VLIZ YOUNG MARINE SCIENTISTS' DAY
20 February 2015

Predictor selection for species distribution modeling in a marine environment.

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

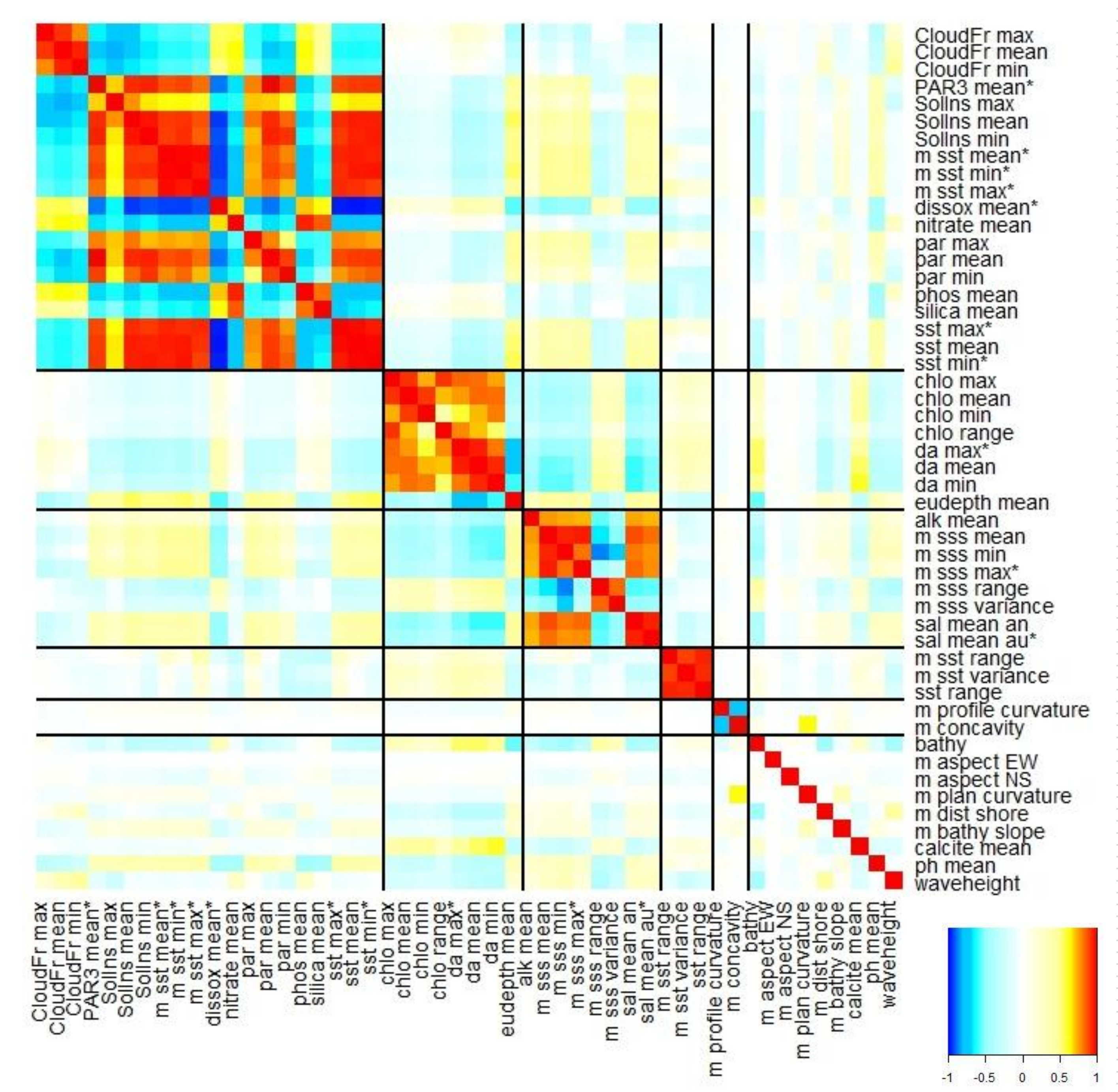
[1] Phycology Research Group, Ghent University, Krijgslaan 281-S8, 9000 Ghent, Belgium. E-mail: samuel.bosch@ugent.be
[2] Flanders Marine Institute (VLIZ), InnovOcean site, Wandelaarskaai 7. B-8400 Oostende.

Over the last few years increasing attention has been given to the development of species distribution models of marine species. New datasets with marine predictors like Bio-ORACLE (Tyberghein et al., 2012) and MARSPEC (Sbrocco & Barber, 2013) have appeared but, with the notable exception of Barbet-Massin & Jetz (2014) addressing bird distribution models, little research has been done on which and how many predictors should be selected in order to obtain a realistic species distribution model (SDM). Predictor selection is, however, a crucial issue as the predictor set directly impacts the performance, the spatial and temporal transferability as well as the interpretation of the model.

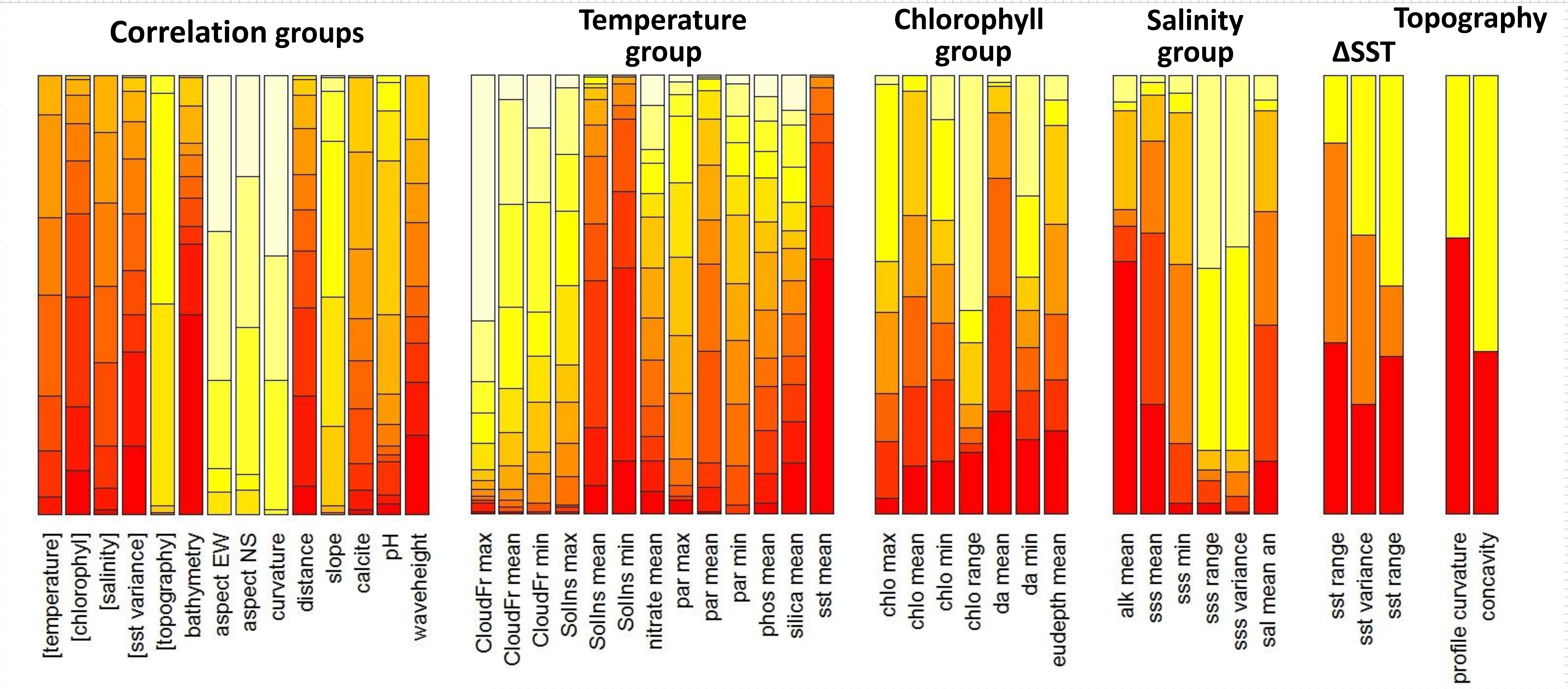
More info and survey
at http://bit.ly/vliz2015

Methods

To find out which predictors are relevant to incorporate in an SDM we created models for a random set of 200 marine species for which quality controlled (Vandepitte et al 2015) distribution records were derived from the international Ocean Biogeographic Information System (OBIS) (www.iobis.org). First we created a pairwise correlation (Pearson's r) matrix for all 53 yearly rasters in BIO-ORACLE and MARSPEC. Based on this correlation matrix we removed 11 variables until all absolute pairwise correlations were smaller than 0.95. Then following the methodology in Dormann et al. (2013) 14 correlation groups were created with a maximum between group absolute correlation of 0.7. Using these correlation groups, SDMs were created for all possible four variable combinations (35031 per species) with MaxEnt on the UGent High Performance Cluster, making sure that two variables from the same correlation group did not occur in one model. From the ranking of the mean, median and maximum test AUC of the predictors from the different correlation groups and within the correlation groups we can infer which predictors are more suitable for building SDMs.



Correlation matrix for all 53 variables from BIO-ORACLE and MARSPEC (prefixed with 'm'). * Variables with an asterix (*) were not included in a correlation group because they had a pairwise correlation greater than 0.95 with another variable.



Representation of the rankings of predictors for every species. [Red = high rank, white = low rank of the average test AUC]. Predictors or correlation groups relevant for most species have a red bar. We can conclude that models which include [temperature], [chlorophyll], [salinity], [ΔSST] and bathymetry and/or distance to shore have higher predictive power.

What's next

With the results of the 4 variable models we have a first indication of which predictors are valuable. Further experiments need to be performed to find out if the results are the same when changing algorithms (e.g. GLM, RF, BRT), number of predictors, set of species and regularization parameters. Also more work has to be done to select the ranking procedure that returns the best predictors to include in an SDM of a species with unknown preferences.

References

Tyberghein, L., Verbruggen, H., Pauly, K., Troupin, C., Mineur, F., & De Clerck, O. (2012). Bio-ORACLE: a global environmental dataset for marine species distribution modelling. Global Ecology and Biogeography, 21(2), 272–281.

Sbrocco, E. J., & Barber, P. H. (2013). MARSPEC: ocean climate layers for marine spatial ecology. Ecology, 94(4), 979–979.

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.

Vandepitte, L., Bosch, S., Tyberghein, L., Waumans, F., Vanhoorne, B., Hernandez, F., … Mees, J. (2015). Fishing for data and sorting the catch: assessing the data quality, completeness and fitness for use of data in marine biogeographic databases. Database, 2015, bau125–bau125.

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.

Acknowledgement

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.