Mixed populations of *Mytilus edulis* and *M.trossulus*: towards to species distribution modelling

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

## Introduction

One of the urgent tasks of the modern marine ecology is the construction of relevant species distribution models (SDM) as numerical tools describing the relationship between species occurrence and environmental parameters (Elith, Leathwick, 2009; ++++). These models allow to predict distribution patterns of ecologically and economically important species (+++) and serve as a basis for prognosis of its distribution in future (+++). The amount of papers devoted to SDMs for marine species has been growing significantly in recent years (Robinson et al 2017).

Although the biological unit for SDM can be either species, genetically distinct subpopulations, or even genetic lineages (Lowen et al 2019), however, in marine ecology SDMs are currently developed for “good” species whose taxonomic status is not in doubt (e.g. Reiss et al., 2011; Lindegren et al., 2022). The taxonomic certainty gives an opportunity to focus on solving actual problems related to the search for modeling algorithms (+++) and/or identification of the most significant environmental predictors in marine realm (+++). However, the increasing evidence for cryptic species among marine animals (+++) makes it urgent to build SDMs in such cases as well. Only few papers have directly addressed the construction of SDMs for marine animals’ within species genetic lineages and cryptic species (e.g. Dennis, Hellberg, 2010; Lowen et al 2019).

The construction of SDM for cryptic species has its own peculiarities. Dealing with cryptic species forming parapatric or sympatric populations both genetic processes (e.g., hybridization) and ecological interactions between closely related species potentially can influence the result of analysis and consequently model prediction quality. Strictly speaking, in the case when SDM is built not for a single species, but for several coexisting ones, the Joint Species Distribution Models (JSDM) framework should be considered (Ovaskainen, Abreg, 2020). Whereas the standard SDM approach assume to use the presence/absence (or abundance) of a species as a response variable, which is under control of numerous environmental predictors, in JSDM framework some correlative assessment of interspecific interactions should be incorporated into the model as well (Ovaskainen, Abreg, 2020).

Blue mussels from “*Mytilus edulis*” complex (+++) provide a good opportunity to work out the approaches to SDM construction in the case of cryptic species. For example, native ranges of atlantic *M. edulis* (thereafter *ME*) and pacific *M. trossulus* (*MT*) are differ (Riginos, Cunningham, 2005), but at North Atlantic coasts they form mixed settlements in several contact zones (Varvio et al., 1988; Hilbish et al., 2000; Mathiesen et al. 2016; Brooks, Farmen, 2013; Katolikova et al. 2016; Vainola, Strelkov 2011; Wenne et al. 2020). In general, it is postulated that the proportion of *MT:ME* in a contact zone change along certain environmental gradients (Väinölä and Hvilsom 1991; Riginos and Cunningham 2005; Väinölä and Strelkov 2011). However, some cases of mosaic distribution of *MT* and *ME* have also been described, when spots of higher concentrations of one or other species are present within the region (Moreau et al., 2006; Katolikova et al., 2016; Wenne et al. 2020). In the area of contact zone the proportion of *MT:ME* can vary in broad limits even in small spatial scales (Katolikova et al. 2016) that allows us to consider the *MT:ME* ratio as a very sensitive parameter responding both to the influence of environmental factors and/or interactions between *MT* and *ME*. Thus the use of *MT:ME* ratio as a dependent variable in the model construction brings it closer to the JSDM framework.

There is no undoubt position bout factors regulating *MT* and *ME* distribution. On a biogeographic scale, the distribution of *Mytilus* species is thought to be regulated by regional temperature and its correlates (Popovic & Riginos, 2019; Wenne et al., 2020). *MT* is believed to be a more northerly species, more tolerant to cold waters inhabiting areas that once existed in glaciation zones. *ME* potentially can distribute in cold waters but tend to form settlements in more temperate latitudes (Wenne et al., 2020).

Salinity was often considered as a key factor regulating the spatial distribution of species *ME*-*MT* contact zones. *MT* are thought to gravitate toward areas of more freshened water, and *ME* toward areas with higher salinity (Riginos & Cunningham, 2005; Ridgway & Naewdal, 2004; Kijewski et al., 2019; Ridgway & Naevdal, 2004). For example, *MT* form virtually single-species settlements in the heavily freshened Baltic Sea (Gosling, 1992; Vainola Strelkov, 2011) but *ME* proportion dramatically increase when crossing the Danish straits toward the North Sea where oceanic salinity occurs (Kijewski et al., 2019). In other contact areas (Gulf of St. Lawrence, Mareau et al. 2005, Bergen area, Ridgway & Naevdal, 2004) a similar pattern is observed: *MT* settlements are represented in the most desalinated areas. Wave action was also considered as an important factor regulated species distribution.

*MT* has been shown to gravitate toward more turbulent habitats and *ME* toward quieter areas (Tam, Scrosati, 2014; Comesaña et al., 1999; Bates, Innes, 1995). The pattern of species segregation by substrate has also been attempted to be linked to surf exposure (Katolikova et al, 2016): *MT* tended toward fucoid thalli, and *ME* toward the surface of bottom stony substrates. The authors (Katolikova et al, 2016) suggested that fucoids act as shock absorbers, mitigating the impact of surf damaging *MT* which possess thinner and more fragile shells (Beaumont et al, 2008; Michalek et al, 2021). Importantly, in the last case, an inverse relationship is assumed: *MT* avoids surf-exposed areas. However, in some studies, when the material collection design was specifically adapted to study of the surf effects, the role of this factor in the regulation of the *MT:ME* ratio has not been proven (Moreau et al., 2005).

Another approach to explanation of patterns in the spatial distribution of *MT* and *ME* can be related to the history of formation of the contact zone. For example, it is supposed that in the Barents and White Seas, where originally only *ME* is thought to have been present, the *MT* settlements appeared as a result of the introduction of this species from the coast of North America by ships performing transatlantic traffic during the World War II (Vainola, Strelkov, 2011). According to this the populations of *MT* is expected to be observed in the ports’ areas (Vainola and Strelkov, 2011; Katolikova et al. 2016), where this species was originally introduced.

We can see a large spectrum of factors (salinity, wave action, proximity to harbor, and substrate type) which were considered as possible agents regulating the distribution of *ME* and *MT*. Moreover, some correlations that seem obvious may have a different interpretation. For example, the relationship of high abundance of *MT* with reduced salinity can be interpreted by another way: the most desalinated areas (river mouths) are often also the quietest areas with minimal wave impact. Port structures, similarly, are often represented in the most sheltered and frequently desalinated areas. Thus, there is no strict evidence for the exclusive role of some leading factor in the regulation of species distribution and consequently no reasons for use it as singular predictor in SDM construction.

The only way to solve this problem is to include all predictors potentially affecting the composition of mixed settlements in one model which is the main methodology in species distribution modelling (+++). Such approach to analysis of of *Mytilus* species distribution have been attempted before (but without direct reference to SDM framework, see Kijewski et al., 2019; Wenne et al., 2020) and allowed to show a clear relationship of the *MT:ME* ratio primarily with salinity and temperature gradients. However, the populations involved in the works cited were scattered over a very wide geographic range, with distances between them reaching hundreds and thousands of kilometers. On the other hand, the gradients of the key factors regulating species distribution can vary on much smaller spatial scales. For example, salinity, as a prime candidate regulating the *MT:ME* ratio, can vary significantly on scales of a few kilometers in river estuaries (Van Damme, 2005). If *MT:ME* ratio is regulated by salinity does this mean that the ratio will show the same pattern at smaller geographic scales as well? In addition, will salinity play a leading role in species distribution amid the inclusion of other predictors in the model as well?

In this paper, we will consider the case where *ME* and *MT* come into contact forming mixed populations in the Kandalaksha Bay of the White Sea (Vainola, Strelkov, 2011; Katolikova et al. 2016; Khaitov et al, 2021). This contact zone found in the relatively small area is extremely convenient for investigation of environmental factors’ role in the regulation of the *MT:ME* ratio and therefore SDM construction.The salinity gradients are well expressed in this area (Filatov et al., 2005) which is determined by the inflow of several large rivers into the sea (Katolikova et al., 2016). Further, the coast of the gulf possesses numerous sheltered areas (inlets and skerries) and areas open to the surf which provide the opportunity to assess the wave impact gradient. There are several active and several abandoned ports (Katolikova et al., 2016) that give an opportunity to trace the role of vessel trafic in species distribution. And finally important is that the small size of the Kandalaksha Bay (Fig. ++) ensures that there are no clear temperature gradients in both warm and cold seasons. Therefore, the influence of temperature, which probably regulates the distribution of *Mytilus* species on a broader geographic scale (Wenne et al. 2020) can be excluded from the discussion. We will consider SDM for *MT* and *ME* coexisting in this contact zone and try to assess the contribution of each predictor included in forming of species distribution pattern.

## Material and methods

### Study area

The 185 km long Kandalaksha Bay in the north of the White Sea is oriented from southeast to northwest and is funnel-shaped (Fig. ++). Climate is continental subarctic with cold winters (4-5 months of ice cover) and relatively warm summers (August SST is 13.8°C) (Filatov et al., 2005). The bay contains hundreds of banks, skerries and islands. The bay’s shoreline is fringed by numerous smaller inlets. Because of this shoreline geometry, different areas even situated close each to other are exposed to wave action in varying degrees. The summer surface salinity is 24 ppm in most of the Bay (“normal” salinity of the White Sea is 24-30 ppm), but lower in the estuary areas (Filatov et al., 2005). A total of 24 rivers and two canals of hydropower plants (ESM STable ++ ) with catchment area of 141 - 12830 km2 (average 240 km2) discharge fresh water into the bay creating local salinity gradients. The top of the Kandalaksha Bay is the most desalinated, due to the influence of the largest river Niva (Katolikova et al., 2016).

Historically (through most of the 20th century), several ports (Fig. ++) were functioning in the area operating high tonnage oceanic vessels (Sailing directions of the White Sea, 1932; Krasavtsev, 2011). Two of these ports, Kandalaksha and Vitino, both in the top of the Bay, are currently functioning while the remaining ports have not been used to serve ocean-going ships in recent decades, but are still visited by coaster ships, according to the conventional knowledge.

Mussels are ubiquitous and abundant along the coasts of the Kandalaksha Bay. On the littoral, they are most abundant at local mussel beds (Khaitov, 2013) and within the fucoid belt formed mainly by *Fucus vesiculosus* and *Ascophyllum nodosum* (Naumov, 2001)

### Field collections

When sampling, we tried to describe littoral populations of mussels in the Kandalaksha Bay in as much detail as possible, also taking into account the heterogeneity of their habitats by salinity, wave exposure and proximity to ports. The material was collected on the littoral of the mainland and the islands of the Kandalaksha Bay (Fig. ++). Collections were carried out in July-August in 2011-2018 at 95 sites.

At each site we collected mussels from two substrates: fucoid’s thallus (thereafter “Algae”) and bottom surface (i.e. stones, pebbles, sand, mud; thereafter “Bottom”). Algae samples were taken as individual bundles of fucoids large enough to contain at least a few dozens of mussels (however, this condition proved to be unfeasible for several samples from sparse populations). Bottom samples were taken using a 0.025 m2 square frame placed on a ground in such a way that at least a few dozens of mussels could be sampled. Three samples (replications) of each type were taken in each site no more than 20 meters apart. Geographic coordinates were determined for each site (i.e., the same coordinates were assigned to each sample within the same site). Each sample was placed in a plastic bag and transported to the laboratory.

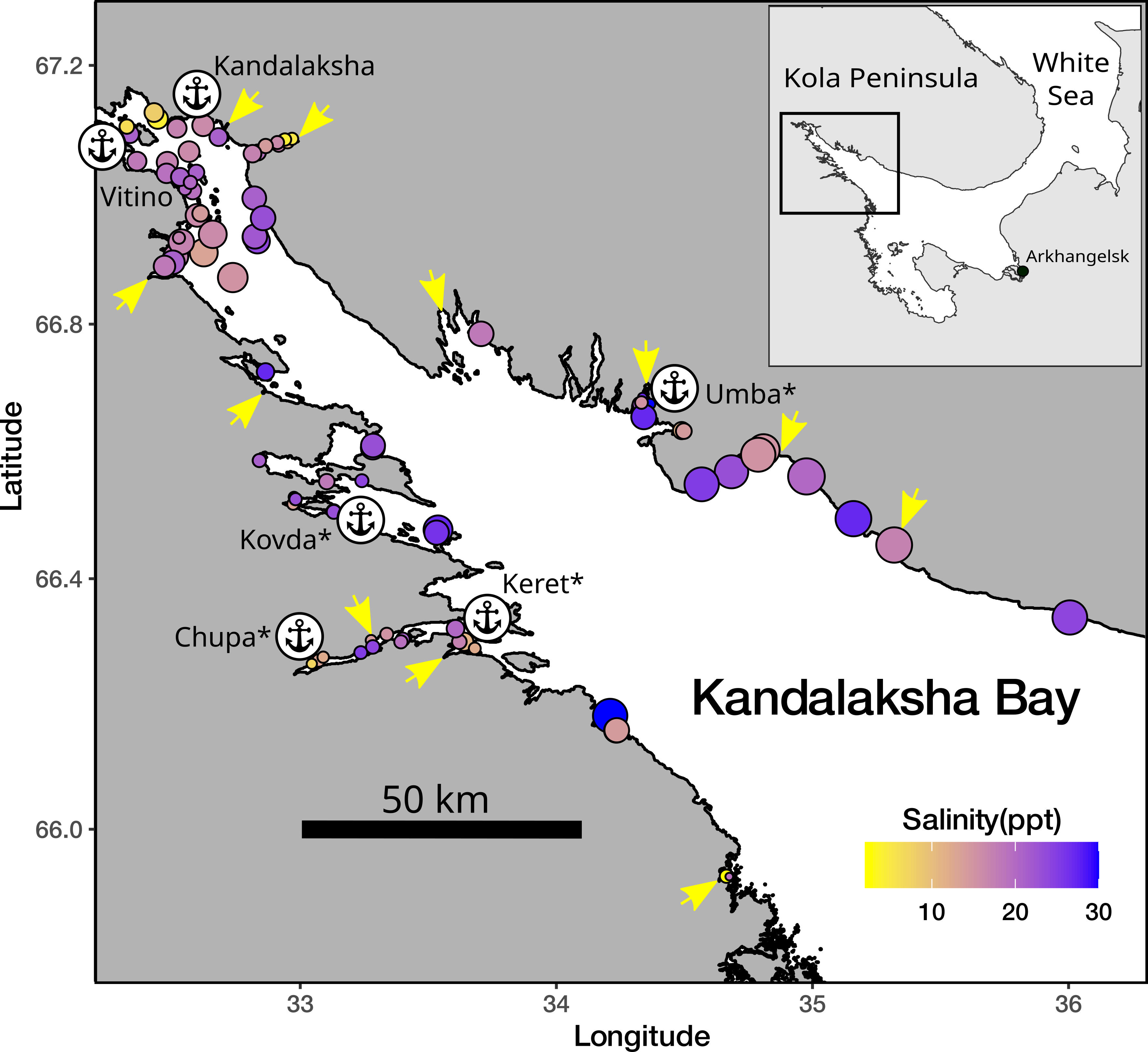


Fig. 1 Sampling sites. The size of points is proportional to average wind fetch. The fill of points represent salinity. Arrows indicate the mouths of the largest rivers. Signs of anchors indicate harbors name of which is marked by asterics if a port is abandoned.

Mussels with a shell length more than 10 mm were cleaned from tissues and used in further analyses. Shell morphotypes (E-morphotype, characteristic to *ME*, and T-morphotype, characteristic to *MT*, see Katolikova et al. 2016) were identified as in Khaitov et al. 2021. In total we processed 5.5516^{4} specimens. The mean number of mussels collected per sample was 97(range 2, - 785).

### Environmental parameters assessment

This work was aimed to examine the effects of substrate, salinity, surf level and proximity to ports. Some of these parameters were measured directly but for others several proxies were considered.

The nature of the substrate was characterized by where the corresponding sample was taken from (*Algae* vs *Bottom*). The degree of proximity to port was estimated as the straight line distance (not accounting shoreline bends) from the sampling site to the nearest of the 6 large ports presented in the area (Fig++, STable ++). Additionally each port was assigned to one of two categories: *“Active”* if at the time of sampling the port was serving heavy-tonnage vessels plying out of the White Sea, or *“Abandoned”* if the port served only small vessels cruising within the White Sea, or was completely abandoned at the time of sampling.

*Wind fetch as a proxy for wave impact*. When selecting sampling sites, we assumed that collections should characterize both surf and sheltered habitats. Each site was visually assigned to one of two categories: *“Exposed”* (sites located on open shores exposed to waves coming from the open part of the bay) or *“Sheltered”* (sites located in bays or skerries). For a more formal assessment, the value of wind fetch was calculated for each site. It was calculated as the unobstructed length of water surface (in km) over which wind from a certain direction can blow over. These values can be obtained from the analysis of geographical maps of the region. To calculate the wind fetch values, the R-package “windfetch” was used (Seers, 2022). The higher fetch values denote the more wave impact could be brought from a certain direction. Wind fetch was measured for each site from all directions (North, East, South and West, STable +) but as a value for the further analysis we used an averaged fetch value.

*Proxies for salinity assessment*. Water samples were taken at each site when mussels were collected, i.e. in the time of the low tide. The salinity was measured using an “Atago S/Mill-E” handheld salinity refractometer (with 1 ppm accuracy). However, salinity measured directly at the sampling site poorly reflects the salinity regime in the site due to its fluctuations during the tidal cycle and due to weather conditions (waving, precipitation *etc.*). Therefore, as an additional estimate of the desalination level in a given site, we used the characteristics of fresh water discharging sources closest to the sampling site (rivers or channels of hydroelectric power plants).

In total, 24 rivers are took place in the region, information about which is present in open sources (STable +). Additionally, two powerful freshwater dischargers (HPP canals) were included in the analysis. For each river, we assessed the catchment basin area and coordinates of its mouth. For HPP canals, this information was obtained by visual analysis of geographical maps. Each of the rivers was assigned to one of two conditional categories: “Large” or “Small” depending on whether its catchment area was more or less than an median catchment area of all rivers included in the analysis (240 km2). To note, this operational classification should not be misleading with classification of rivers size proposed by hydrologists (Khublaryan, 2009).

We calculated the distance (straight line, not accounting shoreline bends) from each sampling site to the mouth of the nearest river. Thus salinity was characterized by three values: directly measured salinity, distance to the nearest river and the nearest river size.

### Statistical analysis

The number of T- and E-morphotype mussels in each sample was counted and proportion of T-morphotype mussels (*PT*) was calculated. Additionally this proportion was assessed for each site (the total number of T-morphotype mussels collected at this site divided by total number specimens collected). The proportion of *MT* in mixed populations in the White Sea is strongly correlated with probability to find *MT* genes in mixed mussels’ populations (i.e. STRUCTURE score, Khaitov et al., 2021) and can be assumed as a proxy for *MT:ME* ratio.

All statistical processing was performed using the statistical programming language R 4.05 (R core Team, 2021). The analysis was conducted in two steps. The first step was to determine which predictors are statistically related to *PT*. The second step was to estimate the importance of each predictor as regulator of high *PT* value in mixed mussel populations.

At the first step, we assumed that the relationship between dependent variable and predictors can be curvelinear that’s why we constructed a generalized additive mixed model (GAMM, +++) with beta-binomial residuals distribution (hereafter *Model 1*). The fixed part of the *Model 1* included predictors as follows:

* *Substrate*: categorical, with two levels “Bottom” vs “Algae”;
* *Salinity*: continuous, the directly measured salinity (ppt);
* *MinDistRiver*: continuous, distance to the nearest river (km);
* *RiverSize*: categorical, with two levels “Small” vs “Large” accordingly to size of the nearest river;
* *MinDistPort*: continuous, the distance to the nearest port (km);
* *PortStatus*: categorical, with two levels “Abandoned” vs “Active”;
* *AverageFetch*: continuous, fetch value (km).

For all categorical predictors, the first mentioned category was used as base level in the analyses. The function gam() from the package “mgcv” (+++) was used to build the GAMM. 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.

To check for the predictors’ collinearaity in the *Model 1* we calculated the variance inflation factor (VIF, Fox & Monette, 1992) for each of them. Additionally we calculated Pearson correlation between continuous predictors.

In the second step, we analyzed the frequency distribution of *PT* values to identify the boundary between *MT*- and *ME*-dominated sites. We hypothesized that these two groups of sites should form separate peaks in the *PT*-distribution, which should have a bimodal shape. In this analysis, we combined all samples from each site and calculated *PT* for each of them (separately for Algae and Bottom). The resulting values were subjected to arcsin transformation (this was necessary to describe the peaks by normal distributions). Using the approach described in Benaglia et al. (2009), we identified two peaks in the *PT*-distribution, which were approximated by separate normal distributions. Two parameters (mean and standard deviation) were fitted for each of the peak using “normalmixEM()” function from the “mixtools” package (Benaglia et al., 2009). The boundary between these distributions was considered as the value separating two statistical populations.

Using this boundary, we assigned samples from each of the sites (separately for Algae and Bottom) to *MT*-dominated (high *PT*) or *ME*-dominated (low *PT*) groups. The former were coded as 1 and the latter as 0. We used this binary variable constructed as dependent one in a mixed logistic regression model (*Model 2*). This model described the dependence of the probability of being in the *MT*-dominated group on the same predictors as in the *Model 1*. Site was considered as a random factor in the *Model 2*. The model was fitted with “glmer()” finction from “lme4” package (Bates et al., 2015).

The fraction of total variability explained by the fixed part of the *Model 2* was estimated using marginal coefficient of determination *R2* (Nakagawa & Schielzeth, 2013). Marginal *R2* can be decomposed into semi-partial coefficients of determination, *semi-partial R2*, that describe the contribution of each individual predictor to the overall coefficient of determination. Using the “partR2” package (Stoffel et al., 2021) we calculated the “inclusive R2” to estimate the total variance explained by each predictors.

To verify that the *Model 1* and *Model 2* meet the assumptions of sampling independence, we examined the presence of residuals’ spatial autocorrelation by means of semivariogram construction (++). Only poor evidences of spatial autocorrelation were revealed for both models.

## Results

## Environmental parameters

Two of six ports presented in the area were active at the date of material collection (Fig. ++, STable ++ SEM). The distance from sampling sites to the nearest ports ranged from 0.11 to 82 km (Median = 18.7 km) (SEM SFig ++, A). Sites visually categorized as “Sheltered” on average were characterized by lower values of wind fetch than those categorized as “Exposed” (ESM SFig. ++, B). 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). Salinity measured at the time of sampling ranged from 2 to 30 ppm (Median = 19 ppm). The most desalinated areas were the top of the Bay and inlets where rivers flow into (Fig. ++1). Salinity at sites situated close to “Large” rivers was lower than at sites closer to “Small” ones (ESM SFig. ++, C). Salinity tended to increase with increasing of distance to the mouth of the nearest river (SEM SFig. ++ D) but no association between distance to the nearest port and salinity was found (SEM SFig. +++ E). Sites situated close to ports tended to have lower wind fetch (SEM SFig. +++ F). All correlations between salinity, distance to the nearest river mouth, distance to the nearest port and wind fetch were rather low (SEM STable +++). The largest correlation revealed was between wind fetch and distance to nearest port (r = 0.46).

## *PT* spatial distribution

The maximum proportion of the T-morphotype was observed in the top of the Kandalaksha Bay itself and in the small bays deeping into the mainland (Fig. ++). At the same time, sites with a lower proportion of the T-morphotype were located in the immediate vicinity of sites where settlements with high *PT* values were represented (Fig. ++), reflecting the high level of patchiness in PT spatial distribution. Settlements with the lowest frequency of T-morphotype were represented in the open areas of the Northwest Coast.

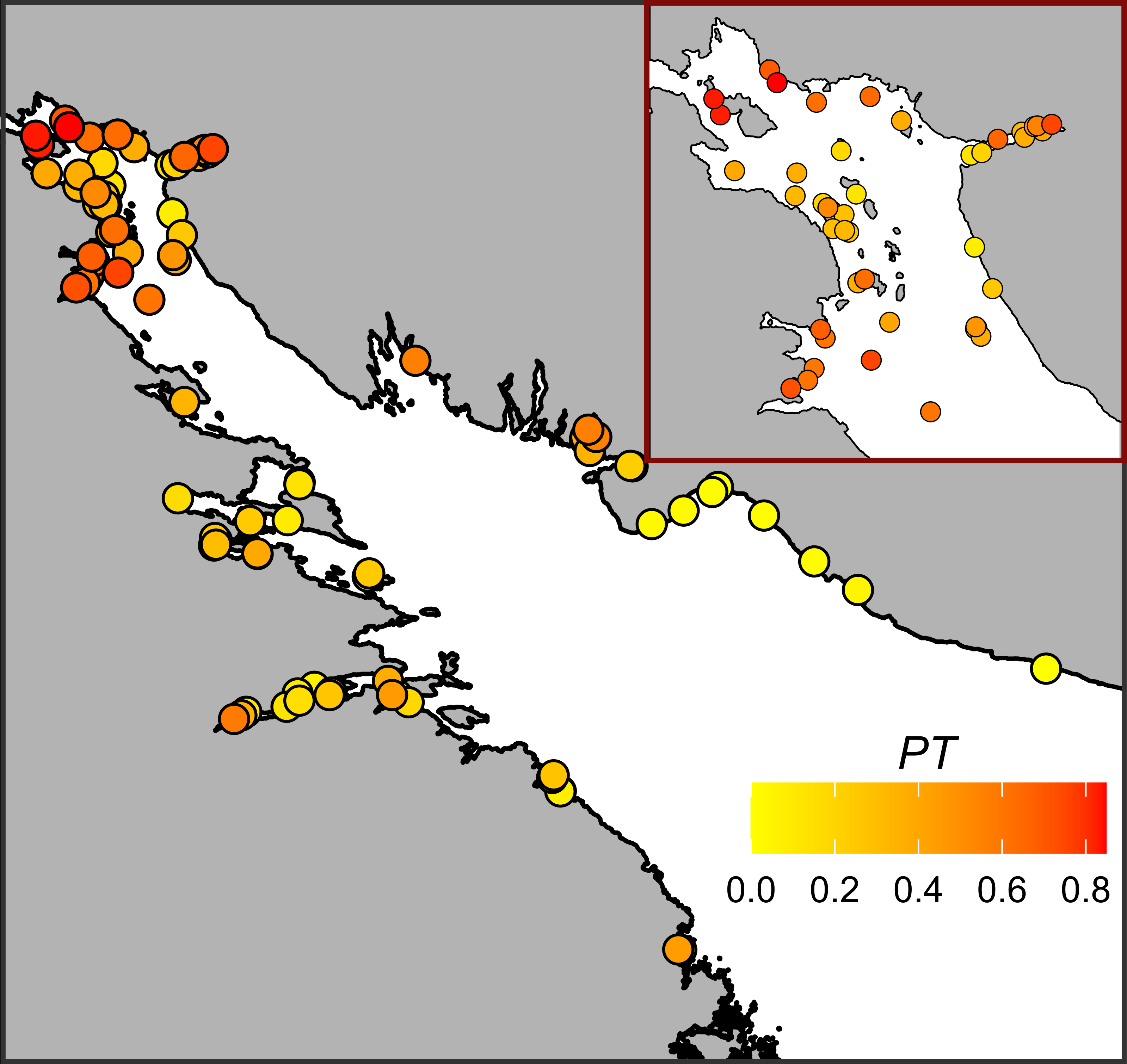


Fig. + Proportion of mussels with T-morphotype on sampling sites. The fill of points represents *PT* value. Insertated plot represents the detailed *PT* distribution in the top of Kandalakshs bay.

## Relationship of *PT* 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 Min\_dist\_river: 1.58). 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 *PT* on all predictors considered we constructed the *Model 1* (Table ++, R2 = 0.65) which explained 85.2% of total deviance. The *Model 1* revealed statistically significant dependency of *PT* on all predictors except distance to the nearest river (*MinDistRiver*).

The effective degrees of freedom (*edf*, Table ++, A) for most smoothers were equal to 1, reflecting the linear dependence of *PT* on such continous predictors as distance to the nearest port (*MinDistPort*) and fetch (*AverageFetch*). The exception was *Salinity*, where a pronounced curvilinear dependence was revealed.

Table + A. Parameters of smoothers for the Model describing dependency of PT on predictors.

| Model term | edf | ref.edf | Chi.sq | p-value |
| --- | --- | --- | --- | --- |
| s(Salinity) | 5.6 | 5.7 | 27.9 | 0.0002 |
| s(MinDistRiver) | 1.0 | 1.0 | 2.4 | 0.1224 |
| s(AverageFetch) | 1.0 | 1.0 | 12.2 | 0.0005 |
| s(MinDistPort) | 1.0 | 1.0 | 5.0 | 0.0249 |
| Random effect s(Site) | 72.0 | 88.0 | 485.0 | 0.0000 |

Table + B. Parametric terms for the Model describing dependency of PT on predictors.

| Model term | Parameter estimate | SE | z-statistic | p-value |
| --- | --- | --- | --- | --- |
| (Intercept) | -2.4 | 0.2 | -11.5 | 0.0000 |
| Substrate(Algae) | 1.1 | 0.1 | 13.7 | 0.0000 |
| RiverSize(Large) | 0.5 | 0.2 | 2.2 | 0.0293 |
| PortStatus(Active) | 0.9 | 0.3 | 3.2 | 0.0012 |

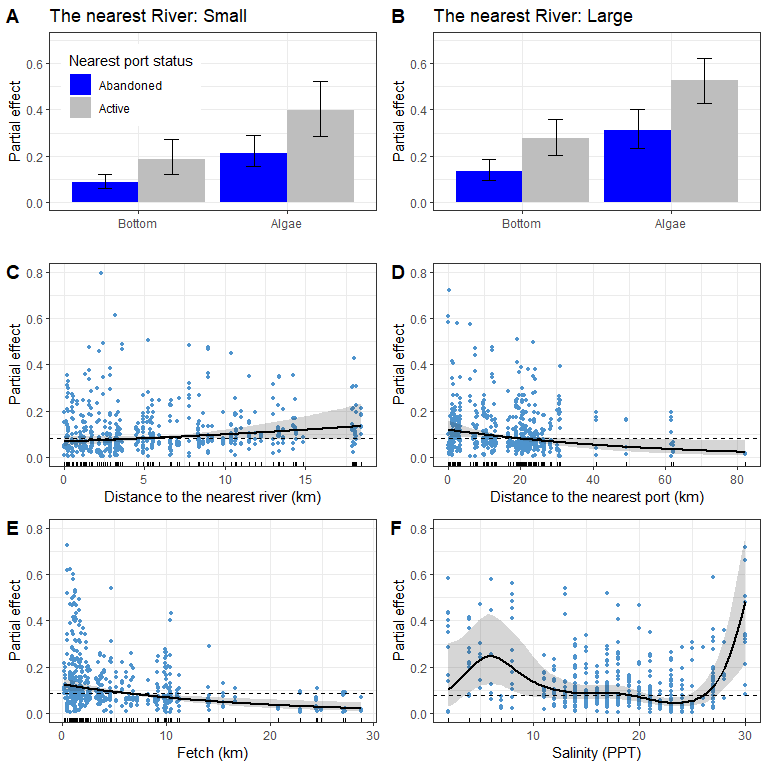


Fig. ++. Predictions of the Model describing dependency of PT on substrate type, nearest port status and size of the largest river (A, B), distance to the nearest river (C), distance to the nearest port (D), average fetch (E) and salinity (E). The wiskers and gray ribbons demonstrate 95% confidence interval. The dashed lines at C-E panels reflect median partial effect value. Points on panels C-F reflect partial residuals but not raw data.

According to the *Model 1* (Table ++, Fig. ++) the value of *PT* is dramatically higher on the Algae substrate than on the Bottom one. This value is higher in those sites which is close to active ports than to abandoned ones. Finally, the *PT* value is slightly higher at sites situated close to large rivers than to small ones. The *PT* has a tendency to decrease with increasing of distance to the nearest port (Fig. ++, D) starting to be in average lower than median partial effect when distance to the nearest port is higher than 20 km. The dependency of *PT* on average fetch (Fig. ++, E) is also negative: *PT* become lower than median partial effect when value of fetch started to be higher of 5 km.

The curvilinear dependence of *PT* on salinity (Fig. ++, F) can be described as follows. The predicted value of *PT* is close to the median partial effect at intermediate salinity range. However, at extremely low (less than 12 ppt) and high (more than 27 ppt) salinity, *PT* becomes higher than the median predicted value. Extreme low salinity was observed at only 15 sites and extreme high salinity at 8 sites. For other 72 sites with intermediate salinity predicted *PT* is close to the median value (Fig. +++, F).

## Assessment of predictors’ importance

## number of iterations= 102

The values of *PT* ranged from 0 to 0.89 at different sites. The frequency distribution of *PT* values displayed a clear bimodal structure both in the case of “Bottom” and “Algae” samples (Fig. ++). Two peaks were recognized in total distribution as well (Fig. ++ lower pannel). Two Gaussian components were fitted into the total *PT* distribution with parameters as follow: = 110.7, = 46 and = 16.9, = 23.53. The boundary between these peaks was 87.7(*PT*=0.48). Thus two types of populations can be recognized in the studied area: populations dominated by *ME* (conditional boundary *PT* < 0.48) and samples with MT prevalence (*PT* > 0.48). In samples from the “Bottom” substrate, samples belonging to the first group were more frequent (the left peak in the distribution predominated). In samples from the “Algae” substrate, both groups were represented in more or less equal proportions.

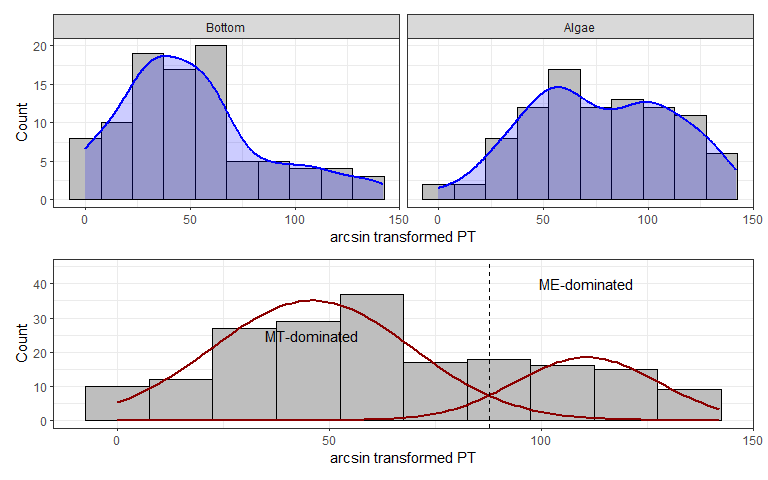


Fig ++. The distributon of arcsin transformed *PT* values at different substrates (upper panels) and in total (lower pannel). Blue lines represent the kernell density smoothing the frequency distributions. Black curves on the lower pannel represent normal distributions fitted, dotted line represent boundary between peaks.

Parameters of the *Model 2* are presented in STable ++. The marginal coefficient of determination of the *Model 2* (R2 = 0.51) was decomposed as follow (Table. +). The highest proportion of total variance explained was associated with *Salinity*. However the influence of other three predictors (*PortStatus*, *RiverSize* and *Substrate*) were close to the former. The minimal influence was determined for *MinDistRiver*. The influence of *AverageFetch* and *MinDistPort* was intermediate.

Table +. Proportion of variation explined by different predictors included into Model 2.

| Predictor | Proportion of variaton explined (%) |
| --- | --- |
| Salinity | 16.2 |
| PortStatus | 15.5 |
| RiverSize | 15.5 |
| Substrate | 11.5 |
| AverageFetch | 9.2 |
| MinDistPort | 6.9 |
| MinDistRiver | 1.2 |

## Discussion

Using the “morphotype test” (Khaitov et al. 2021) as a tool for species identification, we were able to construct a *Model 1* describing the relationship between *PT* frequency and all the factors that are generally considered to be the main regulators of the *MT:ME* ratio in mixed mussel settlements. Our analysis showed that the predictors included in the model, although weakly correlated with each other, do not exhibit high collinearity. This means that we can consider all factors involved as independent entities that do not mask each other in their influence.

We found that all factors, which were considered as crucial for *MT:ME* proportion regulation in previous works, were significant for *PT* value (and consequently proportion of *MT* in mixed populations) in a rlatively small contact zone of the Kandalaksha Bay. This means that the patterns revealed on more extensive contact zones with longer ecological gradients retain their influence in essentially smaller spatial scales.

The proportion of *MT* is clearly governed by three factors in the Kandalaksha Bay: the nature of the substrate, the proximity to active port systems, and the degree of coast exposure to wave action. According to the data we obtained, the highest *MT* concentration should be expected on the fuchcoid tally, in sheltered areas close to active ports.

It is believed that *MT* was introduced into the Barents and 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 Barents and White Seas (Vainola & Strelkov, 2011; Katolikova et al. 2016). Our study confirmed this hypothesis. The proportion of T-morphotype (and hence *MT*) in the studied area increased with approaching to ports. 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 *PT* value in such settlements is higher than in the settlements adjacent to the abandoned harbors. This can be explained in several ways.

It is possible that new *MT* are constantly brought to the ports of the Kandalaksha Bay from the Barents Sea (primarily from the ports of Murmansk in the Kola Bay, where the abundance of *MT* is very high, Vainola&Strelrov, 2013; Marchenko et al., 2023). In the abandoned ports this constant renewation is absent and during the time of port inactivity *MT* have been displaced by *ME*.

On the other hand, active ports tend to have a higher degree of anthropogenic pollution, leading to less stress tolerance 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). However, studies of the immune systems of *MT* and *ME* in the Canadian contact zone suggest a greater suppression of the *MT* immune system in areas prone to pollution (Coray et al., 2006), making them more sensitive to stress than *ME*. The last fact contradicts the hypothesis of better performance of *MT* to pollutant influence allowing them to set populations in active port harbors.

A third explanation is also possible. As it was shown the Bivalvian Transmissible Neoplasia (BTN) is more frequently distributed in the port area (+++). Taking into account the fatality of BTN for mussels (Benabdelmouna &Ledu, 2016; Burioli et al 2019) and asymmetric resistance of mussel species to BTN infection (Yonemitsu et al., 2019, проверить) we can expect that BTN can play certain role in the regulation of *MT:ME* proportion in the harbor area.

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* populations since the closeness to the port harbors was never included into species distribution models developed for other areas. In some cases it is probably an unrealistic aim due to absence of predictor variability. For example in the case of *MT* 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 map of *MT* distribution (Vainola & Strelkov, 2011), 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. The unexpected increase in the number of *MT* in aquaculture plantations in the Scottish fjords (Beaumont 200+) is rather consistent with the hypothesis of the recent invasion as well.

Our analysis revealed that the higher wind fetch is associated with lower *PT* value. That means *MT* avoid open coasts potentially exposed to the surf. The lowest *PT* value (i.e., E-morphotype mussels dominate in the settlements) is observed on open shores of the north-east coast of Kandalaksha Bay (Fig. ++), 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). 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*.

Finally the nature of the substrate is the powerful predictor that has an influence on *PT* variation (Table ++, Figure ++). The proportion of T-morphotype mussels appears to be significantly higher on fucoids 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. It can be assumed that the effect of substrate type on *PT* has the same mechanism that underlies the negative correlation of *PT* with wind fetch. Fucoids may work as shock absorbers, mitigating, to a certain extent, the impact of the surf. Another explanation could also be given. 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 in the thallus of the fucoids. This hypothesis is contradicted by the fact that a comparison of attachment strength in *MT* and *ME* from the contact zone in Canada, showed no significant differences between the species when they were kept under laboratory conditions (Lowen et al., 200+). Additional investigations of attachment strength of both species are needed.

An intriguing result of our study was the absence of clear linear signal of the influence of salinity on the *MT:ME* ratio. GAM fitted reveal a statistically significant curvelinear relationship between *PT* and salinity. From one hand, our *Model 1* revealed expected dependency on salinity: *PT*, in fact, was higher in the lower salinity conditions. However, on the other hand, it was unexpectedly high when salinity was unusually high for the region.

It should be noted that the number of extreme low salinity sites where PT was found to be quite high is relatively small. Similarly, the number of sites where salinity is very high and where *PT* is also high is very small. In total, the number of such sites is about 24.2 % of the total number of sites surveyed. At other sites where salinity intermediate *PT* does not differ significantly from the background value. That is, anomalously high *PT* 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 dominate *ME*. This assumption is supported by the fact that the settlements that we classify as *MT*-dominated are essentially scarcer in the studied area (smaller peak in the frequency distribution, Fig. ++) than the *ME*-dominated sites. Such displacement of weaker competitor into marginal habitat is well known for other systems where inviders met resident species: mosquito larvae (Juliano, 1998), geckos (Case et al 1994) and fishes(Herbold, Moyle, 1986). Usually resident species persist in their optimal habitats if they have not been destroyed but introduced species prevail in marginal habitats (Juliano, 1998).

The analysis of relative importance of different predictors revealed practically equal value of them in regulation of *MT:ME* composition in the Kandalaksha bay contact zone. This means that there is no single leading factor that plays an exclusive role in determining the ratio of species in mixed settlements. Any gradients of factors, whether extended gradients related to salinity or sudden local changes in habitat parameters, as in the case of substrate or surf magnitude, are used by the two mussel species for segregation.

## 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)

### Environmental parameters characteristics

STable +. Ports presented in the Kandalaksha bay

| Port | Status | Latitude | Longitude |
| --- | --- | --- | --- |
| Kandalaksha | Active | 67.13728 | 32.40799 |
| Vitino | Active | 67.07657 | 32.33363 |
| Umba | Abandoned | 66.67797 | 34.35766 |
| Chupa | Abandoned | 66.26996 | 33.06953 |
| Keret | Abandoned | 66.29418 | 33.64066 |
| Kovda | Abandoned | 66.69675 | 32.87540 |

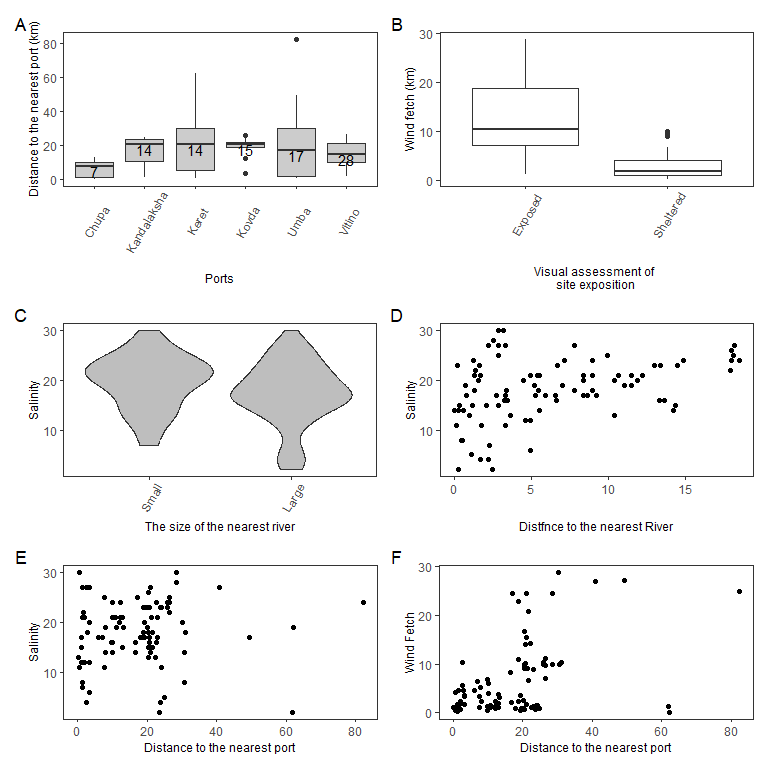
### Associations between environmental parameters

Pearson correlation between environmental parameters

| Predictor | Min\_dist\_river | Min\_dist\_port | Average\_Fetch |
| --- | --- | --- | --- |
| Salinity | 0.368 | 0.049 | 0.261 |
| Min\_dist\_river |  | 0.153 | 0.279 |
| Min\_dist\_port |  |  | 0.462 |

Table +. Parameters of river and other fresh-water discharging source associated with Kandalaksha Bay.

| River | Drainage Area sq. km | River Size | Mouth Latitude | Mauth Longitude | Number of sampling sites near the river |
| --- | --- | --- | --- | --- | --- |
| Chernaya | 141.0 | Small | 66.51896 | 32.94966 | 6 |
| Gridina | 540.0 | Large | 65.92462 | 34.65846 | 2 |
| Keret | 3,360.0 | Large | 66.28172 | 33.58034 | 4 |
| Knyazhegubskaya HPP channel |  | Large | 66.87178 | 32.40742 | 0 |
| Kolvitsa | 1,260.0 | Large | 67.08320 | 32.99110 | 14 |
| Kovda | 26,100.0 | Large | 66.69549 | 32.87103 | 1 |
| Kuzreka | 255.0 | Large | 66.59999 | 34.81303 | 4 |
| Kyatka | 204.0 | Small | 65.88838 | 34.68294 | 0 |
| Lupche-Savino | 146.0 | Small | 67.16291 | 32.34259 | 0 |
| Luvenga | 175.0 | Small | 67.10037 | 32.69313 | 11 |
| Nilma | 167.0 | Small | 66.50577 | 33.13507 | 8 |
| Niva | 12,800.0 | Large | 67.13027 | 32.41825 | 9 |
| Nivskaya HPP channel |  | Large | 67.15676 | 32.37713 | 0 |
| Olenitsa | 403.0 | Large | 66.45276 | 35.30060 | 2 |
| Pila | 313.0 | Large | 66.78056 | 34.15745 | 0 |
| Plavezhma |  | Small | 66.25417 | 33.00065 | 3 |
| Porya | 240.0 | Large | 66.81954 | 33.55116 | 1 |
| Pulonga | 630.0 | Large | 66.30251 | 33.27729 | 6 |
| Ryasanka | 102.0 | Small | 66.73710 | 33.44663 | 0 |
| Salnitsa | 121.0 | Small | 66.37576 | 35.68487 | 1 |
| Shoguy | 103.0 | Small | 66.74761 | 34.02558 | 0 |
| Shushpanikha | 65.2 | Small | 66.85175 | 33.60705 | 0 |
| Sonreka | 167.0 | Small | 66.15655 | 34.22415 | 4 |
| Umba | 6,250.0 | Large | 66.66461 | 34.29766 | 6 |
| Varsuga | 9,840.0 | Large | 66.26487 | 36.94030 | 0 |
| Vypcha | 161.0 | Small | 66.63304 | 34.49202 | 3 |



SFigure +. Associations between environmental factors. (A) Boxplots showing the variation in distances to the nearest port. Numbers insite the box indicate the number of sites for which the corresponding port is the nearest. (B) Boxplots showing the variation of wind fetch for two groups of sites (Exposed vs Sheltered) choosen for material collection. (C) The kernel density violin-plots showing salinity variation for sites near Small and Large rivers. (D) The association of salinity with distance to the mouth of the nearest river. (E) The association of salinity with distance to the nearest port. (F) The association between wind fetch and distance to the nearest port

## Parameters of Model 2

STable ++. Parameters of Model 2

| Model term | Estimate | SE | z-statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | -1.480 | 0.00580 | -255.13 | <0.0001 |
| Substrate(Algae) | 2.238 | 0.00580 | 385.92 | <0.0001 |
| Salinity | -0.125 | 0.00558 | -22.42 | <0.0001 |
| MinDistRiver | -0.009 | 0.00578 | -1.64 | 0.10115 |
| RiverSize(Large) | 2.069 | 0.00580 | 356.72 | <0.0001 |
| AverageFetch | -0.071 | 0.00576 | -12.27 | <0.0001 |
| MinDistPort | -0.034 | 0.00550 | -6.16 | <0.0001 |
| PortStatus | 1.541 | 0.33090 | 4.66 | <0.0001 |
| SD (Intercept) random effect | 0.683 |  |  |  |