become critical for synthesizing the information embedded within these spatio-temporal data.

Empirical Orthogonal Functions (EOFs) represent the keystone method for performing dimension reduction on spatio-temporal data. EOFs were first introduced by Lorenz (1956). Basically, they are Principal Component Analysis (PCA) performed on spatio-temporal data. Since their first formulation, EOFs methods have entailed a rich literature, particularly in meteorology and climate applications (Hannachi, Jolliffe, and Stephenson, 2007). Several extensive overviews of EOFs and derived methods are available in the secondary literature (Preisendorfer, 1988; Von Storch and Zwiers, 2002; Cressie and Wikle, 2011; Hannachi, 2021); they mainly address an audience of climatologists or statisticians.

Recently, these methods have expanded to ecology. They have been applied to biogeochemical variables (Schrum, John, and Alekseeva, 2006; Iida and Saitoh, 2007; Di Lorenzo et al., 2008; Woillez et al., 2010; Le Mezo et al.,

2016; Berkelhammer, 2019) as well as species distribution (Norton and Mason, 2005; Petitgas, Masse, et al., 2006; Petitgas, 2008; Woillez et al., 2010; Petitgas, Doray, et al., 2014; Petitgas, Huret, et al., 2018; Petitgas, Renard, et al., 2020; Grandremy et al., 2023; Van Audenhaege et al., 2022). EOFs have gained wider recognition through the development of the Vector-Autoregressive Spatio-temporal (VAST) package (Thorson, 2019; Thorson, Anderson, et al., 2024) and the application of EOFs to represent the spatio-temporal variability of species communities and ecosystems (Thorson, Ciannelli, and Litzow, 2020; Thorson, Cheng, et al., 2020; Thorson, Arimitsu, et al., 2021). Still, these applications remain mainly focused on marine systems, with very few applications to terrestrial systems (Gedalof, Peterson, and Mantua, 2005; Wu et al., 2023).

Overall, EOFs methods can address a large range of ecological questions from analyzing single species spatio-temporal variability to aiding conservation decisions in an ecosystem perspectives. They allow to detect seasonal signals in species distribution and shifts in phenology (Alglave, Olmos, et al., 2024). Additionally, EOFs are particularly useful in tracking long-term ecological changes, such as ecosystem restructuring or shifts in species composition in response to climate change (Thorson, Arimitsu, et al., 2021; Badger, Large, and Thorson, 2023).

Despite their broad potential, EOFs remain underutilized in ecology,

particularly in terrestrial ecosystems. In this paper, to widen EOFs applications and promote their use in ecology, we provide a thorough and accessible introduction to EOFs. Additionally, we describe two derived methods to address specific ecological questions: (1) a method that sets ecological constraints on EOFs to better disentangle the information of the different dimensions of the EOFs and improve their interpretability, and (2) a multivariate version of EOFs that enables to apply EOFs methodology on species communities or whole ecosystems.

We illustrate these methods based on two real applications. First, we use a terrestrial ecology example based on monthly Normalized Difference Vegetation Index (NDVI) satellite data (MODIS) from 2000 to 2023 (MODIS, 2021). As a second case study, we use spatio-temporal predictions from an Integrated Spatio-Temporal Species Distribution Model (IST-SDM) (Alglave, Vermard, et al., 2023). We analyze the spatio-temporal predictions spanning from 2008 to 2018 for three demersal species in the Bay of Biscay: common sole (*Solea solea*, Linnaeus, 1758), squid (*Loligo vulgaris*, Lamarck, 1798), and whiting (*Merlangius merlangus*, Linnaeus, 1758). In the case of the IST-SDM, when the method is univariate (*i.e.* single species) we focus the analysis on common sole.

Our goal is to provide a clear and concise roadmap for applying EOFs to ecological problems. We emphasize practical guidance through code

snippets and two worked examples, making the approach accessible to a broad audience of ecologists.

2 Basics of EOFs

EOFs were initially introduced by Lorenz (1956) for the analysis and prediction of weather. The method involves extracting the main spatial and temporal patterns that (1) capture the most variance of spatio-temporal data and (2) are orthogonal to each other. In the upcoming subsections, we introduce the notations and outline the principal properties of EOFs through both theoretical explanations and practical applications.

2.1 Continuous formulation of EOFs

We denote a space-time process S such that $S(x, t)$ represents the value of the process S at spatial location $x \in \mathcal{D}$ (with $\mathcal{D}$ a continuous domain) and discrete time step denoted t. When dealing with EOFs, we assume that the spatio-temporal field can be decomposed into a (purely) spatial term and a (purely) temporal term. In its continuous formulation, this decomposition can be expressed as:

$$S(x, t) = \sum_{m=1}^{\infty} \alpha_m(t) \cdot \phi_m(x) \tag{1}$$

where $\phi_m(x)$ denotes the spatial term of the EOFs for dimension m and $\alpha_m(t)$ represents the temporal term of the EOFs for dimension m. Furthermore, $\text{var}\,(\alpha_1(t)) > \text{var}\,(\alpha_2(t)) > \ldots$, and $\text{cov}\,(\alpha_k(t), \alpha_m(t)) = 0$, for all $k \neq m$. A known solution of this problem is obtained through a Karhunen-Loève expansion. Both formulation and practical estimation of the EOFs in the continous case are detailed in Cressie and Wikle (2011) (p.266).

2.2 Discrete formulation of EOFs

The most common and straightforward way to perform EOFs on S is to consider a discretized version of this process. Let's introduce the matrix $\mathbf{S}$ of size $n \times p$, where n denotes the number of discrete spatial positions, and p denotes the number of time steps, with the time stepping denoted by $(t_1, ..., t_p)$.

In practice, to decompose $\mathbf{S}$ through EOFs, these data are centered and then decomposed through Singular Value Decomposition (SVD).

From these data, for a given spatial location x, the temporal average can be calculated as:

$$\overline{S}^t(x) = \frac{1}{p} \sum_{k=1}^{p} S(x, t_k). \tag{2}$$

The time-centered space-time field, also referred to as the anomaly matrix, is then obtained by:

$$\mathbf{S}' = \mathbf{S} - \overline{\mathbf{S}}^t \, \mathbf{1}^\top, \tag{3}$$

where $\overline{\mathbf{S}}^t$ is an $n \times 1$ column vector of row means $\overline{S}^t(x)$ and $\mathbf{1}^\top$ is a $1 \times p$ row vector of ones. Codes to compute the anomaly matrix are available in Supplementary Material (SM S1.1). Examples of both the raw spatio-temporal data and anomalies are available in Figure 1 for satellite NDVI data and in Figure S3.1 (SM3) for the IST-SDM case study.

In this case, the decomposition can be written as:

$$S'(x,t) = \sum_{m=1}^{r} \lambda_m \cdot u_m(x) \cdot v_m(t) \tag{4}$$

Here, r represents the number of dimensions of the EOFs with $r = \min(n, p)$ being the rank of the matrix $\mathbf{S}'$. $u_m(x)$ is a spatial term and $v_m(t)$ is a temporal term; both $\mathbf{u}_m$ and $\mathbf{v}_m$ have unit norm. λ_m quantifies the amplitude of the m^th dimension of the EOFs. The λ_m are called the singular values hereafter (see section 2.2.1).

The key aspect of EOFs analysis lies in the constraints imposed on the spatial terms $u_m(x)$ or the temporal terms $v_m(t)$ because they are crucial for the computation and interpretation of these indices. The most

straightforward and natural constraints involve (1) maximising the variance explained by each spatial pattern $p_m(x)$; (2) setting orthogonal constraints for both the spatial and temporal terms.

2.2.1 EOFs as a solution of Singular Value Decomposition

To obtain the terms $u_m(x)$ and $v_m(t)$ of equation (4), it is possible to decompose $\mathbf{S}'$ through Singular Value Decomposition (SVD), namely:

$$\mathbf{S}' = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{5}$$

with $\mathbf{U}$ a $n \times r$ matrix and $\mathbf{V}^T$ a $r \times p$ matrix, $r = \min(n, p)$ being the rank of the matrix $\mathbf{S}'$. $\mathbf{\Sigma}$ is a $r \times r$ diagonal matrix with non-increasing positive coefficients on the diagonal, denoted as $\mathbf{\Sigma} = \text{Diag}(\lambda_1, ..., \lambda_r)$ and $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r > 0$. This is essentially the same computation that is performed when conducting PCA on any type of multivariate data.

The columns of $\mathbf{U} = (\mathbf{u}_1, ..., \mathbf{u}_r)$ are referred to as the EOFs maps of the anomaly matrix $\mathbf{S}'$. In the following, they will be referred to as **factors**; here the factors are spatial. The columns of $\mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_r)$ contain the terms $v_m(t)$ from Equation (4). They will be called **loadings**; here the loadings are temporal. Finally, $\mathbf{\Sigma}$ contains the singular values of $\mathbf{S}'$ along the diagonal. Note that here, singular values cannot be related to variance because the

centering of $\mathbf{S}$ is realized on rows. Codes to perform the svd are available in SM1 (Section S1.2).

2.2.2 Interpretation of EOFs

The construction of EOFs discussed in the previous subsection implies that the first EOFs map $\mathbf{u}_1$ is the factor that captures the most variance in $\mathbf{S}'$. The second EOFs, $\mathbf{u}_2$ is the next factor that captures most variance while being orthogonal to $\mathbf{u}_1$. This applies to the next j^{th} factors.

The loadings contained in $\mathbf{v}_j$, where $j \in 1, \cdots, r$, are temporal indices. Each is associated with a corresponding factor $\mathbf{u}_j$. When the loading $v_j(t)$ is positive, the anomaly field is distributed according to the factor $\mathbf{u}_j$. Conversely, if $v_j(t)$ is negative, the anomaly field is distributed according to $-\mathbf{u}_j$.

Also it is well known that identifying EOFs dimensions with specific ecological (or physical) processes should be done with care (Monahan et al., 2009). Indeed, because there is no explicit ecological criteria in the computation of EOFs, the signal related to an ecological process could be found on several dimensions. Besides, a single dimension could gather the signal of different ecological processes. As so, methods derived from EOFs

that integrate ecological knowledge would help to better disentangle the signal related to some specific ecological process.

Other details to interpret EOFs are provided in SM2 (*e.g.* computation of the variance related to each EOFs dimension, number of dimensions to retain in the analysis, whether specific dimensions can be related to physical or ecological processes).

3 Applying EOFs to the case studies

In order to illustrate EOFs, we use two spatio-temporal datasets to serve as a case study: (1) NDVI data from MODIS Vegetation Index Products (satellite data on terrestrial ecosystems) and (2) spatio-temporal predictions from the IST-SDM for sole in the Bay of Biscay (IST-SDM). A full description of the case studies is given in SM S3.2 and S3.3. We retain the first two dimensions for the NDVI case study and the first four dimensions in the IST-SDM case study based on the angle identified in the scree plot (left, Figure S4.1).

Both satellite NDVI data as well as the IST-SDM case study emphasize strong seasonality 2. In the case of the NDVI, the first and second dimensions bring out seasonal cycles that can be related to temperature and precipitation. For the IST-SDM, the two first dimensions highlight a signal related to reproduction (onshore-offshore migration) and the second

dimension emphasize a signal related to northward migration. Additional analysis on the EOFs results are available in SM S4.2.

As mentioned previously, ecological signals are mixed between the several dimensions of the EOFs. For instance, both the first and second dimensions for the two case studies capture similar seasonal signals (Figure 2). Also, some loadings seem to mix both long term trends and seasonnal trends (see the second EOFs dimension of the IST-SDM case study). Disentangling the ecological signals from these dimensions would benefit the ecological analysis of these datasets.

4 Informing EOFs with an ancillary temporal variable

As mentioned in previous sections, EOFs may not effectively disentangle the ecological processes that structure the data. To address this limitation, informing EOFs with a temporal variable (or ancillary variable) that is ecologically meaningful can improve the interpretability of EOFs patterns.

This section describes the methodology for extracting EOFs patterns that exhibit the highest correlation with some ancillary temporal variable. This process involves projecting the spatio-temporal data onto the EOFs basis and then performing a Canonical correlation analysis (CCA) between a selected set of EOFs loadings and the ancillary variable. We first introduce

the CCA method, demonstrate how to adapt it to our specific issue, and finally present applications of this technique.

4.1 Basics of canonical correlation analysis

CCA was first proposed by Hotelling (1992). It is a method used to identify and measure the associations among two sets of variables. Consider two space-time processes $S^{(1)}(x,t)$ and $S^{(2)}(x,t)$ at location x and time t. CCA aims at identifying pairs of spatial basis vectors, $\mathbf{w}_1$ and $\mathbf{w}_2$, for the two processes, $S^{(1)}$ and $S^{(2)}$, that maximize the correlation between their projections onto these vectors. Consider the corresponding observed space-time matrices denoted $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ of size $n_1 \times p$ and $n_2 \times p$, respectively.

The CCA maximizes the following correlation coefficient ρ between $\mathbf{y}_1 = \mathbf{S}^{(1)}\mathbf{w}_1$ and $\mathbf{y}_2 = \mathbf{S}^{(2)}\mathbf{w}_2$:

$$\rho = \frac{\mathbb{E}(\mathbf{y}_1\mathbf{y}_2)}{\sqrt{\mathbb{E}(\mathbf{y}_1^2)\mathbb{E}(\mathbf{y}_2^2)}} = \frac{\mathbb{E}(\mathbf{w}_1^T\mathbf{S}^{(1)\,T}\mathbf{S}^{(2)}\mathbf{w}_2)}{\sqrt{\mathbb{E}(\mathbf{w}_1^T\mathbf{S}^{(1)\,T}\mathbf{S}^{(1)}\mathbf{w}_1)\mathbb{E}(\mathbf{w}_2^T\mathbf{S}^{(2)\,T}\mathbf{S}^{(2)}\mathbf{w}_2)}}. \tag{6}$$

Here $\mathbf{y}_1$ and $\mathbf{y}_2$ are the temporal variables that arise from the linear combination of $\mathbf{S}^{(1)}\mathbf{w}_1$ and $\mathbf{S}^{(2)}\mathbf{w}_2$ and whose correlation coefficient is to be maximised. Then, the spatial basis vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ describe which areas contribute to this correlation in the dataset $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ respectively.

In order to find the vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ that maximize ρ, we resort to the method of Lagrange multipliers, leading to a generalized eigenvalue problem (Weenink, 2003).

$$\begin{cases} (\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\mathbf{C}_{22}^{-1}\mathbf{C}_{21} - \rho^2)\mathbf{w}_1 = 0 \\ (\mathbf{C}_{22}^{-1}\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12} - \rho^2)\mathbf{w}_2 = 0 \end{cases} \tag{7}$$

Here, the eigenvalues ρ^2 represent the squared canonical correlations, and the eigenvectors $\mathbf{w}_1$ and $\mathbf{w}_2$ are the canonical correlation basis vectors.

4.2 Coupling CCA with EOFs

To inform the EOFs with a single ancillary variable, it is possible to perform a CCA on the matrix $\mathbf{V}$ (Equation 5) composed of the loadings and an ancillary variable that we denote $\mathrm{An}(t), t \in \{1, \cdots, p\}$. Usually, to apply CCA we only retain the dimensions of the EOFs that capture the process and leave apart the remaining dimensions (see the rule of thumb in SM2). In this case, $\mathbf{w}_1$ are coefficients that allow to rotate the loadings as well as the related factors to obtain:

- one temporal variable that closely fit the ancillary variable. This is the combination of the EOFs loadings that maximizes the correlation with the ancillary variable.

- the related factor (or spatial basis vector) that is a combination of the first EOFs factors. It will capture as much variance as the first pattern of the EOFs but will not be orthogonal to the following EOFs.

Note that the two datasets $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ do not necessarily have the same spatial extent, but they must have the same time stepping.

Such procedure will not necessarily bring out strongly different patterns compared with EOFs; it will rather help to disentangle the processes identified in the EOFs to ease interpretation.

4.3 Illustration

In this section, we inform EOFs of the NDVI satellite data with an ancillary variable representing seasonality. The ancillary variable chosen here is a steady seasonal signal evidencing seasonality with positive peaks in summer (July) and negative peaks in winter (January - red variable in Figure 3). This variable aims at disentangling the common seasonal signal shared by the two first dimensions of the EOFs (Figure 2). The seasonal signal is parameterized as $\mathrm{An}(t) = A \cdot sin(f \cdot t + D)$.

To have a ancillary variable that match seasonal cycles (opposite peaks in January and July) we set $A = 1$ (the amplitude of the signal), $f = 1/12 \times 2\pi$ the frequency of the signal components and the delay component $D = 1$.

Note that this procedure is not invariant to a change in the phase of the seasonal signal and should be fixed with care. We then perform a CCA between the first two loadings of the EOFs (components of the matrix **V** in Section 2.2) and the seasonal signal. Codes to perform the analysis are available in the SM section S1.3.

The loadings obtained from the CCA emphasize a very good fit to the ancillary variable (R2 = 0.98, Figure 3, top right). By comparison, the first two EOFs loadings have positive, but lower correlation coefficient (R2 = 0.82 and R2 = 0.57). The spatial basis obtained through CCA displays a stronger North/South gradient of NDVI compared with the two first EOFs (Figure 2). NDVI are high in orange/red areas during summer (i.e. the center and mountain part of France) which corresponds to high NDVI in mountains and all the southern part and eastern part of France. In these areas, either ice melt and temperatures become favorable for vegetation growth (mountains), either temperature are high and rains remain frequent which increases vegetation coverage (eastern and south western part of France). In winter, these areas are colder and they can be covered by snow which do not favor vegetation growth. In blue areas, NDVI is relatively higher during winter. These areas have a cooler climate during winter and strong precipitation which favors vegetation growth.

A similar analysis is performed on the IST-SDM outputs. It allows to better

evidence the seasonal cycle of sole reproduction and the offshore-onshore gradients that shapes sole migration (Figure S4.3)

5 EOFs for multivariate analysis

When studying entire ecosystems or community dynamics, multiple variables need to be analyzed jointly using multivariate statistical methods (Ovaskainen and Abrego, 2020; Thorson, Ciannelli, and Litzow, 2020). The EOFs framework is flexible and adapts to the multivariate case. It allows to identify the main modes of variability across several variables and to detect shared spatial patterns and temporal trends. In such cases, the underlying EOFs theory remains unchanged (Section 2); only the matrix to be diagonalized $\mathbf{S}'$ is modified.

5.1 Two alternatives for conducting multivariate EOFs

Let $v \in \{1, \cdots, s\}$ denotes the index of the different variables in the analysis, and let $S^{(v)}(x, t)$ denotes the value of the space-time process for location x, time t, and variable v. We denote by $\mathbf{S'}^{(v)}$ the $n \times p$ matrix of the spatio-temporal variable v. Two options are available when conducting multivariate EOFs:

1. Construct the matrix by stacking the matrices row-wise. In this case, the matrix is denoted as $\mathbf{S}'^{(row)}_{multi}$ with dimensions $(n \cdot s) \times p$, and it has the following structure:

$$\mathbf{S}'^{(row)}_{multi} = \begin{pmatrix} \mathbf{S}'^{(1)} \\ \vdots \\ \mathbf{S}'^{(s)} \end{pmatrix} \tag{8}$$

2. Alternatively, one can construct the matrix by stacking the matrices column-wise. In this case, the matrix is denoted as $\mathbf{S}'^{(col)}_{multi}$ with dimensions $n \times (p \cdot s)$ and is structured as follows:

$$\mathbf{S}'^{(col)}_{multi} = \begin{pmatrix} \mathbf{S}'^{(1)} & ; & \cdots & ; & \mathbf{S}'^{(s)} \end{pmatrix} \tag{9}$$

Similar to standard EOFs, SVD enables to compute the factors $\mathbf{U}$ and the loadings $\mathbf{V}$ (Section 2). However, depending on the data stacking of the anomaly matrix, these matrices do not have the same dimensions and do not lead to the same interpretation. The first case is a **temporally synthetic representation** of the multivariate data. For each dimension, there is only one single time series of loadings and there are as many factors as there are variables. The loadings quantify how the different factors evolve over time. This allows to summarize the temporal variation of all the variables in one single temporal variable. A similar approach is adopted in the VAST

package (Thorson, 2019). The second case, is a **spatially synthetic representation** of the data. For each dimension, there is only one factor that summarizes the variability of all the variables and as many time series of loadings as there are variables. The loadings of the different variables quantify how these variables are related to the factor at a specific time step.

Note finally that in some cases, multivariate EOFs can be applied on a single variable with some time-lag. In this case the different variables are the lagged version of the spatio-temporal data. This typically allows for a better identification of temporal patterns such as periodicity or long-term trends. It is referred to as lag EOFs in this paper; it is also often referred as extended EOFs in the literature (Weare and Nasstrom, 1982).

5.2 Illustration

To illustrate multivariate EOFs analysis, we apply the two alternative approaches to the three species in the Bay of Biscay: sole, whiting, and squids. These are coastal species with potentially shared spatio-temporal dynamics *e.g.* same reproduction period (Alglave, Vermard, et al., 2023).

The temporally synthetic multivariate EOFs (EOFs on $\mathbf{S}'^{(row)}_{multi}$) provides seasonal loadings that represent the reproduction phenology for all the species. We only retain the first two dimensions for the analysis (Figure S4.1). For sole, the reproduction areas are generally consistent with those

identified when conducting univariate EOFs (Figure 4). For squids and whitings, reproduction areas are observed along the Landes coast (45°N - 1.5°W), on the Vendée coast (46.5°N - 2°W), and in the north of the Bay of Biscay (47.5°N - 3°W).

The spatially synthetic multivariate EOFs is provided in SM S4.5 as well as the lag EOFs (SM S4.6). The spatially synthetic representation allows to identify the common spatial patterns among each three species as well as synchrony in the spatio-temporal variation of the species. The lag EOFs is applied to the NDVI case study. Compared with standard EOFs (Section 3), it allows to identify an additional periodic signal with a 6-months phase related to an alternance of high growth of vegetation in spring and autumn and lower growth in summer and winter.

6 Discussion

Towards a wider use of EOFs in ecology

EOFs have tremendous potential for ecology. As massive and diverse datasets become available (satellite data, IST-SDM, satellite data and biogeochemical models), these decomposition methods will play a growing role in (1) reducing data dimensionality and (2) extracting ecologically relevant and interpretable information. In this paper, we demonstrate their

potential for both terrestrial and marine systems. We argue that EOFs should be systematically used to analyze spatio-temporal ecological datasets.

We presented two extensions of EOFs, one to better disentangle the processes identified through EOFs and improve ecological interpretability of the factor and another to apply EOFs to multivariate data. Many other methods derived from EOFs could be used to perform spatio-temporal analysis on ecological data and are available in the literature (Preisendorfer, 1988; Von Storch and Zwiers, 2002; Cressie and Wikle, 2011; Hannachi, 2021).

Accounting for spatio-temporal correlation in EOFs

One might argue that EOFs do not explicitly account for spatial or temporal correlations and then are not per se a spatio-temporal method. It is true that EOFs is basically a PCA representation of spatio-temporal data; the spatio-temporal aspects of the method comes from the data rather than from the mathematical formulation of the method per se.

In a spatial context, methods have been developed to better handle spatial correlations and to propose multiscale decompositions of ecological data (P. Legendre and L. Legendre, 2012). Specifically, 'spatial eigenfunction analysis' (e.g., Moran's eigenvector maps, asymmetric eigenvector maps, multiscale ordination) estimate sets of eigenvectors based on spatial

configuration matrices. These methods could potentially be extended to spatio-temporal data to identify distinct spatio-temporal scales that structure ecological data.

More recently, Petitgas, Renard, et al. (2020) and Bez, Renard, and Ahmed-Babou (2023) have proposed a method (Min-max Autocorrelation Factor - MAF - or Empirical Orthogonal Maps - EOM) to provide spatially decorrelated factors. EOM provide maps with stronger orthogonality constraints and aims to better separate the information carried by the factors. Still, such approach is very sensitive to the choice of the spatial lag chosen in the analysis and does not account for temporal correlation.

Non-stationary, non-linear EOFs for spatio-temporal analysis under climate change

Finally, some important hypotheses underlie EOFs, namely stationarity and linearity (Hannachi, Jolliffe, and Stephenson, 2007; Alglave, Olmos, et al., 2024). EOFs can not be used to describe propagating patterns such as range expansion or contraction as is common in species redistribution (Melles et al., 2011; Scheele et al., 2017). Also, as in standard PCA, all non-linear processes will not be detected and captured through EOFs while it is largely recognized that ecosystems are strongly structured by non-linear relationships and sometimes switching dynamics (Scheffer et al., 2001). More generally speaking, handling non-stationarity in a changing

environment is a key challenge for statistical ecology (Litzow et al., 2019; Astigarraga et al., 2020; Bueno de Mesquita et al., 2021).

Although some techniques, such as Hilbert complex EOFs or non-linear PCA (Esquivel and Messina, 2008; Bueso, Piles, and Camps-Valls, 2020), allow to handle non-linear relationship and to evidence propagating patterns, these have been scarcely used in practice to investigate the effect of climate change on ecosystems. Developing and applying approaches to handle non-linearity and non-stationarity is critical in analysing the effect of climate change on ecosystems; it constitues an open research avenue for future study to track the effect of climate change on ecosystems.

Data availability statement

Codes and data are available through the link here.

Acknowledgments

The authors are grateful to Julia Indivero (School of Aquatic and Fishery Sciences, University of Washington) and Nikolai Morokhovich (School of Aquatic and Fishery Sciences, University of Washington) who made the internal review for the NOAA pre-submission procedure, as well as Nicolas

Bez who provided insightful comments on the manuscript. The findings and conclusions of the present paper are those of the authors.

Competing interests

The authors declare that they have no conflicts of interest or competing interests that could influence the content or interpretation of this manuscript.

References

Alglave, Baptiste, Maxime Olmos, et al. (2024). "Investigating fish reproduction phenology and essential habitats by identifying the main spatio-temporal patterns of fish distribution". In: ICES Journal of Marine Science, fsae099.

Alglave, Baptiste, Youen Vermard, et al. (2023). "Identifying mature fish aggregation areas during spawning season by combining catch declarations and scientific survey data". In: Canadian Journal of Fisheries and Aquatic Sciences 80.5, pp. 808–824.

Astigarraga, Julen et al. (2020). "Evidence of non-stationary relationships between climate and forest responses: Increased sensitivity to climate change in Iberian forests". In: Global Change Biology 26.9, pp. 5063–5076.

Badger, JJ, SI Large, and James T Thorson (2023). "Spatio-temporal species distribution models reveal dynamic indicators for ecosystem-based fisheries management". In: ICES Journal of Marine Science 80.7, pp. 1949–1962.

Berkelhammer, Max (2019). "Synchronous modes of terrestrial and marine productivity in the North Pacific". In: Frontiers in Earth Science 7, p. 73.

Bez, Nicolas, Didier Renard, and Dedah Ahmed-Babou (2023). "Empirical Orthogonal Maps (EOM) and Principal Spatial Patterns: Illustration for Octopus Distribution Off Mauritania Over the Period 1987–2017". In: Mathematical Geosciences 55.1, pp. 113–128.

Bueno de Mesquita, Clifton P et al. (2021). "Taking climate change into account: Non-stationarity in climate drivers of ecological response". In: Journal of Ecology 109.3, pp. 1491–1500.

Bueso, Diego, Maria Piles, and Gustau Camps-Valls (2020). "Nonlinear PCA for spatio-temporal analysis of Earth observation data". In: IEEE Transactions on Geoscience and Remote Sensing 58.8, pp. 5752–5763.

Cressie, Noel and Christopher K Wikle (2011). Statistics for spatio-temporal data. John Wiley & Sons.

Di Lorenzo, Emanuele et al. (2008). "North Pacific Gyre Oscillation links ocean climate and ecosystem change". In: Geophysical research letters 35.8.

Esquivel, P and AR Messina (2008). "Complex empirical orthogonal function analysis of wide-area system dynamics". In: 2008 IEEE Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century. IEEE, pp. 1–7.

Gedalof, Ze'ev, David L Peterson, and Nathan J Mantua (2005). "Atmospheric, climatic, and ecological controls on extreme wildfire years in the northwestern United States". In: Ecological Applications 15.1, pp. 154–174.

Grandremy, Nina et al. (2023). "Hydrology and small pelagic fish drive the spatio–temporal dynamics of springtime zooplankton assemblages over the Bay of Biscay continental shelf". In: Progress in Oceanography 210, p. 102949.

Grémillet, David, Damien Chevallier, and Christophe Guinet (2022). "Big data approaches to the spatial ecology and conservation of marine megafauna". In: ICES Journal of Marine Science 79.4, pp. 975–986.
Hampton, Stephanie E et al. (2013). "Big data and the future of ecology". In: Frontiers in Ecology and the Environment 11.3, pp. 156–162.
Hannachi, Abdelwaheb (2021). Patterns identification and data mining in weather and climate. Vol. 10. Springer.
Hannachi, Abdelwaheb, Ian T Jolliffe, and David B Stephenson (2007). "Empirical orthogonal functions and related techniques in atmospheric science: A review". In: International Journal of Climatology: A Journal of the Royal Meteorological Society 27.9, pp. 1119–1152.
Hotelling, Harold (1992). "Relations between two sets of variates". In: Breakthroughs in statistics: methodology and distribution. Springer, pp. 162–190.
Iida, Takahiro and Sei-Ichi Saitoh (2007). "Temporal and spatial variability of chlorophyll concentrations in the Bering Sea using empirical orthogonal function (EOF) analysis of remote sensing data". In: Deep Sea Research Part II: Topical Studies in Oceanography 54.23-26, pp. 2657–2671.
Le Mezo, Priscilla et al. (2016). "Natural variability of marine ecosystems inferred from a coupled climate to ecosystem simulation". In: Journal of Marine Systems 153, pp. 55–66.
Legendre, Pierre and Louis Legendre (2012). Numerical ecology. Vol. 24. Elsevier.
Litzow, Michael A et al. (2019). Nonstationary environmental and community relationships in the North Pacific Ocean.
Lorenz, Edward N (1956). Empirical orthogonal functions and statistical weather prediction. Vol. 1. Massachusetts Institute of Technology, Department of Meteorology Cambridge.
Melles, SJ et al. (2011). "Expanding northward: influence of climate change, forest connectivity, and population processes on a threatened species' range shift". In: Global Change Biology 17.1, pp. 17–31.
MODIS (2021). MOD13Q1 V6. Accessed: 2024-07-11. URL: https://lpdaac.usgs.gov/products/mod13q1v006/.
Monahan, A. H. et al. (2009). "Empirical Orthogonal Functions: The Medium is the Message". In: Journal of Climate 22.24, pp. 6501–6514. DOI: 10.1175/2009JCLI3062.1. URL: https://journals.ametsoc.org/view/journals/clim/22/24/2009jcli3062.1.xml.
Norton, JERROLD G and JANET E Mason (2005). "Relationship of California sardine (Sardinops sagax) abundance to climate-scale ecological changes in the California Current system". In: California Cooperative Oceanic Fisheries Investigations Report 46, p. 83.
Ovaskainen, Otso and Nerea Abrego (2020). Joint species distribution modelling: With applications in R. Cambridge University Press.
Petitgas, Pierre (2008). "Fish habitat mapping with empirical orthogonal functions". In: International Council for the Exploration of the Sea.
Petitgas, Pierre, Mathieu Doray, et al. (2014). "Modelling the variability in fish spatial distributions over time with empirical orthogonal functions: anchovy in the Bay of Biscay". In: ICES Journal of Marine Science 71.9, pp. 2379–2389.
Petitgas, Pierre, Martin Huret, et al. (2018). "Ecosystem spatial structure revealed by integrated survey data". In: Progress in Oceanography 166, pp. 189–198.
Petitgas, Pierre, Jacques Masse, et al. (2006). "Hydro-plankton characteristics and their relationship with sardine and anchovy distributions on the French shelf of the Bay of Biscay". In: Scientia Marina 70.1, pp. 161–172.
Petitgas, Pierre, Didier Renard, et al. (2020). "Analysing temporal variability in spatial distributions using min–max autocorrelation factors: sardine eggs in the bay of biscay". In: Mathematical Geosciences 52, pp. 337–354.
Preisendorfer, RW (1988). Principal Component Analysis in Meteorology and Oceanography.
Scheele, Ben C et al. (2017). "Niche contractions in declining species: mechanisms and consequences". In: Trends in Ecology & Evolution 32.5, pp. 346–355.
Scheffer, Marten et al. (2001). "Catastrophic shifts in ecosystems". In: Nature 413.6856, pp. 591–596.

514 Schrum, Corinna, Mike St John, and Irina Alekseeva (2006). “ECOSMO, a coupled ecosystem model
515 of the North Sea and Baltic Sea: Part II. Spatial-seasonal characteristics in the North Sea as
516 revealed by EOF analysis”. In: Journal of Marine Systems 61.1-2, pp. 100–113.
517 Thorson, James T (2019). “Guidance for decisions using the Vector Autoregressive Spatio-Temporal
518 (VAST) package in stock, ecosystem, habitat and climate assessments”. In: Fisheries Research
519 210, pp. 143–161.
520 Thorson, James T, Sean C Anderson, et al. (2024). “tinyVAST: R package with an expressive
521 interface to specify lagged and simultaneous effects in multivariate spatio-temporal models”. In:
522 arXiv preprint arXiv:2401.10193.
523 Thorson, James T, Mayumi L Arimitsu, et al. (2021). “Forecasting community reassembly using
524 climate-linked spatio-temporal ecosystem models”. In: Ecography 44.4, pp. 612–625.
525 Thorson, James T, Wei Cheng, et al. (2020). “Empirical orthogonal function regression: Linking
526 population biology to spatial varying environmental conditions using climate projections”. In:
527 Global Change Biology 26.8, pp. 4638–4649.
528 Thorson, James T, Lorenzo Ciannelli, and Michael A Litzow (2020). “Defining indices of ecosystem
529 variability using biological samples of fish communities: a generalization of empirical
530 orthogonal functions”. In: Progress in Oceanography 181, p. 102244.
531 Van Audenhaege, Loıc et al. (2022). “Long-term monitoring reveals unprecedented stability of a vent
532 mussel assemblage on the Mid-Atlantic Ridge”. In: Progress in Oceanography 204, p. 102791.
533 Von Storch, Hans and Francis W Zwiers (2002). Statistical analysis in climate research. Cambridge
534 university press.
535 Weare, Bryan C and John S Nasstrom (1982). “Examples of extended empirical orthogonal function
536 analyses”. In: Monthly Weather Review 110.6, pp. 481–485.
537 Weenink, David (2003). “Canonical correlation analysis”. In:
538 Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam. Vol. 25.
539 University of Amsterdam Amsterdam, pp. 81–99.
540 Woillez, Mathieu et al. (2010). “Statistical monitoring of spatial patterns of environmental indices for
541 integrated ecosystem assessment: Application to the Bay of Biscay pelagic zone”. In:
542 Progress in Oceanography 87.1-4, pp. 83–93.
543 Wu, Qiong et al. (2023). “Spatiotemporal variations of water conservation function based on EOF
544 analysis at multi time scales under different ecosystems of Heihe River Basin”. In:
545 Journal of Environmental Management 325, p. 116532.

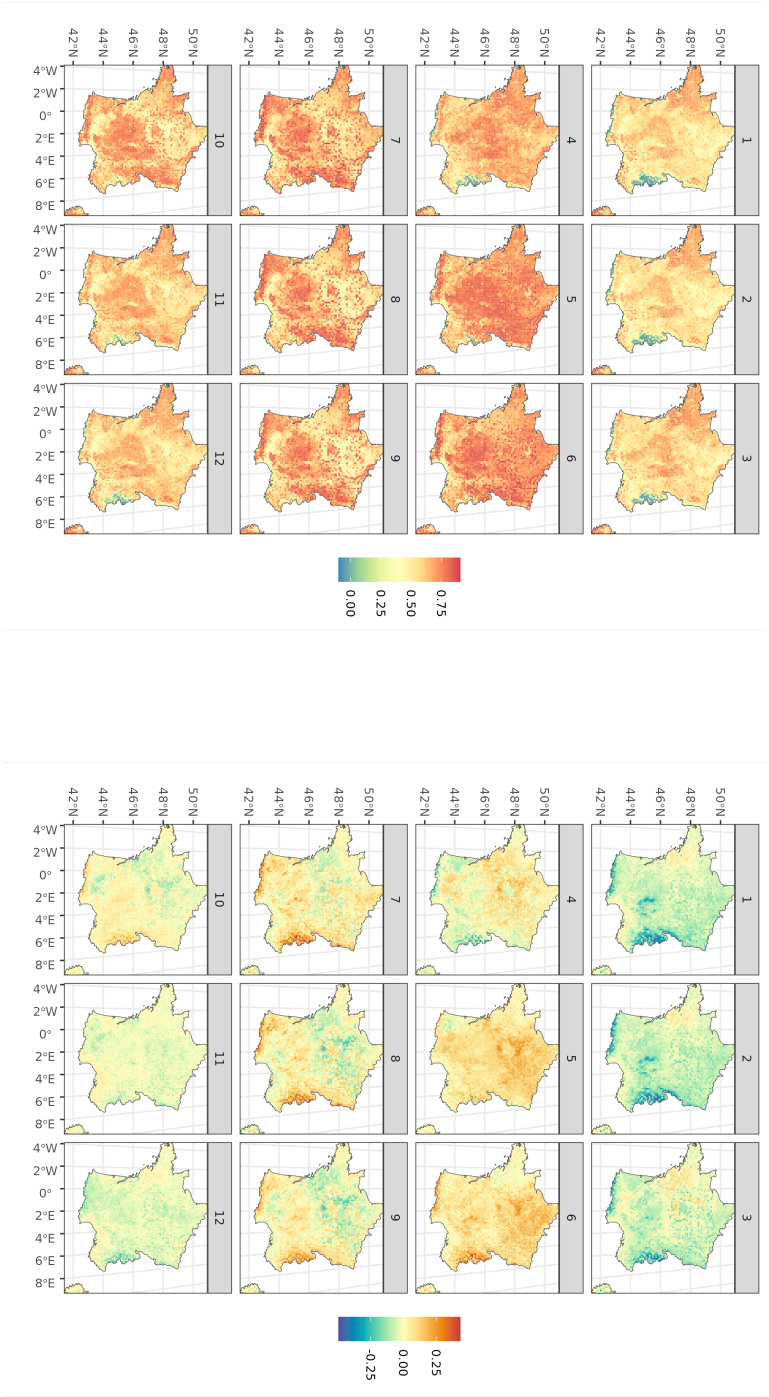

Fig. 1: NDVI data. (left) Monthly spatial data of NDVI $S(x,t)$. (right) Monthly anomalies of the spatial predictions $S^*(x,t)$. Each panel corresponds to the average predictions or anomalies for a month over the period 2000 - 2023.

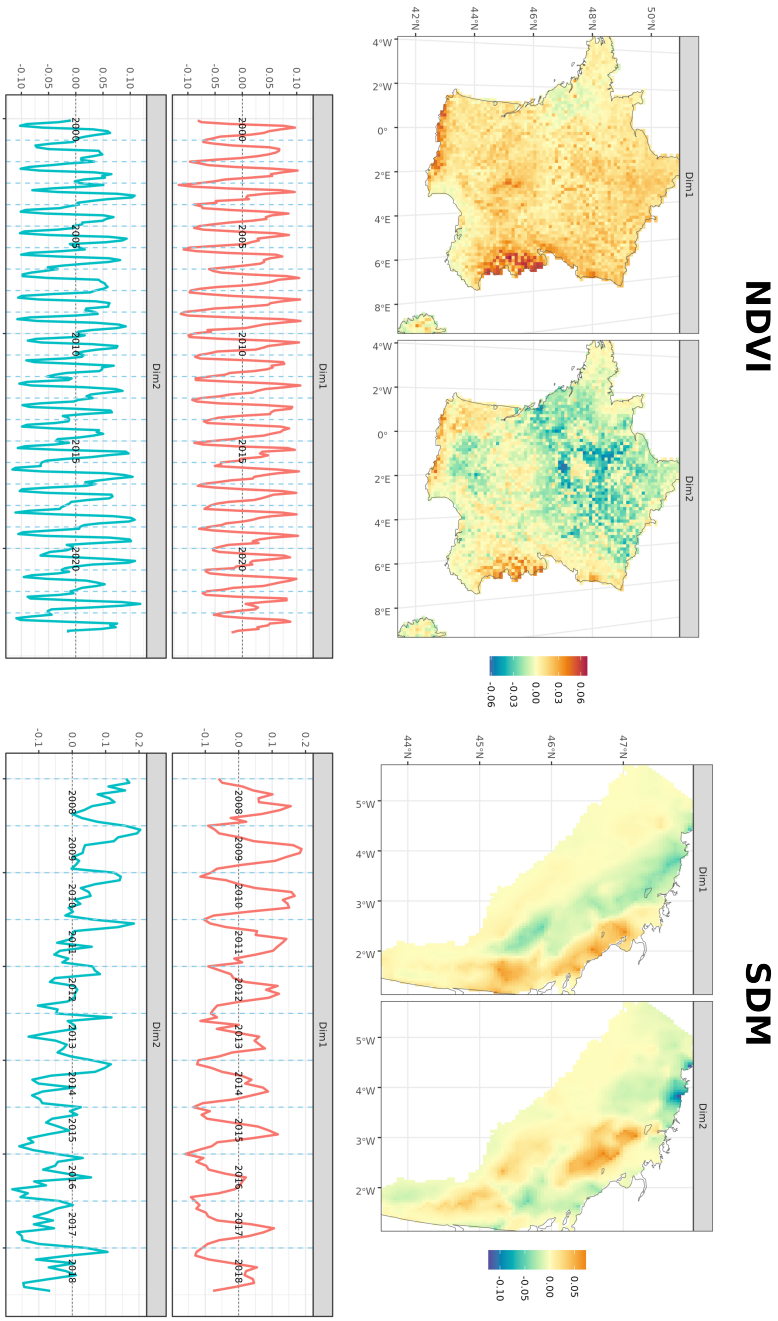

Fig. 2: EOFs results for satellite NDVI data (left) and IST-SDM predictions (right). (Top) Factors for the two first dimensions of the EOFs. (Bottom) Loadings for the two first dimensions of the EOFs. Blue dashed vertical lines correspond to the month of January for each year. The third and fourth dimensions of the IST-SDM case study are presented in Figure S4.2.

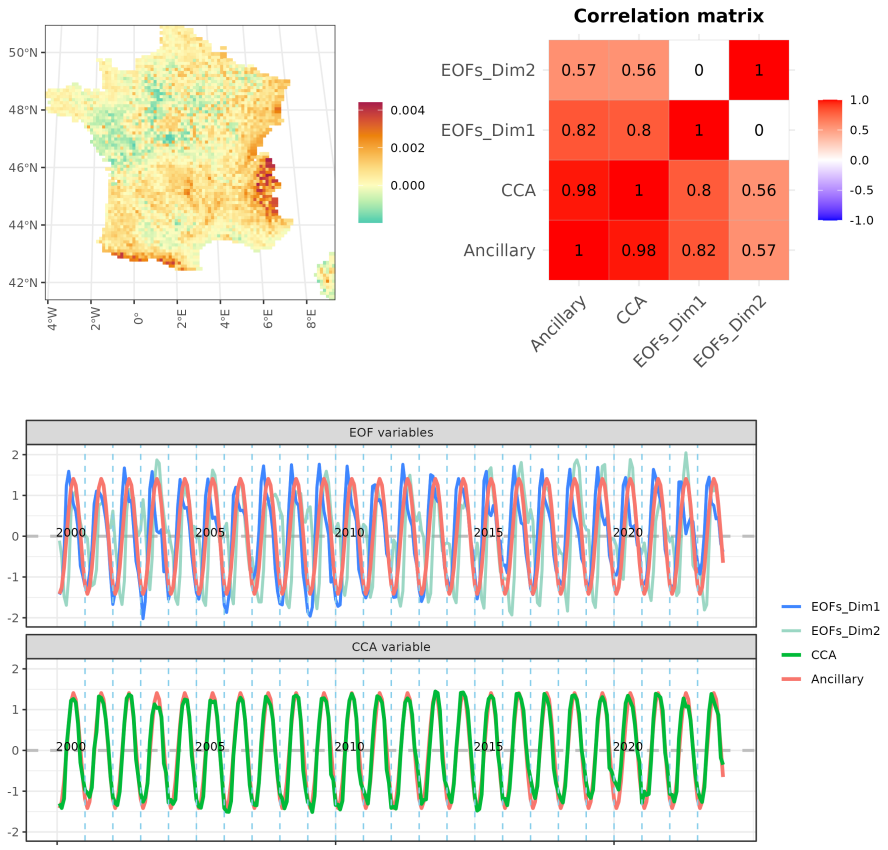

Fig. 3: Satellite NDVI data. Results of the canonical correlation analysis. (Top left) Spatial basis vector that maximize the correlation between the temporal variables $\mathbf{y}_1$ and $\mathbf{y}_2$. (Top right) Correlation matrix between the two first EOFs loadings, the CCA $\mathbf{y}_1$, and the ancillary variable. (Bottom) Comparison of the EOFs loadings with the ecological ancillary variable, and the CCA variable with the ecological ancillary variable. These time series are standardized. Blue dashed vertical lines correspond to the month of January for each year.

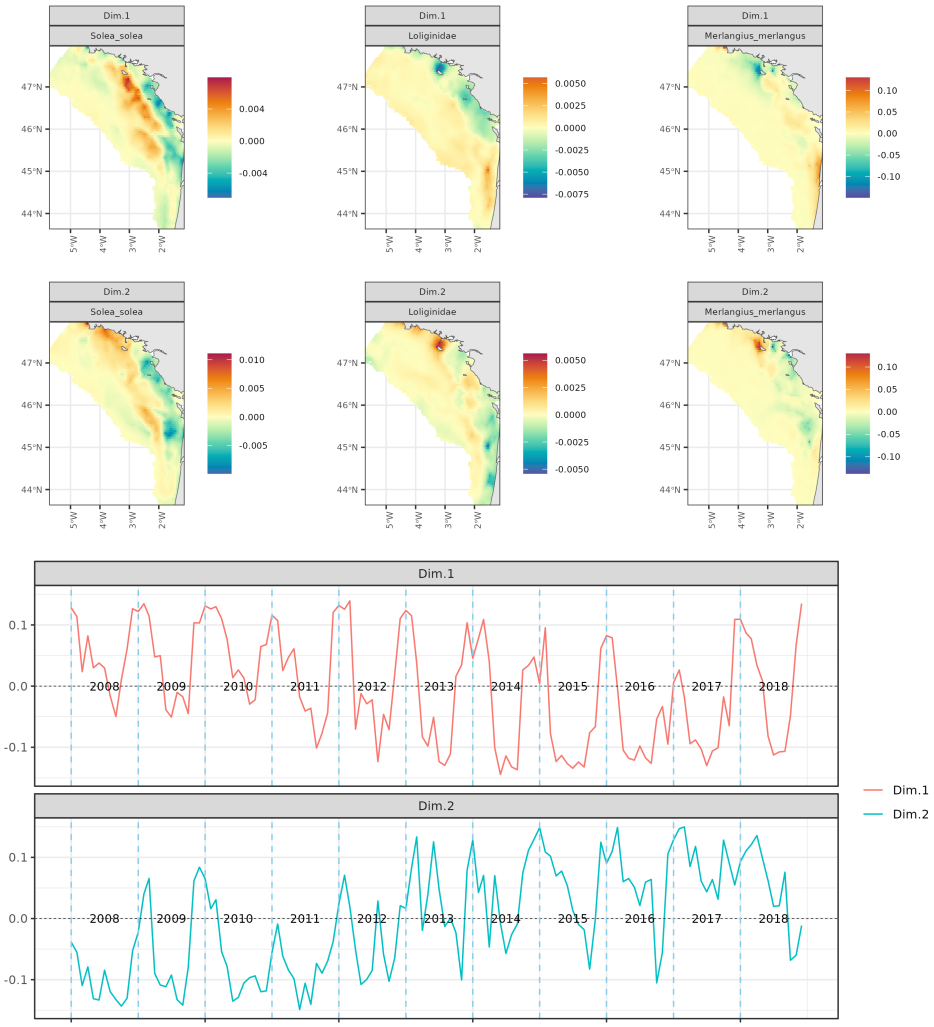

Fig. 4: IST-SDM predictions. Results for the multivariate temporally synthetic EOFs (EOFs on $\mathbf{S}'^{(row)}_{multi}$). (Top) Factor maps for each species and first two dimensions. (Bottom) Loadings for the first two dimensions. Blue dashed vertical lines correspond to the month of January for each year. The loadings emphasize the seasonality of the reproduction for each three species. Orange areas in both dimensions correspond to the reproduction grounds.