Supplementary Material

**Ensemble modelling of Antarctic macroalgal habitats exposed to glacial melt in a polar fjord**

Kerstin Jerosch\*, Frauke K. Scharf, Dolores Deregibus, Gabriela L. Campana, Katharina Zacher, Hendrik Pehlke, Ulrike Falk, H. Christian Hass, Maria L. Quartino

**\* Correspondence:** Dr. Kerstin Jerosch, kerstin.jerosch@awi.de

# Iterative selection of explanatory variables used for the SDMs

The following environmental variables were initially chosen for the analysis (for information that is more detailed also refer to Appendix 2):

* bathymetry (single beam data IAA 2010, multibeam data UKHO 2012) and its derivatives: geomorphology (broad and fine benthic positioning index (BPI) ([Wright et al., 2005](#_ENREF_76)), slope and aspect (direction of slope),
* bed shear stress (caused by wave action) ([Lim et al., 2013](#_ENREF_38)),
* distances to glacier and to coast,
* satellite image indicating suspended particulate matter and mentioned in the following as “SPM” ([DigitalGlobe, 2014](#_ENREF_16)),
* silt, clay and sand content ([Wölfl et al., 2014](#_ENREF_74))
* hard substrate probability, total organic carbon (TOC),
* total sulphur in surface sediments ([Monien et al., 2014](#_ENREF_48))
* turbidity (average 0-10 m water column), and
* bottom-current velocity ([Lim et al., 2013](#_ENREF_38)).

Removing highly autocorrelated environmental variables constrains the complexity of the models based on the study objective, the attributes of the data, and the understanding of how the main abiotic drivers interact with the underlying biological process. It diminishes the resulting variance of regression parameters, improves processing time, reduces errors during processing, and prevents possible misinterpretation of results ([Elith & Leathwick, 2009](#_ENREF_28), [Guisan & Thuiller, 2005](#_ENREF_35), [Merow *et al.*, 2014](#_ENREF_60)). The selection of the explanatory variables used for the SDMs was an iterative process (Supplementary Fig. 1). The ones which represented redundant information (e.g. mean and max bottom-current velocity) as well as highly correlated ones (Pearson correlation coefficient | r | ≥ 0.7) have been subsequently removed, keeping the most relevant variables with regard to macroalgae distribution ([Dormann et al., 2013](#_ENREF_23)). Note that regression-type approaches (e.g., generalized linear models) and machine-learning techniques (e.g. MAXENT) work reasonably well when used under moderate collinearity ([Dormann et al., 2013](#_ENREF_23)). Furthermore, the resulting mean relative importance of all single models run within the modelling process led to an additional exclusion of further environment parameters. Variables with a low mean variable importance value (≤0.1) determined during the modelling process with all models were excluded.

**Supplementary Figure 1** Workflow diagram. Iterative variable reduction by collinearity analysis (data exploration) and mean variable importance analysis (cross-validation) to six remaining variables. In the ensemble model (EM) those alternative realizations are included that scope a True skill statistic (TSS) threshold of 0.7. The EM is further used to identify areas of modelled over and underestimations as well as to predict habitat shifts by simulated SPM variations.



# – Input data

## Hard substrate

The hard substrate presence and absence data compilation was based on several data sources. Since the occurrence of macroalgae is dependent on hard substrate ([Klöser et al., 1996](#_ENREF_34), [Quartino and Boraso de Zaixso, 2008](#_ENREF_53), [Quartino et al., 2013](#_ENREF_54)) the following assumptions were made and applied to the available data sets in Appendix 2: 1. each substrate coarser than “gravelly sand” is assigned as hard substrate. 2. at locations with soft sediment the absence of hard substrate is assumed, 3. macroalgae present locations are also interpreted as hard substrate in all cases where percentage cover of macroalgae on the photograph or the video was up to 100% and the sea floor was not visible.

The hard substrate data samples described in Appendix 2 (spatial distribution visible in Fig. 2B) were interpolated by using indicator kriging. This application of this method results in a probability raster of hard substrate occurrence [%] with the following properties and statistical mean errors: threshold: exceed, threshold value: 0, model type: exponential, anisotropy: true, neighbor type: standard, max.5/min.2 neighbors, 8 sectors, mean: 0.000942197, Root-Mean-Square: 0.147400121, Mean Standardized: 0.001255649, Root-Mean-Square- Standardized: 0.852104884, Average Standard Error: 0.170117036).

## TOC

The TOC [mass %] raster was interpolated by using the top sediment layers (up to 2 cm) of 47 published ([Monien et al., 2014](#_ENREF_48)) and 10 unpublished (Monien unpublished) push core samples taken in 2010. The statistical errors of several interpolation methods (e.g. IDW, Indicator, Ordinary and Co-Kriging) with changing settings have been compared. The Empirical Bayesian Kriging featured the best values with the following configuration: (neighbor type: standard circular, max.15/min.8 neighbors, 1 sector, mean: 0.009099, Root-Mean-Square: 0.240084, Mean Standardized: 0.026216, Root-Mean-Square-Standardized: 0.988577, Average Standard Error: 0.242304). Clipping and bilinear resampling of the environmental raster input to same resolution of 5 m \* 5 m has been processed on a raster stack in an R statistics environment ([R-3.1.2, RCoreTeam, 2014](#_ENREF_56)).

# Model validation

The EM as the mean of 135 SDMs includes the differences between the alternative realizations. The greater the deviations between the realizations, the greater the spatial range of possible macroalgae distributions realized in the EM. While all 20 response curves of the RF runs follow the same behavior, indicating a stable realization of the RF algorithm, the resulting response curves of ANN realizations diverge distinctively (Supplementary Fig. 2) demonstrating inexpediency of the algorithm.

**Supplementary Figure 2** The response curves of 20 Random Forest (RF) runs following the same behavior and inadequate operating Artificial Neuronal Networks (ANN) response curves of each environmental covariate with 20 replicates



**Supplementary Figure 3** Scenario 2 was compared with earlier estimated macroalgae distribution (Klöser et al., 1996) and ground-truthed with video data (Quartino et al., 2005). The glacier fronts (Rückamp et al., 2011) reveal the macroalgae colonization of the newly ice-free areas.