Selecting relevant predictors for presence-only species distribution modelling.

A case study from the marine environment .

Samuel Bosch[1,2], Lennert Tyberghein[2], Sofie Vranken[2], Olivier De Clerck[1]

[1] Phycology Research Group, Ghent University, Krijgslaan 281-S8, 9000 Ghent, Belgium. E-mail: samuel.bosch@ugent.be
[2] Flanders Marine Institute (VLIZ), InnovOcean site, Wandelaarskaai 7, B-8400 Oostende.

Study

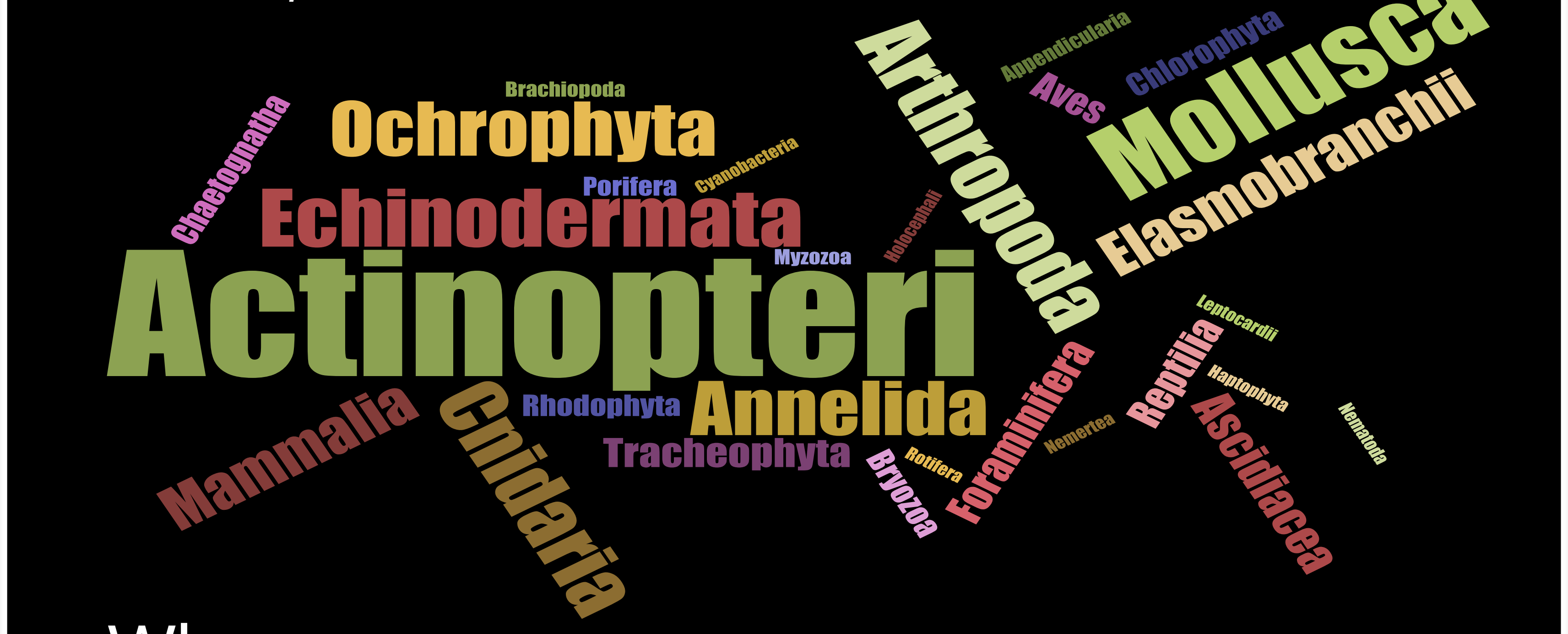
Goals – **1.** Identifying relevant predictors for building presence-only species distribution models of marine species. **2.** Assess the impact of different modelling setups on the relevant predictors. **3.** Case study for the species distribution modelling benchmark dataset MarineSPEED.

Setup – For all 524 species in the MarineSPEED benchmark dataset, distributions where modelled with all combinations of non-correlated environmental variables (Dormann et al., 2013) while varying the number of environmental variables in every one of the models (3,4,7), applying two different sampling bias mitigation techniques (spatial thinning and target-group background) and random or spatial training/test splitting. All setups where compared using four different algorithms (random forests, MaxEnt, GLM and bioclim). Nearly six million models where created and evaluated using the area under the receiver operating characteristics curve (AUC) and the point-biserial correlation (COR) on the UGent High Performance Cluster. Ranking of the predictors within a species was done using rank centrality, rank mean and rank median. Rank differences were analysed using the Friedman test and the Nemenyi post-hoc test. The main R packages used were *marinespeed, dismo, randomForest, scmamp and ggplot2.*

# of predictors	Cross-validation	Sampling bias
3	Random	No thinning
4	Random	No thinning
7	Random	No thinning
7	Random	No thinning
7	Spatial	No thinning
7	Random	Target-group background
7	Random	Spatial thinning

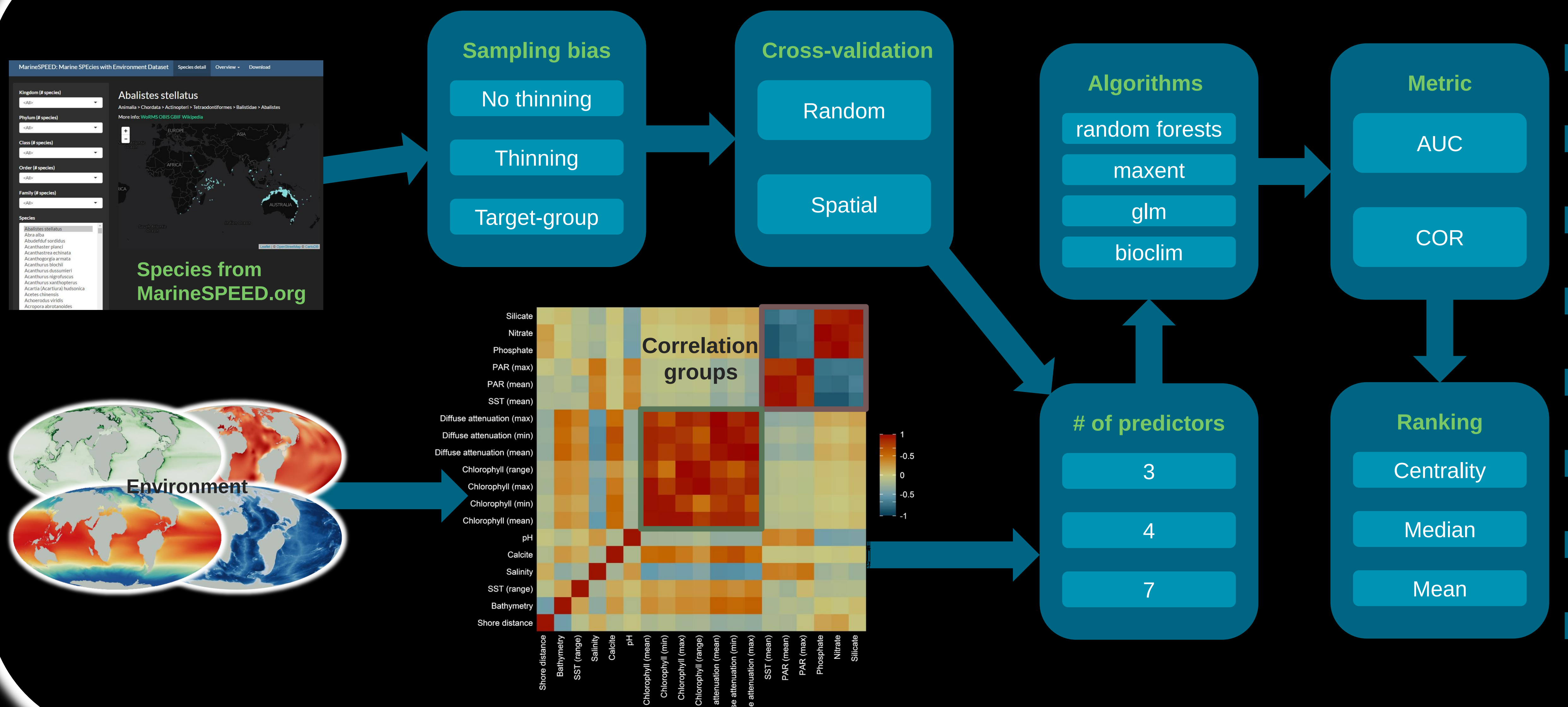
MarineSPEED

What – A quality-controlled dataset of marine species occurrence data from 524 well studied and well identifiable species from all major taxonomic groups with different range sizes and from different ecoregions. Distribution records were downloaded from public repositories such as OBIS, GBIF, EMODNET and Reef Life Survey, filtered on a 25 km² grid and linked to environmental data from Bio-ORACLE and MARSPEC. Additionally taxa are linked to the WoRMS and EOL taxonomy and traits. Different cross-validation and background datasets are included. The R packages used were *robis, rgbif, Reol, raster* and *sdmpredictors.*



Why – A benchmark dataset facilitates the evaluation of SDM algorithms and methods by decreasing the amount of work needed to collect data. Moreover the broad usage of benchmark datasets renders results of different studies comparable.

Where – Website: http://marinespeed.org
R package: https://github.com/samuelbosch/marinespeed



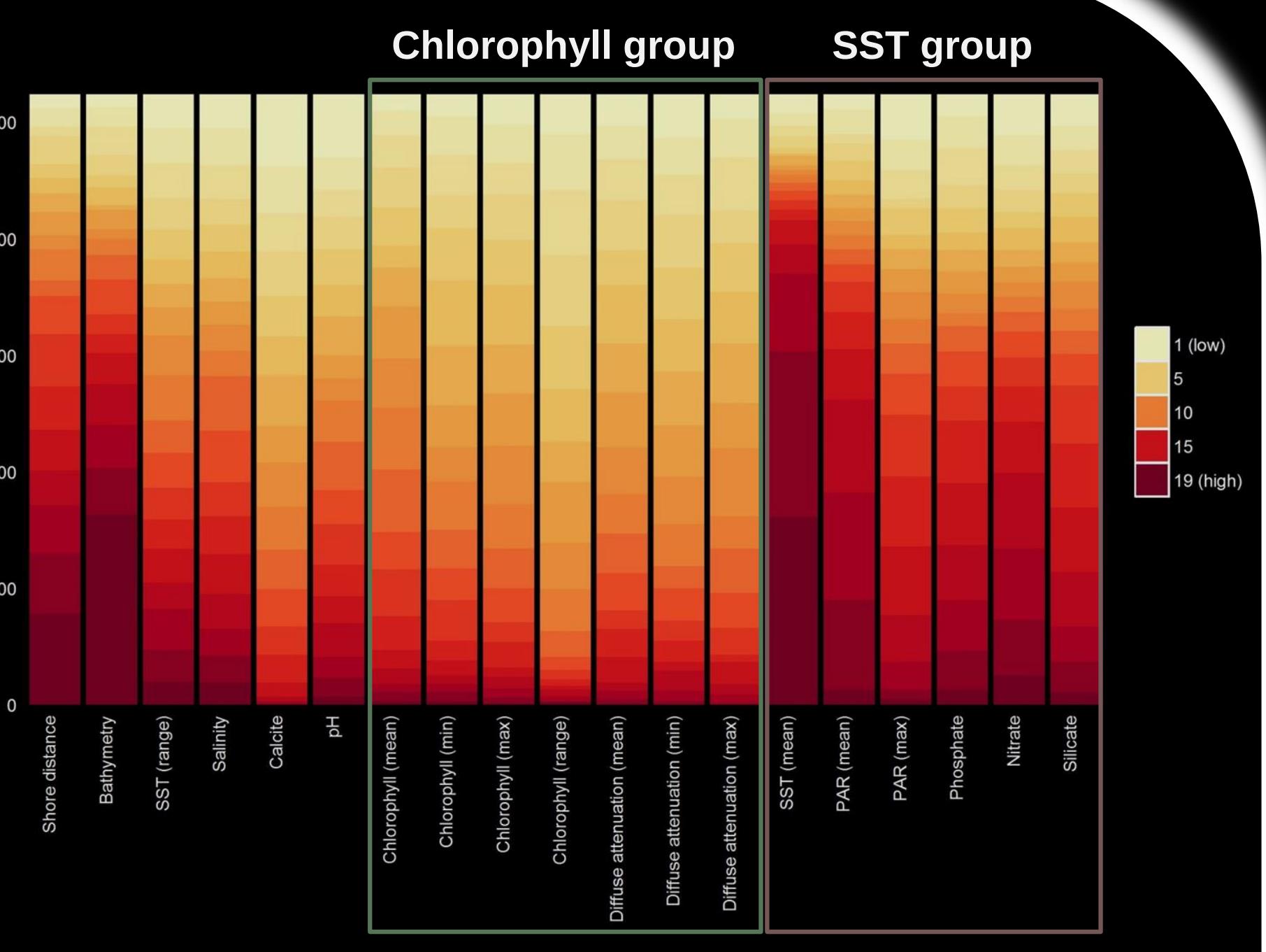


Fig. 1 Count of rankings of the 19 predictors for the 524 species with a specific model setup (7 predictors, random cross-validation, no sampling bias mitigation, bioclim, AUC and rank centrality). Darker colors indicate higher rankings.

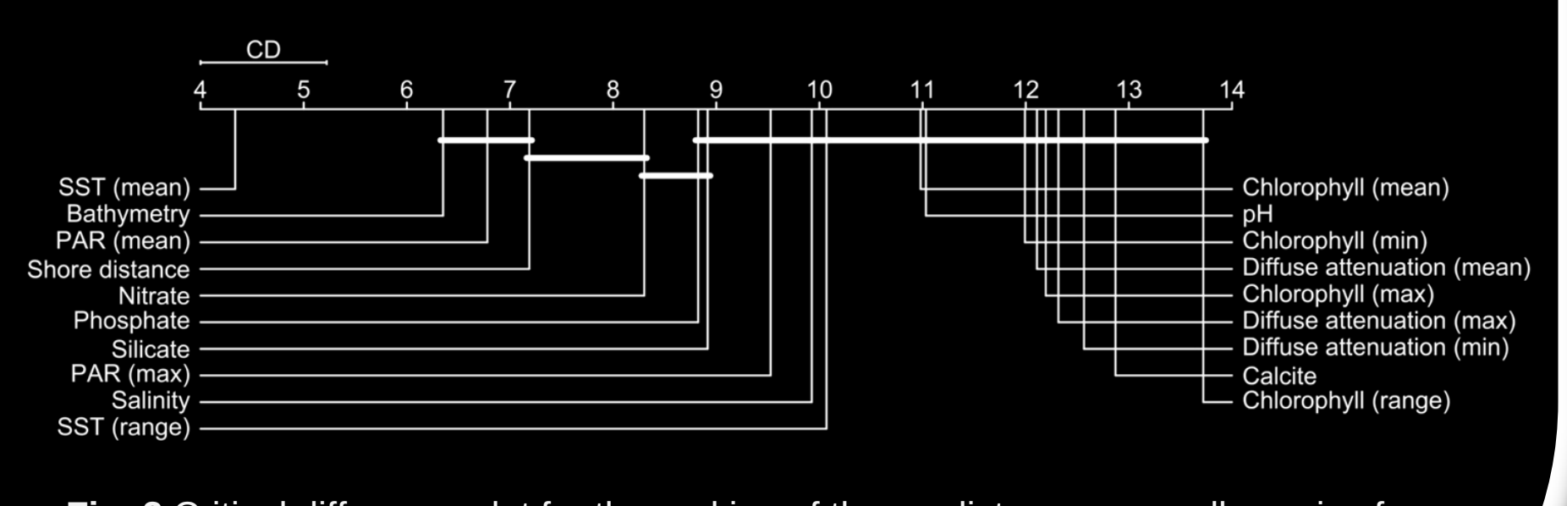


Fig. 2 Critical difference plot for the ranking of the predictors among all species for a specific model setup (7 predictors, random cross-validation, no sampling bias mitigation, bioclim, AUC and rank centrality). Connected lines indicate non-significant differences in ranks.

Results

- Mean sea surface temperature, bathymetry, shore distance, mean photosynthetically active radiation, nitrate and silicate have the highest ranking, highlighting there importance for marine species distribution modelling.
- Diffuse attenuation and SST mean are the most relevant predictors in the chlorophyll and SST group respectively.
- Despite clear overall trends, using different modelling setups has an impact on the predictor rankings.

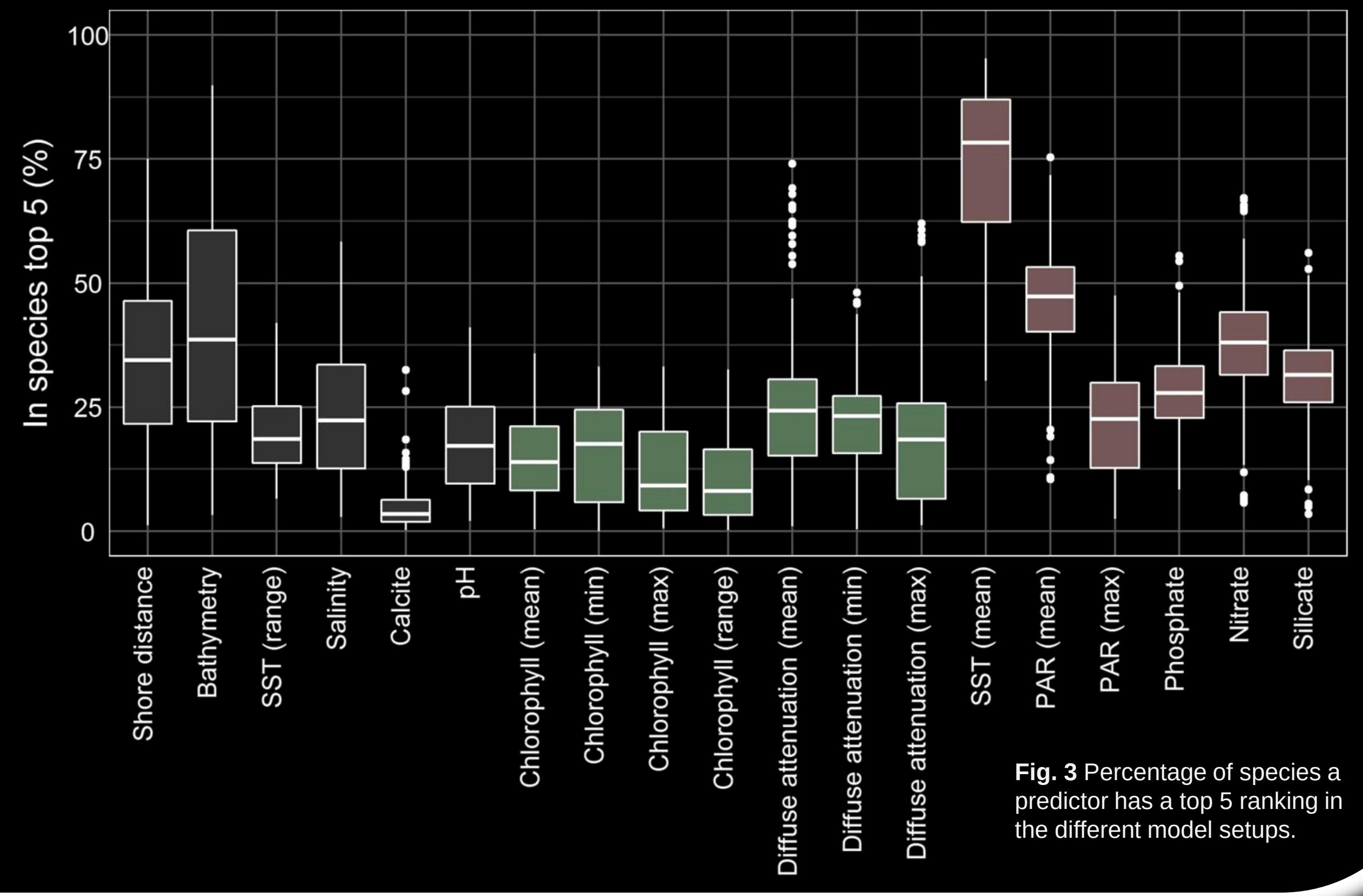


Fig. 3 Percentage of species a predictor has a top 5 ranking in the different model setups.

Conclusion

- MarineSPEED is a valuable dataset for exploring SDM methodologies.
- Some predictors are relevant for a wide selection of marine species but the modelling setup impacts the predictor rankings.
- Further analysis is needed to determine the main drivers of the variation in predictor rankings and to discover possible relationships between species, their traits and the relevant predictors.

References
Barbet-Massin, M., & Jetz, W. (2014). A 40-year, continent-wide, multispecies assessment of relevant climate predictors for species distribution modelling. Diversity and Distributions, 20(11), 1285–1295.
Calvo, B., & Santafé, G. (2015). scmamp: Statistical Comparison of Multiple Algorithms in Multiple Problems, XX, 1–10.
Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., … Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography, 36(1), 027–046.
Hijmans, R. J. (2012). Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. Ecology, 93(3), 679–688. doi:10.1890/11-0826.1

Acknowledgements
The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, the Hercules Foundation and the Flemish Government – department EWI.