

Supplementary material

Using machine learning to link spatiotemporal information to biological processes in the ocean – A case study for North Sea cod recruitment

Kühn, Bernhard; Taylor, Marc H.; Kempf, Alexander

Supplement 1: Details of the SOM analysis

Text S1

SOMs work by projecting a high-dimensional input space to a predefined number of i units (also called neurons) (Pötzlbauer 2004, Sang et al. 2008). The units are arranged in a two-dimensional grid and each unit is associated with a weight vector w_i , which has the same dimension as the input vector. Producing a SOM from the input data is now organised in three steps: 1. Initialisation, 2. Training the SOM iteratively, 3. Final mapping. At first, the weight vectors of each unit are initialised – usually random initialisation as proposed by Kohonen in his original work (Kohonen 1982) or by linear initialisation in the subspace spanned by the first two principal components. Which method to prefer is under discussion (Reusch et al. 2005, Liu et al. 2006, Akinduko et al. 2016). Here we choose random initialisation as Akinduko et al. (2016) showed that random initialisation produces better results in the case of nonlinear input data. During the training process a randomly chosen vector from the high-dimensional input x is selected and the Euclidean distance (= activation function) to each SOM-unit is calculated. The unit closest to the input x_k is selected as the “winner” and its weight vector is updated to better match the presented input, by moving it a fraction of the Euclidean distance towards the presented input vector. The latter is determined by a learning rate α , which decreases as training progresses. Additionally the weights of the neighbouring units of the “winner”-unit are modified according to the specified neighbourhood function ε . This guarantees the topological preservation of the SOM properties as similar states in the input space are mapped to neighbouring units in the lower dimensional output space. The training of the SOM is then proceeded according to the following learning rule:

$$w_i(t+1) = w_i(t) + \alpha(t) \varepsilon(t) [x(t) - w_i(t)] \quad \text{Eq. S1}$$

The training of the SOM is continued until the map units converged. Each observation from the input space is now mapped to its corresponding “winner” also called the best matching units (BMUs) that satisfies the following equation:

$$c(x) = \operatorname{argmin}_i \{\|x - w_i(t)\|\} \quad \text{Eq. S2}$$

As a learning rate, we chose a linear learning rate, decreasing from the initial learning rate $\alpha_0 = 0.05$ to 0.01 during total learning time T_{max} according to:

$$\alpha(t) = \alpha_0 \left(1 - \frac{t}{T_{max}}\right) \quad \text{Eq. S3}$$

Different types of neighbourhood functions can be used, from which “bubble” and “gaussian” are available in the “kohonen” package in R (Wehrens & Kruisselbrink 2018).

$$\varepsilon_{ab}(t) = \begin{cases} F(\sigma_t - d_{ab}), & \text{with } F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases} & \text{bubble} \\ \exp\left(-\frac{d_{ab}^2}{2\sigma_t^2}\right), & & \text{gaussian} \end{cases} \quad \text{Eq. S4}$$

With ε_{ab} defining the neighbourhood relationship between map units a and b with a given neighbourhood radius σ_t at time t and d_{ab} denoting the distance. As both the learning rate and the neighbourhood radius decrease over time, the learning process is guided from a coarse learning with frequent updates in the weights and map to a more fine-tuned adaption of the weights to the input vector. In oceanographic applications, a deterministic batch version of the algorithm is typically used (Liu et al. 2006, Liu, Weisberg, Vignudelli, et al. 2016), where the complete dataset is presented at once and the weights updated accordingly. This has the advantage of being calculated in parallel and has fewer tuning parameters, however, it shows strong dependence on initialisation and problems in organising the data with the possibility of unexpected folds in the mapping (Fort et al. 2002). Therefore, we used the iterative version of the algorithm, described here, and controlled for the stochastic element by repeating the calculation a 100 times. The quality of the SOM-mapping was then assessed with both Quantization error (QE) and Topographic error (TE) (Kiviluoto 1996) and the map with the lowest TE and QE was chosen. QE describes the error, which is made by mapping a high dimensional input space to a lower dimensional one and is simply the average distance of each input vector x_i to the corresponding mapping unit. TE describes the topological preservation properties of the SOM by verifying if the respective best matching unit BMU and second best matching unit $2BMU$ of each observation are neighbours on the map.

$$QE = \frac{1}{N} \sum_{i=1}^N \|x_i - BMU_i\| \quad \text{Eq. S5}$$

$$TE = \frac{1}{N} \sum_{i=1}^N TE(x_i), \quad \text{with } TE(x_i) = \begin{cases} 1, & \text{if } BMU(x_i) \text{ and } 2BMU(x_i) \text{ are not neighbours} \\ 0, & \text{otherwise} \end{cases} \quad \text{Eq. S6}$$

As the result of the mapping is crucially dependent on the size of the grid to which the data is projected, we used the Dynamic Validity index (DVI) (Shen et al. 2005) to obtain an optimal map size. The goal of the DVI is to minimize the intra-compactness of clusters as a measure of homogeneity within the clusters as well as maximising the inter-cluster distances to gain a well separated clustering. To gain an overall index the measures for intra-compactness and inter-separateness are scaled by their maximum value and the DVI is calculated as a weighted sum of the two. The global minimum of the DVI is then taken to represent the optimal value for the cluster partition:

$$DVIndex = \min_{k=1, \dots, K_{max}} \{IntraRatio(k) + \gamma InterRatio(k)\} \quad \text{Eq. S7}$$

$$IntraRatio(k) = \frac{Intra(k)}{\max(Intra)} \quad \text{Eq. S8}$$

$$InterRatio(k) = \frac{Inter(k)}{\max(Inter)} \quad Eq. S9$$

$$Intra(k) = \frac{1}{N} \sum_{i=1}^k \sum_{x \in C_i} \|x - z_i\|^2 \quad Eq. S10$$

$$Inter(k) = \frac{\max_{i,j} (\|z_i - z_j\|^2)}{\min_{j \neq i} (\|z_i - z_j\|^2)} \sum_{i=1}^k \left(\frac{1}{\sum_{i=1}^k \|z_i - z_j\|^2} \right) \quad Eq. S11$$

With x representing an observation in a dataset of N data points, K_{max} the pre-defined upper boundary of clusters, z_i the cluster centre of cluster C_i and γ the tuning parameter allowing to balance between cluster-compactness (*IntraRatio*) and separateness (*InterRatio*). Usually $\gamma = 1$, if there is no/little noise in the raw data. The DVI was calculated for SOMs with grid dimensions $n \times m$ ranging from the lowest possible dimension (1) to \sqrt{N} adjusted downward to the next integer. As larger SOM-maps often produce empty nodes, only Inter-cluster distances to non-empty clusters were considered in the calculation of Inter-separateness. To control for stochasticity in results each SOM, run for 2000 iterations, was build 50 times and the median DVI per grid was calculated. The optimal number of pattern was then chosen to be the minimum of the median DVI.

Supplement 2: Comparison of Extreme Randomized Trees with Random Forest

Comparison Extreme Randomized Trees & Random Forest

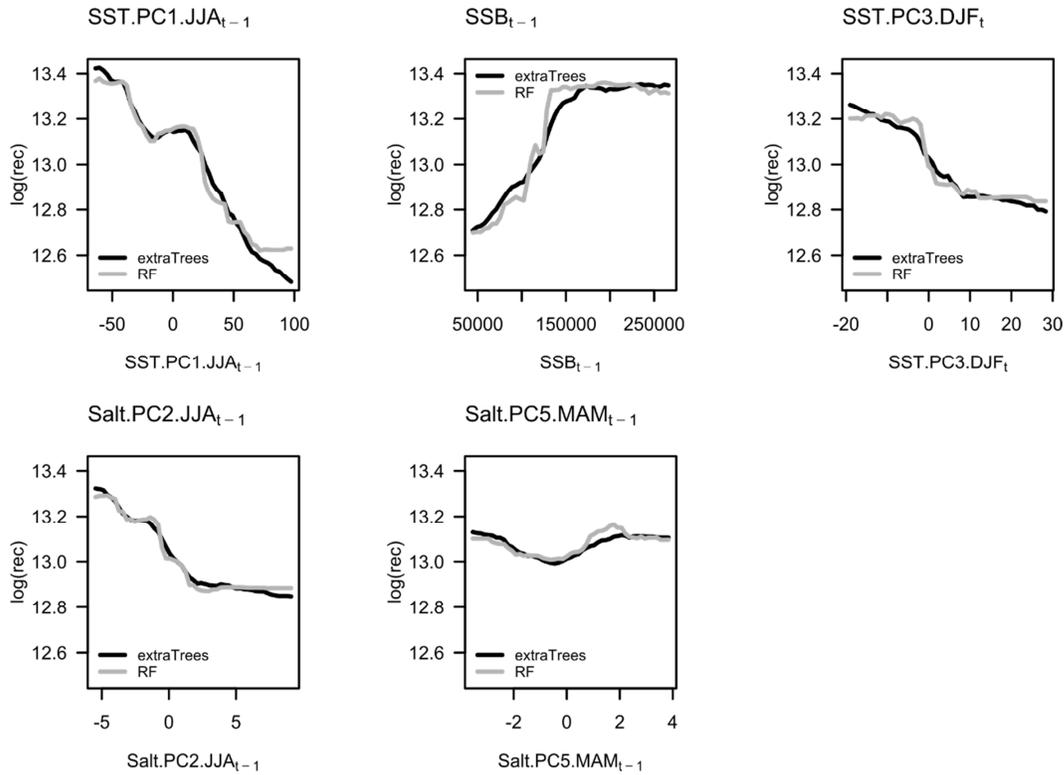


Figure S1: Comparison of the partial effects of each predictor from a Random Forest (RF) and Extreme Randomized Trees (extraTrees) model, which shows a much smoother prediction for Extreme Randomized Trees compared to Random Forest

Supplement 3: Correlation analysis of predictors with local/regional environmental phenomena

Text S2

To understand to which physical phenomena the chosen PC time series in the model correspond to, a correlation analysis with various other environmental variables was performed. Therefore, spatiotemporal fields of SST and Salinity (AHOI, Núñez-Riboni & Akimova (2015)), surface-currents u/v/magnitude (ORAS4, Balmaseda et al. (2013)), 10m-Wind u/v/magnitude (ERA-interim, Berrisford et al. (2011)) and monthly Precipitation over land (E-OBS, Cornes et al. (2018)) were averaged over various spatial and temporal scales (seasonal monthly averages, yearly averages, 0-1 year lags). Seven regions for spatial averaging were chosen (box averaging), spanning the Northwest, Northeast, Southwest, Southeast as well as a smaller region in the Skagerrak as well as the central North Sea and the North Sea region as a whole (Figure S2). Additionally a frontal zone index was derived, by calculating the gradient of the spatial fields of Salinity and SST via the front-detection algorithm of Belkin & O'Reilly (2009) and averaged over the specified regions. This corresponds to an edge-detection algorithm commonly used in image processing. Furthermore important large scale indices

like the North-Atlantic Oscillation (NAO), the East-Atlantic pattern (EA), the Scandinavian pattern (SCAND) (from the NOAA Climate prediction centre, <https://www.cpc.ncep.noaa.gov/data/teledoc/telecontents.shtml>), averaged over different seasons, and the Atlantic meridional oscillation (AMO) (from the NOAA physical sciences laboratory, <http://www.esrl.noaa.gov/psd/data/timeseries/AMO/>) were correlated with the predictors. To account for autocorrelation in the time series, significance of the correlations on the $\alpha = 0.05$ level was estimated by the “modified Chelton method” described in Pyper & Peterman (1998). The five highest significant correlations for each predictor variable are reported in Table S1 and discussed here. To rule out spurious correlations only correlations to environmental variables with a time window of the whole year, the same season or before were considered.

The results of the correlation analysis reveal that SST.PC1.JJA is simply the average of SST-anomalies over the whole North Sea domain. SST.PC3.DJF is positively linked to average SST frontal zone strength in winter, especially in the central and eastern North Sea and SST-anomalies in the northwestern North Sea. Salt.PC2.JJA is strongly positively correlated to the salinity fronts in the northeastern North Sea and negatively associated with salinity anomalies in the Skagerrak, indicating a strong association to the salinity changes driven by the Norwegian Coastal current (NCC) and the Baltic Sea outflow. The high association with frontal zones of the latter two PCs, reveals an interesting aspect of the EOF-analysis, which seems to emphasise small-scale features, especially strong gradients in the data. However, to clarify if these features correspond to actual variability in frontal zone strength, a further validation with data of higher temporal and spatial resolution is needed in future analysis. The Salt.PC5.MAM-signal is the least associated with other environmental variables and weakly correlated to large scale Indices of climate variability (NAO and EA) as well as flow-related (Wind, Currents) variables in winter and spring.

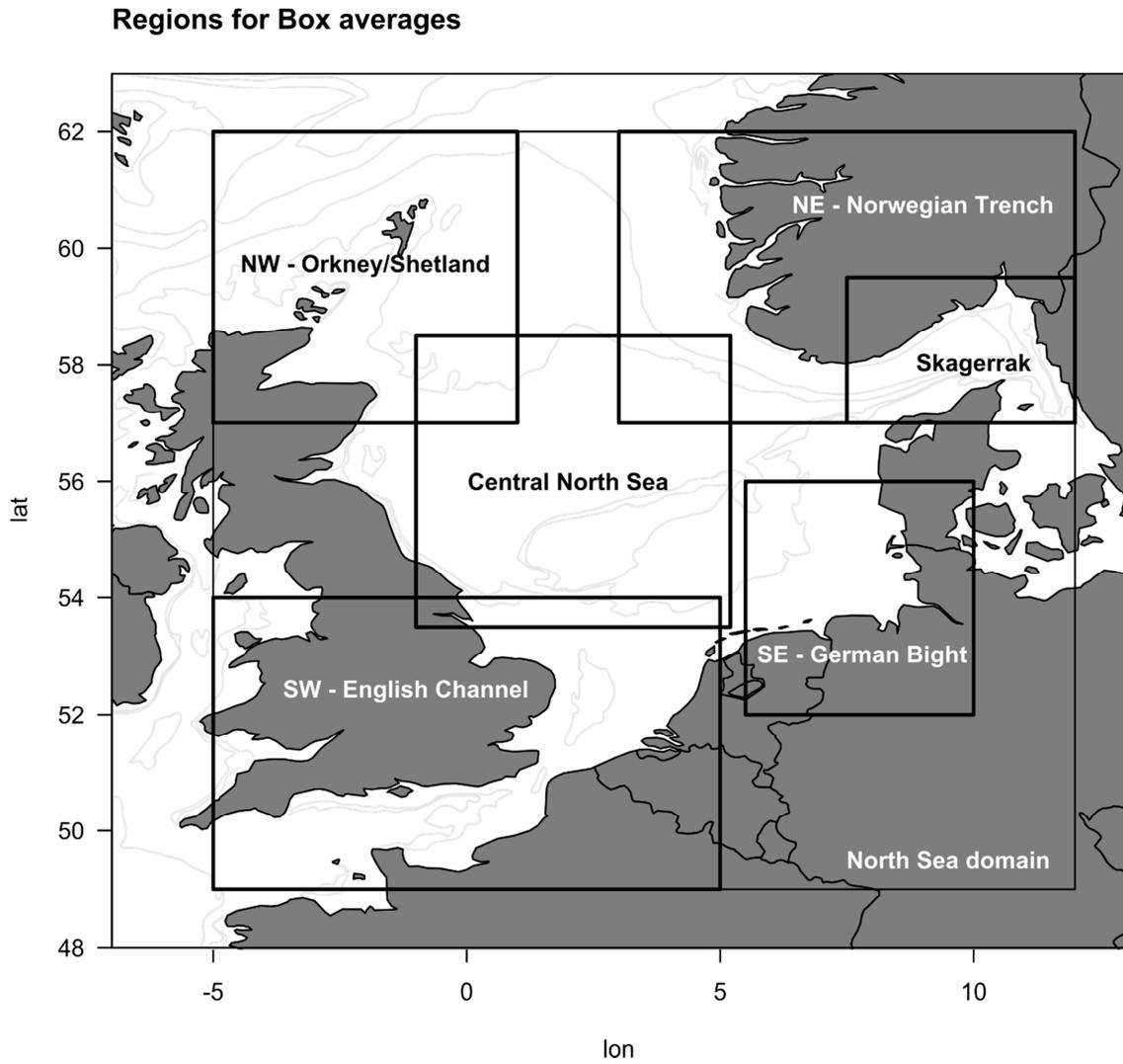


Figure S2: Regions in the North Sea over which environmental variables were averaged for further correlation analysis with the predictors (PC time series) in the final model

Table S1: Five highest significant correlations (r , ($\alpha < 0.05$), method of Pyper & Peterman (1998)) of predictor variables with box-averages of spatio-temporal fields and large scale indices

Variable Name	SST.PC1.JJA	Variable Name	SST.PC3.DJF	Variable Name	Salt.PC2.JJA	Variable Name	Salt.PC5.MAM
Variable, Season, Region, Lag	r	Variable, Season, Region, Lag	r	Variable, Season, Region, Lag	r	Variable, Season, Region, Lag	r
SST-anomaly, JJA, North Sea, lag0	0.996	Fronts-SST, DJF, Central, lag0	0.79	Fronts-Salinity, JJA, NE, lag0	0.78	Wind-v, MAM, SE, lag0	0.341
SST-anomaly, JJA, Central, lag0	0.982	Fronts-SST, DJF, NE, lag0	0.646	Salinity-anomaly, JJA, Skagerrak, lag0	-0.687	NAO, MAM, lag0	0.324
SST-anomaly, JJA, NE, lag0	0.954	Fronts-SST, DJF, North Sea, lag0	0.626	Fronts-Salinity, JJA, Skagerrak, lag0	0.673	EA, MAM, lag0	0.305
SST-anomaly, JJA, SE, lag0	0.943	Fronts-SST, DJF, SE, lag0	0.608	Salinity-anomaly, JJA, NE, lag0	-0.67	Currents-magnitude, MAM, North Sea, lag0	0.299
SST-anomaly, JJA, Skagerrak, lag0	0.914	SST-anomaly, DJF, NW, lag0	0.544	Fronts-Salinity, NE, yearly, lag0	0.639	Currents-v, DJF, Central, lag1	-0.297

Supplement 4: EOF- and SOM-pattern

EOF pattern

SST

SST anomalies DJF

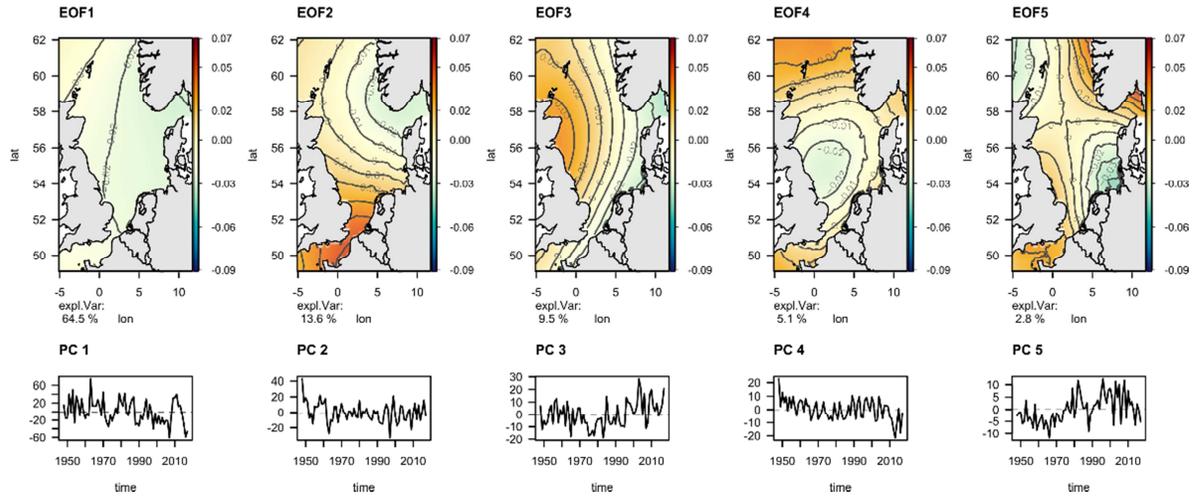


Figure S3: EOF-pattern and corresponding PC time series for SST ($^{\circ}$ C) anomalies in the North Sea in the month December – February (DJF)

SST anomalies MAM

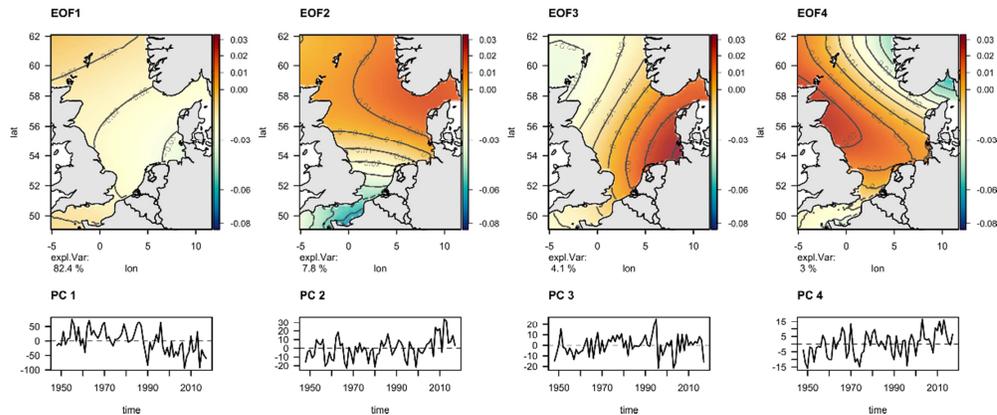


Figure S4: EOF-pattern and corresponding PC time series for SST anomalies ($^{\circ}$ C) in the North Sea in the month March - May (MAM)

SST anomalies JJA

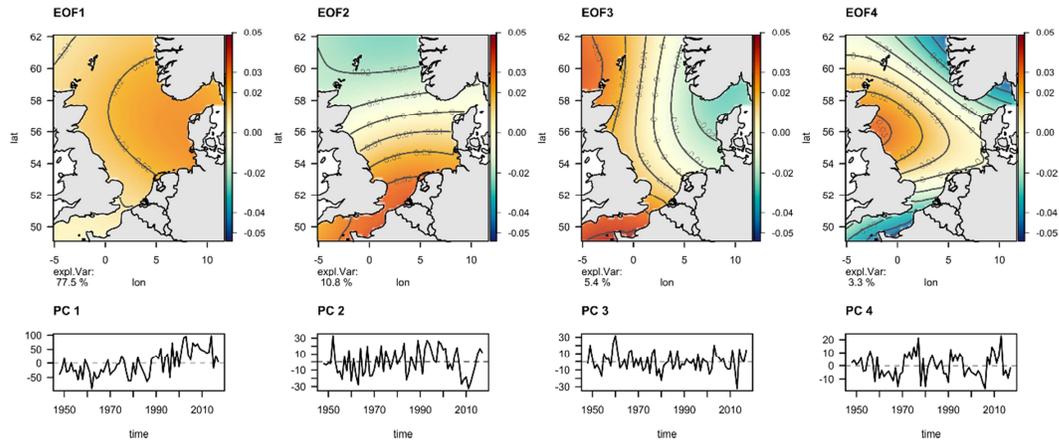


Figure S5: EOF-pattern and corresponding PC time series for SST anomalies ($^{\circ}\text{C}$) in the North Sea in the month June - August (JJA)

SST anomalies SON

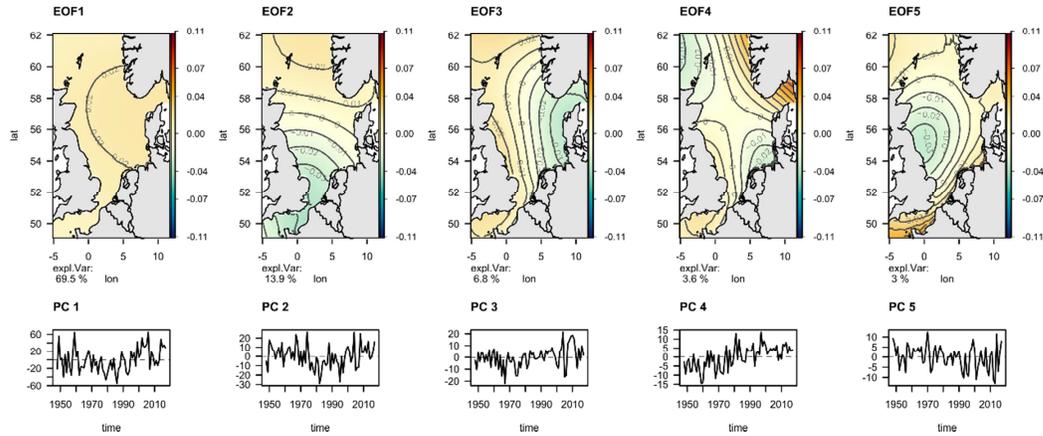


Figure S6: EOF-pattern and corresponding PC time series for SST anomalies ($^{\circ}\text{C}$) in the North Sea in the month September - November (SON)

Salinity

Salinity anomalies DJF

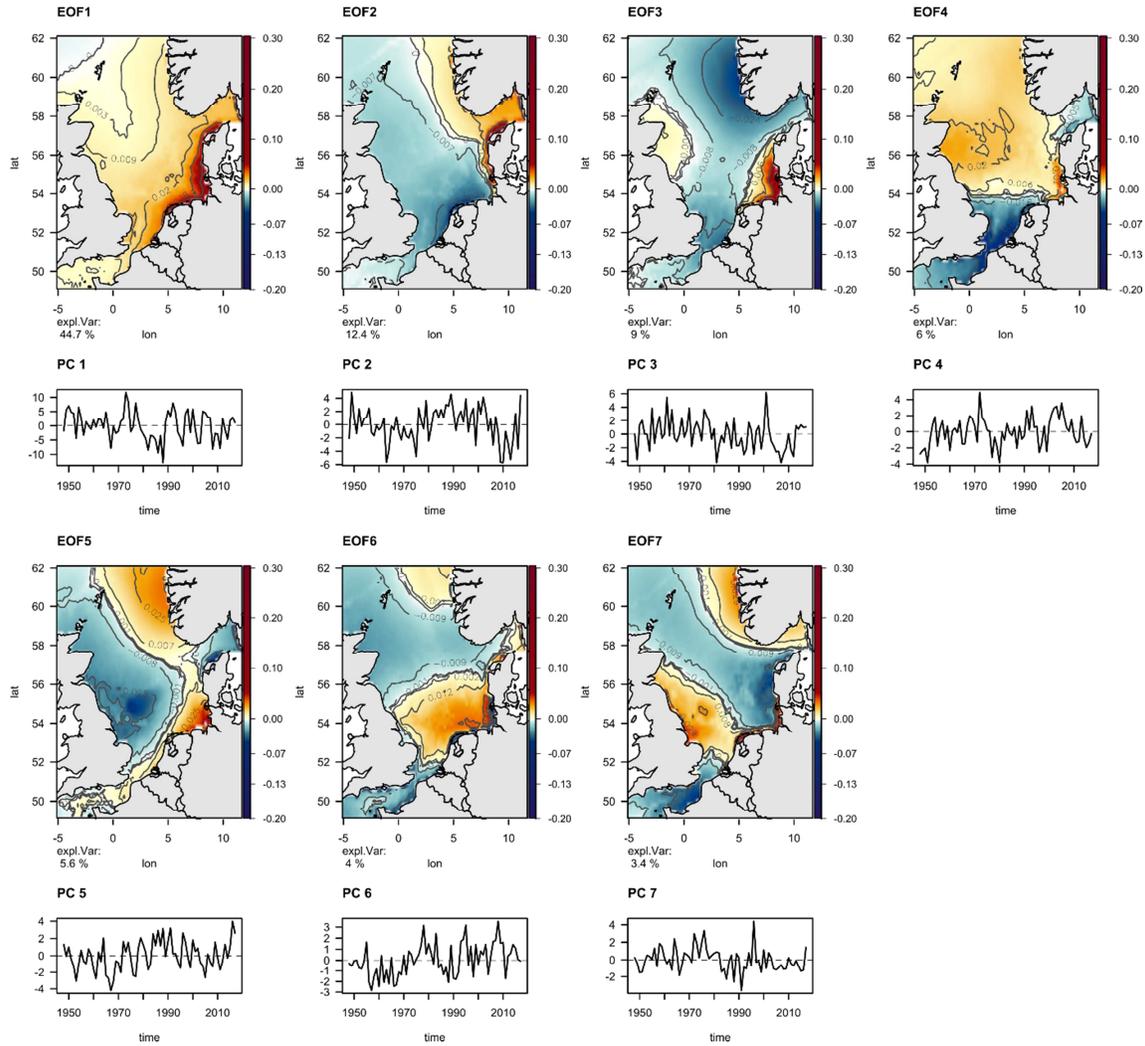


Figure S7: EOF-pattern and corresponding PC time series for salinity anomalies (psu) in the North Sea in the month December - February (DJF)

Salinity anomalies MAM

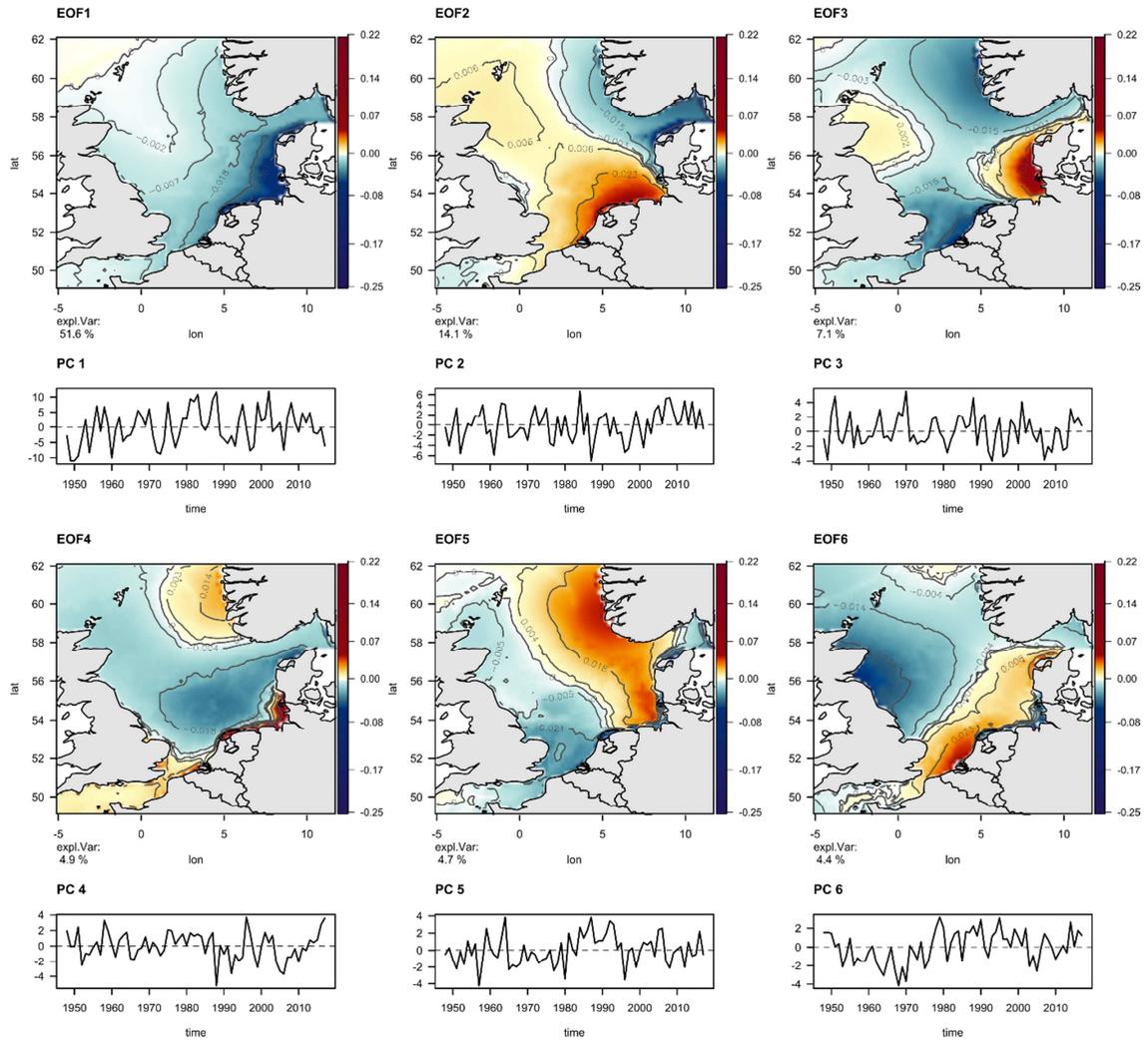


Figure S8: EOF-pattern and corresponding PC time series for salinity anomalies (psu) in the North Sea in the month March - May (MAM)

Salinity anomalies JJA

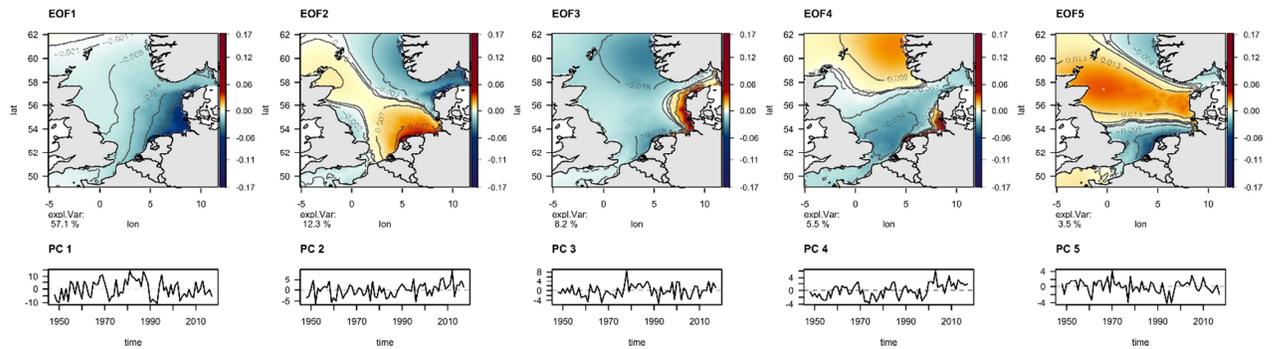


Figure S9: EOF-pattern and corresponding PC time series for salinity anomalies (psu) in the North Sea in the month June - August (JJA)

Salinity anomalies SON

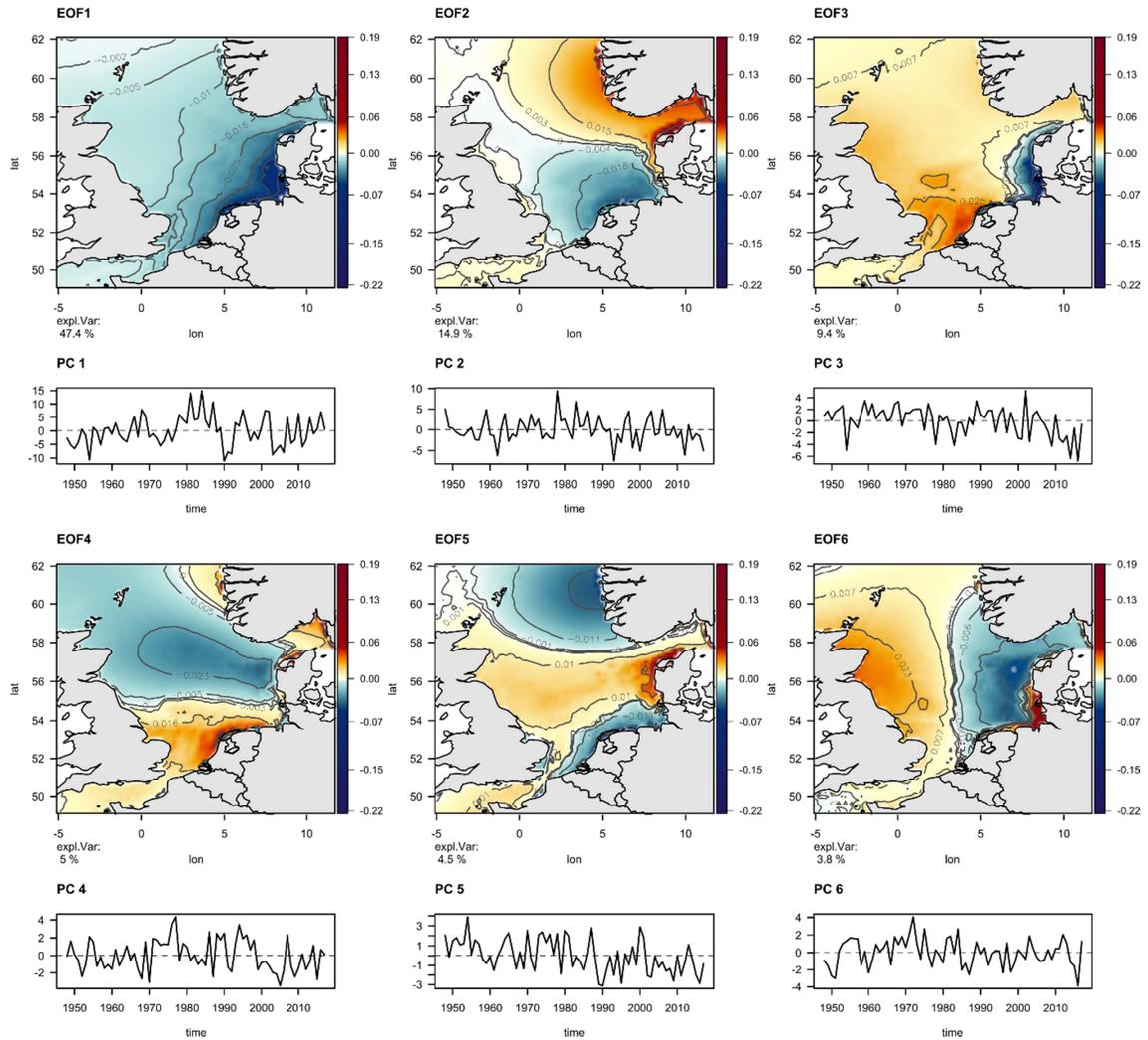


Figure S10: EOF-pattern and corresponding PC time series for salinity anomalies (psu) in the North Sea in the month September - November (SON)

Currents

Current anomalies DJF

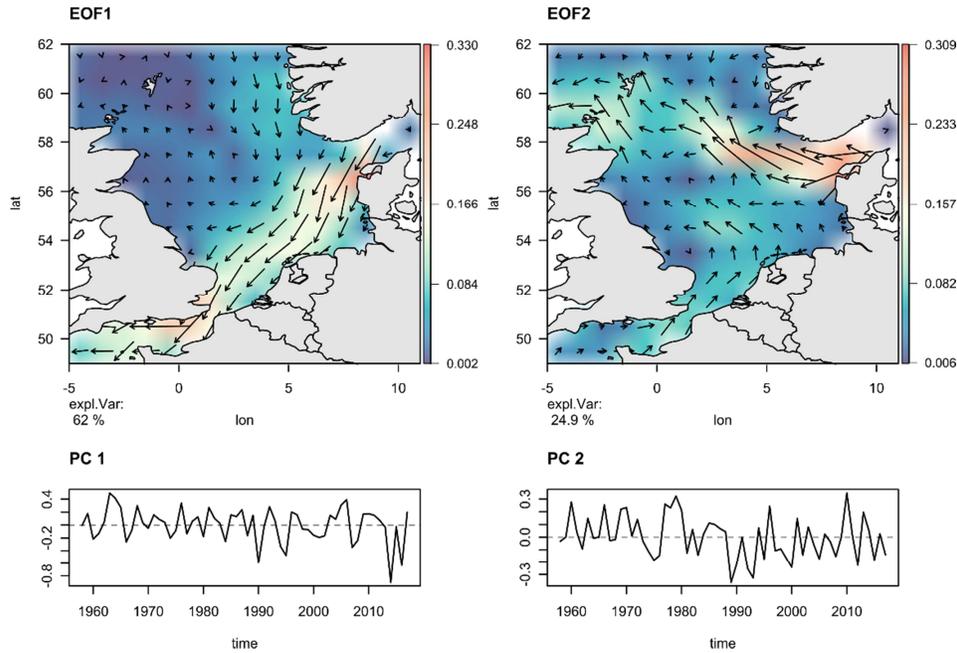


Figure S11: EOF-pattern and corresponding PC time series for Current anomalies (m/s) in the North Sea in the month December - February (DJF), arrows denote the variability in the flow direction and colour represents the magnitude of flows

Current anomalies MAM

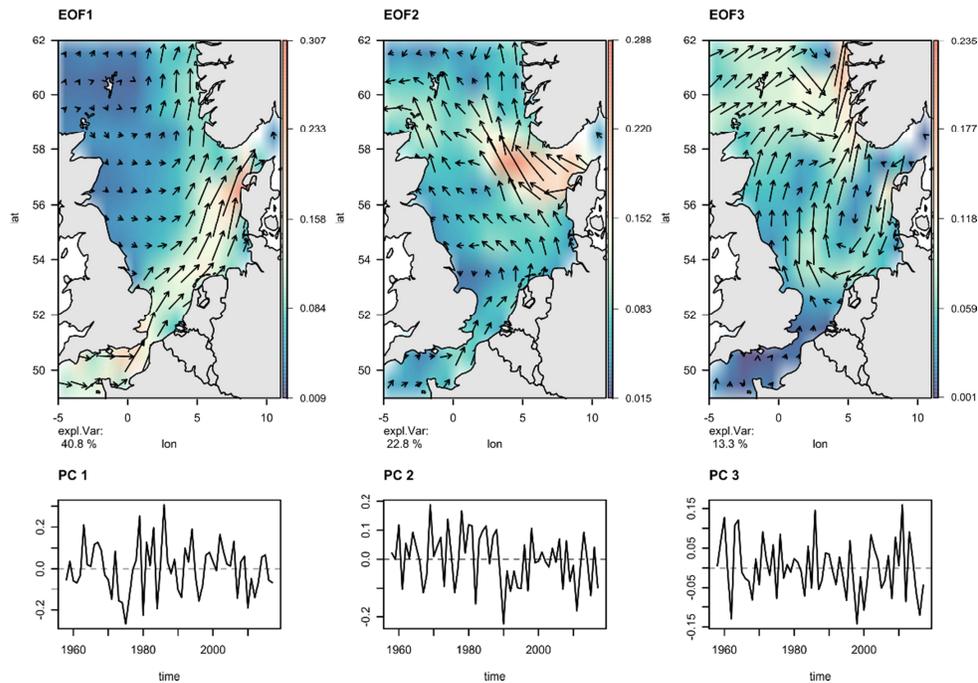


Figure S12: EOF-pattern and corresponding PC time series for Current anomalies (m/s) in the North Sea in the month March - May (MAM), arrows denote the variability in the flow direction and colour represents the magnitude of flows

Current anomalies JJA

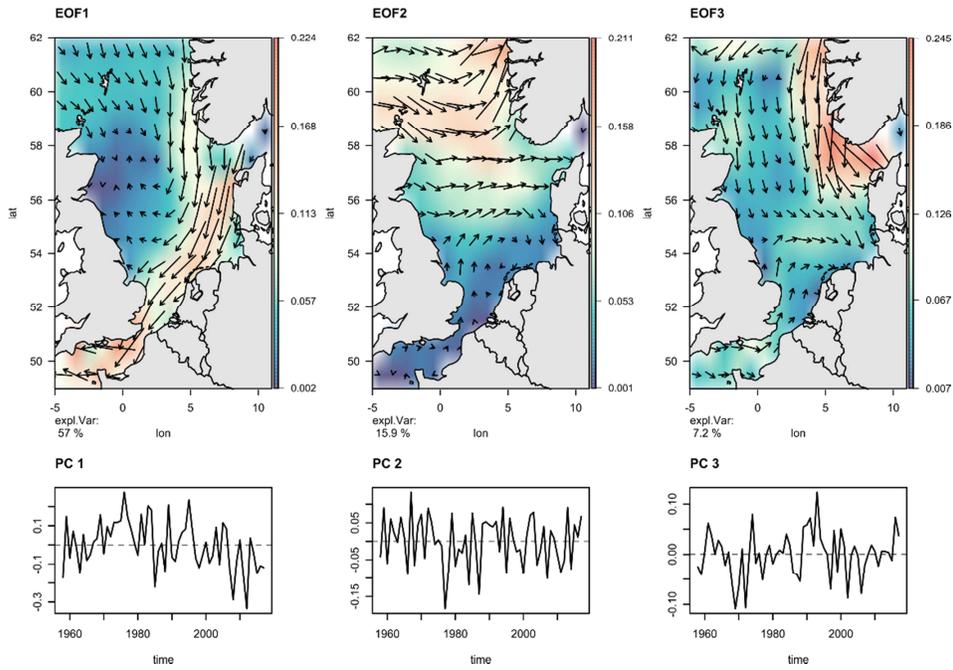


Figure S13: EOF-pattern and corresponding PC time series for Current anomalies (m/s) in the North Sea in the month June - August (JJA), arrows denote the variability in the flow direction and colour represents the magnitude of flows

Current anomalies SON

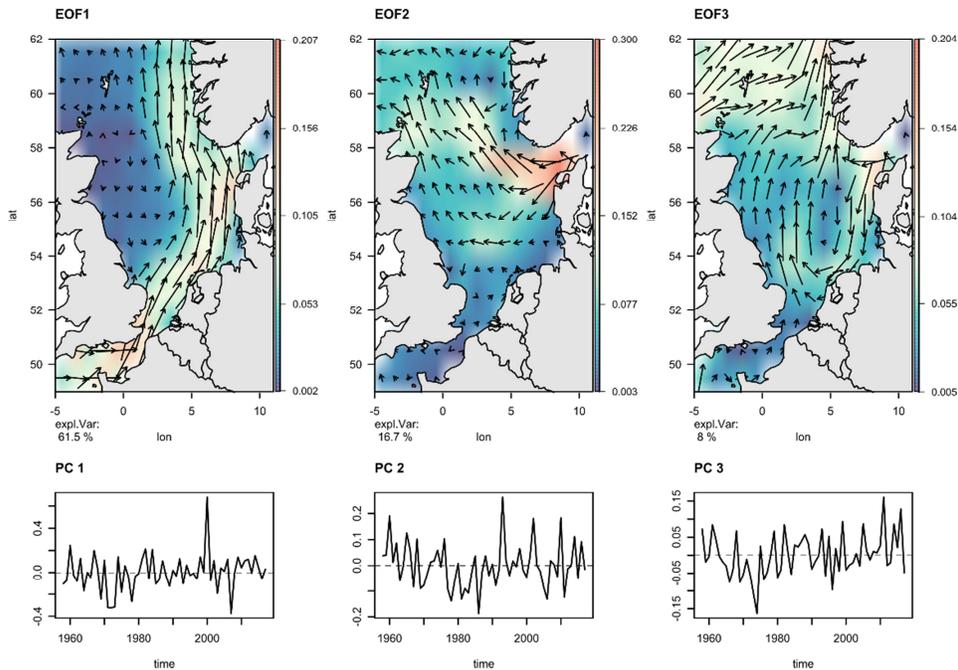


Figure S14: EOF-pattern and corresponding PC time series for Current anomalies (m/s) in the North Sea in the month September - November (SON), arrows denote the variability in the flow direction and colour represents the magnitude of flows

SOM pattern

SST

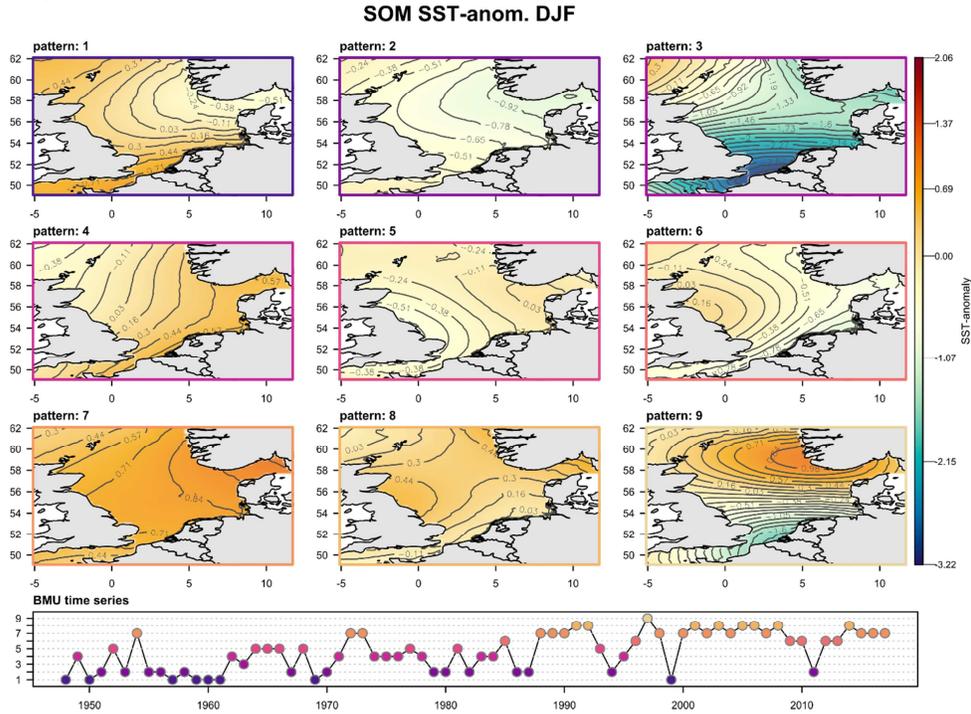


Figure S15: SOM spatial pattern for SST anomalies (°C) with the corresponding BMU-time series, representing the temporal succession of the pattern in the month December – February (DJF)

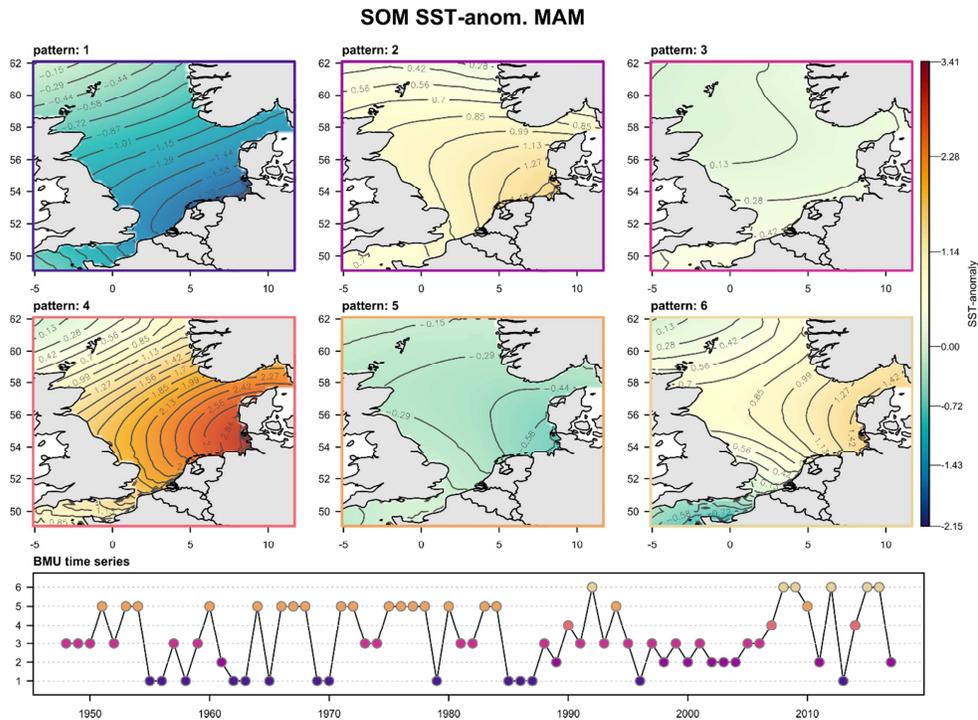


Figure S16: SOM spatial pattern for SST anomalies (°C) with the corresponding BMU-time series, representing the temporal succession of the pattern in the month March – May (MAM)

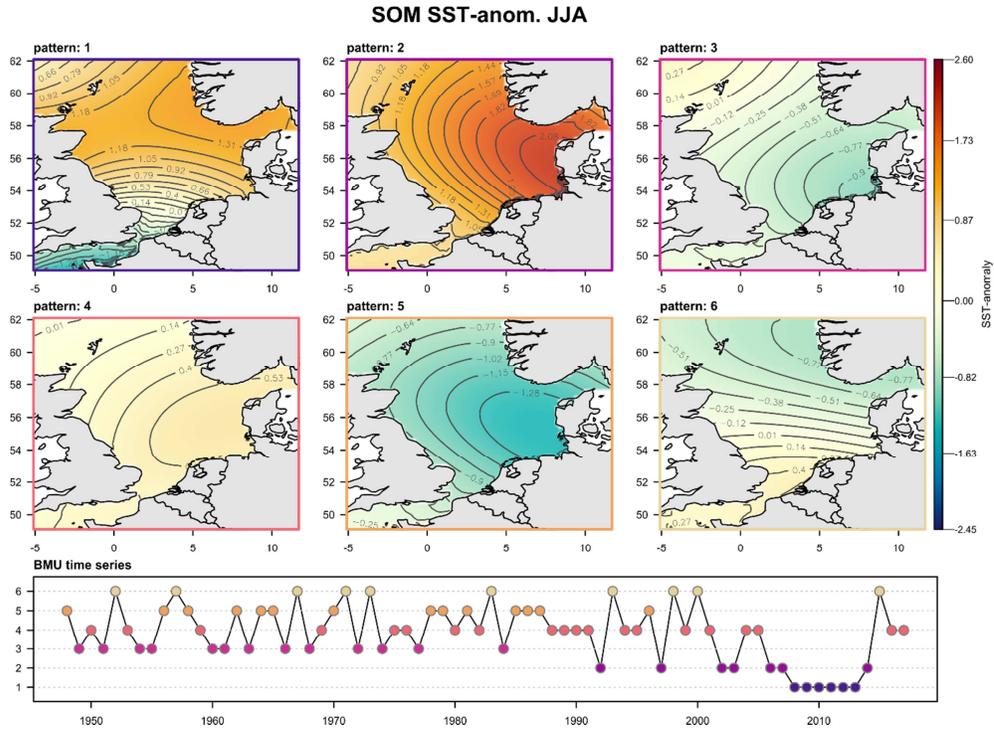


Figure S17: SOM spatial pattern for SST anomalies (°C) with the corresponding BMU-time series, representing the temporal succession of the pattern in the month June - August (JJA)

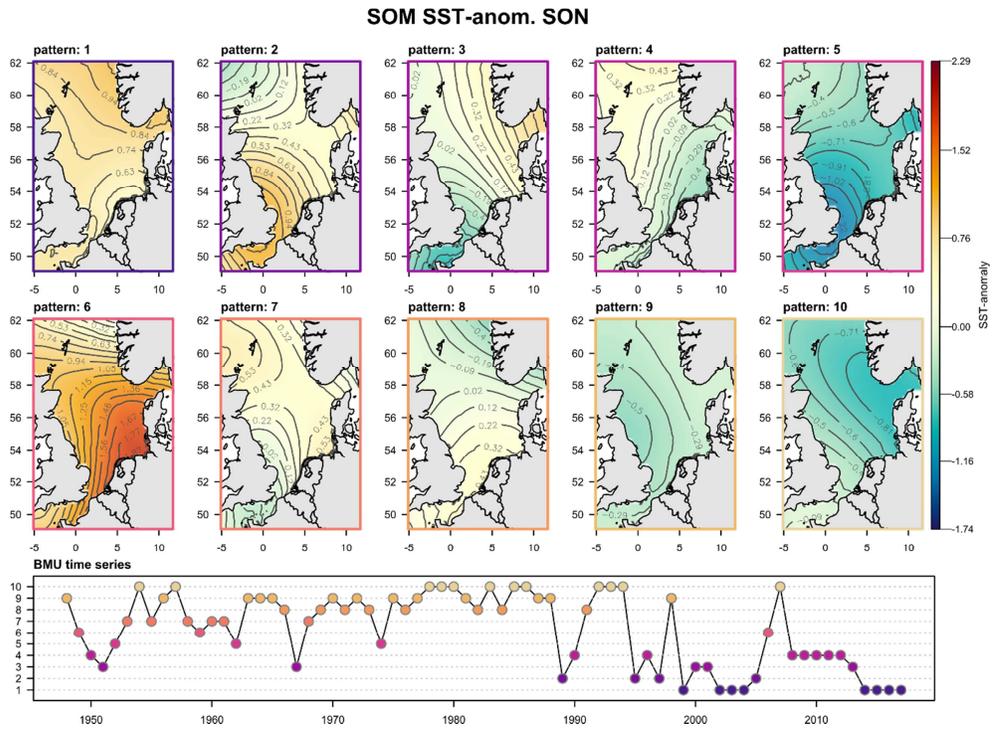


Figure S18: SOM spatial pattern for SST anomalies (°C) with the corresponding BMU-time series, representing the temporal succession of the pattern in the month September – November (SON)

Salinity

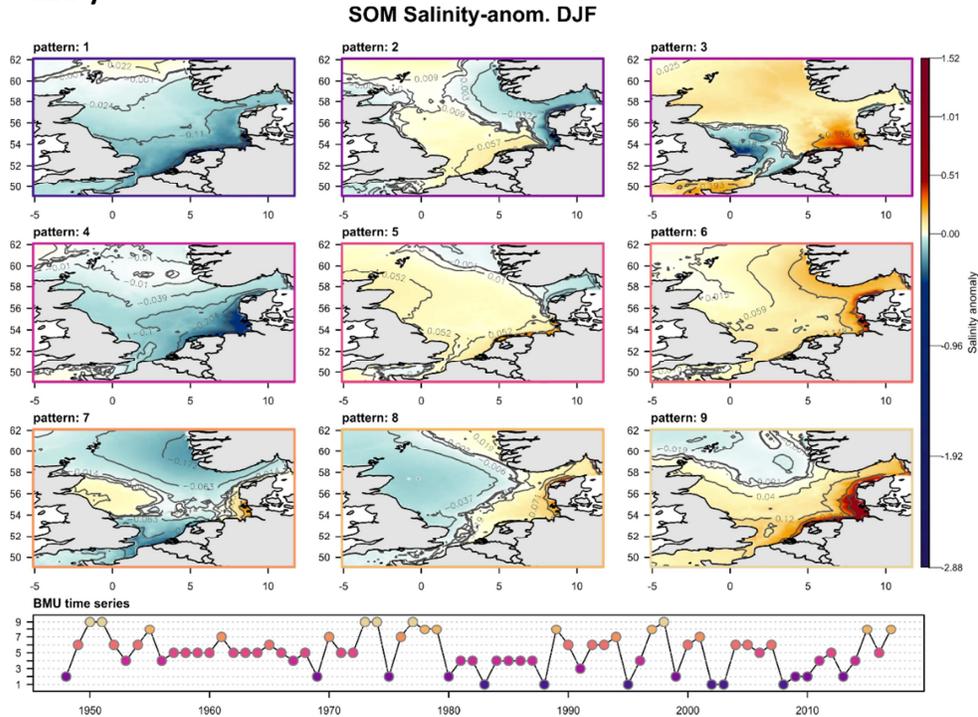


Figure S19: SOM spatial pattern for Salt anomalies (psu) with the corresponding BMU-time series, representing the temporal succession of the pattern in the month December – February (DJF)

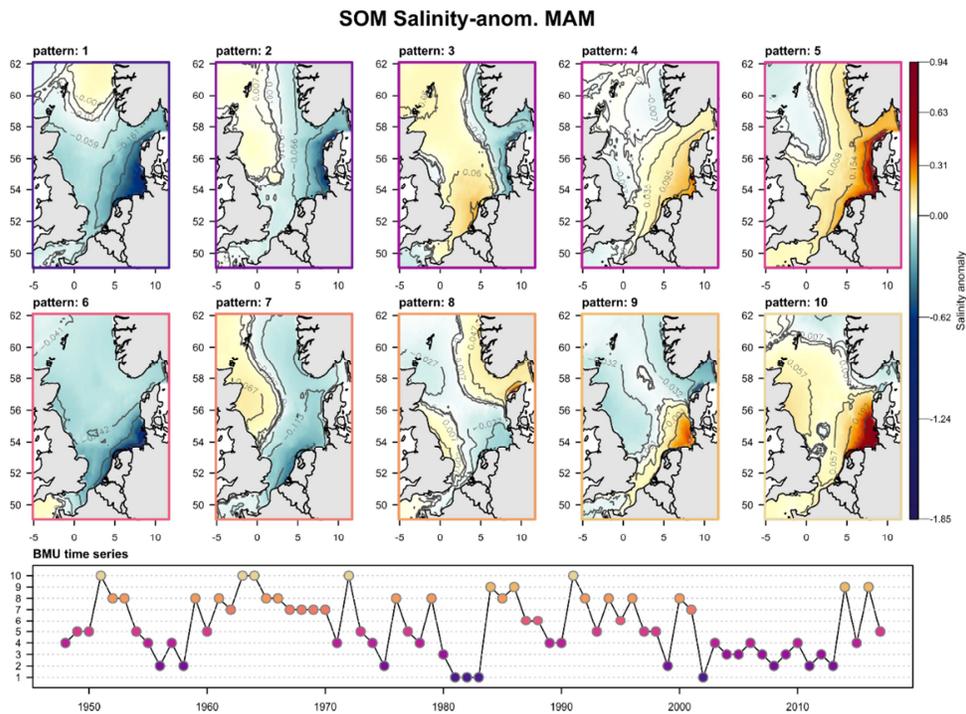


Figure S20: SOM spatial pattern for Salt anomalies (psu) with the corresponding BMU-time series, representing the temporal succession of the pattern in the month March – May (MAM)

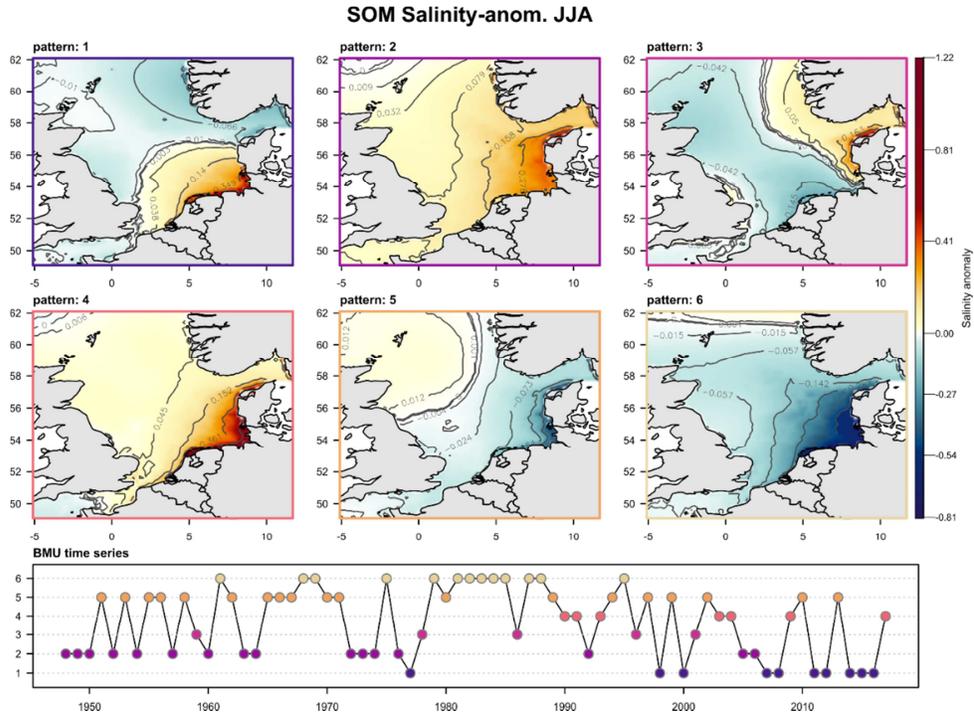


Figure S21: SOM spatial pattern for Salt anomalies (psu) with the corresponding BMU-time series, representing the temporal succession of the pattern in the month June - August (JJA)

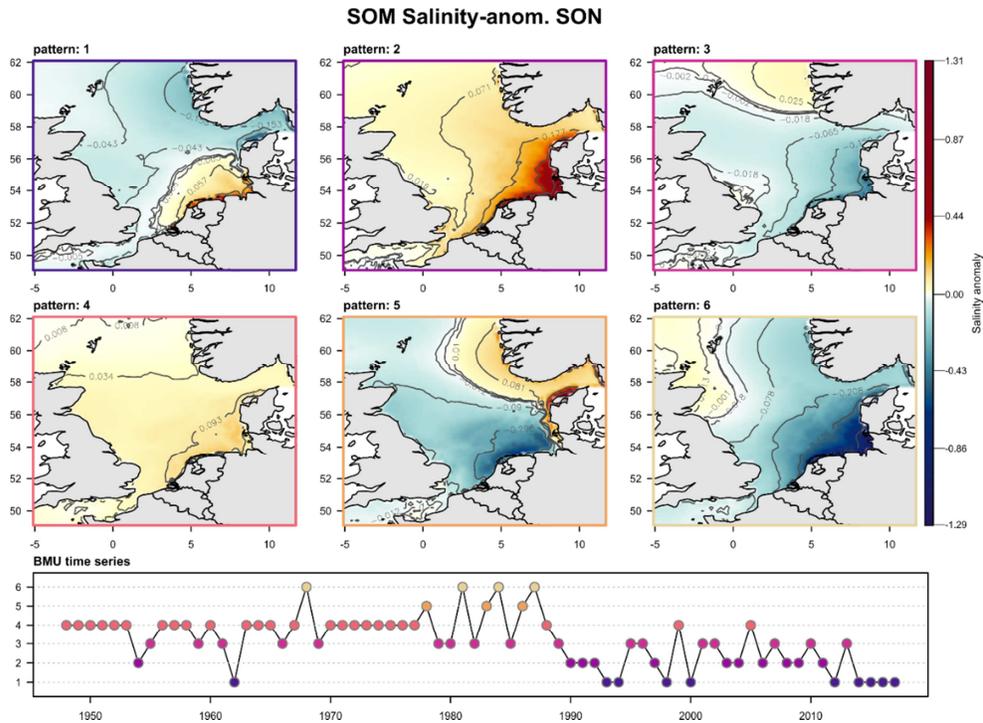


Figure S22: SOM spatial pattern for Salt anomalies (psu) with the corresponding BMU-time series, representing the temporal succession of the pattern in the month September - November (SON)

Currents

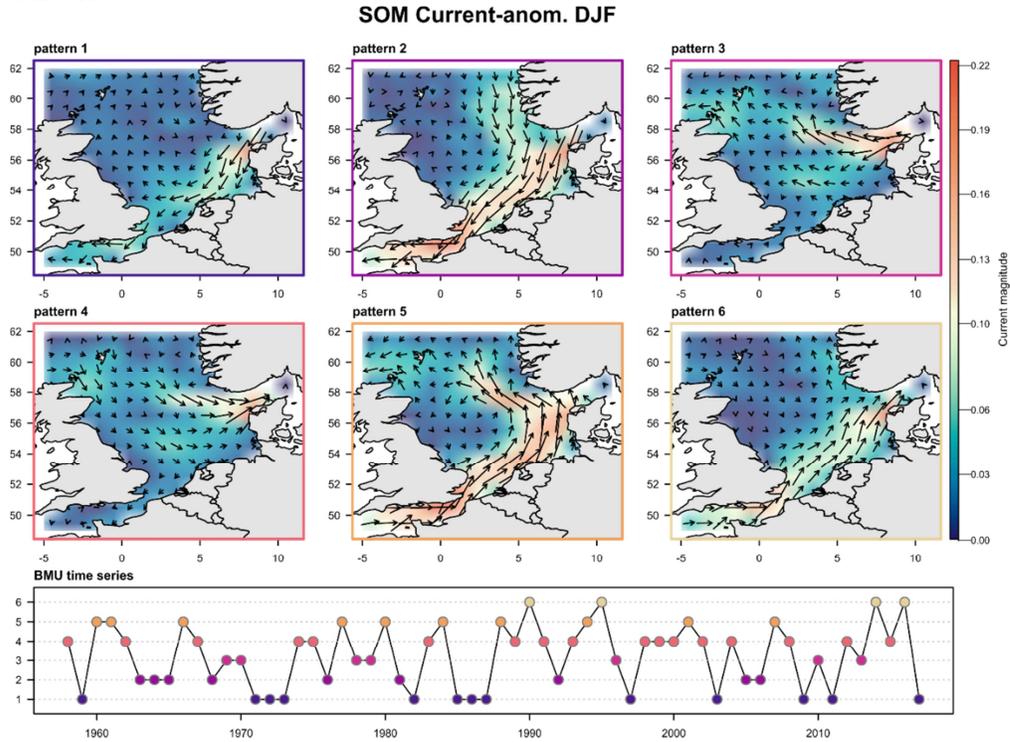


Figure S23: SOM spatial pattern for Current anomalies (m/s) with the corresponding BMU-time series, representing the temporal succession of the pattern in the month December – February (DJF), arrows denote the variability in the flow direction and colour the magnitude of the flow

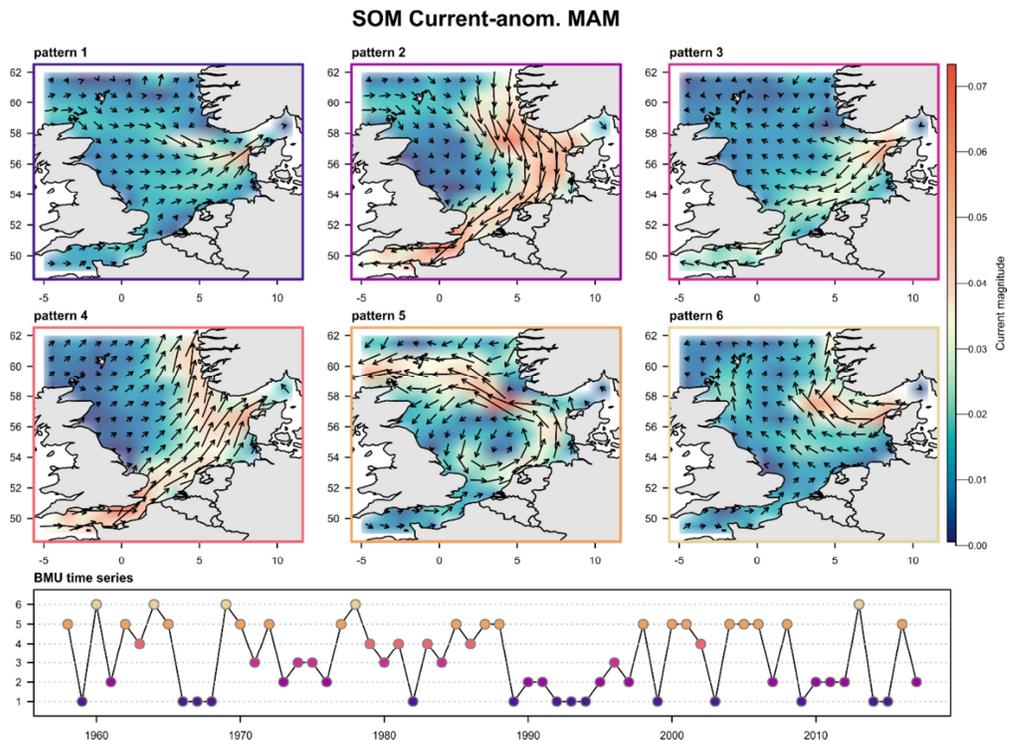


Figure S24: SOM spatial pattern for Current anomalies (m/s) with the corresponding BMU-time series, representing the temporal succession of the pattern in the month March – May (MAM), arrows denote the variability in the flow direction and colour the magnitude of the flow

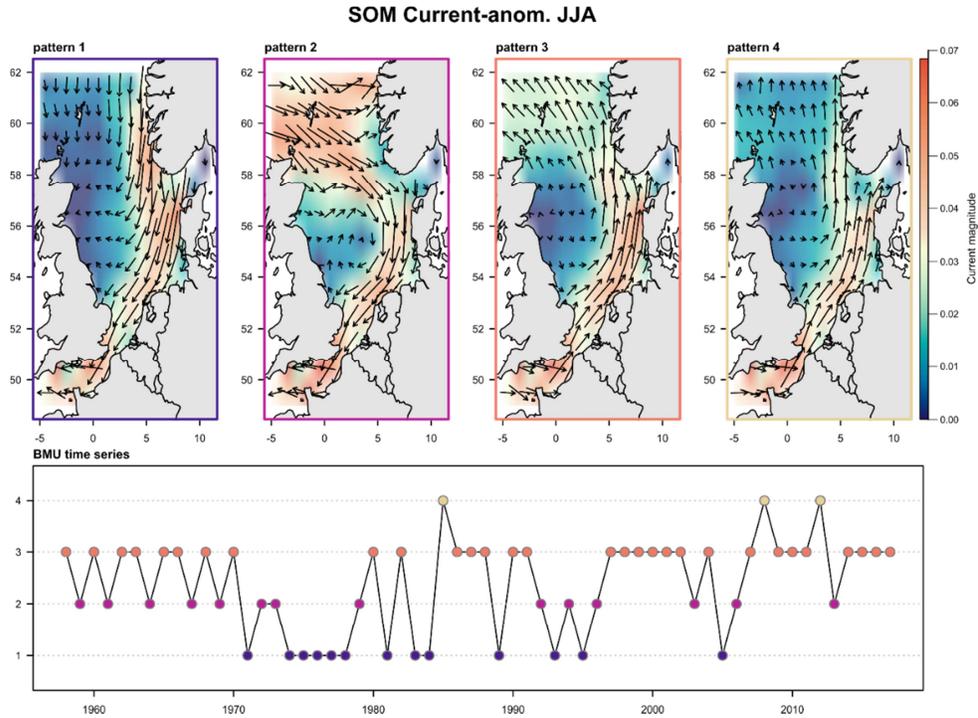


Figure S25: SOM spatial pattern for Current anomalies (m/s) with the corresponding BMU-time series, representing the temporal succession of the pattern in the month June – August (JJA), arrows denote the variability in the flow direction and colour the magnitude of the flow

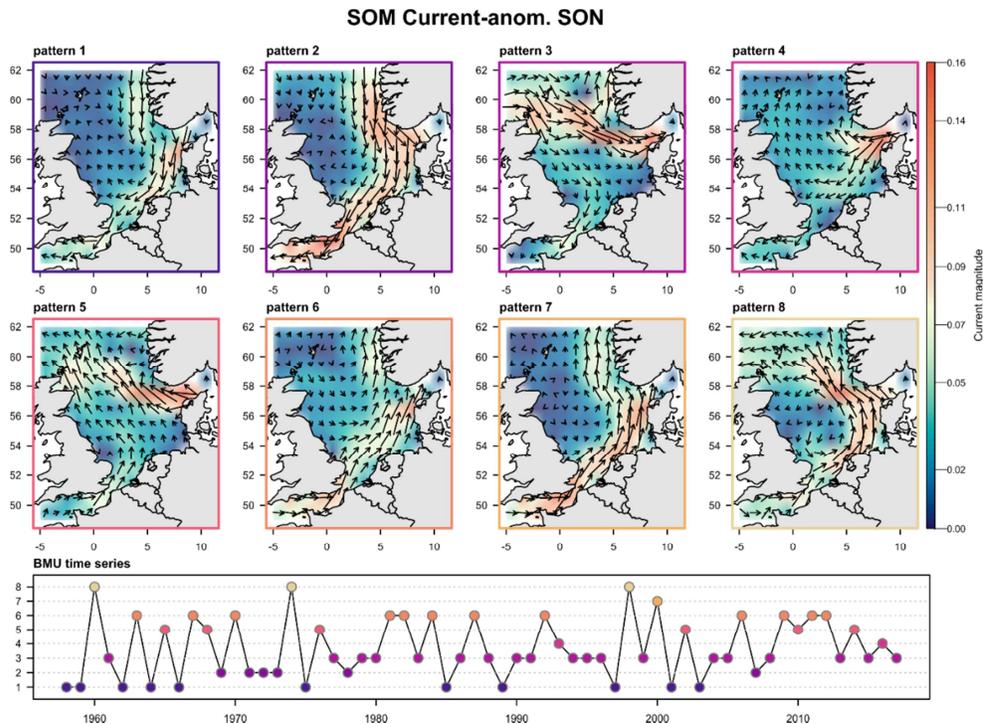


Figure S26: SOM spatial pattern for Current anomalies (m/s) with the corresponding BMU-time series, representing the temporal succession of the pattern in the month September – November (SON), arrows denote the variability in the flow direction and colour the magnitude of the flow

SOM optimal Cluster number

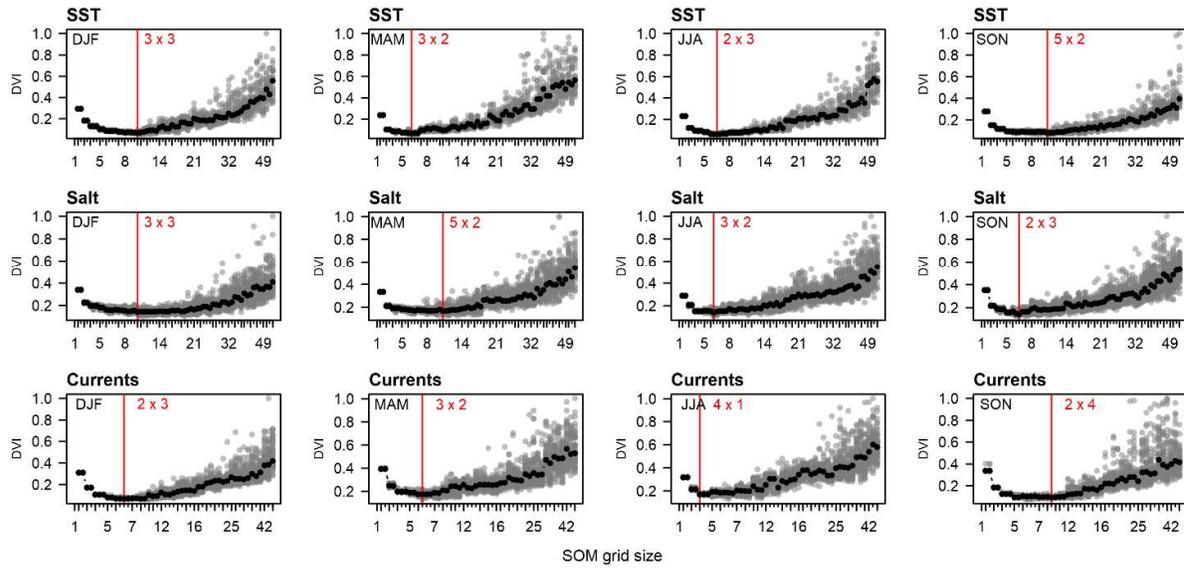


Figure S27: Dynamic Validity index (DVI) (y-axis) for various SOM grid sizes (x-axis). For easier visualisation the x-axis denotes the number of pattern/cluster as the product of the x-y-dimension of the mapping grid (smaller tick marks refer to the same grid size as the subsequent larger tick mark). The final solution, where the median-DVI has its minimum, is demarked with a red line and text, showing the x and y dimension of the map (e.g. 3 x 2); gray: results of 50 stochastic repetitions, black: median DVI over 50 repetitions per grid size

SOM clustering in EOF-space

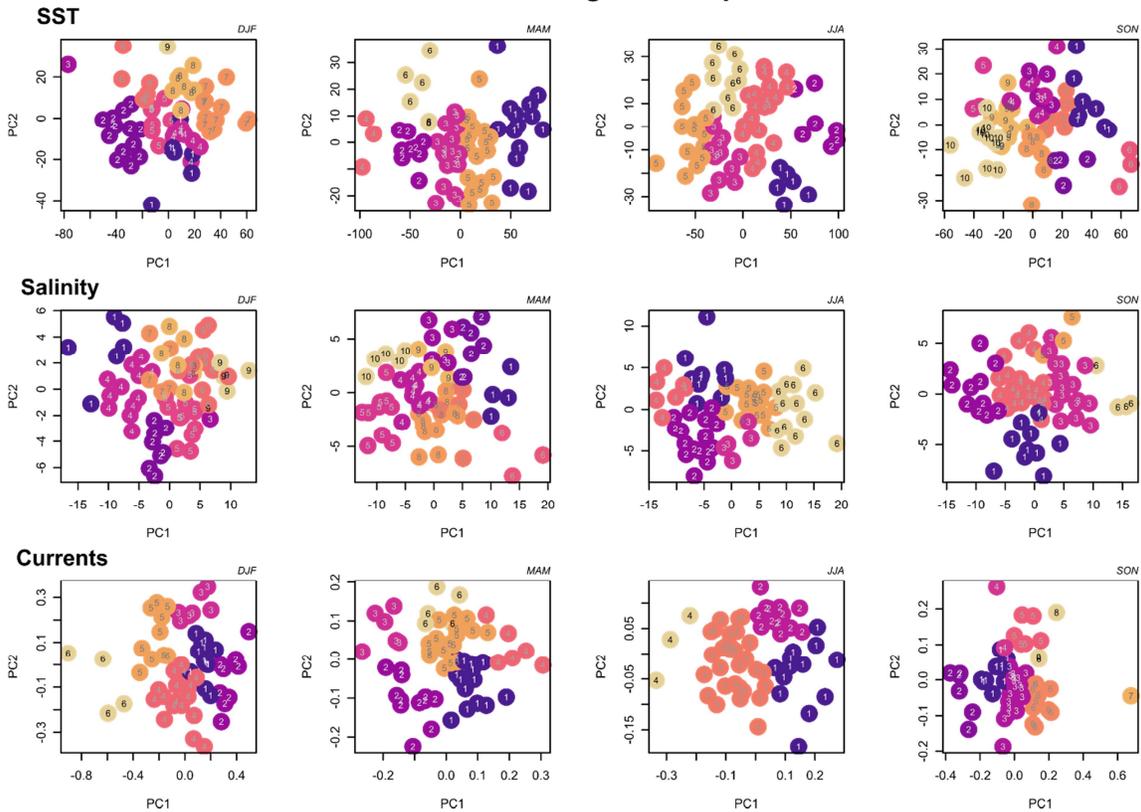


Figure S28: Comparison of the clustering obtained by SOM-analysis (colours and numbers) and the scores from the leading two principal components (PC) (x- and y-axis).

Supplement 5: Results of the model with SOM-derived features

Table S2: Final models as part of the Pareto-Front for the SOM analysis, the chosen solution is highlighted in bold

Covariates	Nr. of Predictors	RMSE _{cv(5x30)}	MAE _{cv(5x30)}	Pseudo-R ² _{cv(5x30)}	Distance to (0,0)
$(SST.SOM.JJA)_{t-1}$	1	0.623	0.495	0.439	0.760
$(SST.SOM.JJA)_{t-1}$ + $(SST.SOM.SON)_{t-1}$	2	0.54	0.437	0.575	0.667
SSB_{t-1} + $(SST.SOM.JJA)_{t-1}$ + $(SST.SOM.SON)_{t-1}$	3	0.51	0.419	0.628	0.636
SSB_{t-1} + $(SST.SOM.JJA)_{t-1}$ + $(SST.SOM.DJF)_t$ + $(SST.SOM.SON)_{t-2}$	4	0.508	0.404	0.628	0.646
SSB_{t-1} + $(SST.SOM.JJA)_{t-1}$ + $(SST.SOM.SON)_{t-1}$ + $(Currents.SOM.JJA)_{t-1}$ + $(Salt.SOM.JJA)_{t-1}$	5	0.502	0.414	0.648	0.662

Supplement 6: Performance evaluation of feature selection

Text S3

We used an artificial example to test if NSGA-II is able to recover true features among a large number of nonsense features with similar properties than the original ones. As a benchmark we compared its performance to other commonly used algorithms namely Boruta (a wrapper feature selection for Random Forest) (Kursa et al. 2010), Recursive feature elimination (RFE) (Guyon et al. 2002) and a Filter method, fitting a generalised additive model (GAM) to each predictor and evaluating its significance (Kuhn 2008). The example consists of four true features of $N = 100$ that mimic typical autocorrelation structure found in environmental data – a first-order autoregressive process (AR(1)), a first-order autoregressive-moving average-process (ARMA(1,1)), a long-wave sine and a cosine with a higher frequency. These four signals were combined to form the response variable by modulating the AR(1) by a time-dependent logistic function, the ARMA(1,1) with a positive quadratic function and an interaction between the sine and the cosine with linear trend and adding normal distributed noise. 200 additional nonsense features were created by the iterative amplitude adjusted Fourier Transform (IAAFT) algorithm (Schreiber & Schmitz 1996, 2000), shuffling the values of the original features, but preserving their inherent autocorrelation. The machine learning model used was Random Forest. NSGA-II was run with the parameter-settings specified in the main methods section. To control for the inherent randomness in the methods, the simulation was repeated for five different seed values.

Results of the simulation showed that NSGA-II showed both highest performance (lowest RMSE) and fewest number of features (6) compared to Boruta, RFE and the Filter (Figure S29). In all of the simulations it recovered three of the four original features (AR(1), ARMA(1,1), Sine) with three additional nonsense features. Notice that the absolute best model in both performance and number of features is the one with three features and not the full model with four true features, indicating that Random Forest has problems detecting the interaction between the Sine and Cosine features. Boruta was the only algorithm able to recover all the four true features, however under the expense of selecting a large amount of nonsense features (17-19).

AR(1):

$$X_t = c + 0.3X_{t-1} + \varepsilon_t$$

Eq. S12

ARMA(1,1):

$$X_t = c + \varepsilon_t + 0.3 X_{t-1} - 0.5\varepsilon_{t-1}$$

Eq. S13

Sine:

$$X_t = \sin\left(\frac{2\pi}{365} t\right)$$

Eq. S14

Cosine:

$$X_t = \cos\left(\frac{2\pi}{73} t\right)$$

Eq. S15

Model:

$$Y_t = \frac{0.6}{1 + \exp\left(-0.2 \left(t - \frac{N}{2}\right)\right)} AR(1) + \frac{\left(t - \frac{N}{3}\right)^2}{\max\left(\left(t - \frac{N}{3}\right)^2\right)} ARMA(1,1) + \frac{0.6t}{N} * Sine$$

* *Cosine* + $\varepsilon_t (\mu = 0, \sigma = 0.124)$; with $N = 100$

Eq. S16

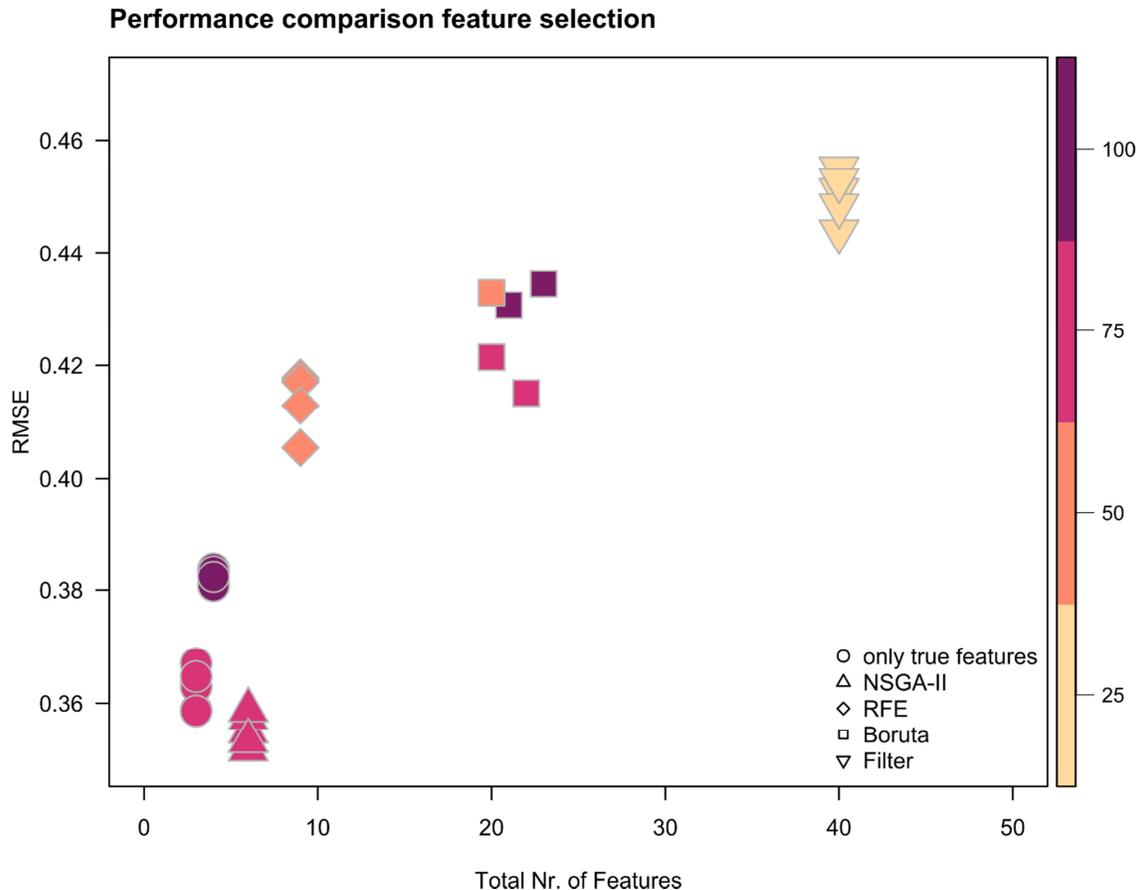


Figure S29: Performance comparison between NSGA-II, Boruta, RFE and a filter method for feature selection in a Random Forest model for different random seeds ($n=5$ for each method), denoting how many features (x -axis) were selected in total in the model (out of 204 possible) and their performance (y -axis). The colour scale denotes the percentage of true features found, and method is denoted by symbols. A Random Forest model with all of the four true features as well as with only three features (most parsimonious model without cosine) is shown for comparison (circles).

Literature cited

- Akinduko AA, Mirkes EM, Gorban AN (2016) SOM: Stochastic initialization versus principal components. *Inf Sci (Ny)* 364–365:213–221
- Balmaseda MA, Mogensen K, Weaver AT (2013) Evaluation of the ECMWF ocean reanalysis system ORAS4. *Q J R Meteorol Soc* 139:1132–1161
- Belkin IM, O'Reilly JE (2009) An algorithm for oceanic front detection in chlorophyll and SST satellite imagery. *J Mar Syst* 78:319–326
- Berrisford P, Dee D, Poli P, Brugge R, Fielding K, Fuentes M, Kållberg P, Kobayashi S, Uppala S, Simmons A (2011) The ERA-interim archive Version 2.0. Shinfield Park, Reading
- Cornes RC, Schrier G van der, Besselaar EJM van den, Jones PD (2018) An Ensemble Version of the E-OBS Temperature and Precipitation Data Sets. *J Geophys Res Atmos* 123:9391–9409
- Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene Selection for Cancer Classification using Support Vector Machines. *Mach Learn* 46:389–422
- Kohonen T (1982) Self-Organized Formation of Topologically Correct Feature Maps. *Biol Cybern*

43:59–69

- Kuhn M (2008) Building Predictive Models in R Using the caret Package. *J Stat Softw* 28:1–26
- Kursa MB, Jankowski A, Rudnicki WR (2010) Boruta – A System for Feature Selection. *Fundam Informaticae* 101:271–285
- Liu Y, Weisberg RH, Mooers CNK (2006) Performance evaluation of the self-organizing map for feature extraction. *J Geophys Res* 111:14pp
- Núñez-Riboni I, Akimova A (2015) Monthly maps of optimally interpolated in situ hydrography in the North Sea from 1948 to 2013. *J Mar Syst* 151:15–34
- Pötzlbauer G (2004) Survey and comparison of quality measures for self-organising maps. In: Proc. 5th Workshop on Data Analysis (WDA 2004).p 67–82
- Pyper BJ, Peterman RM (1998) Comparison of methods to account for autocorrelation in correlation analyses of fish data. *Can J Fish Aquat Sci* 55:2127–2140
- Reusch DB, Alley RB, Hewitson BC (2005) Relative Performance of Self-Organizing Maps and Principal Component Analysis in Pattern Extraction from Synthetic Climatological Data. *Polar Geogr* 29:188–212
- Sang H, Gelfand AE, Lennard C, Hegerl G, Hewitson B (2008) Interpreting self-organizing maps through space-time data models. *Ann Appl Stat* 2:1194–1216
- Schreiber T, Schmitz A (1996) Improved Surrogate Data for Nonlinearity Tests. *Phys Rev Lett* 77:635–638
- Schreiber T, Schmitz A (2000) Surrogate time series. *Phys D Nonlinear Phenom* 142:346–382
- Shen J, Chang SI, Lee ES, Deng Y, Brown SJ (2005) Determination of cluster number in clustering microarray data. *Appl Math Comput* 169:1172–1185
- Wehrens R, Kruisselbrink J (2018) Flexible Self-Organizing Maps in kohonen 3.0. *J Stat Softw* 87