## **Materials and methods**

Study area

The study area was the Kandalaksha Bay, where all previous *ME* and *MT* studies in the White Sea have been conducted (REF). The Bay, 185 km long, is funnel-shaped, with numerous islands and skerries and a 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 ppt in most of the Bay (“normal” salinity for most of the White Sea) and lower in the estuarine areas (Berger ???). Two canals of a hydropower plant and 24 rivers with a catchment area of 141 – 12,830 km2 (Median 240 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).

Six ports operating oceanic vessels were functioning in the area in the 20th century (Fig. 1). Two of them, both at Bay’s top, are still in operation. The other four have been abandoned (Sailing directions of the White Sea, 1932; Krasavtsev, 2011) but are occasionally visited by small ships (our observations/reports of local residents).

Mussels are present everywhere in the shallow waters of the Bay. They are particularly abundant in the littoral fucoid belt (mainly *Fucus vesiculosus* L. and *Ascophyllum nodosum* L.), which is continuous 0.5-1.0 m above mean spring tide depth (Berger et al. 2001). According to the data from 2002–2013, both mussel species were almost ubiquitous in the Bay, but their ratio in settlements varied greatly, with *ME* being generally dominant (Katolikova et al. 2016).

There were 4 data sets used in the work: 1 modeling set and three testing sets

## Modeling data set

Mussel sampling and processing

Mussels were sampled at 95 sites within the littoral fucoid belt in 2011-2018 (Fig. 1). Data for 17 of these sites were included in the study by Katolikova et al. 2016, the other data are new. Sites were chosen to describe littoral populations of the Bay in as much detail as possible and to account for the heterogeneity of their habitat by substrate type, surf level, and distance from rivers and ports. All samples were taken within the fucoid belt to minimize differences in depth. At each site, three samples from fucoid thalli (hereinafter, algal samples) and three samples from bottom substrates (bottom samples) were collected a few meters from each other using 0.25 m2 and 0.025 m2 frames, respectively. The frames were placed not randomly but in such a way as to capture the dense mussel aggregations.

We used mussels with a shell length larger than 10 mm to identify the shell morphotypes reliably (Khaitov et al., 2021). In the bottom samples all mussels from a frame were used. In the algal samples the procedure was different. One bundle of algae, containing at least a few dozens mussels, was chosen and weighed together with the attached mussels. The rest of the algae from a frame were weighted too. Mussels from the bundle were counted and used for further analysis. The ratio between the counted number of mussels and to the bundle weight was applied to the total algal weight to reconstruct the total number of mussels in the sample. For 12 sites the information on total number of mussels in algal samples was lacking, and they were excluded from the analyses which required data on mussel abundance (Model 2, see below).

Shell morphotypes (E-morphotype, characteristic of *ME*, and T-morphotype, characteristic of *MT*) were identified for all selected mussels as in Khaitov et al. (2021). Further, the proportion of morphotypes was converted to the proportion of *MT* (*Ptros*) in each sample, in pooled samples from each substrate from each site (*Ptros*Algae and *Ptros*Bottom) and in pooled samples from each site (*Ptros*Site), using equation

, where *PT* - proportion of T-morphotype.

This equation, derived from the 24 genotyped samples (in total, 1105 multilocus mussel genotypes) from the Kandalaksha Bay, reliably predicts *Ptros* over the entire salinity range in the White Sea (i.e., up to 24 ppt), but may overestimate *Ptros* at higher salinities, as observed in the Barents Sea (Khaitov et al., 2021).

Environmental parameters assessment

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

Table 1. Environmental parameters involved in the study

|  |  |  |  |
| --- | --- | --- | --- |
| 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 | Algal vs bottom |
| *Influence of rivers* | | | |
| *Salinity* | Continuous | Surface salinity (ppt) at the time of sampling, i.e. at low tide. | 2-30 (19) |
| *DistRiver* | Continuous | Straight line distance (km) between the site and the nearest river mouth by map. Log-transformed values were used. | 0-18.5 (4.9) |
| *RiverSize* | Categorical | Rivers are categorized according to whether their catchment area is larger or smaller than the median area for all rivers in the region. | Small vs large |
| *Influence of ports* | | | |
| *DistPort* | Continuous | 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. Log-transformed values of average fetch for four cardinal directions were used. | 0.2-28.8 (3.3) |

Testing datasets

Three datasets were used as testing ones. “Kandalaksha littoral” dataset contained 23 samples from 12 littoral sites in the Kandalaksha Bay. We took only algal samples at four sites, only bottomsamples at four other sites and samples from both substrates at the remaining four sites (STable ++, SFig. 1 B). Environmental parameters were assessed in the same way as for the modeling dataset.

“Tyuva littoral” and “Tyuva sublittoral” testing datasets were extracted from the published data of Marchenko et al. 2023. These authors mapped in detail the distribution of *Ptros* in mussel settlements of the Tyuva Inlet in the Kola Bay of the Barents Sea (**Fig. 1**) sampled in 2009-2010. They provided a number of environmental characteristics including depth, *Salinity*, cover of macrophytes in rank scale, and dominant algal species (usually, kelps in the sublittoral and fucoids on the littoral) for each sampling site. “Tyuva littoral” set contained samples from all 23 littoral sites from the depth range corresponding to the fucoid belt (0.5-1.5 m above mean spring tidal depth, Marchenko et al. 2023; note that the position of fucoid belt in the Barents Sea differs from that in the White Sea due to the different tidal amplitude). “Tyuva sublittoral” contained samples from all 15 sublittoral sites (depth range from -0.5 to -3.5 m). Since the substrate of mussel fouling was not registered during sampling, we classified samples into bottom and algalones by the algal cover in the sites (ranks 1-3 and 4-5, correspondingly). The remaining environmental parameters were assessed as for the modeling dataset, with the nearest port in Ekaterininskaya Gavan Bight considered as Active and the river Tyuva flowing into the inlet as a large one.

*Statistical analysis*

All processing was performed using the statistical programming language R 4.05 (R core Team, 202++)

***Dependency of Ptros on environmental parameters in modeling dataset (Model 1)***

We used GAM (generalized additive model, Wood, 2017) as a modeling technique, which works well for SDM construction (Elith et al., 2006). Importantly, it assumes that the relationship between the dependent variable (in our case *Ptros*) and continuous predictors may not necessarily be linear, as in ordinary regression analysis, but curvilinear (Austin, 2002). GAM fitted (hereafter, *Model 1*) was based on beta-binomial residuals distribution and the restricted maximum likelihood 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 a random factor. The function gam() from the package “mgcv” (Wood 2017) was used to fit the model.

To check for the predictors’ collinearity in the model, we calculated the variance inflation factor (VIF, Fox & Monette, 1992). Additionally, we calculated Pearson correlation between continuous predictors. To verify that Model 1 met 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) and found no evidence of spatial autocorrelation.

#### *Association between Ptros, substrate and mussel abundance*

The ultimate goal of the analysis was to find out how the segregation of *ME* and *MT* between algal and bottom substrates depended on the abundance of each species on each substrate. For each site we calculated the difference between proportion of *MT* in algal and bottomsamples: *Dif* = *PtrosAlgae* - *PtrosBottom*. The obtained *Dif* values were used as a dependent variable in *Model 2*, which was constructed as GAM with Gaussian residuals’ distribution.

Assessing the dependence of *Diff* on *PtrosSite* and mussel abundances, we could not directly operate with *ME* and *MT* densities because they could be calculated otherwise than through *Ptros*, which would have inevitably resulted in the collinearity of the predictors. Therefore, we performed principal component analysis for the abundance matrix of T- and E-morphotypes on algal and bottom substrates and used PC1 and PC2 values as independent variables, along with *PtrosSite*, in Model 2. This means that we had to equate morphotypes with species in this case. However, this assumption should not have crucially biased the results of the analysis, given the proportional relationship between *PT* and *Ptros* in mussel settlements from the study area (Khaitov et al. 2021). We used VIF to control for the level of collinearity of the final set of predictors considering the value less than 3.5 as acceptable (Quinn & Keugh REF).

#### *Assessment of predictive power of Model 1*

We wanted to check whether Model 1 could be used to predict the dominant species in bottom and algal samples at a site with known environmental parameters, *MT* (*Ptros*>0.5) or *ME* (*Ptros*<0.5). To do so, we used all the parameters to predict *PtrosAlgae* and *PtrosBottom* for each site within the modeling dataset and within each of the three testing datasets. The predicted values were categorized into those greater than 0.5 and those less than 0.5 and considered to be classifiers for detecting *MT*- or *ME*-dominated samples. The receiver operating characteristics (ROC) followed by the analysis of the area under the curve (AUC, Fielding, Bell,1997; Fawcett, 2006) were used to evaluate the performance of the models. Function roc() from the package “pROC” (Xavier et al. 2011) was used.