Biogeosciences Discuss., 11, C1539–C1545, 2014
www.biogeosciences-discuss.net/11/C1539/2014/

Biogeosciences
Discussions
Open Access

Interactive comment on "Bio-geographic classification of the Caspian Sea" *by* F. Fendereski et al.

Anonymous Referee #1

Received and published: 9 May 2014

REVIEW: Biogeographic classification of the Caspian Sea F. Fendereski et al. Biogeosciences Discussion

Summary: Fendereski et al. conduct a classification of the Caspian Sea using a combination of a self-organizing map and hierarchical agglomerative clustering. The authors make a compelling case for the necessity of ecoregion classification in the Caspian Sea, an area where little comprehensive oceanographic classification has been conducted and that has recently undergone drastic change as a result of introduction of an invasive Mnemiopsis leidyi. The selection of input physical variables as the basis for classification was done carefully creative reduction of potential autocorrelation between variables is important and represents a step forward in ecoregional classification.

The rationale and general classification methods are sound. However there were sev-

eral methodological and organizational issues that are of concern. The authors use a hierarchical agglomerative clustering but provide no apriori rationale or discussion of spatial hierarchy. I also have serious concerns with how the in situ data were analyzed, and it is unclear whether data treatment resulted in a robust assessment of group identity. As these data were used as a means of independent verification, it calls into question the ecological relevance of the physical classification. The authors should try to more explicitly tie the spatial patterns revealed by a classification based on multivariate physical factors alone (and subsequent bottom-up effects) and the top-down effects alluded to by the patterns of M. leidyi. In other words, explicitly discuss how this classification effort has informed the ecosystem based management needs stated in the introduction. Finally, there are several organizational and clerical errors that need to be addressed. The scientific quality would benefit in many cases by simple elaboration and explicit justification. Thus I would encourage a comprehensive revision. Further detailed suggestions are below.

Introduction: The introduction makes a compelling case for classification in this region and is clearly written. However, there is much focus on the role of the invasive jelly in affecting lower trophic level distributions. Specific predictions of spatial patterns may be merited in the introduction given the top-down control and spatial distribution of Mnemopsis that is alluded to throughout the paper. Inclusion of a distribution map of the species may also be warranted.

Methods: A few more details of the SOM are warranted, e.g. size of the Gaussian neighborhood function and what the final node approximates (e.g. the mean, the maximum likelihood?). The discussion of and subsequent choice of neural map size was appropriate however the choice of subsequent clustering algorithm reflects an underlying assumption about the way the system is organized. Is the Caspian Sea organized in a hierarchical fashion? Or was this chosen with consideration of the scale and flow of management decisions? Perhaps circulation in the different regions coupled with bathymetry dictate a hierarchical framework. If optimizing differences between neu-

rons is the goal, then why not a K-means? If hierarchical organization is presumed, then I would expect a subsequent match up with the larger scale classifications (e.g. North Central and South). In any case, the justification needs to be made clear.

It is unclear in the methods how the input data were standardized prior to classification and whether non-normally distributed data were transformed prior to standardization prior to classification. This is important to understand how total multivariate information was partitioned into the initial neural map.

Agglomerative strategies can be divided into two groups with different objectives: 1) those that optimize some property of a group of entities and 2) those that optimize the route by which the groups are attained. You seem to use both strategies: the SOM and subsequent HAC on the input variables were of the former but the choice of physical variables appeared to be of the latter. However the choice of linkage method and distance metric is critical to how well a particular HAC meets an objective. Please be explicit and state both the linkage method and the distance method for every clustering that is conducted.

Section 2.1.1. Initial resolution of the input variables. When resampled at higher resolution, please acknowledge and or account for the pseudoreplication (spatial autocorrelation) that occurs, thus magnifying any differences between ecoregions. Also, did the different climatological extents affect the patterns? For example, the surface isohaline maps (from pre 1995), the ice coverage (2004-2012) and the ocean color and surface scattering may reflect different mean states if includes anomalous years. State the span and spatial scale over which the climatology was calculated for each variable and then discuss implications if different.

I assume that the absolute value of the correlation coefficients became the linkage function in the clustering algorithm (please explicitly state). The goal of orthogonality between input variables is of great merit. But I wonder why the underlying distributions were not considered. The rank correlation is somewhat robust to this but may be

overly so. The Kappa spatial overlap analysis (Supplementary Section) is important and should be placed in the regular document. Also include what level of hierarchy at which this was assessed (e.g. number of classes).

The cross-validation section needs a little more detail. Based on the logic outlined above Equation S4 (1), I would suspect that error would be calculated at each class (2:15) by k (fold) interaction. This would also allow assessment of the error of classification (mean distance of validation set to classified centroids) for a given number of ecoregions to be assessed for each fold (iteration). While there certainly is some subjectivity involved (e.g. choosing 11 because it represents a drop in error from 6 to 5.9), the spread or variance across iterations would allow objective determination of whether a drop in error is significant. Finally, while this is minor, a 40/400 validation set does not seem a particularly robust test nor is it explicitly stated how the sets were partitioned and whether they were repartitioned between each fold.

The phenology metrics (and the temporal resolution, monthly?) of the chl-a data need to be included in the methods.

In situ clustering: Please include where in situ data were collected on a map. It is unclear whether all in situ samples within an ecoregion were lumped into a single presence column or whether clustering was conducted using multiple samples within an ecoregion (more robust). Please provide justification as to why only 1 and NaNs were included in the classification. The choice of this reflects whether your analysis is robust to errors in omission/commission or both.

Organization: Several sections of the results (see detailed comments) should be relegated to the discussion or introduction. There are also omissions (e.g. the phenology methods are introduced in the results) and redundancies. The results section can be better organized to highlight 1) the robustness/sensitivity of the classification; 2) the spatial distribution of physical variables and 3) the independent verification of chl-a patterns, chl-a phenology and in situ community structure. If hierarchical organization

was the goal as the main ecoregion names imply, then please discuss each of the three above in terms of that organization. Otherwise, explicitly state that a spatial hierarchy is not intended (but again, the choice of HAC belies this) and discuss them North to South.

Discussion: It would be helpful to include a brief discussion about how this classification compares to what is known about the Caspian Sea. Subjective, but expertly informed classifications are dominated by bathymetry and circulation. One might expect an objective hierarchy of surface features to reveal this. Also, please return to the introduction and results regarding differences in ecosystem structure. The inclusion of the multivariate pelagic (and benthic) species distributions is a great asset and the primary validation of ecoregions. Otherwise they are simply (bio)physical regions. Such a discussion may also reveal two types of information: 1) where classified climatological ecoregions "work" and 2) where they don't (e.g. when there is top-down control by an invasive species). Both are critical for the effective management of a region and may help dictate where greater effort is placed for higher frequency and density in situ measurements.

Detailed suggestions:

Chl-a climatology. Why not log10-transform? The Kruskal Wallis test that you employ is good for large outliers, but chl-a is generally treated as a log normally distributed (see Campbell et al. 1995). One can log-transform then conduct a standard parametric ANOVA with multiple comparisons (e.g. the Tukey HSD).

Pg 4412: Lines 25-30: "These authors. . .", "Their results. . .". Please rephrase. It is unclear to which study you are attributing differences in chl-a.

P 4413 Line 5. Suggest that you simply start with " We applied. . ."

P4420 Section 3.1. The spatial coherence is an obvious result, particularly at the scale of the study. I suggest you reframe this in terms of the natural gradients of the

physical variables. Additionally you should discuss (tacitly in results, more completely in discussion) how well these recently classified boundaries compare to any existing boundaries, hydrological or political.

P4421 Section 3.2 Line 8: Reference is redundant to methods, remove. Figure 5 contains redundant information to Figure 6 and is incredibly difficult to read. I would suggest that you simply report the results of your Kruskal Wallis' H to Dunn procedure on Figure 6. Simply annotate with letters above the box plot (with significant differences having unique letters). This might allow inclusion of the principal components analysis shown in the supplementary material (which is a better visualization of overall differences in physical state between ecoregions). Please do something similar with Figure 8. Just report a box-plot with the multiple comparisons results.

P4423 Lines 22-26. "The highest chl-a. . .. " These sentences are unclear and possibly redundant to one another. Please rephrase. For example: "The highest concentrations of chl-a are found in the NCB, specifically the two ecoregions at the mouth of the Volga (NCB-XX and NCB-XX). This river supplies 80

P4424 Section 3.3.2. Lines 18-26. This belongs in the discussion section. Also without maps or more in depth discussion of M. leidyi distribution, this seems quite speculative.

P4424 Section 3.3.2. Lines 27-28. "Significant differences in the date. . .."You did not statistically quantify differences in phenology, please remove the term significant and rephrase in terms of the qualitative differences observed.

Figure 9. In panel D, there appears an error in that the median seasonal cycle is always less than or equal to the annual median. This seems unlikely unless there were strong outliers across multiple months. In panel F, it appears that the bloom onset vertical line is offset from the timing that the median is above the annual median. Also, it is unclear whether phenology was assessed at the climatological level or within each year and whether the grey filled in values refer to temporal variability between years or spatial variability within an ecoregion. Please clarify in appropriate methods and

results sections.

P4425, Section 3.3.3 Polovina et al., 2001? This paper refers to the TZCF in the North Pacific and the convergence of planktonic species (including jellies) and the selective foraging by larger animals. This is not an appropriate reference if one is trying to say that transition zones contains both northern and southern species. Be certain to generalize in the text to transition zones.

P 4428. Line 21-24. " This is in spite of the fact. . ..". Please rephrase. The location of high abundances of M. leidyi may be under physical (circulation) or bottom-up control..

P 4429 Line 1-3. "Due to lack of comprehensive species data. . ." This should be stated upfront in the Methods Section.

Figure 1. Please include locations of in situ benthic and pelagic data

Table 1. Please make certain acronyms are consistent throughout text and supplementary material (e.g. MCB-TR and MCB-T). Figure S.5 is missing an axis label (distance?). Figure S3. Missing units. Also, please define behive. Figure S1is missing the final column and may be more aptly described in Table form. Table S3. It is unclear whether this is mean error or error. Also please state the distance measure (Euclidean). Likewise, please state with Figure S4.

Interactive comment on Biogeosciences Discuss., 11, 4409, 2014.