Divergence in the ecological space between cryptic blue mussel species in the White Sea: towards to species distribution modelling

**V. M. Khaitov, A. A. Zaychikova, P. Y. Safonov, M. V. Katolikova, M. V. Ivanov, P. P. Strelkov**

## Introduction

Species distribution models (SDMs) being a numerical tools describing the relationship between species occurrence and environmental parameters allow to predict distribution patterns of species both in space and time (Elith, Leathwick, 2009). SDMs are also considered as a formal way of species’ ecological niche assessment (Elith, Leathwick, 2009). When the SDMs are applied to a number coexisting species, i.e. community, it allows to describe the ecological niches partitioning between species (Joint Species Distribution Modelling, JSDM, Ovaskainen, Abreg, 2020). SDM/JSDM thus describe how environmental conditions and biological surroundings affect the presence (or abundance) of a species, and allow us to predict under what conditions the presence of the species can be expected. The range of approaches to building SDM/JSDMs is very wide: from regular multiple regressions up to advanced machine learning methods (+++).

The SDM framework is long used in terrestrial ecology (+++) but the amount of papers employing such technique for marine species has been increasing in recent years as well (Robinson et al 2017). SDMs in marine ecology are mostly applied for «good», morphologically distinct species (e.g. Reiss et al., 2011; Lindegren et al., 2022) that can be involved in routine studies which require numerous samples. The growing evidence for cryptic species (Bastrop et al 1998; Fiser et al 2010) and polytypic species (++++REF) among marine animals makes it relevant to search for approaches to analyse their distribution in the case of their sympatric coexistence. The SDM framework seems to be well suited to answer the question of which axes in eclogical space are common for coexisting cryptic species, and along which axes they diverge, occupying different ecological niches. Strictly speaking, by building a distribution model of cryptic species in sympatry, we are describing something like a community and in this sense we come close to the JSDM paradigm. Only few studies have been conducted in this direction in the marine ecology so far (Lowen et al 2019; Dennis, Hellberg, 2010).

In the marine realm, the blue mussel *Mytilus edulis* complex is the longest scientifically known and best-studied group of cryptic species (Knowlton 1993, ++++???). This complex includes several species that are more easily distinguished genetically than morphologically and have the ability to hybridize in sympatry (++++++). Blue mussels are powerful ecosystem engineers in the temperate and subpolar seas playing an important role in coastal communities (+++). They are also important objects of aquaculture (+++). In the North Atlantic, the dominant species are *M. edulis* (thereafter *ME*) and *M. trossulus* (*MT*), which form zones of sympatry (thereafter *contact zones*) in different corners of the region from Scotland and the Gulf of Maine in the south to Greenland and Svalbard in the north (Wenne et al. 2020 and references therein). In contact zones, *ME*, *MT* and their hybrids are often found in the same samples from mussel settlements; such settlements are hereafter referred to as “*mixed*”. Scientists generally agree that *ME* and *MT* are ecologically distinct in sympatry (++++ RC, Katolikova et al.), and have different economic values in aquaculture (+++), but the data on the factors of their ecological segregation i.e. ecological niche partitioning, is fragmentary and frequently contradictory.

On a biogeographic scale, the distribution of *Mytilus* species is thought to be regulated primarily by temperature and its correlates (Hayhurst, Rawson, 2009; Wenne et al., 2020). Both species occur in the Arctic, but *MT* does not penetrate as far south into temperate/boreal seas as *ME* (Wenne et al., 2020), appearing to be a more stenothermic, cold-loving species.

The greatest progress in comparative ecological studies of *ME* and *MT* in sympatry has been made in the contact zones in the Baltic Sea, in the waters of the Kola Peninsula (White and Barents Seas) and in the West Atlantic (mainly, Gulf of Maine and New Scotland). In the Baltic Sea, the brackish areas of the inner part of the sea are inhabited only by *MT*, while the saltier areas closer to the North Sea are inhabited by *ME*. In the middle runs the contact zone, where settlements are usually dominated by hybrids and *MT* gene frequency increases towards the inner Baltic (VS 11, Stuckas et al. 2017). As a result, species distribution is strongly correlated with salinity, against which the role of other factors is negligible (++++).

In the Kola and West Atlantic zones the situation differs from that in the Baltic. Hybrids are always in the minority (less than 20%) in mixed settlements (Wenne et al. 2020; Katolikova et al. 2016). Species are distributed in space in a mosaic fashion, both in regional and local scales (RC, +++++ ). Against the background of the observed mosaic, the relation to salinity is not obvious (+++ RC 05). Some correlation of *MT* proportion with distance to the river’s mouth could be found but no direct evidences of salinity role was provided (Marchenko et al. 2023). In contrast to salinity, a number of other factors of ecological segregation have been proposed for these contact zones.

In the White and Barents Seas, the frequency of *MT* is elevated in port areas, possibly due to introduction of this species into the region in historic times with ship traffic (VS11++++). The only one factor of species segregation explicitly tested in the White Sea was the substrate to which littoral mussels attach (Katolikova et al. 2016). It turned out that *MT* is more common on fucoid algae while *ME* mostly lives directly on the bottom (mud, sand, stones, gravel). The non-random distribution of species across substrates, however, cannot explain the entire local-scale mosaic in their distribution (Katolikova et al. 2016). Parallel studies in the Barents Sea did not consider mussel substrates, but verified the hypotheses of non-random distribution of species with depth. It turned out that on vertical transects, the proportion of *ME* increases with depth and, as a result, *MT* appears to be a more littoral species and *ME* a more sublittoral one (Marchenko et al. 2023). In the western Atlantic, depth, level of anthropogenic pollution and surf exposure has been considered as possible factors in species segregation (++++).

~~It was hypothesized that~~ *~~MT~~* ~~may be more abundant than~~ *~~ME~~* ~~in littoral (Marchenko et al. 2023), polluted (++++), and surf-exposed habitats (+++), but none of these hypotheses have been tested on sufficient material.~~

Summarizing the above, in the non-Baltic *MT*-*ME* contact zones no simple pattern of species distribution along most candidate ecological gradients was revealed and some factors involved into analysis were potentially collinear, masking each other. For example, ports are often located in storm-protected areas, usually in estuaries. This makes it difficult to isolate the effects of shipping (and other anthropogenic factors), surf and salinity. The same could be said about the effects of depth and fouling substrate since the littoral fucoids are rare in the sublittoral where kelps becoming abundant.

This state of knowledge is not surprising given that for most of the time of blue mussel research, scientists have used labor-intensive genotyping methods to identify cryptic species and therefore have been unable to precede large amounts of material (Khaitov et al., 2021). In addition, there were no reliable statistical methods available to model the distribution of sympatric taxa in the space of multiple factors, i.e. no SDM/JSDM approach was realized. To our knowledge, in the history of *ME* and *MT* studies the SDM approach have been applied only twice, by Kijewski et al.(+++) and by Wenne et al. (+++). Both times the machine learning techniques were used to model the macro-geographic distribution of species (technically, of allele frequencies at taxonomically informative genes) in the space of multiple climatic and oceanographic characteristics available from public databases. The conclusions of these studies are summarized by recognizing temperature and salinity as important factors influencing the geographical distribution with *MT* tolerate lower salinities and temperatures than *ME* (Kijewski et al. +++ Wenne et al. +++, see also above).

Finding a simple semi-diagnostic shell trait for *ME* and *MT* - the presence or absence of a continuous prismatic strip under the ligament on the inner side of the shell (Zolotarev +++, Katolikova et al. 2016), allowed us to reliably interpret the taxonomic structure of mixed settlements without genotyping. In the White Sea 74% of *MT*, but only 4% of *ME* have the strip (Katolikova et al., 2016), and using the regression models the frequencies of “striped” morphotypes in samples can be recalculated into proportion of *MT* in population (thereafter *Ptros*, see equations in Khaitov et al., 2021). Note that hybrids are not considered as separate category under this approach.

The use of such extensive data allows the construction of an SDM describing *Ptros* variation from which two outcomes are expected. First, the model will allow us to accurately assess the relationship between the taxonomic composition of mixed mussel assemblages and environmental factors. Second, it will allow to predict in which conditions populations can be expected to be dominated by *ME* (low *Ptros*) and where we should expect elevated concentrations of *MT* (high *Ptros*). Ideally, the second SDM outcome should be expressed as a map constructed using GIS technologies, on which the geolocated predictors’ values are translated into a map of the species presence probability distribution.

In this study, we will construct a model that will describe the relationship between the proportion of *MT* (*Ptros*) in a settlement in the Kandalaksha Bay of the White Sea and four groups of predictors that where considered in the previous works as the most significant environmental gradients regulating species distribution in sympatry (substrate, salinity, proximity to ports and surf level). We will include all the predictors in one model simultaneously to avoid doubts about their collinearity. We will also look in more detail at the influence of substrate type as a possible factor in species segregation. The last of the previously discussed factors, depth, was not studied by us, but was controlled by sampling at the same littoral level. By this way we hope to assess which factors govern the composition of mixed settlements in contact zones and therefore obtain the first type of outcome from SDM (see above). To assess the ability of the model constructed to predict *Ptros* values (the second SDM outcome) we will use test material collected from both the White Sea and neighboring Barents Sea contact zones.

## Material and methods

## Material and methods

### Study area

The 185 km long Kandalaksha Bay is funnel-shaped with numerous islands and skerries and highly indented coastline (Fig. 1). Climate is continental subarctic with 4-5 months of ice cover and the average monthly sea surface temperature in August of 13.8°C. Mean tidal range is about 2 m. Summer surface salinity is 24-25 ppt throughout most of the Bay, but is lower in the estuarine areas (++++). About two dozen of rivers and two hydropower plant’s canals with a catchment area of 141 - 12830 km2 (see Stable ++ ) flow into the Bay, with the largest river, the Niva, entering the Bay at its very top. Due to the complex geometry of the shoreline and numerous rivers, local surf and salinity gradients are pronounced (Filatov et al., 2005).

Mussels are ubiquitous in the littoral zone. They are particularly abundant in the fucoid algal belt (mainly *Fucus vesiculosus* and *Ascophyllum nodosum*), which is continuous 0.5-1.0 m above mean spring tide depth (Berger et al. 2001). Both mussel species (*MT* and *ME*) occur almost ubiquitously in the Bay, but their ratio in settlements varies widely (Katolikova et al. 2016).

Historically (through the 20th century), six ports were functioning in the Bay operating oceanic vessels (Fig. 1). Two of these ports, both in the top of the Bay, are still functioning while the remaining are abandoned (Sailing directions of the White Sea, 1932; Krasavtsev, 2011) but still visited by small ships, according to the common knowledge.

## Modelling data set

### Mussel sampling and processing

Between 2011 and 2018, mussels were sampled at 95 sites within the littoral fucoid belt (materials from 17 sites was published previously as S2 Table in Katolikova et al. 2016). Sites were chosen to describe littoral populations in the Bay in as much detail as possible and to account for the heterogeneity of their habitat in terms of substrate type, surf exposure, and distance from rivers and harbors. All samples were taken in the middle part of the fucoid belt to minimize differences in depth. At each site, three samples from fucoid thalluses (hereinafter, *Algal* samples) and three samples from bottom substrates (*Bottom* samples) were collected in few meters from each other using 0.25 m2 and 0.025 m2 frames respectively. In all cases frame was placed on the bottom area where mussels’ aggregations were visually recognized.

In all samples only mussels with shell size more 10 mm were involved in further analysis. The processing of mussels collected from different substrates was differently organized. From the *Bottom* samples all mussels of appropriated size were selected for further analysis. From *Algal* samples, one bundle of algae (selected so that to contain some dozens of mussels) was taken from the frame and weighted along with attached mussels. The remain algae were taken from the frame and weighed by the same manner (i.e with attached mussels). Only mussels from the bundle selected were used in further analysis. To assess the total number of mussels in the *Algae* sample, the weight of the bundle with the known number of mussels and the weight of the algae rest in the frame were compared. For 12 sites the information on abundance of mussels in algal samples was lacking, and these sites were excluded from those analyses which required this data.

Shell morphotypes (E-morphotype, characteristic to *ME*, and T-morphotype, characteristic to *MT*) were identified for all mussels selected from samples as in Khaitov et al. (2021). As a result, the abundance of mussels of different morphotypes with shell length >10 mm was known for each sample. For further, the proportion of T-morphotypes (*PT*) was converted to the proportion of *MT* (*Ptros*) in each individual sample, total samples from one substrate type in a site (PtrosAlge and PtrosBottom) and in a site in total (PtrosSite ) using the Equation 1 from Khaitov et al. 2021.

### Environmental parameters assessment

In total, we used seven parameters describing possible influence of rivers, ports, surf and type of substrate on mussels. The list of parameters and their description is provided in Table ++. We used three different proxies of salinity and river influence (RiverSize, DistRiver and Salinity), believing that a single estimate of salinity at low tide could be insufficient to characterize salinity per se, and more generally the degree of estuarine conditions of sampling sites. Salinity was measured directly with accuracy of 1 ppm using an “Atago S/Mill-E” refractometer. To classify rivers by size (RiverSize), the data from the ESM +++ was used. To calculate Fetch, the R-package “windfetch” (Seers, 2022) was applied to regional geographic map shape-files.

Table ++. Environmental parameters used for the analysis

| Environmental parameter/ model predictor | Type | Explanation | Range (median) in the data |
| --- | --- | --- | --- |
| *Influence of substrate* | | | |
| Substrate | Categorical | Algal and Bottom samples for each site are treated separately | Algae VS Bottom |
| *Influence of rivers* | | | |
| Salinity | Continuous | Surface salinity (ppm) at the time of sampling, i.e. at low tide. | 2-30 (19) |
| DistRiver | Continuous | The straight line distance (km) between the site and the nearest river mouth by map. The values were log-transformed when used for model fitting. | 0-18.5 (4.9) |
| RiverSize | Categorical | Rivers are categorized according to whether their catchment area is larger or smaller than the median area of all rivers in the region. | Small VS Large |
| *Influence of ports* | | | |
| DistPort | Continuous | The straight line distance (km) between the site and the nearest port by map. Log-transformed values were used. | 0.1-82.2 (18.7) |
| PortStatus | Categorical | Ports are categorized according whether they are active or abandoned | Active VS Abandoned |
| *Influence of surf* | | | |
| Fetch | Continuous | Unobstructed length of water surface (km) over which wind from a certain direction can blow over. Log-transformed values were used. | 0.2-28.8 (3.3) |

### Testing datasets

Two datasets were used as testing ones. The first included mussel samples from 12 sites from the same areas of Kandalaksha Bay (STable ++, SFig. 1 B). At 4 sites only Bottom and at 4 sites only Algae samples were collected with several replications, at 4 sites we collected both Algae and Bottoms substrates but with lesser replications in comparison with the modelling data set (in total 16 units were used). All other procedures were the same as in case of main dataset

The second testing dataset was obtained from published data on mussel populations in the Tyuva inlet (Marchenko et al. 2023, supplementary electronic materials, 44 samples in 2009-2010. This inlet is situated in the Barents Sea where another *MT-ME* contact zone is presented. We used *Ptros* values provided by authors. Predictors were assessed as follow. Since the algae cover was assessed in the sampling area we classified those samples where cover rank was high (4-5 in presented data set) or the samples were taken from kelp forest as samples from *Algae* substrates. When cover ranks were lower (1-3) the samples were classified as *Bottom*. The Tyuva River flowing into the upper part of the inlet having the catchment area of 351 km2 was assessed as a “Large” river (see above). The fetch values for each sites were assessed by the same manner as descrbed above. The nearest port was an active one and was located in the Kola bay (69.19 N, 33.496 E) about 3 km from the mouth of Tyuva inlet.

We split the data from the Tyuva inlet into two parts. The first part included sites located 0.5 - 1.5 m above chart datum, i.e. approximatelly in the fucoid belt zone (as was in the main dataset). The second part included sites located above the fucoid belt (2 m upper chart datum) and in the sublittoral (-0.5 to -3.5 m). We analysed these two parts of dataset separately.

The second testing dataset was taken from the study of Marchenko et al. 2023. It included 44 estimates of *Ptros* from Tyuva inlet in the Kola Bay of the Barents Sea obtained in 2009-2010. For each sampling site a number of environmental characteristics were known including depth, Salinity, cover abundance of macrophytes by rank scale, and dominant algal species (usually, kelps in the sublittoral and fucoids on the littoral). Twenty three sites were from the littoral fucoid belt (depth 0.5-1.5 m; note that the tidal amplitude in the Barents Sea is greater than in the White Sea), 6 - from the littoral above the fucoid belt while the rest were from sublittoral. Since the substrate of mussel fouling was not controlled during sampling, we classified samples into Bottom and Algal ones by the algal cover (ranks 1-3 and 4-5, correspondingly) **but all samples from subtidal kelp forest were classified as Algal**. The remaining environmental parameters were predicted as for the modeling dataset, with the nearest port in Ekaterininskaya Gavan Bight considered as active and the river Tyuva flowing into Inlet as a large one.

### Statistical analysis

All processing was performed using the statistical programming language R 4.05 (R core Team, 2021).

#### Dependency of **Ptros** on predictors in modelling dataset

Assuming that the relationship between the dependent variable (*Ptros*) and continous predictors can be curvilinear (Austin, 2002), we used GAM (generalised additive model , Wood, 2017) as a modelling technique that has worked well for SDM construction (Elith et al., 2006). The GAM fitted (thereafter *Model 1*) was based on beta-binomial residuals distribution and the restricted maximum likelihood (REML) method for parameters estimation. Smoothers for all continuous predictors were fitted using cubic basic splines. Categorical predictors were included as parametric terms in the model. “Site” was considered as random factor in the *Model 1*. The function gam() from the package “mgcv” (Wood 2017) was used to fit the model.

To check for the predictors’ collinearaity in the model we calculated the variance inflation factor (VIF, Fox & Monette, 1992). Additionally we calculated Pearson correlation between continuous predictors. To verify that the *Model 1* meet the assumptions of sampling independence, we examined the presence of residuals’ spatial autocorrelation by means of spline correlogram construction (Bjornstad, Falck, 2001) with the function spline.correlog() from the package “ncf” (Bjornstad, 2022). No evidences of spatial autocorrelation were revealed for the *Model 1*.

#### Association between *Ptros*, subtrate types and mussel abundance

For a more deep analysis of association between *Ptros* and substrate type we calculated the difference (*Dif*) between *PtrosAlgae* and *PtrosBottom* for each site. The obtained *Dif* values were used as dependent variable in the *Model 2* which was constructed as GAM with gaussian residuals’ distribution. The proportion of *MT* in site (*PtrosSite*) and the mean mussel’s abundance in the site per 1 m2 (*MeanN*) were used as predictors in *Model 2*. We used function gam() from the package “mgcv” for the model construction.

#### Assessment of model predictive power

To assess ability of *Model 1* to predict the *MT-ME* distribution we used three data sets: (1) main modelling dataset (the data used for *Model 1* construction), (2) testing dataset from the Kandalaksha bay, (3) the first part of testing dataset from the Tyuva bay, i.e. data collected from fucoid’s belt and (4) the second part of data from the Tyuva inlet, i.e. sites out of fucoid’s belt.

To assess the model’s predictive power all sites included in the analysis were divided into two groups *MT*-dominated (*Ptros* > 0.5) and *ME*-dominated (*Ptros* < 0.5). Using the parameters of *Model 1* we calculated predictions of the model for each of four datasets. Thereafter the predicted values were considered as classifiers for detecting *MT*- or *ME*-dominated groups. The receiver operated characteristics (ROC) followed by analyzing the area under curve (AUC, Fielding, Bell,1997; Fawcett, 2006) was used to evaluate models performance. The function roc() from the package “pROC” (Xavier et al. 2011) was used.

## Results

### Environmental parameters in the Kandalaksha Bay

Two of six ports presented in the area were active at the date of material collection (Fig. 1 A, STable ++ SEM). The most exposed sites (maximum values of wind fetch) were located on the western coast of the Kandalaksha Bay, and on the open shores of the islands at the top of the bay (Fig. 1 A). The most desalinated areas were the top of the Bay and inlets where rivers flow into (Fig. 1 A). Salinity at sites situated close to “Large” rivers was lower than at sites closer to “Small” ones (SFig. ++, C). Salinity tended to increase with increasing of distance to the mouth of the nearest river (SFig. ++ D) but no association between distance to the nearest port and salinity was found (SFig. +++ E). Sites situated close to ports tended to have lower wind fetch (SFig. +++ F). All correlations between salinity, distance to the nearest river mouth, distance to the nearest port and wind fetch were rather low (STable +++). The largest correlation revealed was between wind fetch and distance to nearest port (r = ++).

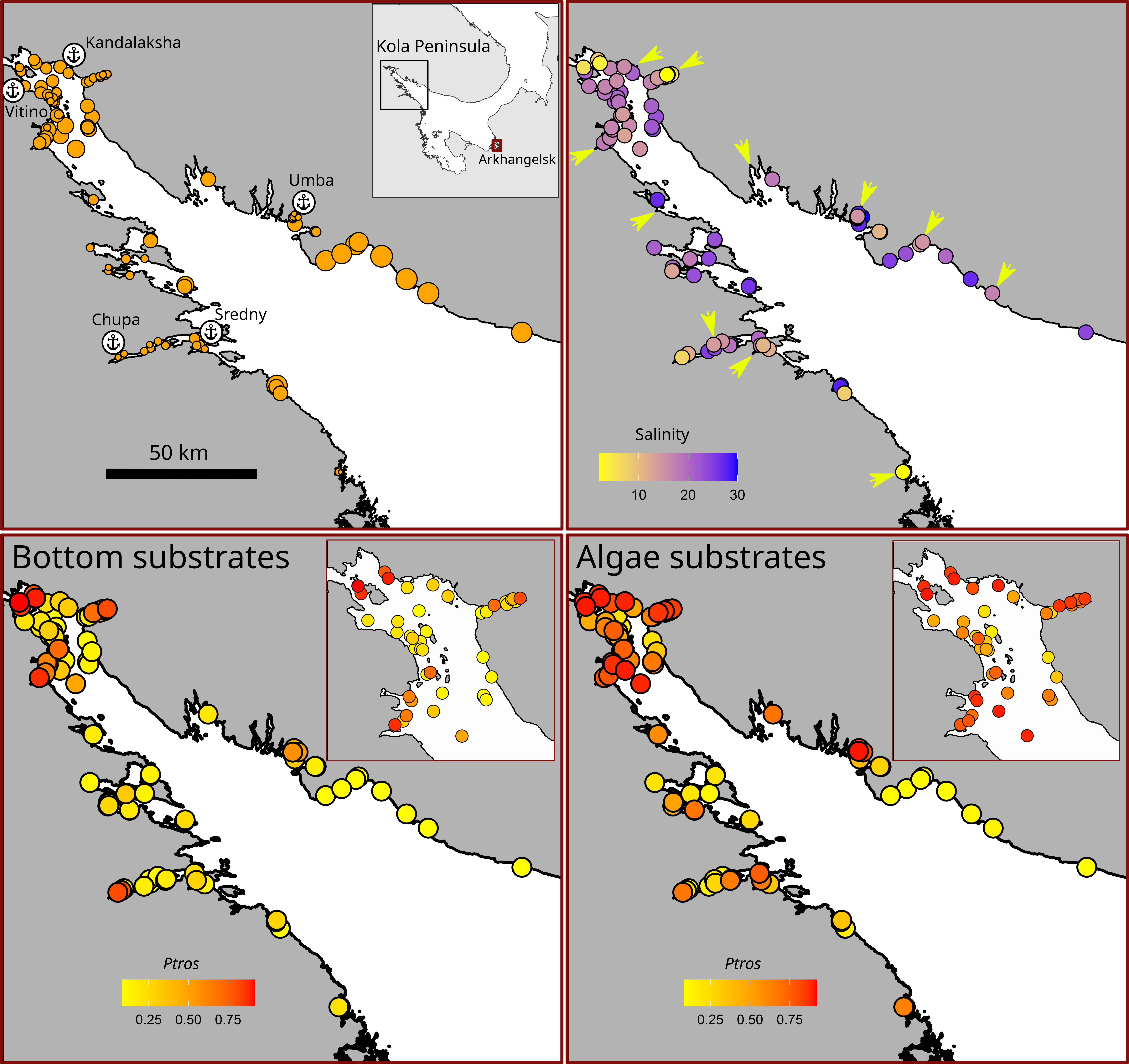


Figure +. Characteristics of sampling sites in the Kandalaksha bay. (A) The surf condition in the sampling sites: point size is proportional to Fetch value. Anchor signs mark position of harbours. Abandoned ports are marked by asterisks. (B) Salinity in the sampling sites. Arrows mark large river’s mouths. (C) and (D) *Ptros* in Bottom and Algae samples, respectively. Inset maps are placed to show detailed points in the upper parts of the bay.

### *Ptros* spatial distribution in the Kandalaksha Bay

The maximum *Ptros* was observed in the top of the Kandalaksha Bay itself and in the small bays deeping into the mainland (Fig. +, C, D). At the same time, sites with a lower *Ptros* were located in the immediate vicinity of sites where high *Ptros* values were represented (Fig. +, C, D), reflecting the high variation of taxonomical structure in small spatial scale. Settlements with the lowest *Ptros* both in *Bottom* and *Algae* samples were represented in the open parts of the North-West coast.

### Relationship of *Ptros* to environmental parameters

The values of VIF calculated for the predictors’ set included in the analysis were rather small (the maximal VIF was detected for Fetch: 1.76). Although some pairwise correlation between environmental factors were found (see above) the low VIF values can be interpreted as a negligible collinearity between predictors (i.e. they don’t mask influence of each for other).

To assess the dependency of *Ptros* on all predictors considered we constructed the *Model 1* (Table +, R2 = 0.71) which explained 77% of total deviance. The *Model 1* revealed statistically significant dependency of *Ptros* on all predictors except one.

The effective degrees of freedom (*edf*, Table ++) for most smoothers were close to 1, reflecting the linear dependence of *Ptros* on such continuous predictors as distance to the nearest port (*DistPort*) and fetch (*Fetch*). The exception was *Salinity*, where a pronounced curvilinear dependence was revealed.

Table 1 . Parameters of smoothers and coefficients of parametric terms for the Model describing dependency of Ptros on predictors.

| Smoother terms | edf | ref.edf | Chi.sq | p-value |
| --- | --- | --- | --- | --- |
| s(Salinity) | 2.4 | 9 | 396.7 | 0.0033 |
| s(DistRiver) | 0.0 | 9 | 0.0 | 0.6724 |
| s(Fetch) | 0.9 | 9 | 88.2 | 0.0417 |
| s(DistPort) | 1.0 | 9 | 276.2 | 0.0016 |
| Random effect s(Site) | 74.4 | 92 | 453.6 | 0.0000 |

| Parametric terms | Parameter estimate | SE | z-statistic | p-value |
| --- | --- | --- | --- | --- |
| (Intercept) | -1.7 | 0.1 | -11.8 | 0.0000 |
| Substrate(Algae) | 0.9 | 0.1 | 14.6 | 0.0000 |
| RiverSize(Large) | 0.4 | 0.2 | 2.6 | 0.0091 |
| PortStatus(Active) | 1.0 | 0.2 | 5.7 | 0.0000 |

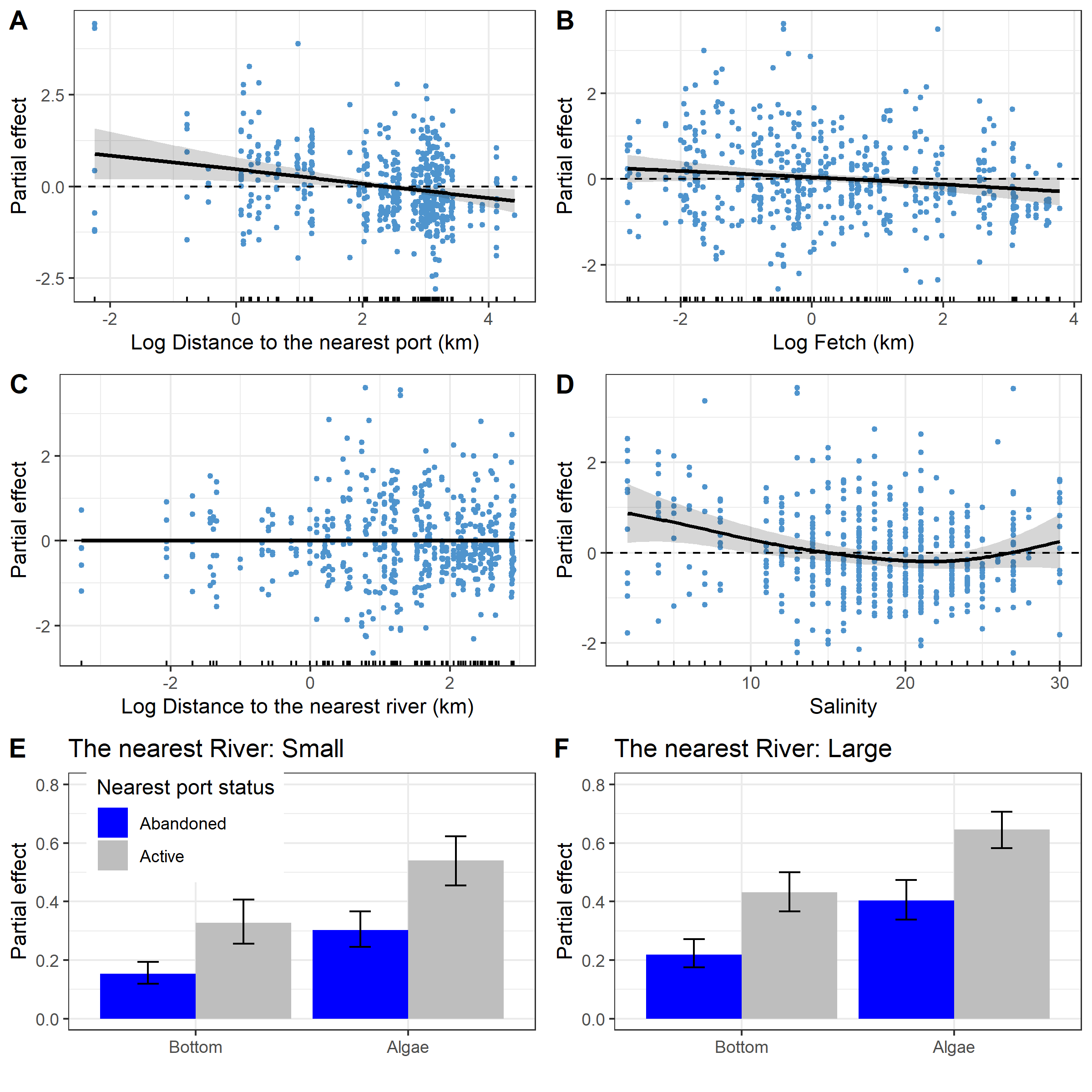


Fig. 3. Predictions of the *Model 1* describing the dependency of *Ptros* on distance to the nearest port (A), average fetch (B), distance to the nearest river (C), salinity (E) and substrate type, nearest port status and size of the largest river (E, F). The gray ribbons and wiskers demonstrate 95% confidence interval. The dotted lines in panels A-D represent the median predicted value, this value is given to track the trend in the change in model predictions. Points on panels A-D reflect partial residuals but not raw data for each sample.

According to the *Model 1* (Table 1, Fig. 3) *Ptros* has a tendency to decrease with increasing of distance to the nearest port (Fig. 1 A). The nearest port status had also significant influence on *Ptros*: the value was higher when port status was *Active* (Fig. 3 E, F). The dependency of *Ptros* on fetch (Fig. 3, B) was negative. No significant association between *Ptros* and the distance to the neares river was was found (Table 2, Fig 3 C).

The curvilinear dependence of *Ptros* on salinity (Fig. 3, D) can be described as follows. The predicted value of *Ptros* decrease when salinity moving from minimal values to apr 20-22 ppt when it reached minimum. When salinity increases after this point the *Ptros* value increase as well.

According to the *Model 1* the value of *Ptros* is dramatically higher on the *Algae* substrate than on the *Bottom* one (Table 1, Fig. 3 E, F). This value is higher in those sites which is close to *Active* ports than to *Abandoned* ones. Finally, the *Ptros* value is slightly higher at sites situated close to Large rivers than to Small ones.

### Dependency of *Ptros* on substrate type

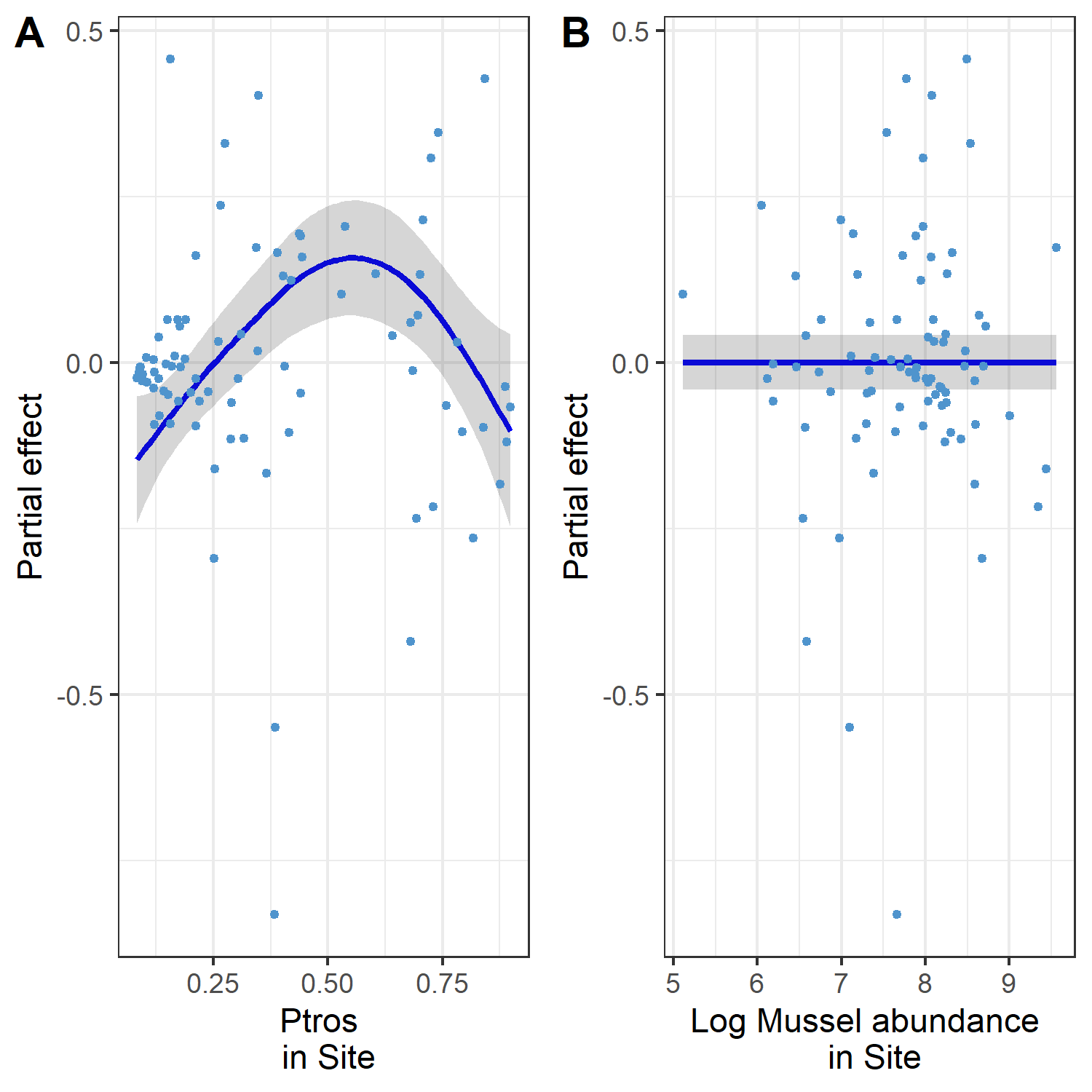


Fig. 4. The dependence of difference between *PT* on Algae and Bottom subatrates on PtrosSite (A) and total mussel abundance (B).

The difference between *PtrosAlgae* and *PtrosBottom* substrates was significantly dependent on *PtrosSite*, i.e. prevalence of *MT* (STable +, Fig. 4 A). The dependency revealed was curvelinear (edf = 2.7, STable +) with minimal *Dif* values in sites with minimal and maximal values of *PtrosSite*. The maximal *Dif* values were associated with intermediate *PtrosSite*. This pattern reflect the prevalence of *MT* at *Algae* substrate and prevalence of *ME* at *Bottom* substrate in those sites where both species are presented in equal amount. No significant dependency was revealed with total absolute mussel abundance (STable +, Fig. 4 B).

### Assessment of model predictive power

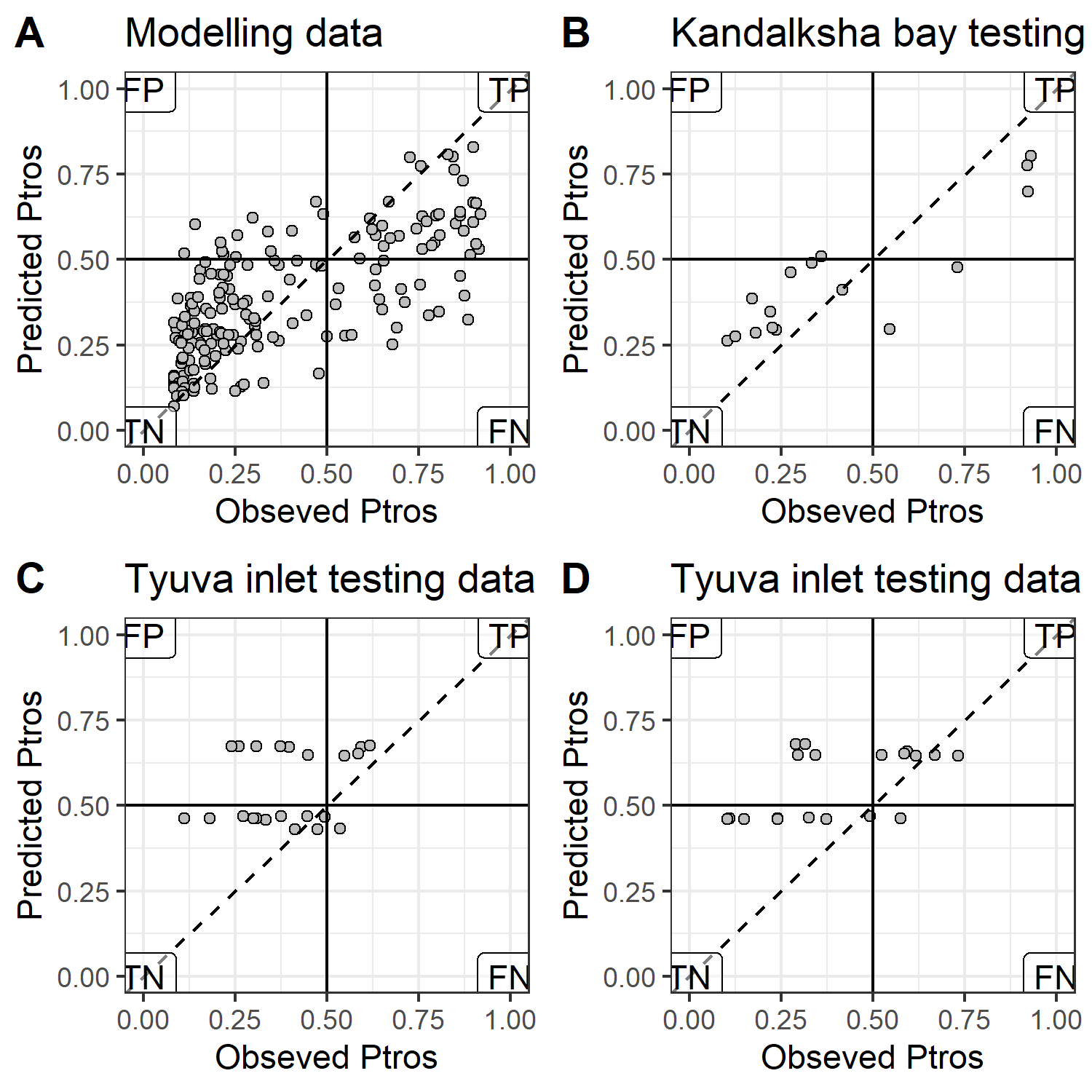


Fig. 5. Comparison of *Ptros*-values predicted by *Model 1* and obsreved *Ptros* for training data used for the model fitting (A), testing data from the White sea (B) and testing data from the Tyuva inlet colleced in and out of the fucoid’s belt (C and D respectively). The horizontal and vertical solid lines reflects the boundary between *MT*- and *ME*-dominated sites. Dotted lines represent the position of perfect fit when observed and predicted values are equal. Labels in the plot corners mark the quadrants with false positive (FP), true positive (TP), true negative (TN) and false negative (FN) predictions.

The *Ptros* values predicted by the *Model 1* and observed ones (Fig. 5, A) were in rather good agreement for modelling dataset (AUC = 0.87). The most of 190 units (joined Algae and joined Bottom samples from 95) were correctly classified as MT-dominated (37 units in true positive zone) or ME-dominated (120 units in true negative zone). A few points are situated in the false positive zone (the *Model 1* predicts presence *MT*-dominated populations, but *ME* - dominated ones were observed, 13 units). A part of points were presented in the false positive zone (the *Model 1* predicts presence *ME*-dominated populations, but *MT* - dominated ones were observed, 19 units).

In the case of testing dataset from the Kandalaksha bay (Fig. 1 B) the most sites were correctly classified by the *Model 1* (AUC = 0.84). Only 2 points were located in false negative and only 1 in false positive zone.

The *Model 1* being applied to the first part of testing dataset from the Tyuva inlet (i.e. collections from fucoid’s belt) display lesser but still rather high predictive power (AUC = 0.71). Six samples were classified as false positive, but only one as false negative. The predictive power of the *Model 1* being applied to the second part of the Tyuva testing dataset was similar to the previous case (AUC = 0.71). Four samples were classified as false positive and only one as false negative. In total in Tyuva data set 32 of 44 samples were correctly classified as MT-dominated of ME-dominated populations using predictions of *Model 1*.

## Discussion

Processing a large number of samples from various sites across a wide range of environmental conditions in the Kandalaksha Bay contact zone, we developed the *Model 1*. This model can be considered both as SDM describing the association with environmental predictors and JSDM reflecting the joint distribution of the two mussel species, *Mytilus edulis* and *M. trossulus*. It is important to acknowledge that the model we have constructed represents only one of the initial efforts to characterize the relationship between cryptic mussel species and their environment. Similar to previous studies by Kijewsky et al.(+++) and Winne et al. (+++), our analysis focused on the distribution of mussels in their contact zones exclusively. The broader pattern of species distributions, including their relationships with various factors across the entire range of the cryptic species, falls outside the scope of this study. Nevertheless, the most significant events related to species interactions with environmental factors and with each other (including intraspecific relationships) occur primarily in contact zones.

The set of environmental predictors included in the *Model 1* obtained most of those factors that were generally considered as the important regulators of the taxonomic structure of mussel’s populations in *MT-ME* contact zones (+++++). Being included in one SDM these factors possessed significant influence on *Ptros* (i.e. proportion of *MT* ). Thus, our analysis provided the consistent evidence that factors such as salinity, proximity to ports, level of surf and substrate type influence the taxonomic composition of mixed mussel populations. In addition, we found that the predictors included in the model are not collinear. This means that we can consider all these factors as relatively independent ecological axis that do not mask each other, which is important for SDM construction (De Marco and Nobrega, 2011). In addition we considered environment variation in a relatively small contact zone in the Kandalaksha Bay. We can state that the patterns revealed on more extensive contact zones (e.g. in the Baltic - North Sea contact zone, see Kijewsky et al, 2019) with longer ecological gradients retain their influence in essentially smaller spatial scales. This fact allow to optimistically expect the ability to construct working SDM describing distribution of *MT* and *ME* in contact zones of different sizes.

The SDMs are frequently considered as models reflecting structure of species’ ecological niche (Elith, Leathwick, 2009). Although the very concept of ‘niche’ is under the criticism (McInerny, Etienne, 2012 a,b,c) analysing the relationship of cryptic species to environmental factors and investigation of mechanismes of their distribution along environmental gradients is undoubtedly an important task. In this sense, any statistically significant association of *Ptros* with some environmental predictor would reflect the divergence of *MT* and *ME* in ecological space (i.e. partitioning of niches).

From all environmental variables included into analysis the distance to the nearest port and its status (*Abandoned* VS *Active*) may have dual interpretation. We revealed higher *MT* proportion close to active ports. On the one hand, active ports tend to have a higher degree of anthropogenic pollution, leading to stress of mussels (Hellou & Law, 2003). Different mussel species are known to interact differently with pollutants (Beyer et al., 2017). For example, in the Halifax area *MT* are more abundant in the most polluted habitats than *ME*, but show less stress tolerance than *ME* (Hellou & Law, 2003).

Another explanation may be that there is some foci of disease associated with port areas that affect *MT* and *ME* diferently. For example we can look for the reason in such infections as the bivalve transmittable neoplasia which is associated with harbors (Hammel et al., 2023) or in other parasite diseases for which ports are hot-spots for invasion (Feis et al., 2019). This could be a perspective direction in future attempts to construct more reliable SDM.

From the other hand the dependency of *Ptros* on distance to the nearest port and its status can reflect not true axis in ecological space but some trace of historical processes. It is believed that *MT* was introduced into the Barents and the White Seas due to ship traffic during World War II (Vainola & Strelkov, 2011). This hypothesis was based primarily on the fact that high concentrations of *MT* were found in the vicinity of currently active or abandoned ports in the region (Vainola & Strelkov, 2011; Katolikova et al. 2016). The results of our study are inline with this hypothesis. The *Ptros* value increased with approaching to harbors in the studied area. Apparently, the degree of port activity also plays an important role. According to our data, if the port nearest to the mixed settlement is active, i.e. it serves the ships coming from outside the White Sea (in our case it is “Kandlaksha” and “Vitino” ports), then the *Ptros* value in such settlements is higher than in the settlements adjacent to the abandoned harbors. The observed relationship may be related to the fact that the currently active ports were the largest and most economically sustainable systems in the region. This may be based on the higher shipping traffic, which has led to a more massive invasion of *MT* from other regions (primarily from the ports of Murmansk in the Kola Bay, where the abundance of *MT* is high, Vainola&Strelrov, 2013; Marchenko et al., 2023). In the abandoned ports, the introduction of the alien species occurred during the times of their active operation. However, during several decades of inactivity of abandoned ports *MT* could be displaced by native *ME*.

In any case, our data are in line with earlier stated hypothesis that in the case of the White Sea *MT* is an invasive species (Vainola&Strelkov 2011). Unfortunately, only few data could be provided to check this hypothesis for other European *MT* -*ME* populations. However the ship traffic was successfully included into SDMs describing invasive species in European waters (Lindegren et al., 2022) this factor was never included into models describing *MT*-*ME* distribution previously. In some cases it is probably an unrealistic aim due to absence of predictor variability. For example in the case of *MT*-*ME* populations in the Baltic Sea, where there is extremely active ship traffic and vessels visiting almost all parts of the Baltic Sea, it is not possible to identify a correlation between the distribution of *MT* and the presence of ports. At the same time, if we look at the more broad map of *MT* distribution (Vainola & Strelkov, 2011) around European coasts, the confinement of *MT* to large European ports out of the Baltic Sea (e.g. Bergen, Murmansk), rather speaks in favor of the hypothesis of *MT* introduction due to ship traffic. Thus, while proximity to ports is not an environmental factor in a direct sense, it is a parameter that should be kept in mind when building a more general SDM describing *MT*-*ME* distribution in contact zones.

Our analysis revealed that the higher wind fetch is associated with lower *Ptros*. That means *MT* avoid open coasts potentially exposed to the surf. The lowest *Ptros* was observed on open shores of the north-east coast of Kandalaksha Bay (Fig. 1 A), exposed to wave action due to southeast and south-west winds which are frequent in the White Sea region (Berger, Naumov, 2001). This result corresponds well with the data on the thickness of *MT* and *ME* shells. It is known that *MT* have thinner, more fragile shells (Beaumont et al., 2008; Michalek et al, 2021; our unpublished observations of mussels from the White Sea are in line with this data). It is logical to expect that mollusks with such characteristics should avoid places with high wave loading. In such habitats, *ME* with stronger shells are expected to benefit. A similar pattern can be seen in other contact zones. For example, in Greenland, an increased frequency of *MT* was noted at the sheltered top of the fjord (GLS and GLD sites, Fig. 1 in Wenne et al. 2016), while the population at a more open site located near fjord mouth (GLL site in Wenne et al. 2016) is dominated by *ME*. We believe that when building a more general SDM, based on the study of many contact zones, it is necessary to take into account this environmental parameter. It is possible that some more reliable way to estimate the degree of wave impact will be proposed, but fetch has already been shown as a convenient parameter on which mussel settlement characteristics depend (LaBarre et al., 2023).

The nature of the substrate was discovered as the powerful predictor regulating *Ptros* variation in the scale of several meters. The *Ptros* appears to be significantly higher on algal than on bottom substrates (silty-sandy sediment, pebbles, boulders and rock surfaces) located in close proximity (not more than 20 m, within the same fucoid belt). This pattern was first detected in the analysis of a significantly smaller volume of material (Katolikova et al., 2016). Bringing in new data collected in a wider range of conditions confirmed the presence of this pattern.

However, a closer examination of the relationship with substrate type (Fig. 4 A) showed that the maximum *Ptros* divergence between Algae and Bottom substrates occurred only when both mussel species were present in approximately equal numbers at a given location. This could be interpreted as divergence of realized niches as a result of interspecific competition. This, however, is contradicted by the lack of correlation of *Dif*-value with the absolute abundance of mussels at the site. If the mentioned divergence of realized niches were a consequence of competition, one would expect that the strength of this interaction would increase with increasing settlement density, which is not observed. Thus another explanation should be considered.

As possible explanation we can propose some difference in byssus attachment strength. Perhaps *MT* are capable of attaching to underwater substrates more strongly than *ME* do. This should cause less strongly attached molluscs to break off and remain on the bottom under constant fluctuations of fucoid’s thallus. Additional investigations of attachment strength of both species are needed.

The divergence of the two species by substrate is undoubtedly significant evidence of ecological niche partitioning between creyptic species, but the value of substrate type as an SDM parameter for predicting the conditions under which an increased concentration of a species can be expected in mixed settlements may not be as high. Austin (2002) considers two types of environmental gradients: distal and proximal ones. The proximal gradient is a causal environmental variable determining the direct organism response. The response to distal gradients is mediated by a longer chain of reactions. For example, fetch or proximity to ports has no direct effect on mussel viability and they should be considered as distal gradients. Whereas can a mussel attach itself to the present substrate in given location seems to be dependent primarily on substrate type. In this sense the later should be classified as proximal gradient. Since it is very difficult to provide GIS coverage for proximal variables their value for predictive mapping of species distribution probably impractical (Austin, 2002). However the type of substrate from which mussels were collected should be obviously registered during the sampling session.

An intriguing result of our study was the absence of clear linear dependency between taxonomic structure of mixed populations and salinity that was revealed in other contact zones (Kijewsky, 2019). However, surprisingly, in spite of the fact that salinity is considered as a leading factor at least in Baltic contact zone only a few works have been devoted to comparison of ecophysiological responses of *MT* and *ME* to salinity (Gardner&Thompson 2001; Qiu et al., 2002; Knobel et al. 2021). Moreover, the direct physiological investigations did not reveal unequivocal evidence that adult *MT* are better adapted to reduced salinity than adult *ME* (Gardner&Thompson 2001; Qiu et al., 2002). It has only been shown that larval stages of *MT* survive lower salinity better than the same *ME* stages (Qiu et al., 2002). That is the prevalence of *MT* in lower salinity could be explained by early selective pressure during passive larval drift across salinity gradient (Knobel et al. 2021).

In our investigation the GAM fitted reveal a statistically significant curvelinear (close to U-shape) relationship of *Ptros* to directly measured salinity. From one hand, our *Model 1* revealed higher *Ptros* in lower salinity which is typical for Baltic contact zone. On the other hand, *Ptros* was inflated when salinity was high. We can propose three explanation to this fact.

Firstly, although directly measured salinity seems to be a natural ecological factor, this may not be the case for littoral mussel settlements. The salinity levels significantly vary during the tidal cycle and the range of the variation is highly dependent on the distance from the fresh water source (Attrill, 2002; Marchenko et al., 2023). We do not know which salinity is more important for littoral mussels living in the estuaries: high during the ebb time or reduced in the low tide. Further, the salinity measured once during the sampling process may occasionally be anomalously low or high by accident (e.g. due to precipitation or upwelling of deep salted water, respectively). If this explanation is correct, then the observed increase in *Ptros* at high salinity may be a consequence of incorrect estimates of the salinity regime..

A second explanation for the *Ptros* peak in high salinity sites may be related to the lower reliability of the morphotype-test for species identification under high salinity conditions. Khaitov et al (2021) suggested that in cold Arctic waters with high salinity, *ME* more often exhibit T-morphotype, which reduces the sensitivity of the method. However this is contradicted by the fact that in the same locations where high salinity and high *Ptros* values were detected (e.g. in the Umba area, Fig. 1 A, B) genetic analysis of mussel populations revealed an increased frequency of *MT* genes (Katolikova et al.2016).

–>

And finally, it should be noted that sites with high *Ptros* and low (<15 ppt) or high (>27 ppt) salinity are not numerous (compare concentration of points in Fig ++, D). In locations, where salinity is intermediate, *Ptros* does not differ significantly from the background value. That is, anomalously high *Ptros* values are observed in some marginal habitats. Such pattern can be expected if we assume that *MT* is a weaker competitor, which on average cannot displace *ME*. This assumption is supported by the fact that the settlements that we classify as *MT*-dominated are essentially scarcer in the studied area (TP + FN part of the Fig. 5 A) than the *ME*-dominated sites (FP + TN part of the Fig. 5 A). The roughly U-shape association between *Ptros* and salinity can be turned into bell-like dependency of proportion of *ME* on salinity. This means that the majority of mussels presented in the sites within intermediate salinity range are mostly *ME*. Such displacement of weaker competitor into marginal habitat is well known for other systems where inviders met resident species: mosquito larvae (Juliano, 1998), fishes(Herbold, Moyle, 1986) and geckos (Case et al 1994). Usually resident species persist in their optimal habitats if they have not been destroyed but introduced species prevail in marginal habitats (Juliano, 1998). To note, actively used harbors appear to be the very marginal habitats as well.

In any case, we should not ignore the role of salinity in regulating of taxonomic composition of mixed population when constructing SDMs. The general desalination regime in the contact zone should also probably be taken into account. For example, in our case, higher *MT* concentrations were observed when the nearest river was a Large one. The relationship with the status of the nearest river can be explained not only by the fact that a large river can be a powerful source of desalination. It may be a consequence of the fact that in the immediate vicinity of large rivers there may be areas of *MT* concentration. These refugiums could serving as a source of larvae of *MT* for the surrounding regions.

The strongest argument that the constructed model is useful is its ability to predict phenomena on the dataset that was not used to build the model. We used two testing datasets. One was represented by material collected in the same area as the training samples (Fig. 5 B). The ROC-analysis showed that *Model 1* has a high ability to predict the assignment of the sample to one of the two categories *MT*-dominated vs *ME*-dominated. The AUC-values calculated for both modelling and testing data sets from the White Sea were higher than useful amount of discrimination, empiricaly assessed as AUC >0.75 (Elith et al., 2006).

When the *Model 1* was used for predictions in a different region, the Tyuva inlet, where an independent contact zone is presented, the quality of the prediction decreased. However, the AUC values were close to the above threshold and were well within the average predictive power values for GAM as a method of SDM constructing (Elith et al., 2006). At the same time, it should be kept in mind that a key predictor such as substrate type was estimated very roughly when analyzing this dataset. This may have reduced the predictive power of the model. In addition, it should be taken into account that Tyuva Bay is located in a different climatic zone than the White Sea and differences in, for example, temperature regime may be affected. We have not included predictors with latitudinal variation in the *Model 1*. Perhaps in the future, when building a more universal SDM, some factors with latitudinal variation (e.g. temperature, precipitation, etc.) should be included.

The important result of the *Model 1* applying was the high number of false positive predictions in the case of the Tyuva inlet. There may be some ecological reasons for this pattern. If a suitable habitat is not colonized by a target species which population is small (i.e. the area is not saturated) or the population is collapsed due to displacement as a result of inter-specific competition, all this reasons could produce false positive result when a model-classifier is applied (Fielding & Bell, 1997). Indeed the *MT* population in the Tyuva inlet displayed pronounced decreasing associated with increased *ME* population (Marchenko et al 2023). We may expect the Tyuva inlet conditions are highly suitable for *MT* population but once appeared it demonstrated decrease probably due to competitive displacement by *ME*.

We expect that alternative scenario may take place in the case of the Kandalaksha Bay. Let’s remember that false negative predictions were rather frequent in the area, i.e. MT-dominated populations are frequently present in unsuitable conditions. This could be interpreted as evidence of the oversaturation of the region by MT-populations. This may be observed during mass invasion of the species into new habitats.

## Conclusion

In our analysis, we did not aim to construct some perfect SDM it is the task for further investigation when information on different contact zones would be combined. That is why we did not try to build some final model (for example, by procedure of backward selection of optimal predictors’ set) and did not try to find the most appropriate modelling method (see Elith et al 2006). It was important for us to understand whether in principle a working model can be built based on those predictors that have come to the attention of researchers studying contact zones of cryptic mussel species. Our results are optimistic about this possibility but since the contact zones between *MT* and *ME* are scattered over a very wide range of latitudes and situated in various biotic surroundings the appropriate form of the “general” model should be well considered. It is obvious that some additional predictors (e. g. temperature or its correlates, the position in the intertidal-subtidal gradient, *etc*) should be included in a future model. It is also clear that the “general” model should include some biotic interactions (see Wisz et al 2013) and be developed within the JSDM framework (see Ovaskainen & Abrego, 2021 for a theory). The dependent part of this model should include not only proportions of two coexisting mussel species but a matrix describing the community where mussel’s species coexist with other species. Our study has shown that algae play an important role in the local distribution of species. Perhaps the spatial variation in algae abundance (and probably algae species composition) can play a significant role in the distribution of the cryptic mussel species. Thus algae characteristics in region should be somehow incorporated in the model. Perhaps some role may also be played by predators, which can regulate the proportion of mussel species (Khaitov et al.2018; Khaitov et al.2023 ). However the first step to construct the working SDM is the deep comparison of ecological features of different *MT-ME* contact zones. It may be that the different contact zones are not as disparate as previously thought (RC 2005 ++++).

## Literature cited

Лоция Белого моря / Гидрографическое управление ; общ. ред. А. Н. Рождественского. – 5-е изд. – Ленинград : Издание Гидрографического управления, 1932. – LXXII, 566 с., 22 вкл. л. ил., карт., схем.

# Supplementary electronic materials (SEM)

–>

–>