Methodology to estimate continuous spatial environmental layers from point data for the Caspian Sea

Background

As the Caspian Sea is not connected to the world ocean data is lacking from global oceanic datasets as Bio-ORACLE (Tyberghein et al., 2012) and MARSPEC (Sbrocco and Barber, 2013). This document describes the methodology used to generate comparable environmental data for the Caspian Sea, in line with the Bio-ORACLE dataset. We utilized spatial interpolation techniques on point data measurements, to create raster layers of environmental variables that can subsequently be used for species distribution modeling (Raes and Aguirre– Gutiérrez, 2018). Finally, we resampled our dataset to create an effective extension of Bio-ORACLE for the Caspian Sea.

The most important variables in the marine environment that determine the distribution of Pontocaspian taxa include seawater temperature, salinity, and oxygen content. For these variables, we interpolated the minimum, mean and maximum temperatures at the surface and the bottom of the Caspian Sea, totaling 18 environmental variables (Fig. 1).

Pipeline

Our data processing pipeline consisted of the following steps:

1. Point data measurements were downloaded from the World Ocean Dataset (NOAA, 2018) – for the time period from 1914 to 2011 for the Caspian Sea extent including 27315 data points.
2. Checking and removing errors in the dataset (data out of the Caspian Sea boundaries, missing decimal punctuation etc.)
3. Raster generation of metrics over the indicated time period (e.g. raster mean, minimum and maximum values) by Universal Kriging with Automap package in R (Hiemstra, 2015). We used the GEBCO bathymetric raster dataset at 1 arc-minute spatial resolution as initial template (GEBCO, 2014).
4. Clipping raster layers to Caspian Sea bounding box.
5. Filling empty cell values derived by mismatching with the Caspian Sea bounding box with the mean values in a 3*3 focal window, package Raster in R (Hjimans, 2018)
6. Conversion of raster to GeoTiff format.
7. Calculation of evaluation measures to estimate the interpolation performances
8. Finally, the environmental layers were resampled at 5 arcmin to match the Bio-Oracle dataset

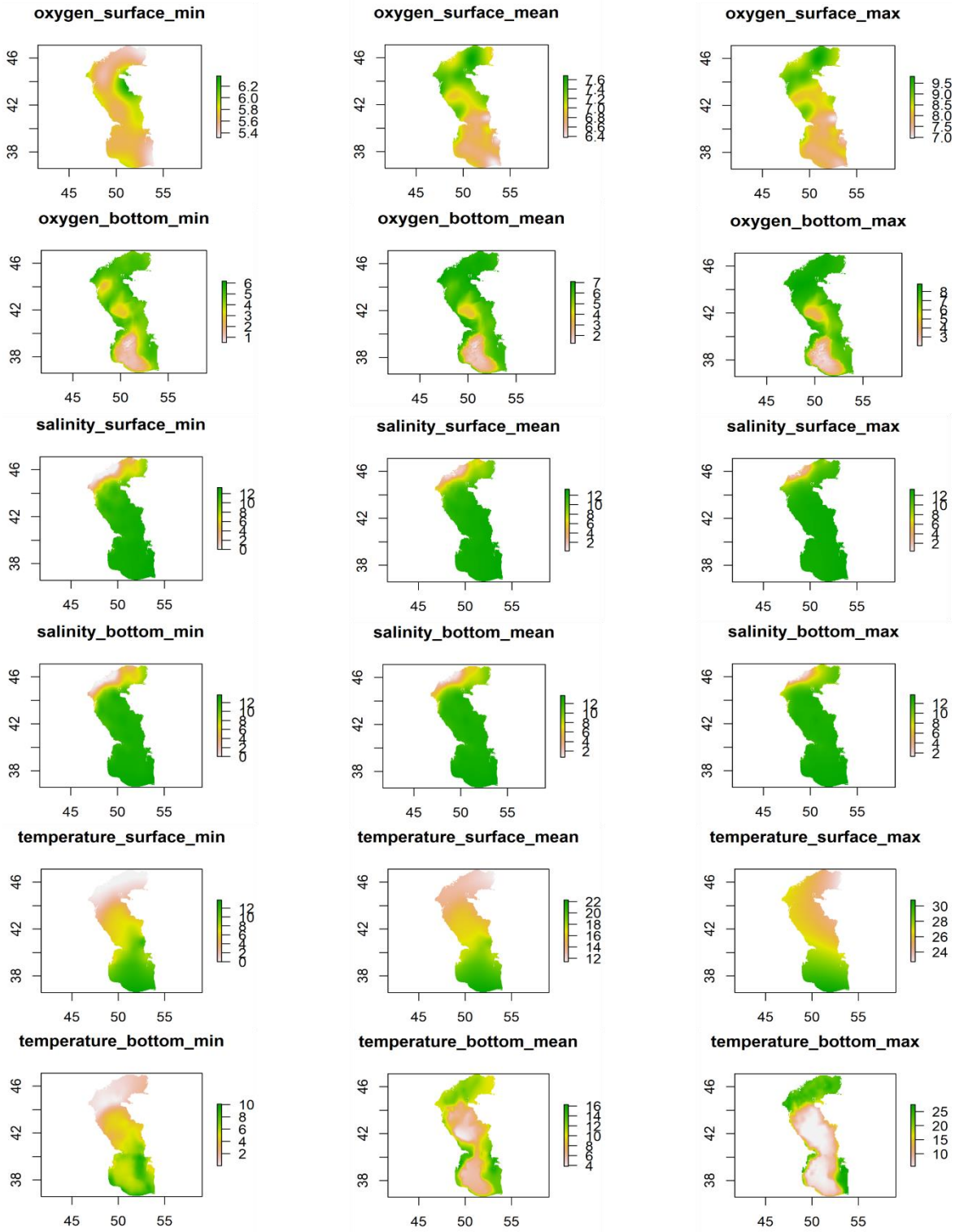

Figure 1: Overview of the environmental layers

Results

Salinity (PSS) – max, min, mean at surface and bottom

Maximum and minimum salinity are interpolated for the months December – March, and May – July, respectively. These two– time ranges correspond to the lowest and highest freshwater quarterly discharge from rivers into the Caspian Sea. December was added to extend the dataset to improve the distribution of point data because in winter fewer measurements are taken due to ice coverage. Mean salinity is the median of the maximum and minimum layers. The average could not be

calculated as data for coldest months are few for the Northern part pf the Caspian Sea as it is normally frozen and sampling effort was inconsistent.

The interpolation of surface values was computed with a universal cokriging algorithm with a variable representing the river influence in a specific grid cell. This value is given by the maximum (or minimum) river discharge divided by the log of distance from the river mouth. The bottom values were interpolated with bathymetry as an explanatory variable in universal cokriging.

Temperature (°C) – max, min, mean at surface and bottom

Maximum and minimum temperature are interpolated for the months December – March, and July – September, respectively. These two– time ranges correspond to the lowest and highest quarterly mean temperature. December is added to extend the dataset to improve the distribution of point data because in winter are taken fewer measurements due to ice coverage. Mean temperature is the median of the maximum and minimum layers. The average could not be calculated as data for coldest months are few for the Northern part pf the Caspian Sea as it is normally frozen and sampling effort was inconsistent.

To predict the surface values, we used simple universal kriging. For the bottom values, bathymetry was added as an explanatory variable.

Dissolved Oxygen (µmol/kg) – max, min, mean at surface and bottom

For the minimum dissolved Oxygen estimation, we selected the warmest four months of the year, June to September, as the mean dissolved Oxygen was the lowest, due to higher water temperature. On the opposite, for the maximum Oxygen, we selected the data in the coldest months, December – March. Mean dissolved Oxygen is the median of the maximum and minimum layers. The average could not be calculated as data for coldest months are few for the Northern part pf the Caspian Sea as it is normally frozen and sampling effort was inconsistent.

We predicted unknown dissolved Oxygen interpolating the surface measurements with simple universal kriging and the bottom measurements adding bathymetry as an explanatory variable in the universal cokriging algorithm.

1. Layer evaluation

We calculated evaluation measures for the minimum and maximum layers (see Tab. 2). RMSE was calculated on the original interpolation, whereas the other metrics were obtained with a cross–validation with k– fold = 10 (Tab. 2). In Fig. 2 it is shown the prediction variance of the interpolation.

Tab. 1: evaluation metrics. RMSE: Root Mean Square Error, ME: Mean Error, MSPE: Mean Squared Prediction Error, MSNE: Mean Square Normalized Error, Corr_PO: Correlation predicted values ~ observed, Corr_PR: Correlation predicted values ~ residuals

Layer	RMSE	Mean_error	MSPE	MSNE	Corr_pred_obs	Corr_pred_res
oxygen_bottom_max	2.207	0.008	2.397	0.878	0.761	0.016
oxygen_bottom_min	1.152	0.003	2.388	0.923	0.762	0.017
oxygen_surface_max	0.575	0.002	0.802	1.078	0.543	0.029
oxygen_surface_min	0.41	0.002	0.524	1.053	0.221	– 0.033
salinity_bottom_max	4.293	– 0.005	0.556	0.855	0.857	0.104
salinity_bottom_min	0.672	0.004	1.53	0.958	0.929	– 0.009
salinity_surface_max	1.166	– 0.008	1.292	1.336	0.679	0.038
salinity_surface_min	1.768	0.001	1.855	0.858	0.915	0.013
temperature_bottom_max	1.227	0.051	20.936	0.791	0.835	0.072
temperature_bottom_min	0.909	– 0.012	2.392	0.845	0.823	0.047

temperature_surface_max	0.691	0.004	1.934	1.283	0.649	0.016
temperature_surface_min	1.379	0.031	2.624	0.865	0.844	0.031

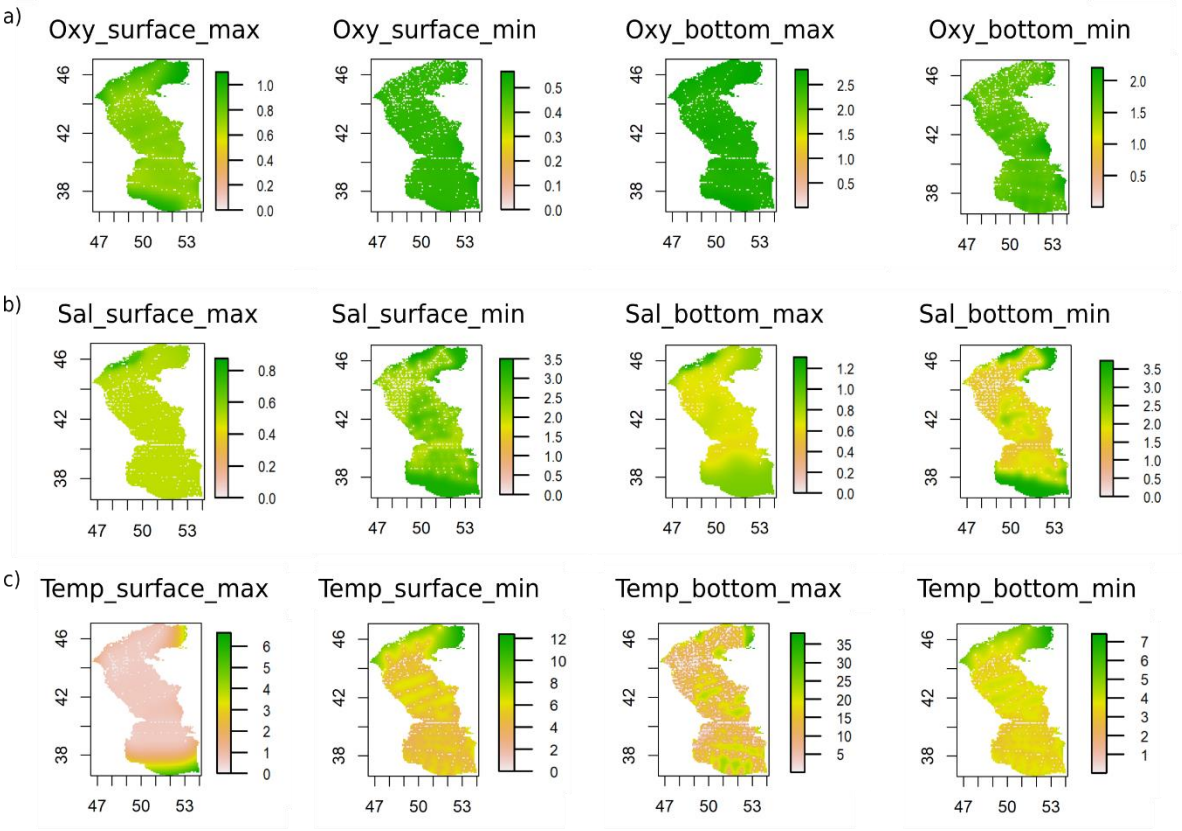

Figure 2: Prediction variance in for the different environmental parameters interpolated. a) Oxygen, b) Salinity, c) Temperature.

References

GEBCO (2014). Gridded Bathymetry Data.

Hiemstra, P. (2013). Automap.

Hjimans, R. (2018). Raster.

NOAA (2018). World Ocean Database.

Raes, N., and Aguirre– Gutiérrez, J. (2018). A Modeling Framework to Estimate and Project Species Distributions in Space and Time. In Mountains, Climate and Biodiversity: A Comprehensive and up–to– Date Synthesis for Students and Researchers, (Wiley Blackwell), p. 13.

Sbrocco, E.J., and Barber, P.H. (2013). MARSPEC: ocean climate layers for marine spatial ecology. Ecology *94*, 979–979.

Tyberghein, L., Verbruggen, H., Pauly, K., Troupin, C., Mineur, F., and Clerck, O.D. (2012). Bio–ORACLE: a global environmental dataset for marine species distribution modelling. Glob. Ecol. Biogeogr. *21*, 272–281.