Detecting changes in essential ecosystem and biodiversity properties- towards a Biosphere Atmosphere Change Index: BACI

Deliverable 8.4: Improving species distribution models using novel, high-resolution satellite data

Project title:	Detecting changes in essential ecosystem and biodiversity properties- towards a Biosphere Atmosphere Change Index
Project Acronym:	BACI
Grant Agreement number:	640176
Main pillar:	Industrial Leadership
Topic:	EO- 1- 2014: New ideas for Earth-relevant space applications
Start date of the project:	1st April 2015
Duration of the project:	48 months
Dissemination level:	Public
Responsible for the deliverable:	Signe Normand (Aarhus University) Phone: +4587154345, Email: signe.normand@bios.au.dk
Contributors:	Robert Buitenwerf[1], Jens-Christian Svenning[1], Signe Normand[1] [1] Aarhus University
Date of submission:	28.09.2017

Contents

Aim

Deliverable D8.4 aims at assessing the utility of high-resolution description of ecosystem properties based on EO-data for enhancing fine-scale species distribution models for a wide range of plant and bird species. The final goal is to test the value of a framework where species distributions can be continuously updated as new high-resolution remote sensing layers become available. For D8.4 we specifically assessed the value of available high-resolution remote sensing data for the projection of plant and bird species distributions in Denmark. We use Synthetic Aperture Radar (SAR) to investigate if we can improve models of the geographical ranges of species by characterising attributes of the physical environment that affect organisms at higher spatial resolutions than previously possible.

Introduction

Effective conservation requires knowledge of where species occur. Charting species ranges using empirical data is an expensive and time-consuming activity and for the vast majority of species only very limited empirical data is available. However, tools have been developed to infer species (potential) distributions using available data on species occurrences and environmental correlates. These species distribution models (SDM), also known as environmental niche models (ENM), exploit the relationships between a sample of known species occurrences (and sometimes known absences) and the environmental conditions at those locations in order to estimate the probability of occurrence in un-surveyed locations.

Data quality and choices related to SDM implementation can affect model performance and the accuracy of projections. Many of these issues have been discussed thoroughly in the literature, including the quality of presence and/or absence data (Guillera-Arroita *et al.*, 2015), the choice of modelling technique (Elith *et al.*, 2006; Brewer *et al.*, 2016), how best to implement popular modelling techniques (Morales *et al.*, 2017), how to interpret model outputs (Araújo & Peterson, 2012), how to select background points when absence data is not available (Barbet-Massin *et al.*, 2012) and the value, or lack of value, of correlative SDMs (Araújo & Guisan, 2006; Kearney, 2006; Higgins *et al.*, 2012).

Data quality

One aspect that has received comparatively little attention is the uncertainty associated with predictor variables. Many studies rely solely on gridded climate variables. The most commonly used climate data in SDM studies are gridded climate surfaces based on interpolation from weather-station data. Popular examples of this type of data include worldclim[1] and CRU[2]. Other sources for gridded climate data include process-based climate models, climate re-analyses, where climate models are used in conjunction with observations to project climate across space and (backwards) in time (e.g. NCAR[3]) and satellite-based observation of e.g. land surface temperature (MODIS[4]) or precipitation (TRMM[5]).

Uncertainty in gridded climate data has also been extensively discussed in the climate science literature, particularly in relation to the very uneven distribution of weather stations around the globe. Even over short distances in comparatively densely sampled regions (e.g. Europe), there can be considerable climate variation, particularly in areas with steep topographic gradients. The finest

[1] http://www.worldclim.org/

[2] http://www.cru.uea.ac.uk/data

[3] https://climatedataguide.ucar.edu/

[4] https://lpdaac.usgs.gov/dataset_discovery/modis

[5] https://trmm.gsfc.nasa.gov/

resolution at which global climate data are currently available is 1 km^2 [6][7]. This is relevant to SDM because most individual plants occupy areas ranging from a few square centimetres to a few square meters. Micro-climatic conditions may vary substantially over very short distances. For example, plant communities on stream banks are often completely different from plant communities only several meters away, but this soil moisture gradient could not be detected using 1 km resolution rainfall data.

The consequence is that typical SDMs predict species occurrences at a spatial resolution that is often too coarse for landscape planning and conservation practices. Moreover, if SDM projections are used to plan survey programs, any reduction of the search area below 1 km^2 could save considerable time and resources. Here we explore whether high resolution remotely-sensed surface information can be used to further constrain projected species distributions.

Recent improvements

There have been some recent efforts at improving SDM performance by using better or different predictor variables. For example, in remote areas where weather stations are scarce, remotely sensed climate variables may provide better estimates of local climates (Deblauwe *et al.*, 2016). In other regions species distributions may not be sufficiently explained by bottom-up environmental processes, but top-down processes such as fire may affect species ranges. This idea was tested by for plant communities in California, but fire occurrence did not improve model performance (Crimmins *et al.*, 2014). However, it is possible that the fire occurrence data were too coarse to add predictive power. A final example is a study in which proxies for plant productivity were derived from MODIS satellite data to predict species richness in North American bird communities (Hobi *et al.*, 2017). In summary, EO data is being increasingly used to assess species ranges and overall biodiversity metrics.

Novel EO data and problem statement for D8.4

Although the previous examples have shown some potentially promising avenues for improving SDMs, the coming years will likely see a steep increase in methods and studies that will utilise EO data to: 1) characterise attributes of the physical environment that affect organisms but are difficult to quantify using traditional environmental data, 2) increase the spatial resolution of established environmental predictor variables and 3) exploit the temporal (i.e. dynamic) dimension that EO data increasingly offers.

In this study we aim to address all three aims by using synthetic aperture radar (SAR) improve geographical ranges of bird and plant species. SAR, as implemented on the Sentinel 1 platform, is an active remote sensing technique, where radar waves are transmitted toward the Earth surface and reflectance of these waves is registered by a sensor. Reflectance is transformed to backscatter, which in simple terms provides in indication of the proportion of radar waves that return to the satellite after being reflected off of the surface. Backscatter therefore contains information on surface roughness, e.g. a smooth surface will result in most waves returning to the satellite sensor, and the moisture content of the surface. Backscatter is therefore also related to three-dimensional vegetation structure, which in turn is likely to affect the probability that a plant or bird species can occur in a given location. For example, grassland specialists would not be likely to occur in forests or vice versa.

There are several advantages of Sentinel 1 SAR compared to older SAR sensors, or compared to optical sensors. First, SAR is not hindered by cloud cover, which is a major advantage in many parts of the world. Secondly, Sentinel 1 SAR has very high spatial resolution (between 5 and 40 m), which allows better delineation of sharp boundaries (e.g. a hedgerow along a field) and smaller objects (e.g. tree

[6] www.worldclim.org

[7] www.chelsa-climate.org

clumps in a savanna). Third, Sentinel 1 provides a high temporal frequency, which enables rapid change detection (e.g. grazing in rotational or migratory systems) but also better descriptions of seasonal cycles (e.g. leaf phenology). Finally, the data is freely available.

Methods

Study area

All work in D8.4 focusses on Denmark, which is one of the BACI focus areas. We use Denmark as a test case in this study because we have access to relevant high-quality ground data, as will be described in the following sections.

Plant data

For this study we selected 123 plant species that significantly increased or decreased cover, as measured from vegetation cover, between 2004 and 2010 in Danish Natura 2000 protected areas (Timmermann *et al.*, 2015). This set of species is therefore relevant to nature conservation and management, and potentially as indicators of biodiversity dynamics under ongoing global change.

Distribution data for the species was taken from two large vegetation plot data sets. First, we used data from the Danish NOVANA environmental monitoring scheme. Data on 31,660 plots were collected between 2003 and 2010. Secondly, we used vegetation plot data from sPlot[8], a massive data repository for vegetation plot data around the world. These data are described in more detail in Deliverable 3.4. Here we extracted data for Denmark, and only for plots surveyed since 2003 to match with the NOVANNA starting date. This resulted in 24,139 individual plots. After merging the NOVANA and sPlot data we excluded species with fewer than 50 records and species with problematic taxonomic resolution. The final data set contained presence and absence data on 110 species in 53,378 plots.

Bird data

Distribution data on bird species in Denmark were obtained from the national Danish bird survey, further described in Deliverable 3.4. Briefly, the data are presence records from survey locations that were visited once a year from 2000 to 2016. We scored presences for the entire study period and therefore ignored temporal change. In total there were 7,013 distinct survey locations. After excluding very rare species, i.e. with <10 records, the data set contained presence and absence data for 165 species.

Environmental data

We used four environmental variables as predictors in the SDMs, including two climate variables and two environmental variables representing local surface and topographic variation.

The climate variables were so called bioclimatic variables taken from worldclim 2.0[9]. We used mean annual precipitation (MAP) and mean minimum temperature (Tmin) of the coldest month. Although Denmark occupies a relatively small area and only has minor topographic variation, there is a substantial range in these two climate variables, primarily driven by dominant wind direction and proximity to the coast. MAP ranges approximately from 550 to 900 mm, while Tmin ranges from approximately -7 to 0°C. Since physiological tolerance of plant species to available moisture and

[8] www.idiv.de/splot

[9] http://www.worldclim.org/

minimum temperatures varies strongly, these climatic variables are expected to represent the dominant climate effects on species ranges in Denmark.

The surface variables included clay content of the soil and a topographic wetness index (TWI) calculated from a digital terrain model. Clay content for the whole of Denmark is estimated on a regular 30 m grid, based on soil samples and spatial interpolation. The TWI is calculated from a digital terrain model, which was derived from a 2006 aerial lidar scan of Denmark. The TWI is based on the slope and aspect of a pixel in relation to neighbouring pixels and provides an estimate of how much rainfall runoff is expected to pass through each pixel. TWI was a regular grid of 10 m resolution.

EO data

To test the potential for novel space data to be used in SDM we used synthetic aperture radar (SAR) from the Sentinel 1 platform. Raw SAR data were pre-processed by WP2 and delivered as individual scenes for each acquisition data. A total of four variables were provided, including ascending and descending scans, both with vertical-send-horizontal-receive (VH) and vertical-send-vertical-receive (VV) polarisations. Initial data exploration showed that these variables are highly correlated in the study area, so we selected a single variable: ascending VH. Furthermore, to further reduce data complexity for SDM testing, we selected data from one single year (2014), which was the first full year for which Sentinal 1 SAR data were available. A total of 96 scenes collected on different dates throughout the year were stacked. Initial data exploration showed that simple metrics calculated from the intra-annual time series were highly informative of land cover. In some small test areas, unsupervised classifications using the annual mean and annual amplitude of SAR backscatter could recover the main land cover and land use types. We therefore select annual mean and annual amplitude in SAR backscatter as the two EO-based variables in our SDMs.

The SAR grids have a spatial resolution of 0.00032 × 0.00018 degree, which is approximately 20×20 m.

Modelling

Correlative species distribution modelling (or environmental niche modelling) has been extensively used, tested and reviewed in the recent scientific literature. There are many pitfalls and challenges in implementing and interpreting these models appropriately. Here we take a pragmatic approach by selecting two modelling algorithms on opposite sides of the spectrum from simple (and easy to interpret) to a complex (and difficult to interpret) machine learning algorithm. We used standard techniques and indices to assess model performance, which is largely justified as we do not compare performance statistics between species and we do not project outside the current temporal domain (i.e. no past or future predictions) (Lobo *et al.*, 2008).

Specifically, we used a bioclimatic envelope (bioclim) algorithm as implemented in the *dismo* package for R (Core Team, 2017) and boosted regression trees (Elith *et al.*, 2008) as implemented in the *gbm* package in R. For each species we used a subsampling approach in which 30% of the data for each species is held out for testing the fitted model. This step is repeated five times for each species. To assess model performance we used the area under the receiver-operator curve (AUC). We also used two statistics calculated from the confusion matrix: the sensitivity score, (i.e. the number of correctly predicted presences as a fraction of the total number of observed presences) and the specificity (i.e. the number of correctly predicted absences as a fraction of the total number of absences).

All modelling procedures were implemented using the *sdm* package (Naimi & Araújo, 2016) for R(Core Team, 2017) using the default settings unless otherwise specified in the text.

For the plant SDMs we used all environmental and EO layers at their native resolution (between 10 and 1000 m), as plants occupy a fixed location in the landscape at which they are exposed to environmental conditions. However, birds are mobile and will often utilise multiple areas within a landscape, depending on the species, the landscape characteristics itself, the time of year, behaviour etc. We therefore chose to upscale environmental and EO predictors to a 1/120 degree resolution (~ 1km). This particular scale is arbitrary to an extent and the suitability of this choice may vary among species. Nonetheless, by aggregating to 1 km grid cells we incorporate information on environment and surface texture at a scale that should be more appropriate for birds than the very high native resolution of some environmental layers. To keep the number of predictor variables in the SDMs manageable we simply averaged across finer pixels, but there is room to explore the explanatory power of other texture metrics e.g. measures of variance among finer pixels within larger pixels.

Results

Plants

The results of 1100 SDMs are summarised in Figs. 1-3. The performance of models with only SAR as predictors was substantially worse than models with only climate variables or only surface variables. This pattern was consistent across modelling methods. Furthermore, models in which SAR was added to climate predictors performed no better than models with only climate. These results strongly suggest that the SAR-derived variables we used did not capture relevant aspects of the species' environmental niche. Interestingly, the high-resolution surface variables (clay content and TWI) performed much better, with model performances close to models with only climate variables.

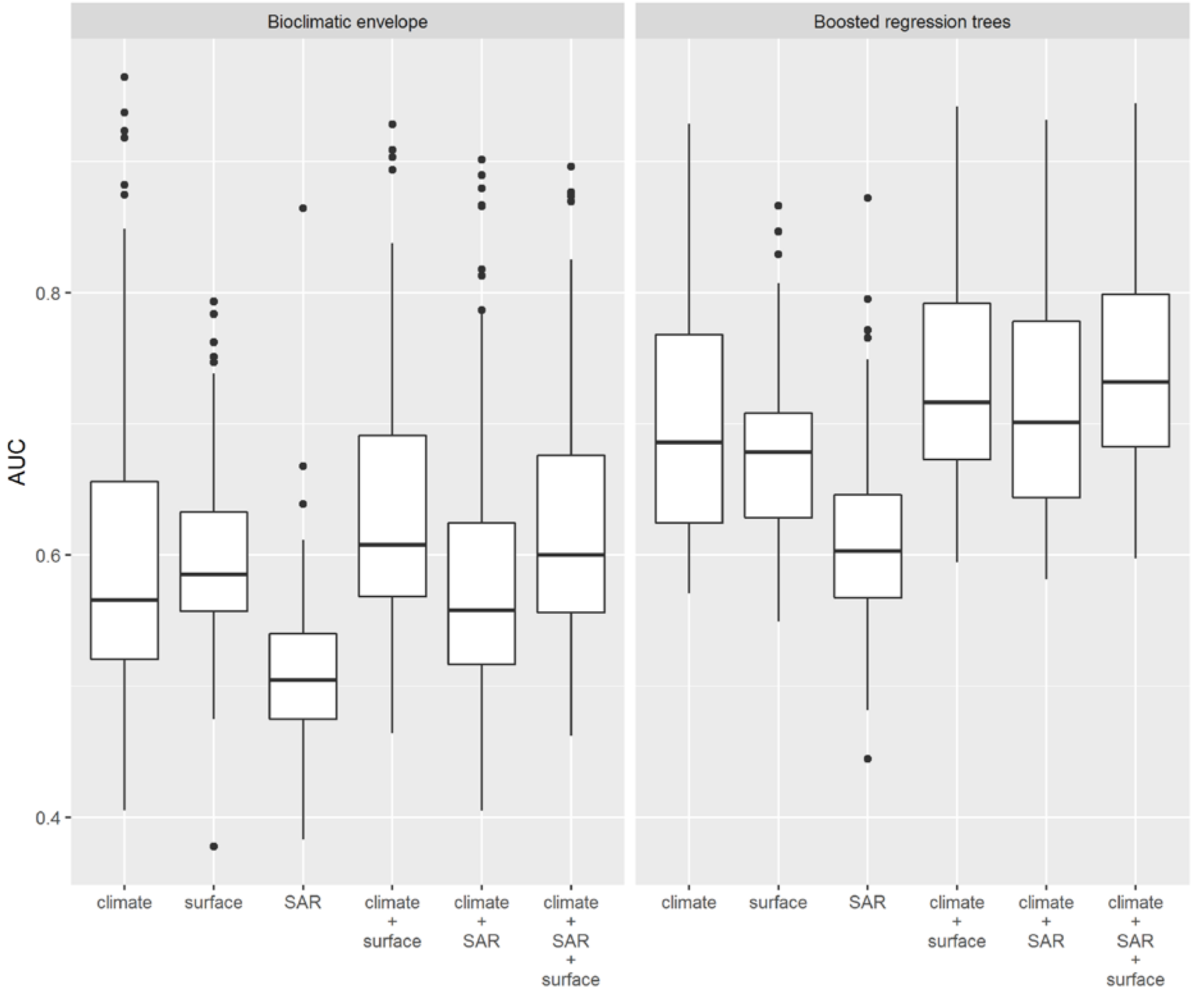

Figure 1 Model performance of species distribution models on 110 plant species in Denmark. Model performance is measured as the area under the receiver-operator curve (AUC). The x-axis shows the different combinations of predictor variables, where climate includes mean annual precipitation (MAP) and mean minimum temperature of the coldest month (Tmin), surface includes clay content and topographic wetness index (TWI) and SAR includes mean annual backscatter and annual backscatter amplitude (for 2014).

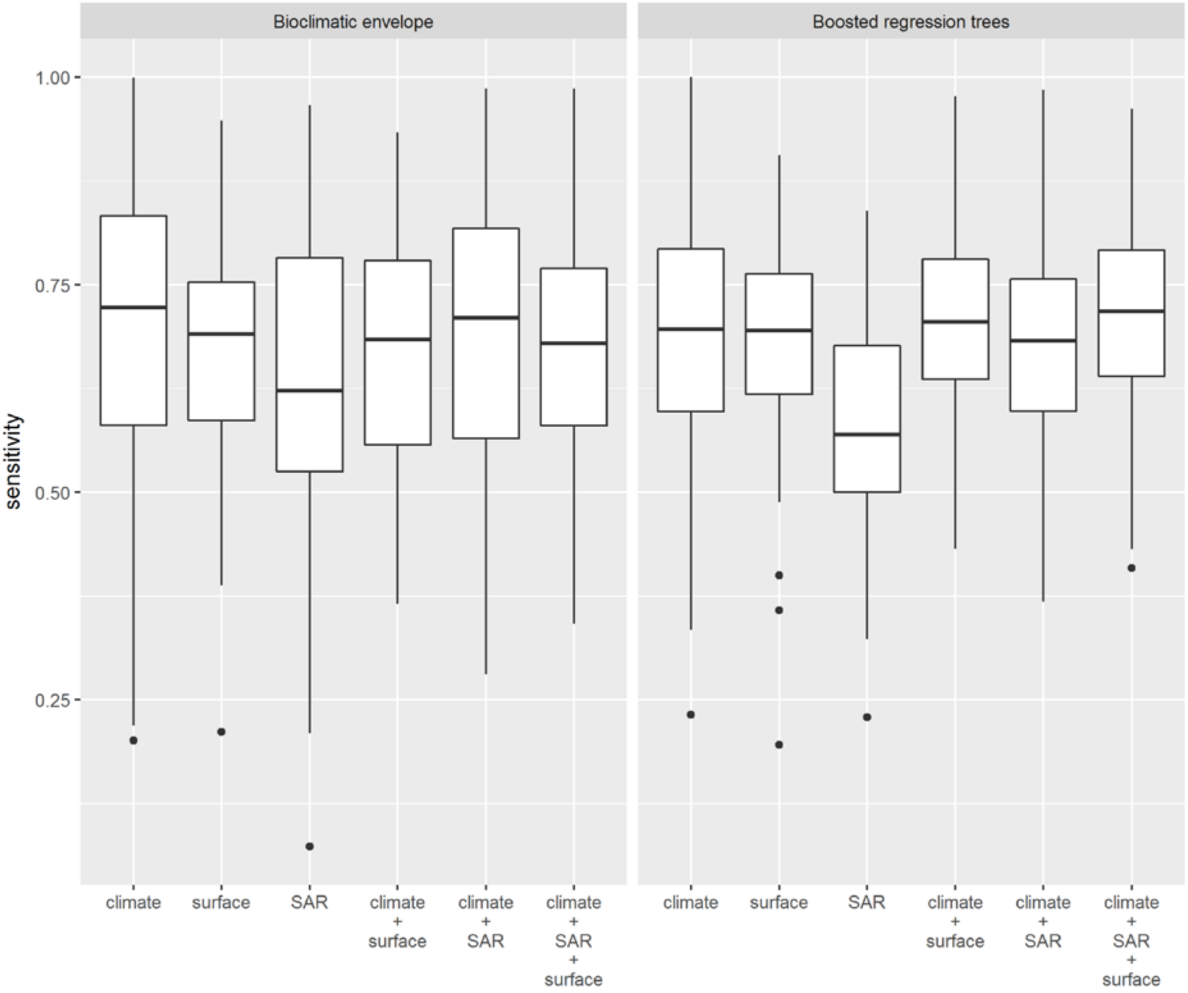

Figure 2 Model performance of species distribution models on 110 plant species in Denmark. Model performance is measured as the sensitivity. The x-axis shows the different combinations of predictor variables, where climate includes mean annual precipitation (MAP) and mean minimum temperature of the coldest month (Tmin), surface includes clay content and topographic wetness index (TWI) and SAR includes mean annual backscatter and annual backscatter amplitude (for 2014).

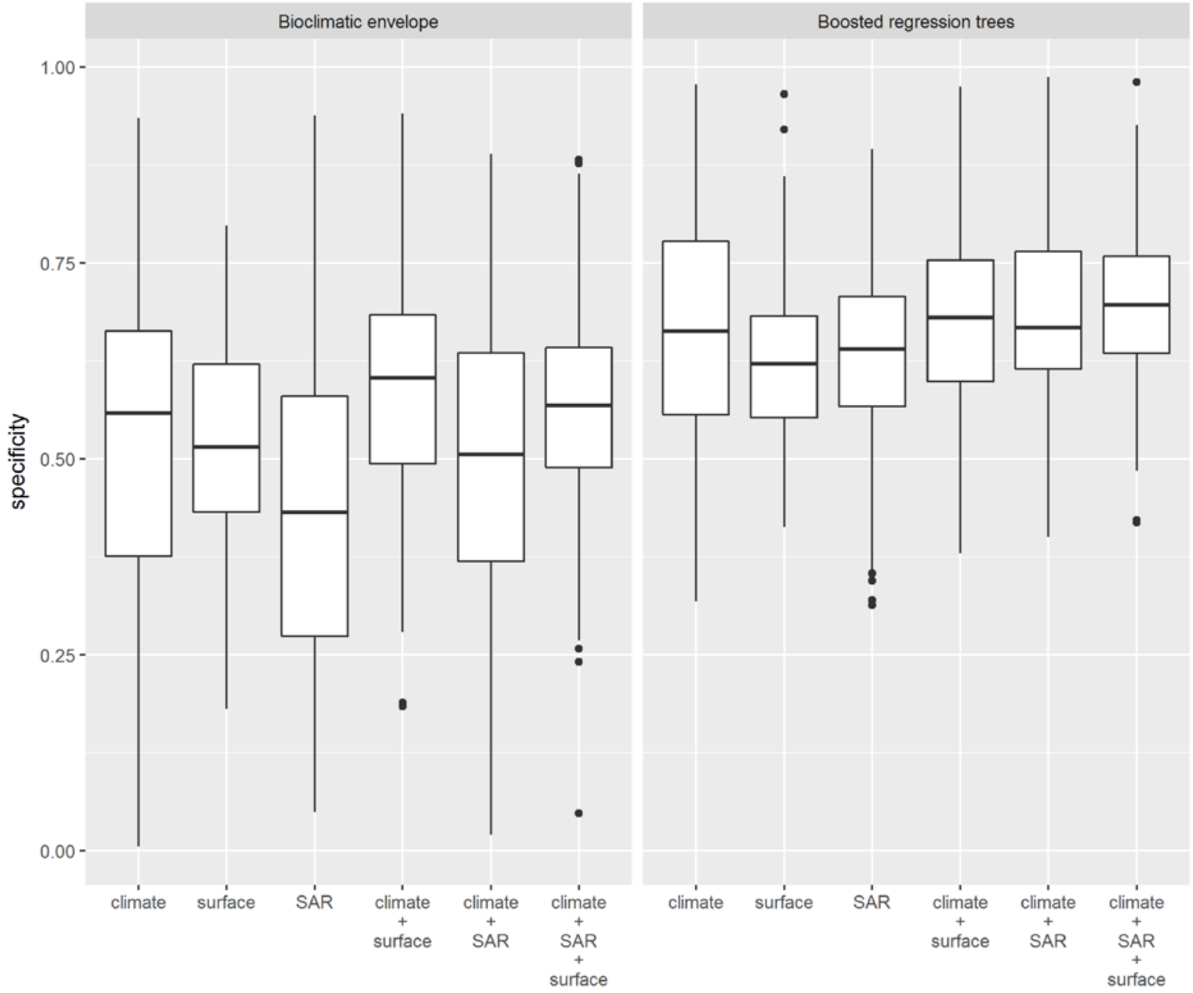

Figure 3 Model performance of species distribution models on 110 plant species in Denmark. Model performance is measured as the specificity. The x-axis shows the different combinations of predictor variables, where climate includes mean annual precipitation (MAP) and mean minimum temperature of the coldest month (Tmin), surface includes clay content and topographic wetness index (TWI) and SAR includes mean annual backscatter and annual backscatter amplitude (for 2014).

Birds

The results of 1650 SDMs are summarised in Figs. 4-6. Across modelling methods and model performance statistics, SAR variables generally performed better as predictors for bird occurrences than both climate and surface variables. The differences between different sets of predictor variables were not very large, but these results suggest that the 1 km SAR-derived variables that we used capture a substantial part of the variation in the multidimensional environmental niche of bird species. Moreover, the models that included both climate and SAR performed better than either climate or SAR alone, suggesting that these variables capture different aspects of the environmental niche.

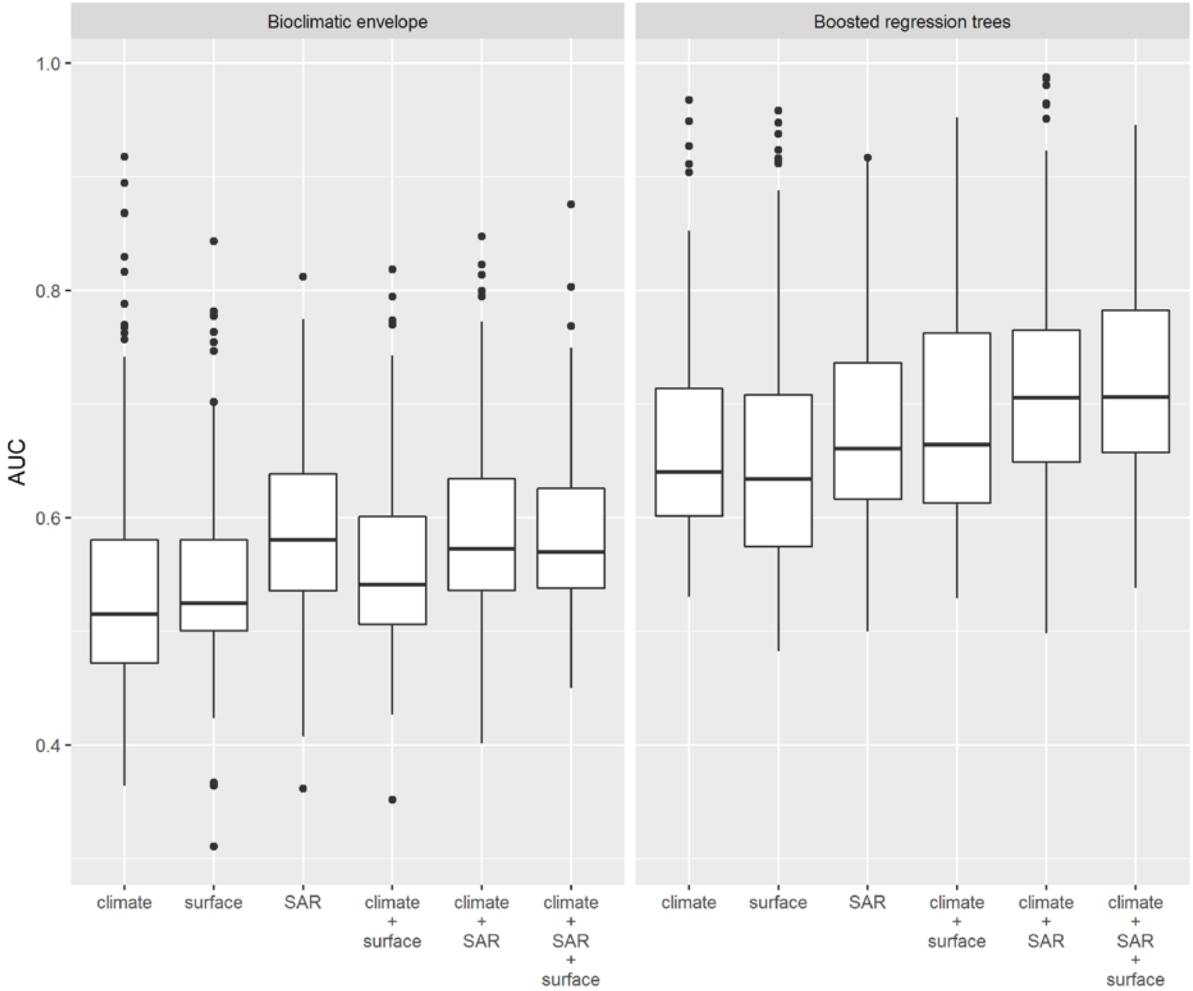

Figure 4 Model performance of species distribution models on 165 bird species in Denmark. Model performance is measured as the area under the receiver-operator curve (AUC). The x-axis shows the different combinations of predictor variables, where climate includes mean annual precipitation (MAP) and mean minimum temperature of the coldest month (Tmin), surface includes clay content and topographic wetness index (TWI) and SAR includes mean annual backscatter and annual backscatter amplitude (for 2014).

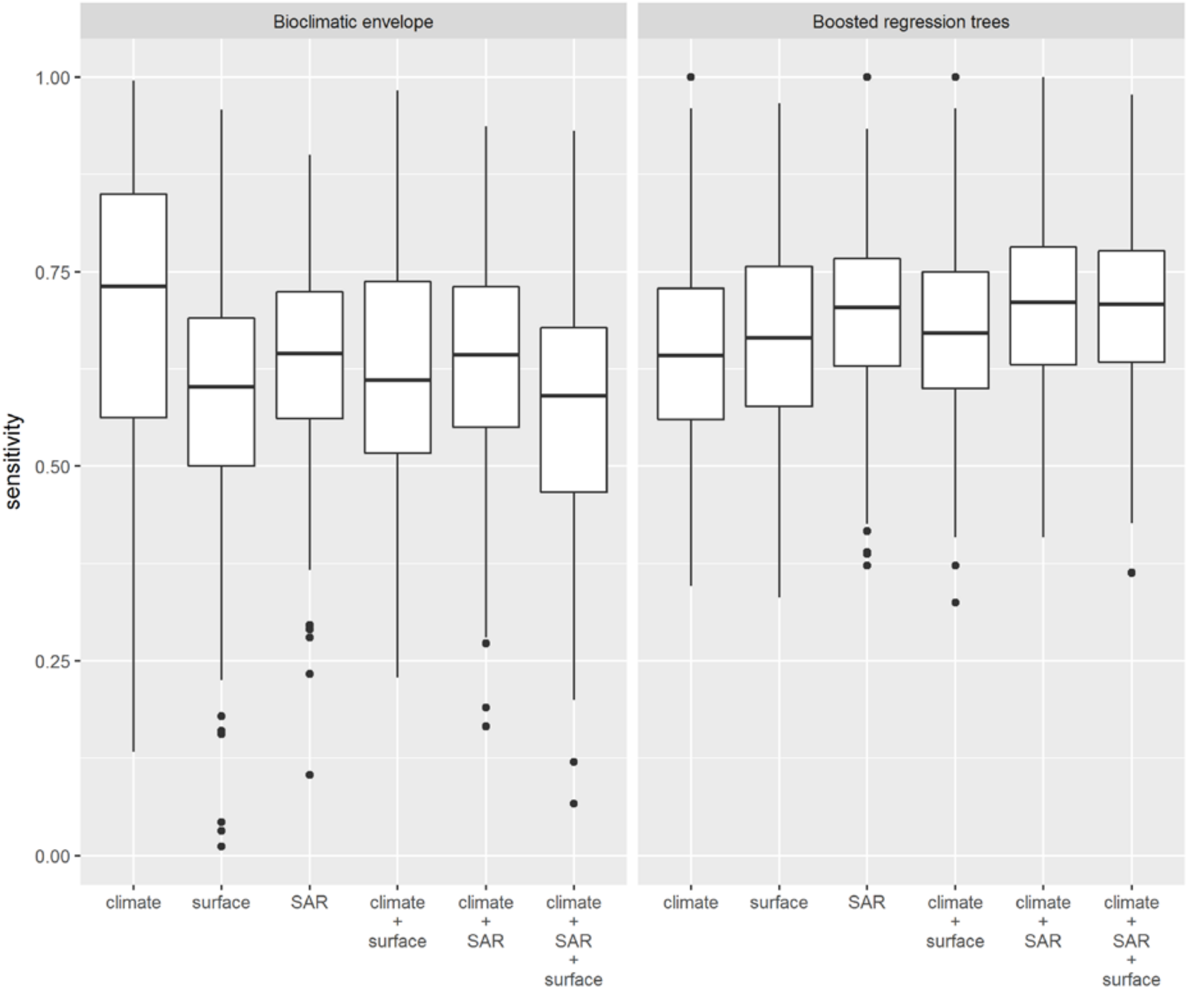

Figure 5 Model performance of species distribution models on 165 bird species in Denmark. Model performance is measured as the sensitivity. The x-axis shows the different combinations of predictor variables, where climate includes mean annual precipitation (MAP) and mean minimum temperature of the coldest month (Tmin), surface includes clay content and topographic wetness index (TWI) and SAR includes mean annual backscatter and annual backscatter amplitude (for 2014).

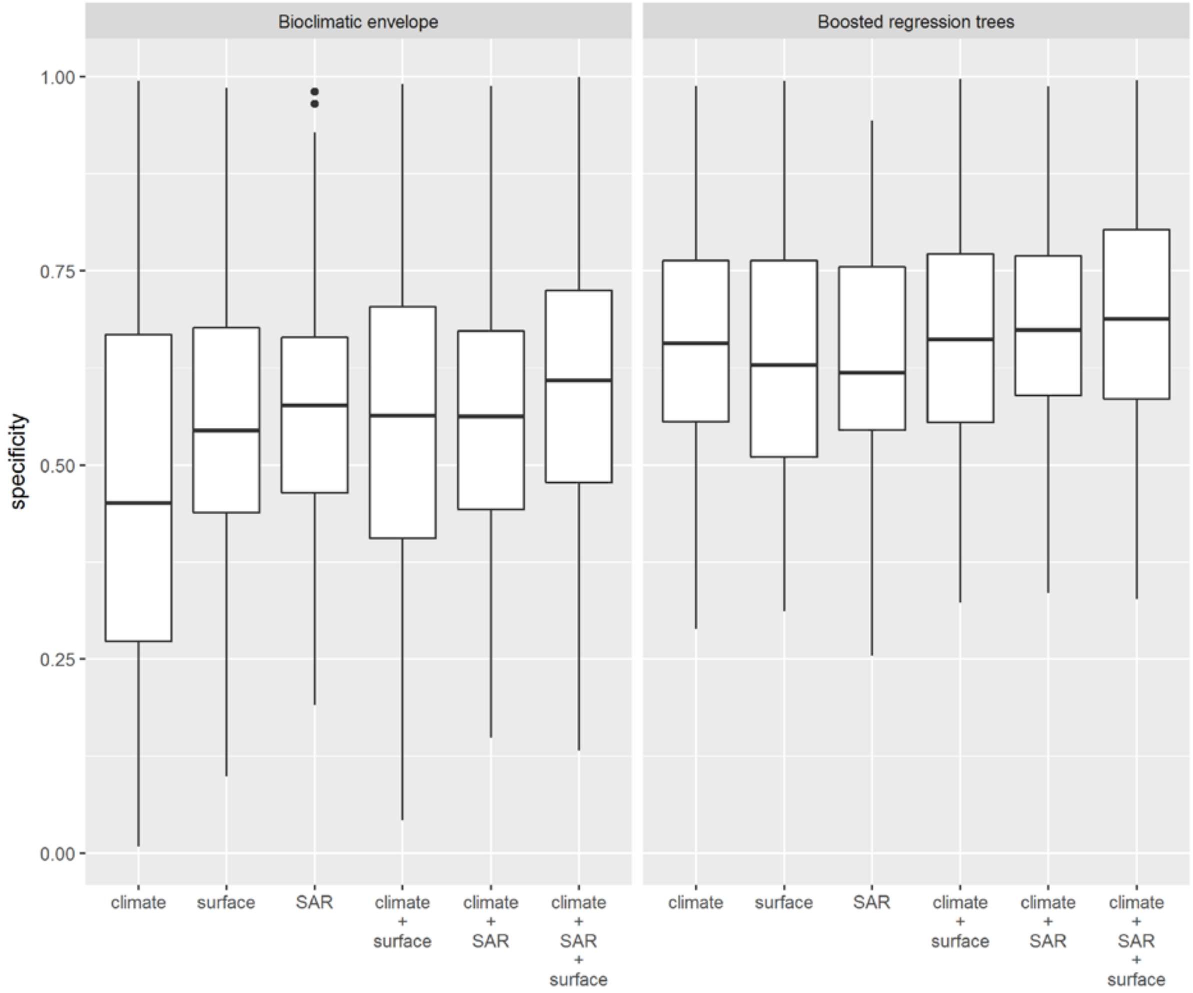

Figure 6 Model performance of species distribution models on 165 bird species in Denmark. Model performance is measured as the specificity. The x-axis shows the different combinations of predictor variables, where climate includes mean annual precipitation (MAP) and mean minimum temperature of the coldest month (Tmin), surface includes clay content and topographic wetness index (TWI) and SAR includes mean annual backscatter and annual backscatter amplitude (for 2014).

Discussion

Overall, SAR data improved SDM performance for birds, but not for plants. We briefly discuss potential explanations for why the results diverged for these different taxonomic groups. We will then lay out how these results may be used to direct future research.

The poor performance of SAR-derived predictors in plant SDMs suggests that these variables do not accurately describe axes of plant niches. Since initial analyses showed that the SAR variables were strongly correlated to land cover, this result was unexpected. However, we did not formally relate our SAR variables to independent land cover classifications and it is possible that the SAR variables accurately describe large contrasts (e.g. between a forest and a crop field), but not more subtle contrasts, such as between different types of deciduous forest. Nonetheless, a lot of information on land cover was contained in our variables and it is therefore likely that there are additional issues.

First, a simple requirement for using high resolution environmental data in SDM is that the species occurrence data has a high spatial accuracy. Although spatial accuracy is generally a valid and substantial concern in extracting species occurrences from observational databases (principally GBIF[10]), the plot databases that we used here may be assumed to have accurate geographic coordinates.

If the SAR variables contain information that is relevant for ecological processes and species occurrence data is spatially accurate, additional explanations for the poor performance of SAR variables must be sought. It is well known that SAR data can be highly variable from pixel to pixel due to partly stochastic processes, resulting in speckled noise or “salt-and-pepper noise” on maps. Such effects can be alleviated by spatial smoothing or aggregation, but we found that our temporal averaging also substantially reduced variability between neighbouring pixels, at least within seemingly homogenous land cover.

Finally, it is possible that the spatial scale of the SAR information is not at the correct scale, i.e. not at the scale at which the organism experiences its environment. For example, a 20×20 m glade in a forest may generate equivalent backscatter to a larger grassland, but these are not ecologically equivalent situations. This implies that surrounding pixels should be taken into account.

This is precisely what we did for the bird data, aggregating from the native 20×20 m resolution to 1×1 km. In the bird SDMs the SAR data performed better than climate, suggesting that the coarser scales better represented niche axes. It is possible that plant SDMs would also benefit from coarser SAR data. At coarser resolution the SAR data would represent landscape scale land cover and heterogeneity, rather than very fine scale land cover.

It is worth considering that using SAR, or any other (EO) variable that is directly affected by the species on the ground, might introduce circularity into SDM models. This might be especially true in the case of large mono-dominant stands, such as spruce/fir forest in the boreal zone. Whether this type of circularity is problematic depends entirely on the purpose of the modelling. If the purpose is to project species ranges into space, i.e. determine where in the landscape a species may be found, this is not a concern. However, if the purpose of the SDM is to describe the environmental niche of a species and use the niche description to project species ranges through time (e.g. future ranges under different climates), including variables like SAR would not be appropriate.

Additional avenues of exploiting SAR data for SDM can and should be explored. Specifically, we would explore 1) additional metrics calculated from the intra-annual time series, 2) inter annual variability

[10] www.gbif.org

and change and 3) different spatial scales. However, a full sensitivity analysis of all potentially important climate, SAR-derived and other environmental variables (including the surface variables used here) is beyond the scope of this study, and most likely both unfeasible and undesirable. Given the study organisms, a first selection of potentially important environmental axes should be made based on knowledge of ecological processes.

Conclusion

In this study we included those climate and other environmental variables that are most likely to set hard limits on the physiological tolerance of species (i.e. tolerance to frost, drought and nutrient poor and waterlogged soils). Within the potential niche delineated by these environmental variables, we asked whether high-resolution SAR data could refine predictions of where in the landscape species should occur. We found promising applications of novel SAR data in improving SDMs for birds, while for plants further work is needed to establish the potential for SAR to improve SDMs.

In conclusion, we demonstrate the value of using high-resolution estimates of ecosystem properties, as quantified using EO data, in estimating species ranges of birds and plants. We thereby provide the first step in developing a modelling framework in which species distributions of birds can be updated continuously (on annual time steps) as new high-resolution EO data becomes available.

References

Araújo M, Guisan A (2006) Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, **33**, 1677–1688.

Araújo M, Peterson T (2012) Uses and misuses of bioclimatic envelope modeling. *Ecology*, **93**, 1527–1539.

Barbet-Massin M, Jiguet F, Albert C, Thuiller W (2012) Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, **3**, 327–338.

Brewer M, O'Hara R, Anderson B, Ohlemüller R (2016) Plateau: a new method for ecologically plausible climate envelopes for species distribution modelling. *Methods in Ecology and Evolution*, **7**, 1489–1502.

Core Team R (2017) R: A Language and Environment for Statistical Computing.

Crimmins S, Dobrowski S, Mynsberge A, Safford H (2014) Can fire atlas data improve species distribution model projections? *Ecological Applications*, **24**, 1057–1069.

Deblauwe, Droissart, Bose et al. (2016) Remotely sensed temperature and precipitation data improve species distribution modelling in the tropics. *Global Ecology and Biogeography*, **25**, 443–454.

Elith J, Graham CH, Anderson RP et al. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Elith, Leathwick, Hastie (2008) A working guide to boosted regression trees. *Journal of Animal Ecology*, **77**, 802–813.

Guillera-Arroita G, Lahoz-Monfort J, Elith J et al. (2015) Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography*, **24**, 276–292.

Higgins S, O'Hara R, Römermann C (2012) A niche for biology in species distribution models. *Journal of Biogeography*, **39**, 2091–2095.

Hobi M, Dubinin M, Graham C, Coops N, Clayton M, Pidgeon A, Radeloff V (2017) A comparison of Dynamic Habitat Indices derived from different MODIS products as predictors of avian species richness. *Remote Sensing of Environment*, **195**, 142–152.

Kearney (2006) Habitat, environment and niche: what are we modelling? *Oikos*, **115**, 186–191.

Lobo J, Jiménez-Valverde A, Real R (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.

Morales N, Fernández I, Baca-González V (2017) MaxEnt's parameter configuration and small samples: are we paying attention to recommendations? A systematic review. *PeerJ*, **5**, e3093.

Naimi B, Araújo M (2016) sdm: a reproducible and extensible R platform for species distribution modelling. *Ecography*, **39**, 368–375.

Timmermann A, Damgaard C, Strandberg M, Svenning J (2015) Pervasive early 21st-century vegetation changes across Danish semi-natural ecosystems: more losers than winners and a shift towards competitive, tall-growing species. *Journal of Applied Ecology*, **52**, 21–30.