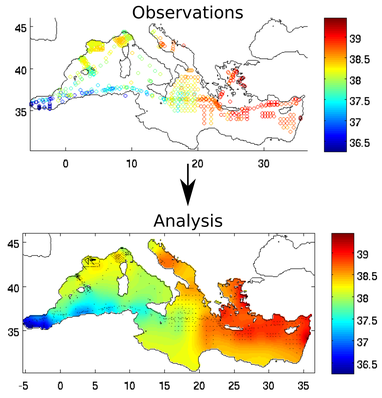
Combining variational interpolation (DIVAnd) and neural networks to generate ocean climatologies from in situ observations

Alexander Barth[(1)] Peter Herman[(2)], Charles Troupin[(1)], Aida Alvera-Azcárate[(1)], Jean-Marie Beckers[(1)]

[(1)] GHER, University of Liege, Belgium. [(2)] Delft University of Technology, The Netherlands. Contact: A.Barth@uliege.be

Data-Interpolating Variational Analysis

- DIVAnd: Data Interpolating Variational Analysis in n-dimensions
- Objective: **derive a gridded climatology from in situ observations**
- The variational inverse methods aim to derive a continuous field which is:
  - **close to the observations** (it should not necessarily pass through all observations because observations have errors)
  - "**smooth**" (i.e. small first and second derivatives)
- DIVAnd is essentially a monovariate reconstruction method
- How can it be extended to use other related variables?



Multivariate analysis

- Multivariate analysis with covariables with a list of covariable $\mathbf{z}_1, \mathbf{z}_2, ...$

$$\mathbf{x} = \mathbf{x}' + f(\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{W}_1, \mathbf{b}_1, \mathbf{W}_2, \mathbf{b}_2, ...)$$

where f is a non-linear function of the known covariables and unknown parameters $\mathbf{W}_1, \mathbf{b}_1, \mathbf{W}_2, \mathbf{b}_2, ...)$

- The structure of the function f is given here by a neural network (multilayer perception).
- The field $\mathbf{x}'$ is also unknown. Its spatial structure is constrained by DIVAnd.

Neural network

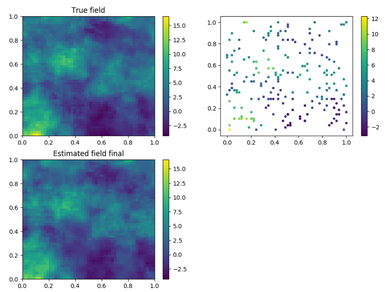
- For every location j, initially the value of vector $\mathbf{v}^1$ are the co-variables at the location j.
- This vector is linearly transformed by a weight matrix $\mathbf{W}_k$ and an offset vector $\mathbf{b}_k$ and then a non-linear activation function (here RELU) is applied to each element element of the resulting vector (except for the last step).

$$\mathbf{v}_j^{(k+1)} = g^{(k+1)}(\mathbf{v}_j^{(k)}\mathbf{W}_k + \mathbf{b}_k)$$

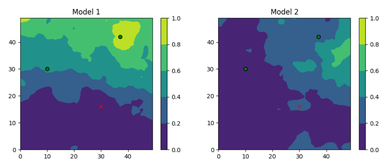
- Here, the weights $\mathbf{W}_k$ and $\mathbf{b}_k$ do not dependent on space but the longitude and latitude are one of the covariables.

Experiments with synthetic observations

- Create a series of random field which are the "covariables"
- Create synthetic observations by combining these covariables: true field
- Sample these fields at random location: synthetic observations
- Perturb these covariables as these covariables are not perfectly known in practice
- Try to recover the true field from the synthetic observation using the imperfectly known covariables using the neural network



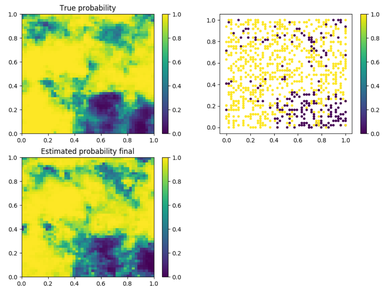
Logistic regression problem



$$p(\text{Model 1} \,|\, \text{obs.}) = 0.9 \cdot 0.7 \cdot (1 - 0.1) = 0.57$$

$$p(\text{Model 2} \,|\, \text{obs.}) = 0.3 \cdot 0.1 \cdot (1 - 0.3) = 0.02$$

- Similar as the previous case, but the true field is a probability of occurrence
- synthetic observation are binary (occurrence or not)
- cost-function to minimize the based on the negative log-likelihood (i.e. find the model which maximize the probability of the observations)

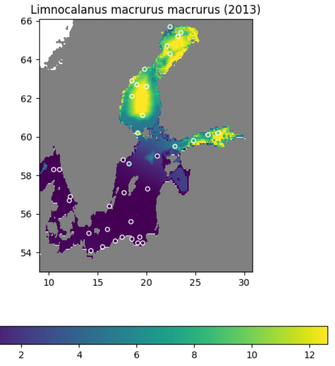


Application

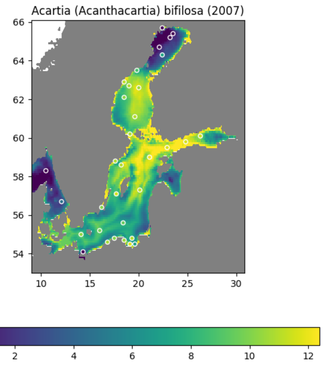
- A gridded data product for 40 zooplankton species using DIVAnd and the neural network library Knet in Julia.
- The neural network uses the following variables as input:
  - dissolved oxygen
  - salinity
  - temperature
  - chlorophyll concentration
  - bathymetry
  - the distance from coast
  - the position (latitude and longitude) and the year
- Abundance values are expressed in number per m^2 and transformed by the function log(x/a + 1) where a is 1 m^{-2}.
- The covers the area from 9°E to 30.8°E and 53°N to 66.1°N at a resolution of a tenth of a degree.
- Gridded data product for the years 2007, 2008, 2010, 2011, 2012 and 2013 have been made. No observations were available for the year 2009.
- The fields represent the yearly average abundance.
- For every specie the correlation length and signal to noise ratio is estimated using the spatial variability of the observations.

Results

- Interpolated field show good agreement with the observations and the cross-validation data points
- Complex spatial dependencies could be learned from the covariable



Limnocalanus macrurus macrurus (year 2013)



Acartia (Acanthacartia) bifilosa (year 2007)

Conclusions

- Neural network can extract non-linear relationships useful to generated gridded data products
- We present essentially a multivariate extension to DIVAnd where the dependency to other variables ("covariables") are estimated from the observations
- Tests with synthetic data show that the underlying true field can be reconstructed from observations, even when the covariables are not perfectly known
- The technique was also applied to abundance of 40 zooplankton species in the Baltic
- The gridded dataset for all 40 species is available at `http://www.emodnet-biology.eu/`.

Acknowledgments

F.R.S.-FNRS (Belgium), EMODnet Biology and SeaDataCloud are acknowledged for their funding and support.