Supplementary Material: Bio-geographic classification of the Caspian Sea

F. Fendereski[1,2],M. Vogt[2], M. R. Payne[2,3], Z. Lachkar[2], N. Gruber[2], A. Salmanmahiny[1] and S. A. Hosseini[1]

(1) Gorgan University of Agricultural Sciences and Natural Resources, Gorgan, Iran
(2) Institute of Biogeochemistry and Pollutant Dynamics-Department of Environmental Sciences, ETH Zürich, Switzerland
(3) Centre for Ocean Life, National Institute of Aquatic Resources (DTU-Aqua), Technical University of Denmark, Charlottenlund, Denmark

S1. Selection of input variables

Variables were sorted into groups of dependent variables and represented in a dendrogram, where the vertical axis represents the degree of similarity (σ) between variables based on their correlation coefficient (Fig. S1). The 'y' axis values (σ) of the dendrogram were calculated as below:

$$\sigma(i, j) = 1 - C(i,j)$$

where i and j are the variables grouped in the lowest branches of the dendrogram, C(i,j) is the correlation coefficient between them and σ (i,j) is the correspondent 'σ' value for each two connecting variables.

For higher hierarchies, σ is obtained as follow:

$$\sigma(i,j,h) = ((1-C(i,h))+(1-C(j,h)))/2$$

Where i and j are the two variables from the first cluster, and h is a new variable that can be added to this set at a higher level.

S2. Swapping dependent variables

To estimate the effect of the choice of representatives from the two groups of dependent variables on the classification output (Fig. 2, section 2.3), several sensitivity tests were run

where we exchanged one or more of the input variables with another representative from the same group (i.e. variable with high degree of correlation; Table S1). The output of each experiment (i.e. different maps of ecoregions) was then compared with the final classification vector. The agreement between the two classification outputs was quantify using the Kappa index of agreement. Kappa is a chance-adjusted measure of the relative percentage of similarity between the two classification results that considers the location and value of each two correspondent pixels (Gregr and Bodtker, 2007). Possible values of Kappa range from -1 to 1, with 1 indicating perfect agreement and -1 representing worse than random agreement (Sim and Wright, 2005).

The Kappa score, ranged from 0.6 to 0.8 (Table S1), indicated a substantial agreement between the results of the experimental and the reference maps (Landis and Koch, 1977). It suggested that little variation occurs in the classification output when the dependent variables were swapped. However, the percentage of agreement differed depending on which variables had been swapped. The variation in the Kappa index was a direct function of degree of correlation between the exchanged dependent variables in most cases.

S3. Number of neurons

In order to define the optimal number of neurons, a series of SOM training runs were performed with different number of neurons. The corresponding quality of each SOM experiment (goodness of the map) was assessed on the basis of average quantization and topological errors (Uriarte and Martin, 2005). Quantization error (QE) is the average distance between each observation vector and its best matching unit (BMU) on neuron map while topological error (TE) measures topology preservation of the SOM (Kohonen, 2000). Increasing the number of neurons decreased the quantization error and increased the topological error (Uriarte and Martin, 2005). Total error was calculated by summing up normalized quantization and topological errors (Table S2). A 20×20 map size provided the lowest number of neurons after which increase in the number of neurons would not lead to a significant decrease in total error anymore (Fig. S2). Hence, we chose a 20×20 neuron map as our standard map.

S4. Number of classes

A 10-fold cross validation approach was conducted to determine the optimum of number of classes (De' ath and Fabricius, 2000). The data were divided into two unequal parts of 90% (the training set) and 10% (the validation set). The training data set was reduced to 400 classes (20×20 neurons) using SOM. Theses 400 prototypes were further agglomerated using the HAC algorithm. The clustering of the validation data set was performed using the following procedure:

1) Each observation from the validation set was compared to the 400 neurons.

2) The closest neuron on the map (using the Euclidean distance), also called best matching unit (BMU), was identified.

3) The class of the BMU was attributed to the observation.

For each cross validation experiment (k = 1,…,10), the optimal classification was the one that minimizes the average distance of the validation observations to the center (average) of their respective classes (E_k):

$$E_{k(k=1:10)} = \frac{1}{6} \sum_{j=1}^{6} \left(\frac{1}{nv} \sum_{n=1}^{nv} \sqrt{\left(v_{ijk} - t_{ijk} \right)^2} \right) \quad (1)$$

where v_{ijk} is the validation observation associated with variable j and cross validation experiment k, t_{ijk} the average of the corresponding class in training data set assuming a number of class i (i=2:15) and where nv is the number of points in the validation set.

Data were normalized in advance to solve the problem of the inconsistency of the variables units. The final cross validation error was computed based on the errors for all the 10 given cross validation folds (Table S3):

$$E = \frac{1}{10} \sum_{k=1}^{10} E_k \quad (2)$$

Fig. S4 shows the amount of final cross validation error plotted against the number of classes. The error decreased monotonically with increasing the number of classes. Hence, the lowest number of classes after which the increase in the number of classes no longer led

to a substantial reduction in the error was considered as the optimal number of class in this study. This gained saturated at a number of 11 classes, and we chose the point where less than 5% of the first decrease was gained by the addition of further classes as our cut-off.

S5. Species composition data

The Dice coefficient (*DC*) is defined as two times the volume of overlap between two sets belonging to different groups (A and B) divided by the sum of volumes of the two groups, given by

$$DC = \frac{2|A \cap B|}{|A|+|B|}$$

S6. (Dis) Similarity between ecoregions

In order to establish the degree of dissimilarity of the resulting clusters, we employed the annual mean climatologies of the physical input variables and visualized the relationships between the six-dimensional observational points within/between ecoregions using the Non-metric Multidimensional Scaling (NMDS) method (Clarke, 1993). NMDS is a data reduction technique that projects n-dimensional data onto a space of lower dimensionality based on a distance matrix between data points (Quinn and Keough, 2002). We performed NMDS on a similarity matrix obtained using pair-wise Euclidean distances between 11760 observations of the six standardized input variables (SST, DSSS, Depth, TSM, DTSM and ICE; Reich et al., 1999; Quinn and Keough, 2002).

The Non-metric Multidimensional Scaling of the environmental conditions in the different ecoregions reveals that the 10 ecoregions tend to group into two clusters: one cluster groups the ecoregions of the NCB (Fig. S6; NCB-UF, NCB-WS, NCB-ES and NCB-RO, red circle) while the other groups those of the MCB and SCB (Fig. S6; MCB-OS, SCB-OS, MDB-C and SCB-C, green circle). The two sub-clusters are separated by NCB-T and MCB-T, which act as a "transition zone". Points in MCB-T (red circles) are more similar to the group from the MCB and SCB and those of NCB-T (pink circles) are more similar to the group from the NCB. NCB-ES was most different from the other ecoregions in terms of environmental conditions (Fig.

S6; light orange circles, upper right). The larger spread between points in ecoregions in the NCB suggested a smaller degree of bio-geophysical homogeneity within the NCB. In the SCB and MCB, SCB-OS (dark blue circles) and MCB-OS (light blue circles) were more similar to one another than the other ecoregions in this area (Fig. S6; lower points on the left). The degree of similarity was reflected in the hierarchical sequence in which ecoregions were formed (Fig. S7).

References

1. Clarke, K. R.: Non-parametric multivariate analyses of changes in community structure. AUST J ECOL., 18, 117–143, doi: 10.1111/j.1442-9993.1993.tb00438.x, 1993.
2. De'ath, G. and Fabricius, K. E.: Classification and regression trees: A powerful yet simple technique for ecological data analysis. ECOLOGY, 81, 3178–3192, doi: 10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2, 2000.
3. Gregr, E. J. and Bodtker, K. M.: Adaptive classification of marine ecosystems: Identifying biologically meaningful regions in the marine environment. DEEP-SEA RES PT I, 54, 385–402, doi:10.1016/j.dsr.2006.11.004, 2007.
4. Kohonen, T.: Self-Organizing Maps, Springer, 3rd edn., 2000.
5. Landis, J. R. and Koch, G. G.: The measurement of observer agreement for categorical data. BIOMETRICS, 33, 159-174, 1997.
6. Quinn, G. P. and Keough, M. J. (Eds.): Experimental Design and Data Analysis for Biologists, Cambridge University Press, New York, 2002.
7. Reich, P. B., Ellsworth, D. S., Walters, M. B., Vose, J. M., Gresham, C., Volin, J. C., and Bowman, W. D.: Generality of leaf trait relationships: a test across six biomes. ECOLOGY, 80, 1955–1969, 1999.
8. Sim, J. and Wright, C. C.: The kappa statistic in reliability studies: use, interpretation, and sample size requirements. PHYS THER., 85, 257-268, 2005.
9. Uriarte, E. A. and Martin, F. D.: Topology preservation in SOM. International Journal of Mathematical and Computer Sciences, 1, 19-22, 2005

Supplementary Tables

Table S1. Kappa Index of Agreement (0-1) for comparisons of the classification outputs after swapping dependent input variables and the final classification output used in this study as the reference.

Input variables	Kappa Index of Agreement
SST, ICE, DSSS, Depth, TSM, DTSM	1.0000
SST, SSS, DSSS, Depth, TSM, DTSM	0.6535
SST, PAR, DSSS, Depth, TSM, DTSM	0.6018
SST, DSST, DSSS, Depth, TSM, DTSM	0.6350
Windsp3, ICE, DSSS, Depth, TSM, DTSM	0.8010
DPAR, ICE, DSSS, Depth, TSM, DTSM	0.6332
Windsp3, SSS, DSSS, Depth, TSM, DTSM	0.7386

Table S2. Quantization and topological errors with different number of neurons

Number of neurons	Quantization error	Topological error	Total error
5×5	0.94	0.06	1
10×10	0.55	0.15	1.5
15×15	0.41	0.14	1.28
20×20	0.3	0.11	0.73
25×25	0.25	0.11	0.67
30×30	0.21	0.11	0.68
35×35	0.19	0.11	0.6
40×40	0.16	0.1	0.49

Table S3. Final cross validation error for different number of classes ($\times 10^{-3}$)

number of classes	2	3	4	5	6	7	8	9	10	11	12	13	14	15
error	16	14	12	11	9	8	7	7	6	6	5.9	6	5.5	5.4

Table S4. Mean ± std for annual mean of each physical input variable in each ecoregion. Higher std is seen for Depth in the MCB and SCB while for TSM and its seasonal amplitude (DTSM) higher std is seen in NCB and shallow continental shelf of SCB (SCB-C).

Ecoregion	SST (C°)	DSSS (ppt)	Depth (m)	TSM (g/m^3)	DTSM (g/m^3)	ICE (%)
NCB-RO	13.16±1.92	-2.21±0.78	13.63±5.46	12.49±3.51	-8.4±5.43	21.43±5.32
NCB-WS	12.63±0.67	-1.8±0.65	9.25±4.9	12.89±5.74	0.2±4.06	20.99±5.13
NCB-UF	12.1±0.28	-0.52±0.23	4.83±1.38	3.85±2.5	0.37±1.78	22.6±3
NCB-ES	11.68±0.69	-0.72±0.21	9.16±6.81	29.61±11.35	10.17±10.97	28.33±1.84
NCB-T	13.58±0.5	-1.61±0.38	8.81±4.57	3.27±2.56	-1.4±1.32	7.82±5.42
MCB-T	14.34±0.92	-0.69±0.29	24.67±15.72	2.19±2.53	-1.47±2.02	2.17±4.08
MCB-C	15.56±1.05	0.06±0.24	60.93±36.72	1.03±0.6	-1.01±0.49	0
MCB-OS	14.84±0.32	0.1±0.18	388.11±191.25	0.77±0.15	-0.82±0.16	0
SCB-C	18.68±1.04	0.12±0.12	42.86±52.35	1.66±1.83	-0.08±1.05	0
SCB-OS	18.09±0.82	0.1±0.12	542.13±207.8	0.82±0.11	-0.7±0.19	0

Absolute values of Depth (m) have been shown rather than its logarithm

Table S5. Monthly mean climatologies and descriptive statistics on annual mean climatologies of Chl-a concentration (2003-2010) (mg/m^3) in ecoregions. Due to a non-normal distribution of Chl-a in marine environments median of monthly mean climatologies in each ecoregion is shown instead of mean (Nezlin, 2005).

	Ecoregion									
	NCB-RO	**NCB-WS**	**NCB-UF**	**NCB-ES**	**NCB-T**	**MCB-T**	**MCB-C**	**MCB-OS**	**SCB-C**	**SCB-OS**
month										
1	4.9	4.18	1.33	2.05	1.78	1.1	1.15	0.94	1.29	1.23
2	5.46	6.1	1.56	2.23	2.1	1.14	1.15	0.91	0.98	0.96
3	5.3	5.3	1.1	1.38	1.6	1.03	1.1	1.19	0.77	0.98
4	4.5	4.93	0.83	1.12	1.34	0.74	0.8	0.98	0.71	0.96
5	5.75	6.67	1.14	1.47	1.94	0.65	0.69	0.77	0.55	0.96
6	5.84	5.92	1.72	1.91	2.81	1.03	0.61	0.76	0.48	0.89
7	7.7	6.74	2.33	2.17	3.61	1.24	0.83	0.92	1.45	1.31
8	7.1	6.16	2.18	2.43	3.31	1.33	1.17	1.1	1.82	1.66
9	6.5	5.63	2.06	2.41	3.41	1.59	1.36	1.5	1.38	1.57
10	6.38	6.13	1.88	1.88	2.66	2.01	1.77	1.78	1.56	1.82
11	6.96	6.68	2.16	1.98	2.19	1.84	1.68	1.52	1.71	1.77
12	4.19	4.42	1.9	2.09	1.78	1.59	1.77	1.45	1.61	1.45
annual										
median	5.88	5.29	1.68	2.41	2.48	1.2	1.11	1.14	1.18	1.32
mean	5.79	5.14	1.77	2.72	3.04	2.01	1.28	1.16	1.27	1.34
std	0.98	1.66	0.3	1.02	1.66	1.46	0.32	0.11	0.34	0.14

Lower std in open ocean and higher std in NCB except for the Ural Furrow in NCB-UF.

Table S6. Presence information on the 27 marine species in ecoregions. '1' represents presence '-' represents lack of observation of species in the given ecoregion. Points located near/on the boundaries of ecoregions were also assigned '-' (Source: CEP, 2002).

Species	Ecological taxonomic group	Ecoregion									
		NCB -RO	NCB -WS	NCB -UF	NCB -ES	NCB -T	MCB -T	MCB -C	MCB -OS	SCB-C	SCB -OS
Rhizosolenia fragilissima	Phytoplankton	-	-	-	-	1	1	1	1	1	1
Eurytemora grimii	Zooplankton	-	-	-	-	1	1	1	1	1	1
Mnemiopsis leidyi		-	-	-	-	1	1	1	1	1	1
Stenodus leusichtys	Pelagic fish	-	1	1	1	1	1	1	-	-	1
Liza aurata		-	1	1	1	1	1	1	1	1	1
Liza saliens		-	1	1	1	1	1	1	1	1	1
Salmo trutta caspius		-	-	-	-	1	1	1	1	-	1
Alosa kessleri kessleri		-	1	-	-	1	1	1	1	1	1
Alosa saposchnikowii		-	1	1	-	1	1	1	1	1	1
Atherina boyeri caspia		-	1	1	1	1	1	1	1	1	1
Clupeonella cultriventris caspia		1	1	1	-	1	1	1	-	1	-
Clupeonella engrauliformis		-	-	-	-	1	1	1	1	-	1
Rutilus rutilus		1	1	1	-	1	1	1	-	1	1
Rutilus frisii kutum		-	1	-	-	1	1	1	-	1	1
Cyprinus carpio		1	1	1	1	1	1	1	-	1	-
Abramis brama		-	1	1	-	1	1	1	-	-	-
Acipenser gueldenstaedtii	Demersal fish	-	1	1	-	1	1	1	1	1	1
Acipenser persicus		-	1	1	-	1	-	1	1	1	1
Huso huso		-	1	1	1	1	1	1	1	1	1
Neogobius melanostomus		-	1	1	-	1	1	1	1	1	1
Benthophilus stellatus		1	1	1	1	1	1	1	1	1	1
Acipenser stellatus		-	1	1	1	1	1	1	1	1	1
Acipenser nudiventris		-	-	1	1	-	1	1	1	1	1
Caspiastacus pachypus	Benthic invertebrata	-	-	-	-	-	1	1	-	1	-
Pontastacus eichwaldi		1	1	1	1	-	1	1	-	1	-
Abra (Syndesmya) ovata		-	1	1	-	1	1	1	-	1	-
Hypanis angusticostata		1	1	1	1	1	1	1	-	-	-

Supplementary Figure Captions

Fig. S1 Spearman correlation matrix between input variables

Fig. S2 Sum of normalized quantization and topological errors as a function of number of neurons

Fig. S3 The SOM component plane indicates the distribution of each variable across the map and those neuron "bee hive" plots showing the variability between neurons in the individual input variables.

Fig. S4 Cross validation error against the number of classes

Fig. S5 HAC dendrogram showing the hierarchy of the ecoregions. HAC successively agglomerates pairs of classes based on their similarity. The bottom-up clustering procedure. starts with each neuron being considered as a single class. The iteration ends when all the neurons have been merged into a single class (Frades and Matthiessen, 2010). Dashed red line shows the levels of the hierarchy for classifications with 11 number of ecoregions.

Fig. S6 Nonmetric multidimensional scaling (NMDS) ordination plot of the individual 0.1 degree pixels in the study area. Different colors of the points represent the correspondent ecoregion for that point. The distance between points reflects their underlying similarity/dissimilarity, i.e. their distance in 6-dimensional environmental variable space. Ecoregions in the NCB (red circle on the right) are distant from ecoregions in the MCB and SCB (green circle on the left).

Fig. S7 hierarchical sequence (from upper left, levels 1 to 9) of ecoregions formation (2 to 10 ecoregions).

Supplementary Figures

SST										
0.83	SSS									
-0.35	-0.54	TSM								
0.92	0.93	-0.43	PAR							
0.53	0.46	-0.65	0.5	Depth						
-0.76	-0.78	0.74	-0.77	-0.68	ICE					
-0.91	-0.73	0.32	-0.85	-0.5	0.68	WNDSP3				
-0.78	-0.86	0.63	-0.85	-0.62	0.77	0.69	DSST			
0.64	0.57	-0.34	0.61	0.5	-0.71	-0.53	-0.58	DSSS		
0.25	0.27	0.07	0.26	-0.06	-0.03	-0.27	0	-0.07	DTSM	
-0.85	-0.68	0.16	-0.81	-0.32	0.6	0.9	0.59	-0.43	-0.27	DPAR
0.9	0.7	-0.2	0.84	0.42	-0.66	-0.92	-0.56	0.56	0.29	-0.9

Fig. S1

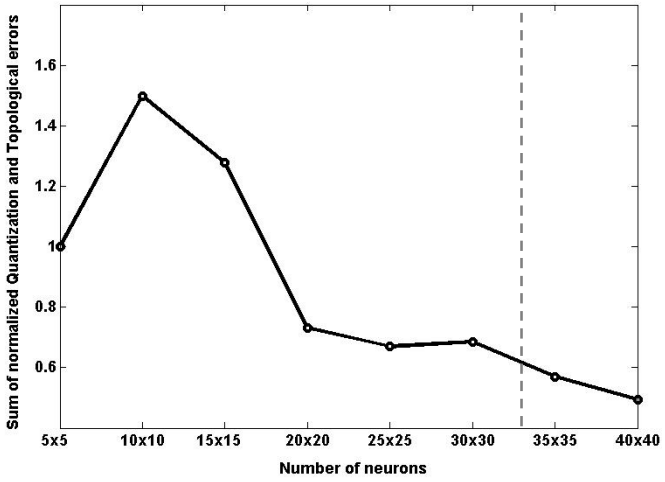
Sum of normalized Quantization and Topological errors
1.6
1.4
1.2
1
0.8
0.6
5x5
10x10
15x15
20x20
25x25
30x30
35x35
40x40
Number of neurons

Fig. S2

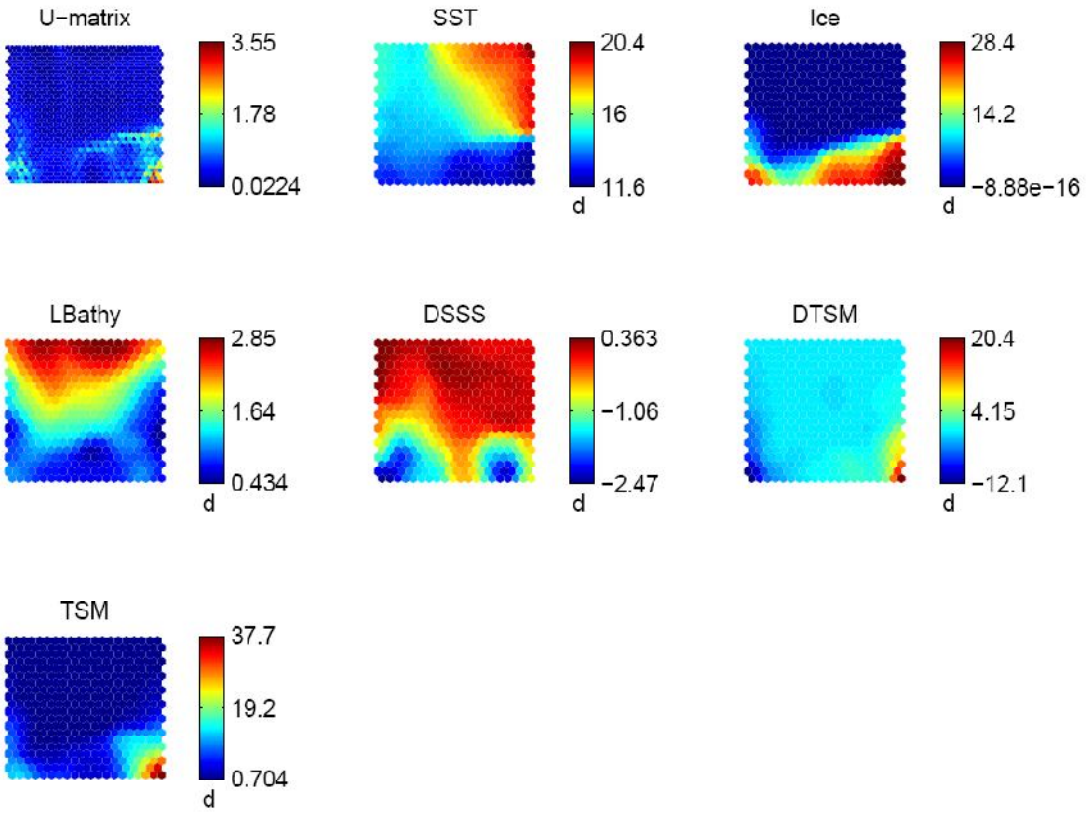
U−matrix
3.55
1.78
0.0224
SST
20.4
16
11.6
d
Ice
28.4
14.2
−8.88e−16
d
LBathy
2.85
1.64
0.434
d
DSSS
0.363
−1.06
−2.47
d
DTSM
20.4
4.15
−12.1
d
TSM
37.7
19.2
0.704
d

Fig. S3

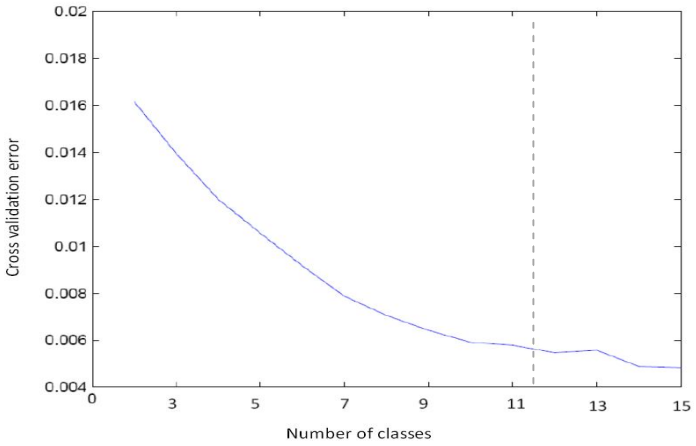
0.02
0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
Cross validation error
0
3
5
7
9
11
13
15
Number of classes

Fig. S4

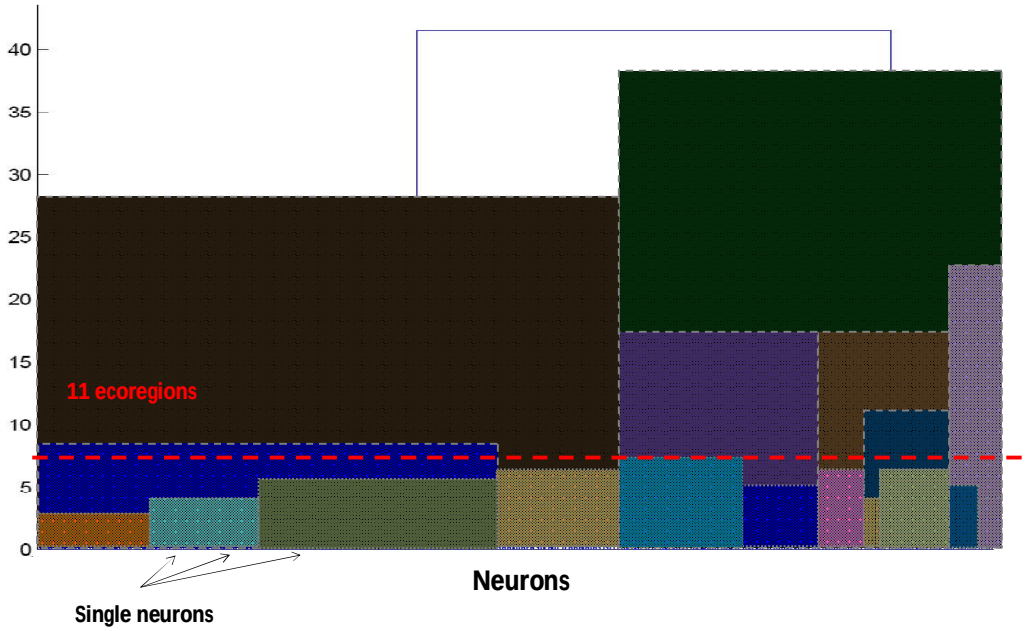

Fig. S5

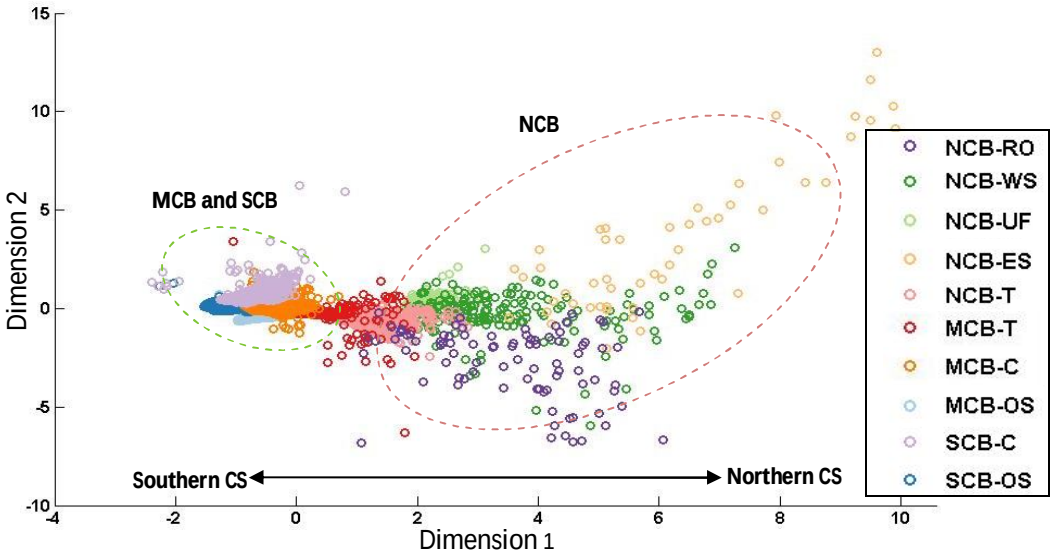
NCB
MCB and SCB
Southern CS
Northern CS
Dimension 1
Dimension 2
NCB-RO
NCB-WS
NCB-UF
NCB-ES
NCB-T
MCB-T
MCB-C
MCB-OS
SCB-C
SCB-OS

Fig. S6

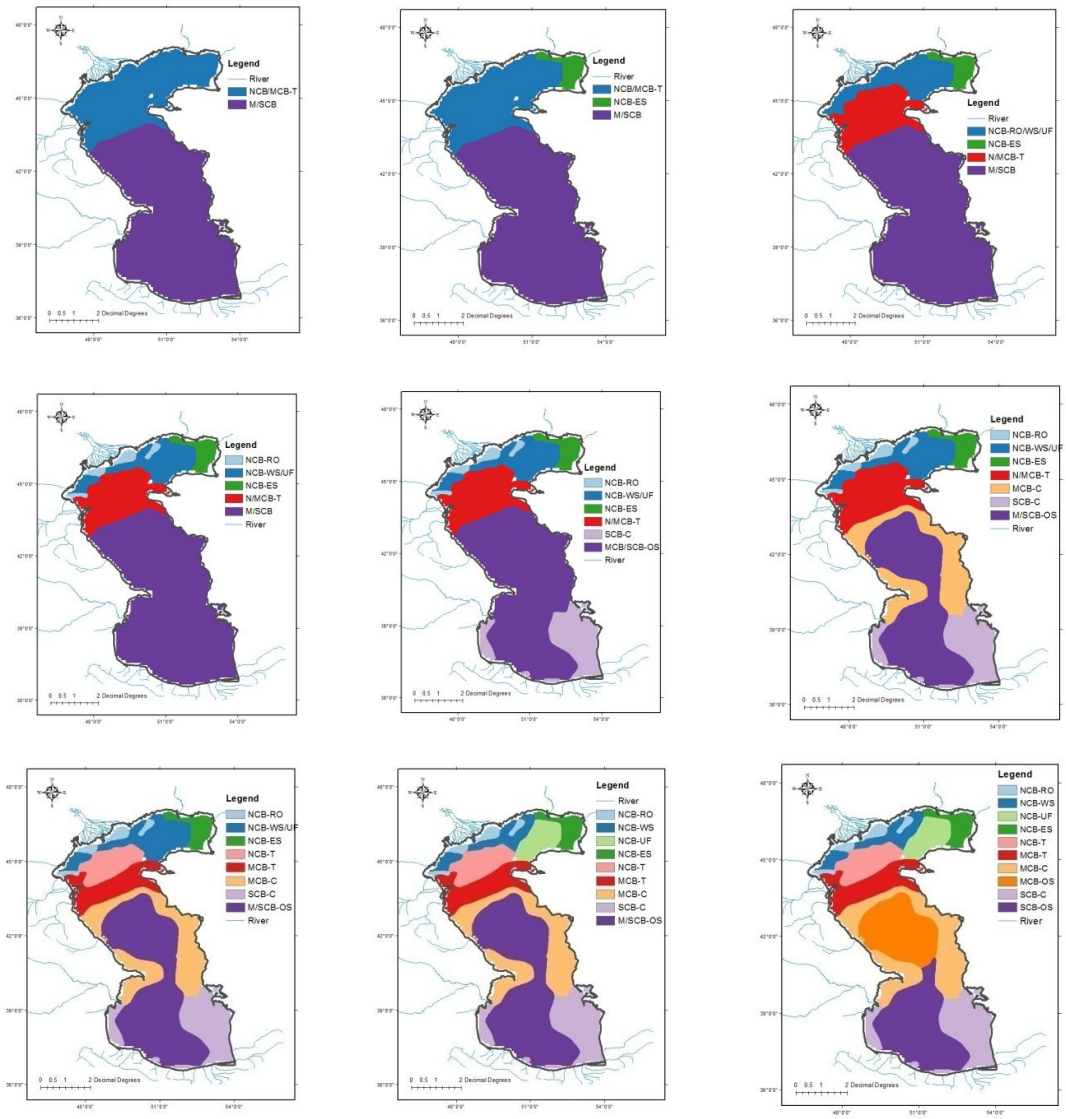
Legend
River
NCB/MCB-T
M/SCB
Legend
River
NCB/MCB-T
NCB-ES
M/SCB
Legend
River
NCB-RO/WS/UF
NCB-ES
N/MCB-T
M/SCB
Legend
NCB-RO
NCB-WS/UF
NCB-ES
N/MCB-T
M/SCB
River
Legend
NCB-RO
NCB-WS/UF
NCB-ES
N/MCB-T
SCB-C
MCB/SCB-OS
River
Legend
NCB-RO
NCB-WS/UF
NCB-ES
N/MCB-T
MCB-C
SCB-C
M/SCB-OS
River
Legend
NCB-RO
NCB-WS/UF
NCB-ES
NCB-T
MCB-T
MCB-C
SCB-C
M/SCB-OS
River
Legend
River
NCB-RO
NCB-WS
NCB-UF
NCB-ES
NCB-T
MCB-T
MCB-C
SCB-C
M/SCB-OS
Legend
NCB-RO
NCB-WS
NCB-UF
NCB-ES
NCB-T
MCB-T
MCB-C
MCB-OS
SCB-C
SCB-OS
River
0 0.5 1 2 Decimal Degrees

Fig. S7