Cryptic mussels in a contact zone: 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 (SDM) 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 usually considered, as well, as a formal way of species’ ecological niche assessment (Elith, Leathwick, 2009) and when the modelling is applied to several coexisting species, i.e. community, this approach (Joint Species Distribution Modelling, JSDM) allow to discover the ecological niches partitioning between species (Ovaskainen, Abreg, 2020). The SDM/JSDM framework is very popular in terrestrial ecology (+++) but the amount of papers devoted to SDMs for marine species has also been increasing in recent years (Robinson et al 2017). The range of approaches to building such models is very wide: from regular multiple regressions up to advanced machine learning methods (+++).

SDMs 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 in the case of marine ecology require numerous samples. The increasing evidence for cryptic species (Bastrop et al 1998; Fiser et al 2010) and polytipic species (++++REF) among marine animals makes it urgent to find out approaches to build SDM (more precisely JSDM) for sympatric cryptic species ~~, as well as to infraspecific taxa such as subspecies or phylogenetic lineages within polytypic species~~. Only few studies have been conducted in this direction so far (Lowen et al 2019; Dennis, Hellberg, 2010).

In the marine realm, the longest scientifically known and best studied complex of cryptic species is the blue mussel “*Mytilus edulis*” complex (Knowlton 1993, ???), which includes a number of species that are better distinguished genetically than morphologically and 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 sector, the dominant species are *M. edulis* (thereafter *ME*) and *M. trossulus* (*MT*), which form zones of sympatry (thereafter *contact zones*) in different parts 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 partly 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* being more stenothermic, cold-loving species does not penetrate as far south to the temperate/boreal seas as *ME* (Wenne et al., 2020).

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. In the Baltic Sea, only *MT* inhabits the brackish areas of the inner Sea parts, but the more saline areas closer to the North Sea inhabited by *ME* only. In the contact zone hybrids usually predominate in mixed settlements and the frequency of *MT* genes 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 (++++). However in the Kola region and in West Atlantic zones the situation differs from that in the Baltic. Hybridization is limited with less than 20% of hybrids 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, +++++ ). The relationship between distribution of species and salinity is nowhere obvious (RC 05, Katolikova et al. 2016; Marchenko et al. 2023), but a number of other factors of ecological segregation have been proposed.

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 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. 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 (++++), 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 clear pattern of species distribution along 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. 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. 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 (but without direct reference to this) 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 a discribed strip (Katolikova et al., 2016), and using the regression models the frequencies of “striped” morphotypes in samples can be recalculated into proportion of *ME* and *MT* (thereafter *ME:MT* ratio; see equations in Khaitov et al., 2021). Note that hybrids are not considered as separate category under this approach.

In this study, we used the SDM approach to simultaneously assess the effects of substrate, salinity, proximity to harbors, and surf on the *ME:MT* ratio in the Kandalaksha Bay of the White Sea. All these factors have been proposed in previous studies as determinants of species’ ecological niche partitioning in sympatry. The last of the previously discussed factors, depth, was not studied by us, but was controlled by sampling at the same littoral level. To assess the predictive value of the obtained model, we also analyzed independent testing data sets from the same Kola contact zone.

## 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 (Filatov et al., 2005). 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 of the rivers (++++). 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 (Fig 1 a, b).

Mussels are ubiquitous and abundant on the littoral, particularly in the fucoid algal belt, consisting mainly of *Fucus vesiculosus* and *Ascophyllum nodosum* (Berger et al. 2001).

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.

### Mussel sampling and processing

Between 2011 and 2018, mussels were sampled at 95 sites within the littoral fucoid belt. 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 (T- and E-morphotype, i.e. mussels with and without of a continuous prismatic strip under the ligament, respecively) were identified as in Khaitov et al. (2021). For further, the frequencies of morphotypes in each sample was converted to the proportion of *MT* (*Ptros*) using the *Eq 1* (Khaitov et al., 2021) describing the association of *Ptros* and proportion of mussel with T-morphotype (*PT*). The value of *Ptros* was calculated for each sample and additionally for each site in total (thereafter *PtrosSite*).

### Environmental parameters assessment

In total, we used seven parameters describing possible influence of rivers, ports, surf and type of substrate on mussels (Table ++). The measures of river influence was conducted by three ways. We directly measured salinity in the time of sampling with accuracy of 1 ppm using an “Atago S/Mill-E” refractometer. However, believing that a single estimate of salinity at low tide may not be sufficient to characterize salinity per se, and more generally the extent of estuarine conditions at the sampling sites, we measured the distance to the nearest river and classified neares rivers by its size on the basis of their catchment areas (ESM STable +++). To calculate Fetch as a proxy of surf exposure, the R-package “windfetch” (Seers, 2022) was used.

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

Additionally we included into analysis two other data arrays which were used as testing datasets. The first one consisted of samples from 23 sites collected in the Kandalaksha bay (STable ++, Fig. 1 B). These sites were sampled using a different protocol than in the case of main dataset (i.e. only Bottom or only Algae samples were collected per each site). All predictor’s values and *Ptros* for these samples were assessed as for main dataset.

The second testing dataset was obtained from published data on mussel populations in the Tyuva inlet described in 2009 and 2010 (Marchenko et al. 2023, supplementary electronic materials, 25 samples in 2009 and 19 in 2010). This inlet is situated in the Barents Sea where another *MT-ME* contact zone is presented. We used *Ptros* values provided by authors in supplementary materials. 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 using the geographic map. The nearest port is an active one and is located in the Kola bay (69.19 N, 33.496 E) about 3 km from the mouth of Tyuva inlet. Since the taxonomic structure in the Tyuva inlet is unstable in time (Marchenko et al. 2023) we considered the 2009 and 2010 samples separetely.

### Statistical analysis

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

Assuming that the relationship between the dependent variable (*Ptros*) and 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 was used. 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 build 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.

For a more deep analysis of association between *Ptros* and substrate type we calculated the difference (*Dif*) between *Ptros* on Algae and Bottom substrates 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.

To assess ability of *Model 1* to predict the *MT-ME* distribution we used three data sets: (1) main training dataset (the data used for *Model 1* construction), (2) testing dataset from the Kandalaksha bay and (3) testing dataset from the Tyuva bay.

To assess the model’s predictive power all sites included in the analysis were divided into two groups *MT*-dominated and *ME*-dominated using *Ptros* = 0.5 as a boundary between them. Using the parameters of *Model 1* we calculated predictions of the model for each of three 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 = ++).

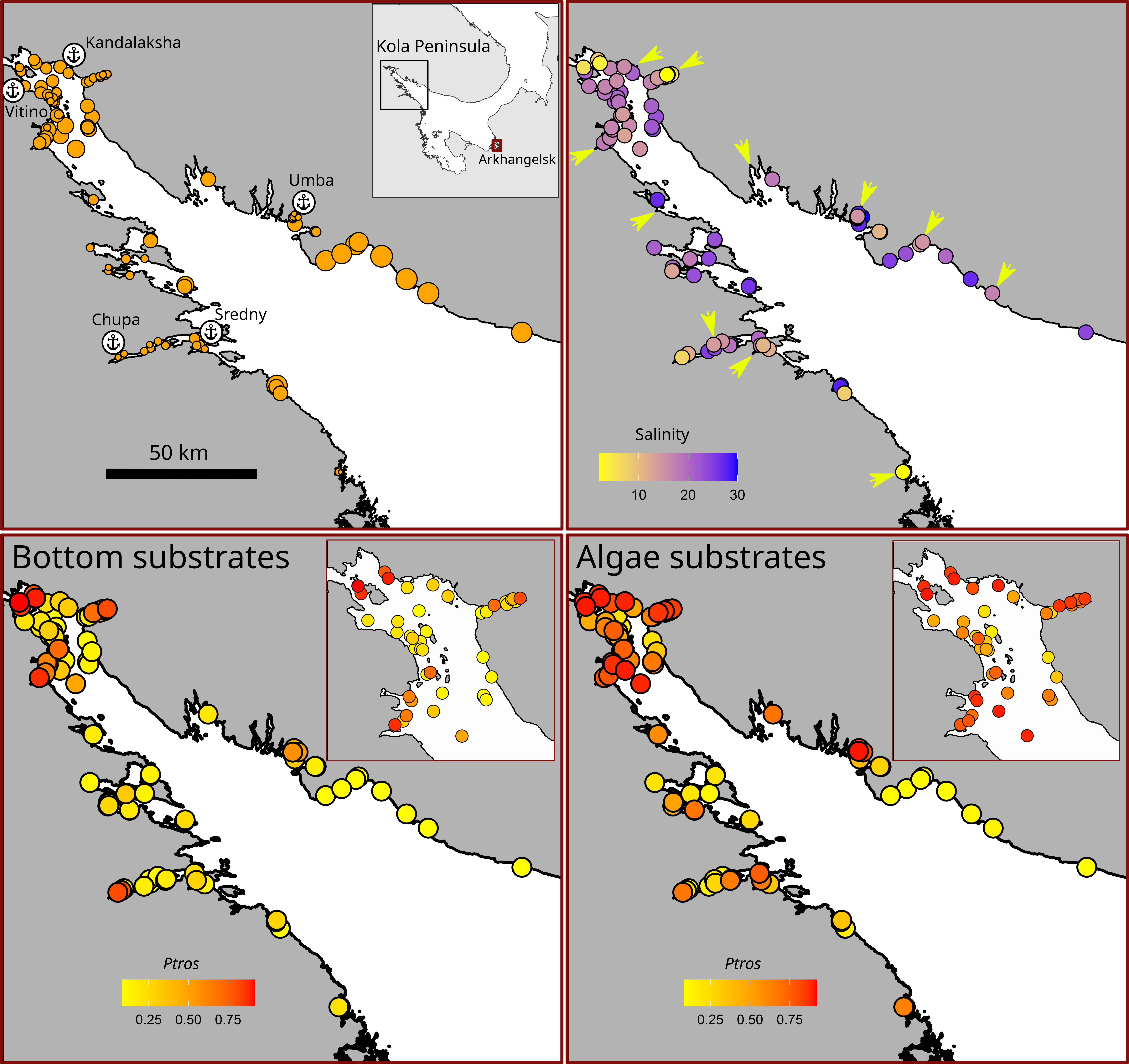


Figure +. Characteristics of sampling sites in the Kandalaksha bay. (A) The surf condition in the sampling sites: point size is proportional to Fetch value. Anchor signs mark position of harbours. Abandoned ports are marked by asterisks. (B) Salinity in the sampling sites. Arrows mark large river’s mouths. (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 1, 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 equal to 1, reflecting the linear dependence of *PT* 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 |

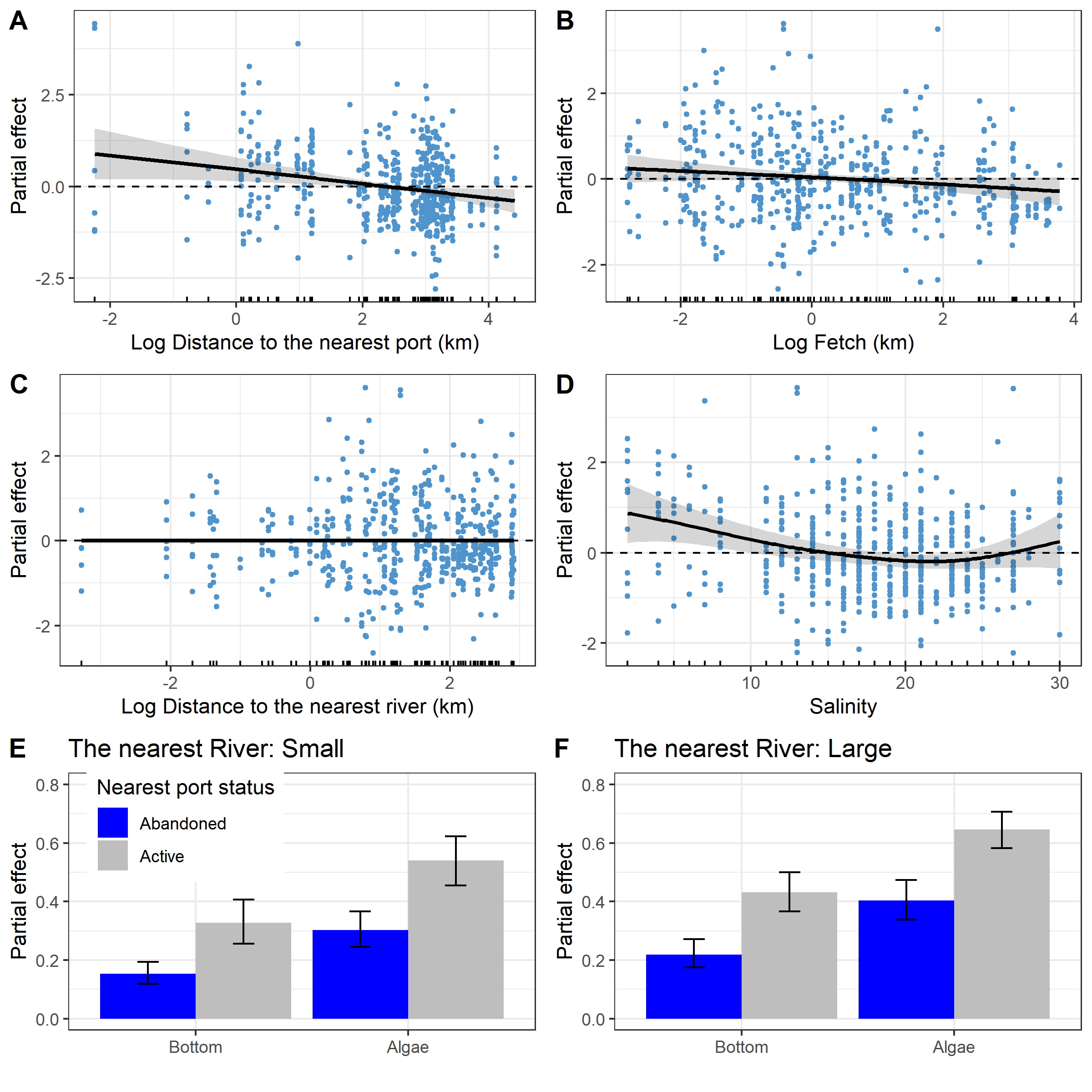


Fig. 3. Predictions of the *Model 1* describing the dependency of *Ptros* on distance to the nearest port (A), fetch (B), distance to the nearest river (C), salinity (E) and substrate type, nearest port status and size of the largest river (E, F). The gray ribbons and wiskers demonstrate 95% confidence interval. The dotted horizontal lines in panels A-D represent the zero smoother’s value. 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

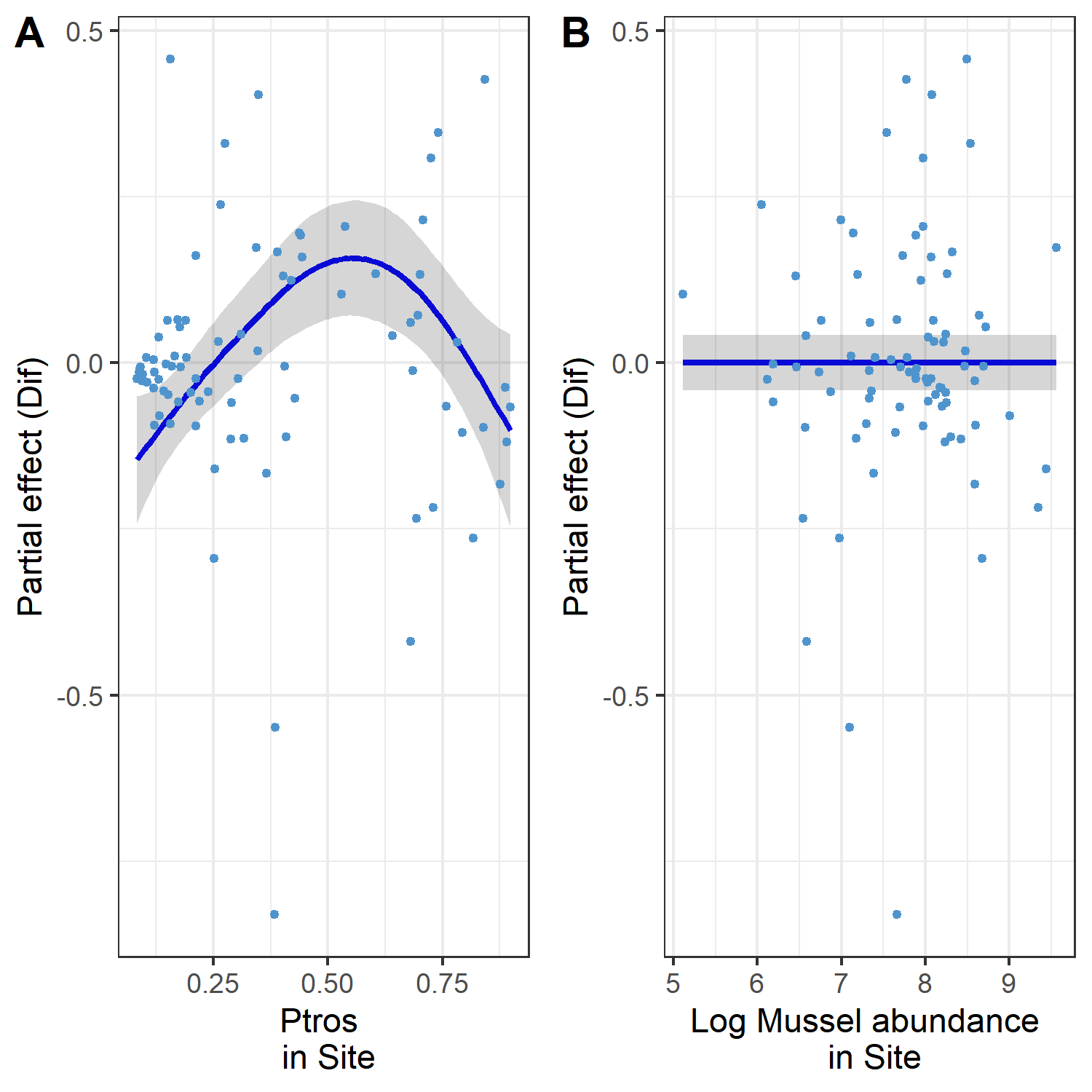


Fig. 4. The dependence of difference between *PT* on Algae and Bottom subatrates on PtrosSite (A) and total mussel abundance (B).

The difference between *Ptros* on Algae and Bottom substrates was significantly dependent on *PtrosSite* in site, i.e. prevalence of *MT* (STable +, Fig. 4 A). The dependency revealed was curvelinear (edf = 2.7, STable +) with minimal *Dif* values close to zero 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

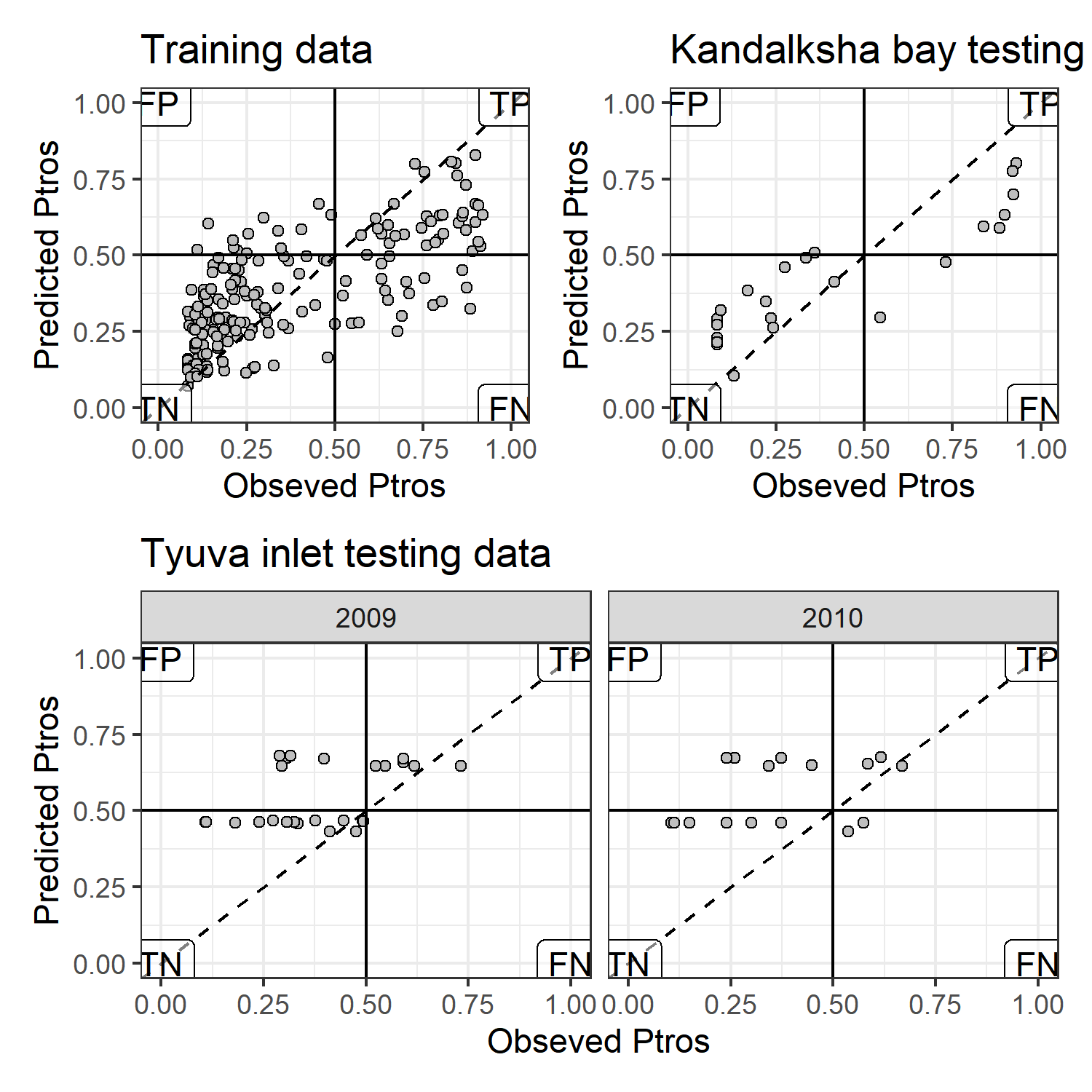


Fig. 5. Comparison of *Ptros*-values predicted by *Model 1* and obsreved *Ptros*. The horizontal and vertical solid lines reflects the boundary between *MT*- and *ME*-dominated sites. Dotted lines represent the position of perfect fit when observed and predicted values are equal. Labels in the plot corners 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) were in rather good agreement for training dataset (AUC = 0.87). In the case of testing dataset from the Kandalaksha bay (Fig. 1 B) the most sites were correctly classified by the *Model 1* (AUC = 0.92) indicating high predictive ability of the *Model 1*. This model being applied to testing dataset from the Tyuva inlet in 2009 display lesser but still rather high predictive power (AUC = 0.76). Only five samples were classified as false positive and no one were classified as false negative. The lowest predictive power of the *Model 1* was revealed in analysis of Tyuva testing dataset from 2010 (AUC = 0.65). Five samples were classified as false positive and two as false negative.

–>

–>

–>

## 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)

–>

–>