SLU Artdatabanken
Jörg Stephan, Sofie Wikberg,
Sebastian Sundberg

RAPPORT 2022-05-05
SLU.dha.2021.5.2-87

Spatial modelling of habitat suitability for *Calypso bulbosa*, a protected plant species in boreal moist forest

[Rumslig modellering av habitatlämplighet för norna i norra Sverige]

Summary

Calypso bulbosa is a rare orchid listed in the Habitats Directive's annex 2 and 4. Since the species occupies rather common habitats, mesic and moist forest in the true boreal region, a major dark figure of hitherto undetected occurrences of the species is probable. These unknown occurrences are important to discover for species protection and to get a better estimate of the probable population size and distribution of the species in Sweden. Therefore, we modelled and mapped the species' potential sites of occurrence at the hectare level, based on presence/ pseudo-absence of *Calypso* and in relation to an initial set of 113 environmental and habitat variables, including e.g. land cover, land use, forest type, climate, soil moisture, soil and bedrock type. We used a forward model selection, using a Bayesian species distribution model, which ultimately resulted in 11 explanatory variables that best explain the presence/absence of the species and had the highest predictive power. Our final model explained typical variation (17%) in the species occurrence given these macro scale environmental variables and could very well discriminate between presences and absences (AUC = 0.9). The resulting habitat suitability map indicates that there may be many undiscovered *Calypso* sites in spruce forests in the far north, especially in the alpine region in the northwest. The probability map may be used as a guide for finding undiscovered sites/hot spots. After further sampling the accuracy of the model could be tested as the number of false negative and positive would be available. If reliable, the model may also be used to calculate dark figures for the distribution and populations size of *Calypso bulbosa*.

Content

Aim

The aim of this analysis is to provide “hotspots” where the forest-plant species *Calypso bulbosa* could occur. It further describes the opportunities and limitations of using opportunistic reporting of species occurrences (Artportalen) and large-scale GIS based variables in order to predict species occurrences in new areas.

Background

All plant species have ecological niches within which they can persist. These niches are given by environmental conditions like light availability, microclimate, humidity, water availability, pH, or nutrients. In addition, the possible presence of the species in a locality is affected by biotic factors like competition or facilitation with/by other species. Populations are also not static but constantly change, with this change depending e.g. on the proximity to (historical) occurrences of the species and the ability to disperse and establish.

The aim of this report was to develop a species distribution model for *Calypso bulbosa* [norna], a plant species in moist forests that is listed in annex 2 and 4 of the Habitats Directive.

The first objective was to identify variables that are important for predicting the occurrences of *C. bulbosa*.

The second objective was to build a model that can be used to predict occurrences given these environmental variables over new areas allowing the identification of areas where new populations could be discovered.

Methods

The used approach to identify important variables for predicting the species occurrences is explained in detail in this section. This includes intermediate results that led to decisions on how to proceed, while the main output of hot spots of occurrences is stored as layers in ArcGIS and in a project in Artportalen. The spatial scale of the data output is quadrats of one hectare (100 ×100 m).

Preparation of data and first selection of variables

Species observations

The presences of *C. bulbosa* in each quadrat were extracted from Artportalen. In case of several observations within the quadrats each quadrat was counted as "present".

The absences were estimated by using 70 plant species that are frequently associated with *C. bulbosa*. A quadrat was recorded as species present if *C. bulbosa* was found within and recorded as absent if any of the 70 associated plants were found but not *C. bulbosa*. Species occurrences of the associated plant species were limited to northern Sweden, from the provinces of Dalarna and Gästrikland and northwards during the period May 15-July 1 (when *C. bulbosa* is most likely discovered).

Environmental variables

Available data consisted of the presence of *C. bulbosa* (norna_presence), the coordinates (point_x, point_y), and 113 variables, with many of them offering very sparse information (Fig. 1). The variables were summarized to the used scale (e.g. mean, proportion of area). The environmental variables were extracted using ArcGIS (provided by Sofie Wikberg).

Climate

Precipitation and temperature data were obtained from SMHI maps with a resolution of 4 × 4 kilometres. Data from the period 1991-2013 were used (most recent available) and we tested both annual (precip_ann, temp_ann) and seasonal data (spring (precip_mam, temp_mam), summer (precip_jja, temp_jja), fall (precip_son, temp_son), winter (precip_djf, temp_djf)). From the same source we also got the start and the length of the vegetation period (veg_start, veg_per).

Exposure

Elevation data for the quadrats were based on the Swedish National Elevation Model with a resolution of 2 meters. The 2-meter cells were first aggregated into 10-meter cells and mean elevation was calculated. From the 10-meter cells minimum, maximum and mean values of elevation, slope and aspect were calculated for each of the one hectare quadrats (hojd100m_min, hojd100m_max, hojd100m_mean, slope100m_min, slope100m_max, slope100m_mean, aspect100m_min, aspect100m_max, aspect100m_mean). From the aspect values the minimum, maximum and mean percent south of 100 m^2 cell was calculated (if x larger than 180-> x=360-x) (perc_south_min, perc_south_max, perc_south_mean).

Bedrock

Information about the bedrock material in the one-hectare quadrats was obtained from the SGU bedrock map, with a scale between 1:50 000 and 1:250 000. There were 58 different bedrock types, divided into calcareous (Kalkberg) and not calcareous types, as well as into 14 different types of chemical composition of the bedrock. Proportions of each type in the one-hectare quadrats were calculated.

Soil types

Soil type maps from SGU were used to calculate the proportion of different soil types in the one-hectare quadrats. The maps come in different scales, from 1:25 000 to 1:750 000, but the one with the most detailed scale available was always used in any location. The soil types were grouped into eight classes (Sväm- eller älvsediment, Organisk jordart, Lera-silt, Isälvssediment, Grov silt-finsand, sand eller grus, Morän, moränlera eller lerig morän, Sedimentärt berg, Berg).

Soil chemistry

To get an estimate of the soil chemistry in the quadrats, we used data from the Swedish Forest Soil Inventory. The sample points of this inventory are at least 12.5 km apart and interpolation using an inverse distance weighted (IDW) technique was used to obtain a raster surface with cells corresponding to the one hectare quadrats. We used pH, hydrogen ion concentration and base saturation data from both the humus layer (0-30 cm depth) as well as from the mineral soil (>65 cm depth) (idw_ph_h30, idw_hconc_h30, idw_base_sat_h30, idw_ph_m2065m, idw_base_sat_m265).

Soil moisture

Soil moisture data were obtained from the SLU Soil moisture map with a resolution of 2 meters. The 2-meter cells were first aggregated into 10-meter cells and mean soil humidity was calculated. From the 10-meter cells minimum, maximum and mean values soil humidity were calculated for each of the one hectare quadrats (soil_moisture100m_min, soil_moisture100m_max, soil_moisture100m_mean).

Land cover

We used the National Land Cover map with a resolution of 10 meters to obtain the proportions of 24 different land cover types in the one-hectare squares. The same source also provided us with proportions of three forest productivity classes (Ej_skogsmark, Improduktiv_skogsmark, Produktiv_skogsmark).

From all available variables several where selected to be individually tested as explanatory variables for the *C. bulbosa* occurrences (Table 1). As can be expected, many of these were highly correlated (Fig. S1). We therefore selected several uncorrelated extracted variables and calculated several new variables. Hence, this first selection was based on (i) discussions of the relevance of certain explanatory variables for the species occurrences among Sofie Wikberg, Jörg Stephan, and Sebastian Sundberg, (ii) that the predictor needed to offer some variability, and (iii) that the predictors cannot be correlated among each other. If correlated one, biological more meaningful, explanatory variable was selected.

The variables aimed to represent larger groups of important conditions: exposure, soil, climate, forest type, ground type (Table 1). Three variables were tested with the aim to evaluate if the variability within a hectare quadrat affected the species occurrences. The reasoning behind was that the species could occur at quite different heights/slopes/percent souths if there was great variability within the quadrat, meaning the association with the quadrat mean could be very weak. To avoid large computational efforts the range, not the standard deviation was used. Two variables were tested, if they would provide a more meaningful measure of pH.

The final data set with all environmental variables consisted of 9 636 observations across Sweden (species absences = 8 983; species presences = 653).

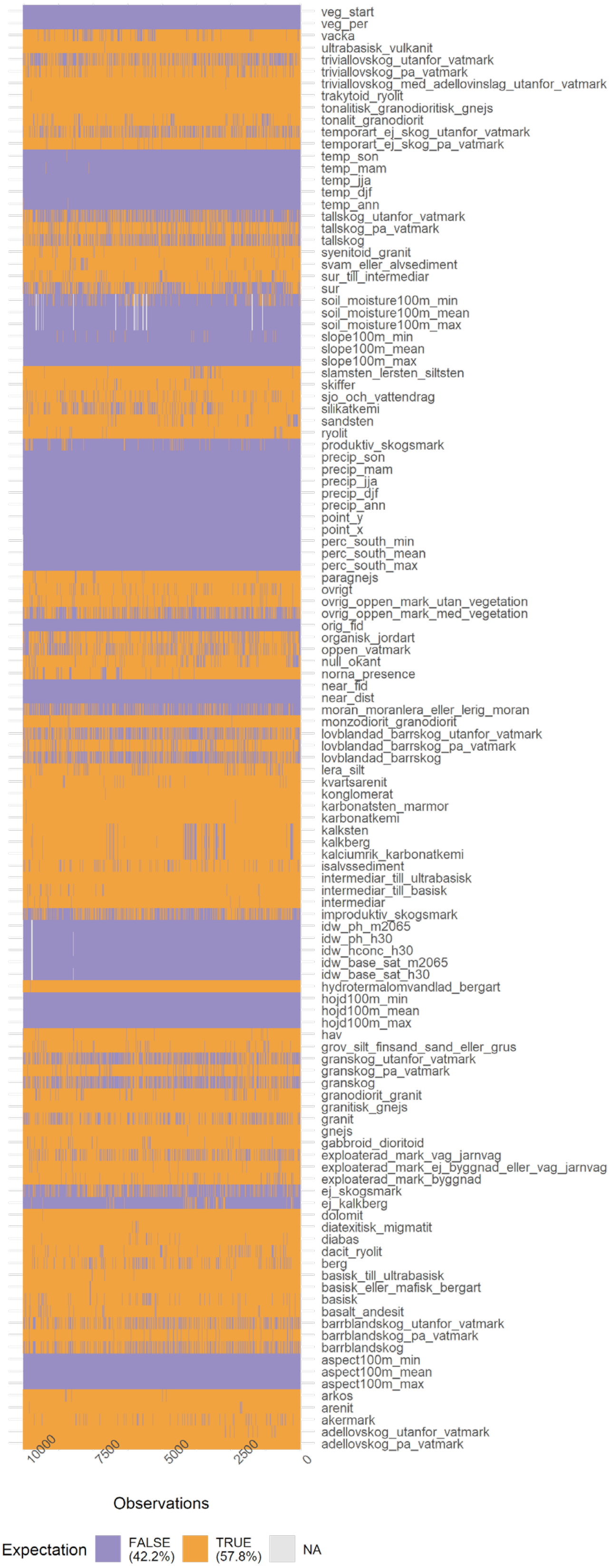


Fig. 1: Variability of the 113 environmental variables. Shown are each observation site (hectare quadrat) with respect to each extracted variable. Orange indicates that the value is 0, Grey indicates that the value is any number other than 0, and white indicates the data are not available

Table 1: Overview of variables that were uncorrelated, assumed to be important for the species occurence, and further tested individually. Variables in italics were calculated from GIS-extracted variables.

Group	Predictor	Explanation
Exposure	hojd100m_mean	mean height of 100 m^2 cell
	slope100m_mean	mean slope of 100 m^2 cell
	perc_south_mean	mean percent south of 100 m^2 cell (from aspect, if x larger than 180-> x=360-x)
Soil	idw_ph_h30	pH in the top humus layer
	soil_moisture100m_mean	mean soil moisture in 100 m^2 cell
	organisk_jordart	proportion of 100 m^2 cell with organic soil
	berg	proportion of 100 m^2 cell with surface rock
Climate	precip_ann	mean annual precipitation
	temp_ann	mean annual temperature
Forest type	*granskog*	proportion of 100 m^2 cell with spruce forest (=granskog_pa_vatmark + granskog_utanfor_vatmark)
	tallskog	proportion of 100 m^2 cell with pine forest (=tallskog_pa_vatmark + temporart_ej_skog_utanfor_vatmark)
	barrblandskog	proportion of 100 m^2 cell with mixed conifer forest (=barrblandskog_pa_vatmark+ barrblandskog_utanfor_vatmark)
	lovblandad_barrskog	proportion of 100 m^2 cell with mixed forest (=lovblandad_barrskog_pa_vatmark+ lovblandad_barrskog_utanfor_vatmark)
Ground type	silikatkemi	proportion of 100 m^2 cell with silicate in soil
	kalkberg	proportion of 100 m^2 cell with limestone
	granit	proportion of 100 m^2 cell with granite
Quadrat variability	hojd100m_range	range of heights within 100 m^2 cell (=max-min)
	slope100m_range	range of slopes within 100 m^2 cell (=max-min)
	perc_south_range	range of percent south within 100 m^2 cell (=max-min)
pH-alternatives	idw_base_sat_m2065	other measure of pH (least correlated with idw_ph_h30)
	idw_hconc_h30	hydrogen concentration

Modelling framework used and general model set up

Here we used a species distribution model to predict the occurrences. We used R (R Core Team 2020) and Hierarchical modelling of species communities (Hmsc) using the Hmsc package (Tikhonov et al. 2019, 2021). This joint species distribution model (Ovaskainen & Abrego 2020) is a Bayesian multivariate, hierarchical generalized linear mixed model. The response variable can be constituted by the matrix of presence-absence (or abundance, percentages) of each species at each site. Here we modelled only one species. This model type offers options to be applied to spatially explicit data and large data sets over extensive areas (Tikhonov et al. 2020).

All models had a Bernoulli likelihood and probit link function. All explanatory variables were centred (subtracted by mean) and scaled (divided by standard deviation). The resulting z-scores are on the same scale and their effect sizes can be compared.

The model's explanatory power were quantified using Tjur's R^2, which is the average predicted occurrence probability among the sites where the species occurs, minus the average where the species does not occur (Tjur 2009). The model's predictive power was compared using AUC and WAIC for the individual models and further quantified for the final model using a four-fold cross-validation with Tjur's R^2 and AUC averaged over the folds (cTjur's R^2, cAUC). We further plotted the ROC curve and estimated several model performance measures (Fig. S4). To evaluate the importance of the variables for the species occurrence, we performed variance portioning. Default prior distributions were used and model convergence was examined using the potential scale reduction factors (Gelman & Rubin 1992) and the effective number of samples.

Identifying variable that are important for *C. bulbosa* occurrences with one model for each variable

We fit one model for each of the variables in Table 1 and one model each with the northing and southing coordinates (Fig. 2). For most of the variables, we found a negative estimate, indicating that with increasing value of the variable the occurrence probability decreases. Neither the variability within the quadrat nor the pH-alternatives had strong effects. Individual variables explained between zero and 6.79% of the occurrences (Table 2).

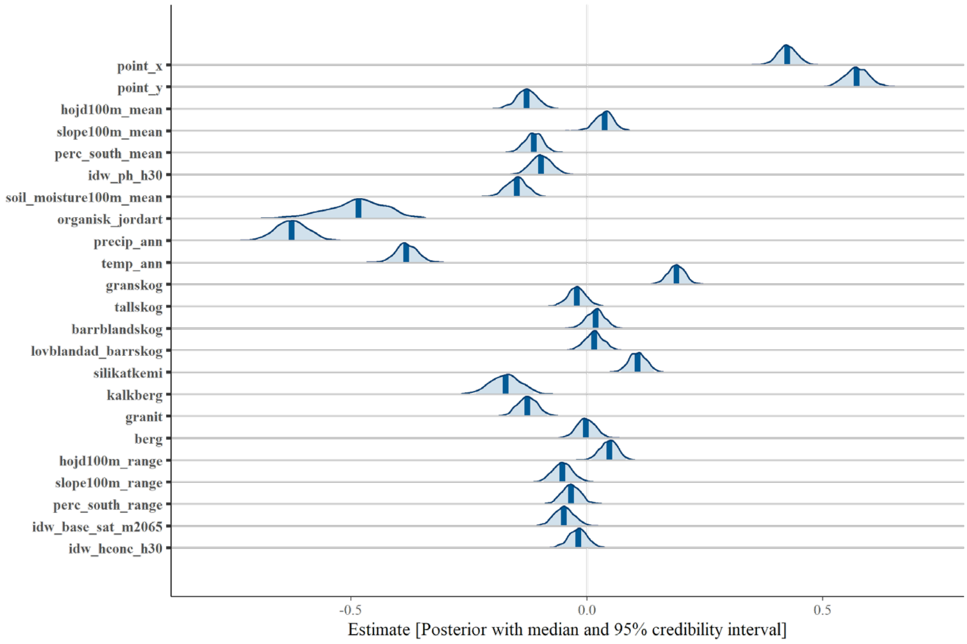
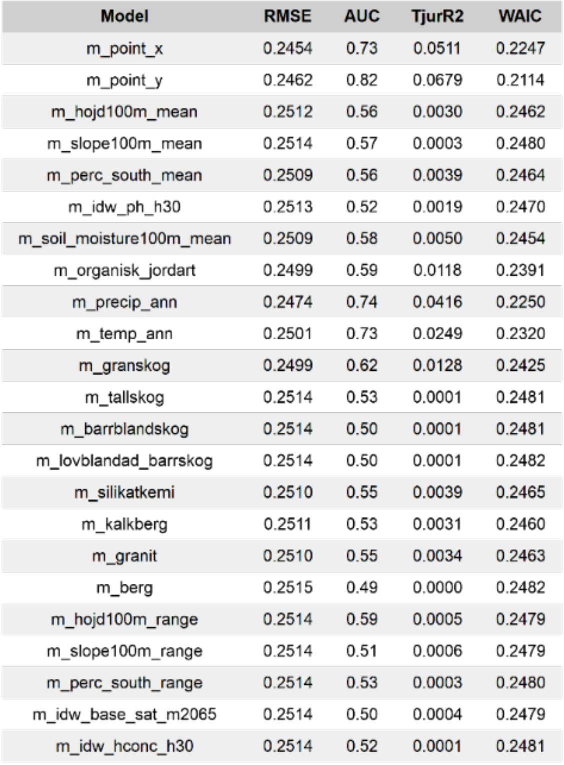


Fig. 2: Posterior distribution of estimates for each explanatory variables in models with only one explanatory variable.

Table 2: Variability in species occurrence explained by individual explanatory variables in models with only one explanatory variable. For example, in a model with only annual precipitation (m_precip_ann) this predictor explained 4% of the variability in species occurrences (Tjur's $R^2 = 0.0416$).



Model	RMSE	AUC	TjurR2	WAIC
m_point_x	0.2454	0.73	0.0511	0.2247
m_point_y	0.2462	0.82	0.0679	0.2114
m_hojd100m_mean	0.2512	0.56	0.0030	0.2462
m_slope100m_mean	0.2514	0.57	0.0003	0.2480
m_perc_south_mean	0.2509	0.56	0.0039	0.2464
m_idw_ph_h30	0.2513	0.52	0.0019	0.2470
m_soil_moisture100m_mean	0.2509	0.58	0.0050	0.2454
m_organisk_jordart	0.2499	0.59	0.0118	0.2391
m_precip_ann	0.2474	0.74	0.0416	0.2250
m_temp_ann	0.2501	0.73	0.0249	0.2320
m_granskog	0.2499	0.62	0.0128	0.2425
m_tallskog	0.2514	0.53	0.0001	0.2481
m_barrblandskog	0.2514	0.50	0.0001	0.2481
m_lovblandad_barrskog	0.2514	0.50	0.0001	0.2482
m_silikatkemi	0.2510	0.55	0.0039	0.2465
m_kalkberg	0.2511	0.53	0.0031	0.2460
m_granit	0.2510	0.55	0.0034	0.2463
m_berg	0.2515	0.49	0.0000	0.2482
m_hojd100m_range	0.2514	0.59	0.0005	0.2479
m_slope100m_range	0.2514	0.51	0.0006	0.2479
m_perc_south_range	0.2514	0.53	0.0003	0.2480
m_idw_base_sat_m2065	0.2514	0.50	0.0004	0.2479
m_idw_hconc_h30	0.2514	0.52	0.0001	0.2481

Building a model with all relevant explanatory variables

From the 23 tested explanatory variables 11 were selected given their higher explanatory power. We further aimed to have one variable for each group (Table 1). The selected variables were included in one model (Fig. 3), and were not correlated (Fig. S2). Hence, each variable explained different variability. The coordinates were not included as we aimed to include these within the random part of the model to make large-scale predictions. Furthermore, it is not the coordinates themselves that are biologically meaningful, but rather the environmental conditions that are changing along these coordinates.

The model explained 16% out of the occurrences (Tjur's $R^2 = 0.16$; Table 3). Also the predictive power was good as cross-validated measures of fit (averaged over the four folds) were very close to the measures of fit. Individual variables explained between 0.4% and 29.4% of the variability, based on variance partitioning.

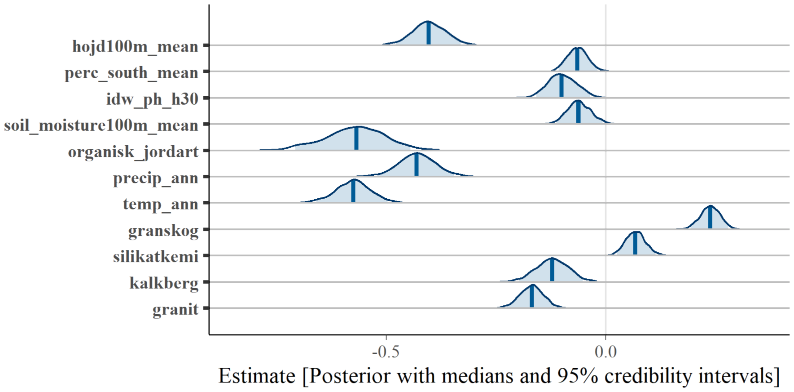
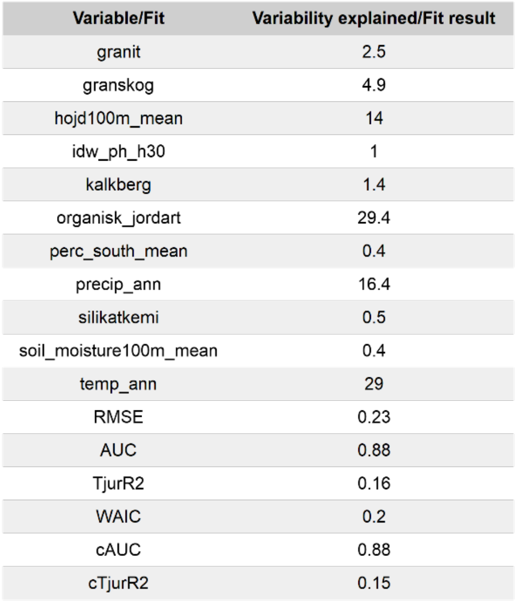


Fig. 3: Posterior distribution of estimates for each variables within the model.

Table 3: Variability in species occurrence explained by individual explanatory variables in models with 11 explanatory variables. For each variable, the percent variability explained based on variance partitioning is shown. The last rows show the explanatory power (Tjur's R^2) and the predictive power of the model (AUC), was well as the average of these over the four-fold cross-validations (cTjur R^2, cAUC).



Variable/Fit	Variability explained/Fit result
granit	2.5
granskog	4.9
hojd100m_mean	14
idw_ph_h30	1
kalkberg	1.4
organisk_jordart	29.4
perc_south_mean	0.4
precip_ann	16.4
silikatkemi	0.5
soil_moisture100m_mean	0.4
temp_ann	29
RMSE	0.23
AUC	0.88
TjurR2	0.16
WAIC	0.2
cAUC	0.88
cTjurR2	0.15

Building a spatial explicit model

The last step of the model building was to additionally include a spatial random effect in the model, which enables us to predict occurrence probability for the used hectare squares as well as over new areas in Sweden. The large number of observations makes it not computationally feasible to use a simple spatial structured random effect with each observations as random level. Hence, we implemented a model with Gaussian Predictive Process (GPP) to account for the spatial structure. This method assumes that the information on the spatial structure can be summarized with a small number of so called 'knot' locations (Fig. S3). The explanatory power slightly increased (Tjur's $R^2 = 0.17$; Table 4) and the importance of three predictors further decreased to nearly zero.

Table 4: Variability in species occurrence explained by a model with 11 explanatory variables and a spatial explicit random effect. For each variable the percent variability explained based on variance partitioning is shown. The last rows show the explanatory power of the model (Tjur's R^2, AUC).

Variable/Fit	Variability explained/Fit result
hojd100m_mean	14
perc_south_mean	0
idw_ph_h30	1
soil_moisture100m_mean	0
organisk_jordart	30
precip_ann	16
temp_ann	28
granskog	5
silikatkemi	0
kalkberg	2
granit	2
Random: norna_id	0.01
RMSE	0.23
AUC	0.90
TjurR2	0.17
WAIC	0.2

With this model we predicted the probability of species presence (posterior mean), which fairly well matches the observed presences (Fig. 4, Fig. S4).

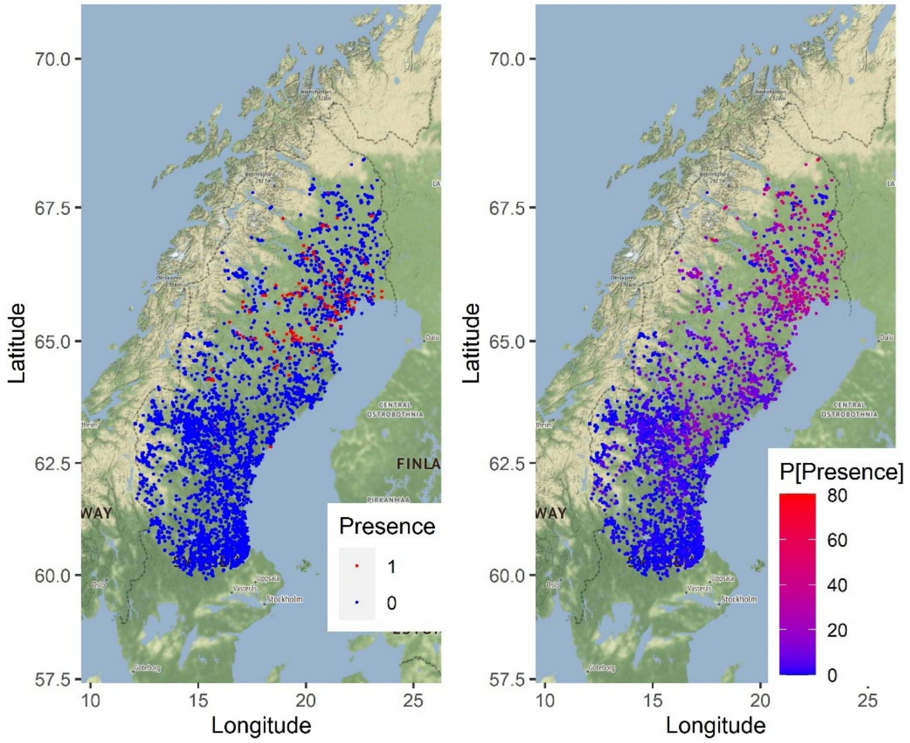


Fig 4: The left shows recorded species presence (red) and inferred species absence (blue). The right shows the occurrence probability in percent.

Predicting over new areas with a spatial explicit model

Using the extracted variables for 22 155 548 quadrats over Sweden, we used the developed model to predict the species occurrence. In order to do so, we standardized the new data using the mean and SD from the data the model was built with. We included only quadrats where the species could exist. These were the ones containing some type of forest but not forests in wetland areas.

For each quadrat we estimated a posterior predictive distribution of the expected values (between 0 and 1), rather than a posterior predictive distribution of the data (0 or 1). From these we calculated the posterior mean, representing the probability of occurrence (with 95% credibility intervals), and the median, representing the most likely value (1 = occurs, 0 = does not occur). These predictions can be viewed in the GIS layer/Artportalen as probability of finding *Calypso* (Fig. S5-S7, Electronic appendix). To increase computation speed the ~22 million quadrats were split into 22 data sets for which occurrence probabilities were estimated separately.

Results and discussion

Using derived variables and pseudo-absences of the species, we were able to build a species distribution model that explained typical amount of variability (~17%) with a good predictive power (AUC = 0.9).

While we accomplished to select potential variables to predict the presence of *C. bulbosa* and could predict its presence over a vast area, one has to bear in mind that this may only still be a rough estimation (e.g. see Fig. S4). Our method of generating pseudo-absences using frequently associated species is considered rather robust (Phillips et al. 2009). Still, we only modelled the potential distribution that could be reached if unlimited by dispersal ability or dispersal barriers (De Kort et al. 2020). However, our aim was to model the potential distribution over new areas and in such a case it is recommended to exclude dispersal-limited absences (Hattab et al. 2017). We further included, as recommended (De Kort et al. 2020), climate and land-use variables. Here, organisk_jordart may be seen as an indicator of soil fertility and nutrient level, which often is affected by human agri- and silviculture.

The first challenge was that only presences are recorded for the species, meaning pseudo-absences had to be inferred, which is highly critical as it can influence the modelling outcome (VanDerWal et al. 2009). It has been shown that only presence data from opportunistic reporting are especially useful at larger spatial scales and can be seen as complementation of systematic collection of data (Henckel et al. 2020). A common bias of opportunistic data is that records more reflect where the reporters are, rather than where the species is, which may lead to an incomplete evaluation of the habitats the species persist in. One approach to address this bias is to use the number of records (e.g. per quadrat) as additional predictor (Andersson et. al. 2015, Stephan & Toräng 2021). It is further possible to improve the opportunistic collected species reports by using check-lists or questionnaires for the reporters (Bradter et al. 2021). Another common method to infer absences is to use background sites, like 1000 forest sites at varying distances from each recorded presence (Greiser et al. 2020). Here we used the presence of 70 plant species that are associated with the focus plant as a method of using background sites. We assumed that focus and associated species have similar habitat niches. Hence, pseudo-absences can be inferred to the location where the focus species is absent, but the associated species is present since there is a good chance it can exist at this site, but was not recorded. A drawback may be that, consequently, the range of the environmental variables is restricted *a priori*. However, our data to build the model are very similar to the data used to predict over new areas (Table S1).

The resulting map shows that the predicted probability of occurrence of *C. bulbosa* is highest in northern Sweden, especially in forested parts of the alpine region and in the northernmost part of the boreal region (Fig. S5). In Norrbotten county and in the alpine biogeographical region, there is a rather high chance of finding new sites since about one quarter of the forested area is predicted to have at least a 2 % probability of hosting *Calypso* (Table 5). At a more detailed scale we get clear indications of where the probability of finding the plant is highest (Fig. S6). However, the contrast between the predicted probabilities and the known presences and pseudo-absences are rather large (Fig. S7, Fig. S4). Hence, we could exclude many sites where the species will most probably not be found and provide first indications where new occurrences could be discovered. Since the model does not include the dynamic predictor forest age, one has to take that into account when searching for *Calypso* because the species appears to be generally disfavoured by clearcuts. Hence, efforts to find the plant should also use the most resent estimation of forest age and future species distribution modelling should include more information on forest characteristics.

The computational effort of this project deserves some discussion also. The estimations and predictions of the models in R and the computations in ArcGIS were very time and resource intensive. For example, predicting over one set of the 22 sites across Sweden took around 15 hours of High-Performance Computing. On the other hand, the (computation intensive) Bayesian estimation enables to account for all uncertainty in the data. Alternative methods may require less computational effort, but we decided to use the model type with the best predictive performance currently available to model species distributions (Norberg et al. 2019).

Lastly, we have done *in silico* model validation (cross-validation, AUC, WAIC) but the planned future sampling may offer one of the rare opportunities of in situ model validation (Williams et al. 2009). During the new monitoring, sampling will be performed in model-predicted hot spots that show highest likelihood of *C. bulbosa* presence. This validation would be possible if also sites are sampled where the model predicted no presence, but the experience in the field or other biological indicators make it still likely that it is present. Hence, false positives and false negatives could be evaluated.

Table 5: The predicted area of hot spots in different modelled probability classes for finding *Calypso bulbosa*, and proportion of forest area according to the land cover map, per county, biogeographical region and in protected (Natura 2000) and not protected areas in northern Sweden. Only areas considered forest on not-wetland in the Swedish landcover map are included.

Prob. class	**5-7 %**		**4-5 %**		**3-4 %**		**2-3 %**		**Sum**	
County	ha	%	ha	%	ha	%	ha	%	ha	%
Dalarna	0	0	0	0	0	0	2 214	0.11	2 214	0.11
Gävleborg	0	0	0	0	0	0	2 066	0.14	2 066	0.14
Västernorrland	0	0	0	0	130	0.01	52 006	2.95	52 136	2.96
Jämtland	0	0	0	0	198	0.01	58 950	1.96	59 147	1.96
Västerbotten	3	0.00	1 465	0.04	29 075	0.81	281 179	7.85	311 722	8.71
Norrbotten	1 403	0.03	17 892	0.39	142 081	3.10	1 040 526	22.72	1 201 902	26.25
SUM	1 406	0.01	19 357	0.12	171 484	1.05	1 436 941	8.77	1 629 187	9.94
Alpine region	1 335	0.06	10 787	0.45	89 326	3.72	565 548	23.54	666 996	27.76
Boreal region	71	0.00	8 570	0.06	82 159	0.59	871 390	6.23	962 190	6.88
Protected	1 247	0.09	11 313	0.78	82 185	5.64	418 760	28.72	513 505	35.22
Not protected	159	0.00	8 043	0.05	89 299	0.60	1 018 181	6.82	1 115 682	7.47

Acknowledgements

The project was funded by Swedish EPA's biogeographical monitoring programme. The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

References

Bradter, U., A. Ozgul, M. Griesser, K. Layton-Matthews, J. Eggers, A. Singer, B. K. Sandercock, P. J. Haverkamp, and T. Snäll. 2021. Habitat suitability models based on opportunistic citizen science data: Evaluating forecasts from alternative methods versus an individual-based model. Diversity and Distributions 27:2397–2411.

De Kort, H., M. Baguette, J. Lenoir, and V. M. Stevens. 2020. Toward reliable habitat suitability and accessibility models in an era of multiple environmental stressors. Ecology and Evolution 10:10937–10952.

Gelman, A., and D. B. Rubin. 1992. Inference from iterative simulation using multiple sequences. Statistical Science 7:457–472.

Greiser, C., K. Hylander, E. Meineri, M. Luoto, and J. Ehrlén. 2020. Climate limitation at the cold edge: contrasting perspectives from species distribution modelling and a transplant experiment. Ecography 43:637–647.

Hattab, T., C. X. Garzón-López, M. Ewald, S. Skowronek, R. Aerts, H. Horen, B. Brasseur, E. Gallet-Moron, F. Spicher, G. Decocq, H. Feilhauer, O. Honnay, P. Kempeneers, S. Schmidtlein, B. Somers, R. Van De Kerchove, D. Rocchini, and J. Lenoir. 2017. A unified framework to model the potential and realized distributions of invasive species within the invaded range. Diversity and Distributions 23:806–819.

Henckel, L., U. Bradter, M. Jönsson, N. J. B. Isaac, and T. Snäll. 2020. Assessing the usefulness of citizen science data for habitat suitability modelling: Opportunistic reporting versus sampling based on a systematic protocol. Diversity and Distributions 26:1276–1290.

Norberg, A., N. Abrego, F. G. Blanchet, F. R. Adler, B. J. Anderson, J. Anttila, M. B. Araújo, T. Dallas, D. Dunson, J. Elith, S. D. Foster, R. Fox, J. Franklin, W. Godsoe, A. Guisan, B. O'Hara, N. A. Hill, R. D. Holt, F. K. C. Hui, M. Husby, J. A. Kålås, A. Lehikoinen, M. Luoto, H. K. Mod, G. Newell, I. Renner, T. Roslin, J. Soininen, W. Thuiller, J. Vanhatalo, D. Warton, M. White, N. E. Zimmermann, D. Gravel, and O. Ovaskainen. 2019. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. Ecological Monographs 89:1–24.

Ovaskainen, O., and N. Abrego. 2020. Joint species distribution modelling: With applications in R. Cambridge University Press.

Phillips, S. J., M. Dudík, J. Elith, C. H. Graham, A. Lehmann, J. Leathwick, and S. Ferrier. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications 19:181–197.

R Core Team. 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Tikhonov, G., L. Duan, N. Abrego, G. Newell, M. White, D. Dunson, and O. Ovaskainen. 2020. Computationally efficient joint species distribution modeling of big spatial data. Ecology 101:e02929.

Tikhonov, G., Ø. H. Opedal, N. Abrego, A. Lehikoinen, M. M. J. de Jonge, J. Oksanen, and O. Ovaskainen. 2019. Joint species distribution modelling with the R-package Hmsc. Methods in Ecology and Evolution:2041–210X.13345.

Tikhonov, G., O. Ovaskainen, J. Oksanen, M. de Jonge, O. Opedal, and T. Dallas. 2021. Hmsc: Hierarchical model of species communities.

Tjur, T. 2009. Coefficients of determination in logistic regression models – a new proposal: The coefficient of discrimination. The American Statistician 63:366–372.

VanDerWal, J., L. P. Shoo, C. Graham, and S. E. Williams. 2009. Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? Ecological Modelling 220:589–594.

Williams, J. N., C. Seo, J. Thorne, J. K. Nelson, S. Erwin, J. M. O'Brien, and M. W. Schwartz. 2009. Using species distribution models to predict new occurrences for rare plants. Diversity and Distributions 15:565–576.

Appendix

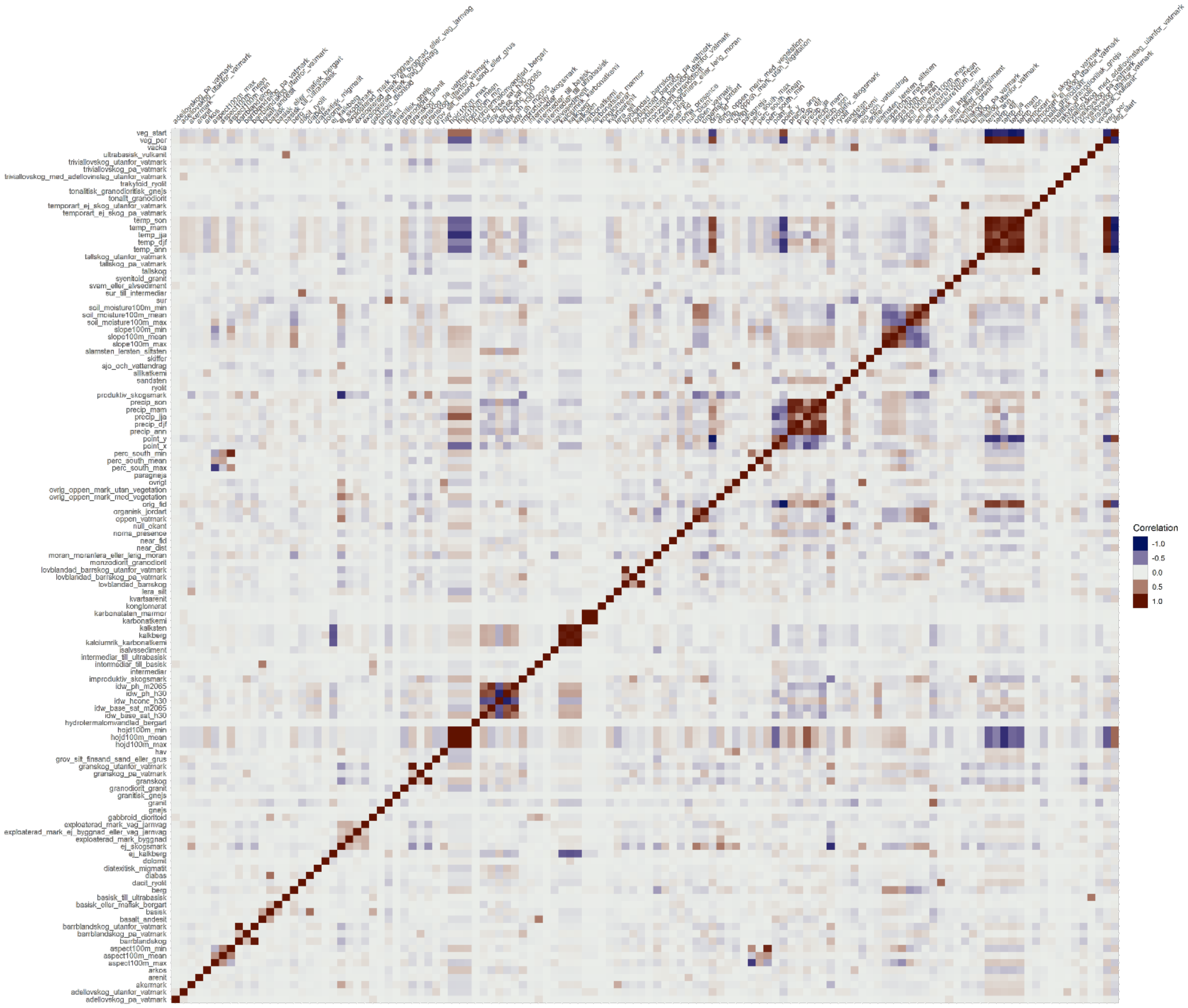


Fig. S1: Pearson correlations among all extracted variables.

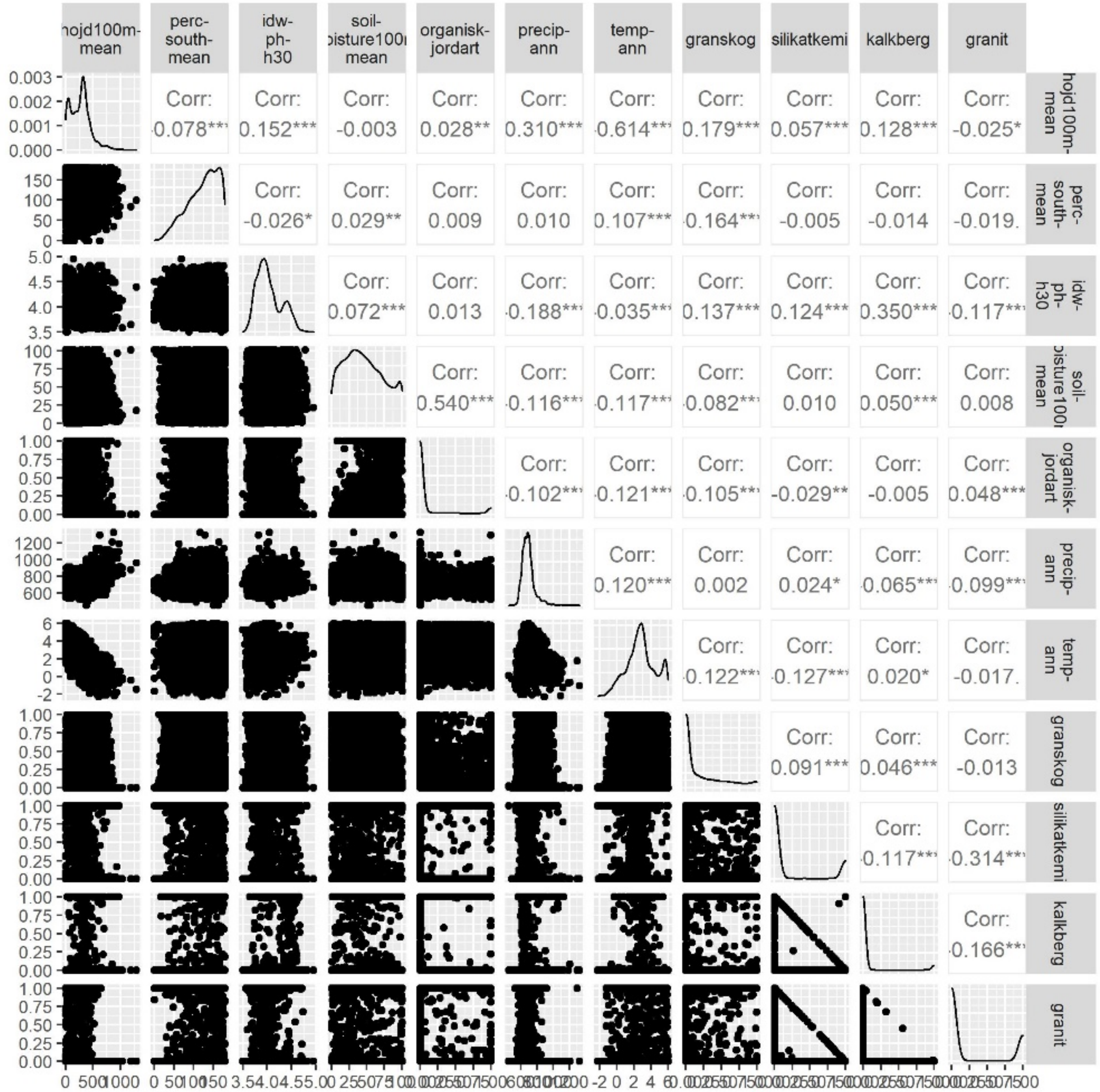


Fig. S2: Correlation among explanatory variables included in one model on original scale (spread can be seen on the y-axis).

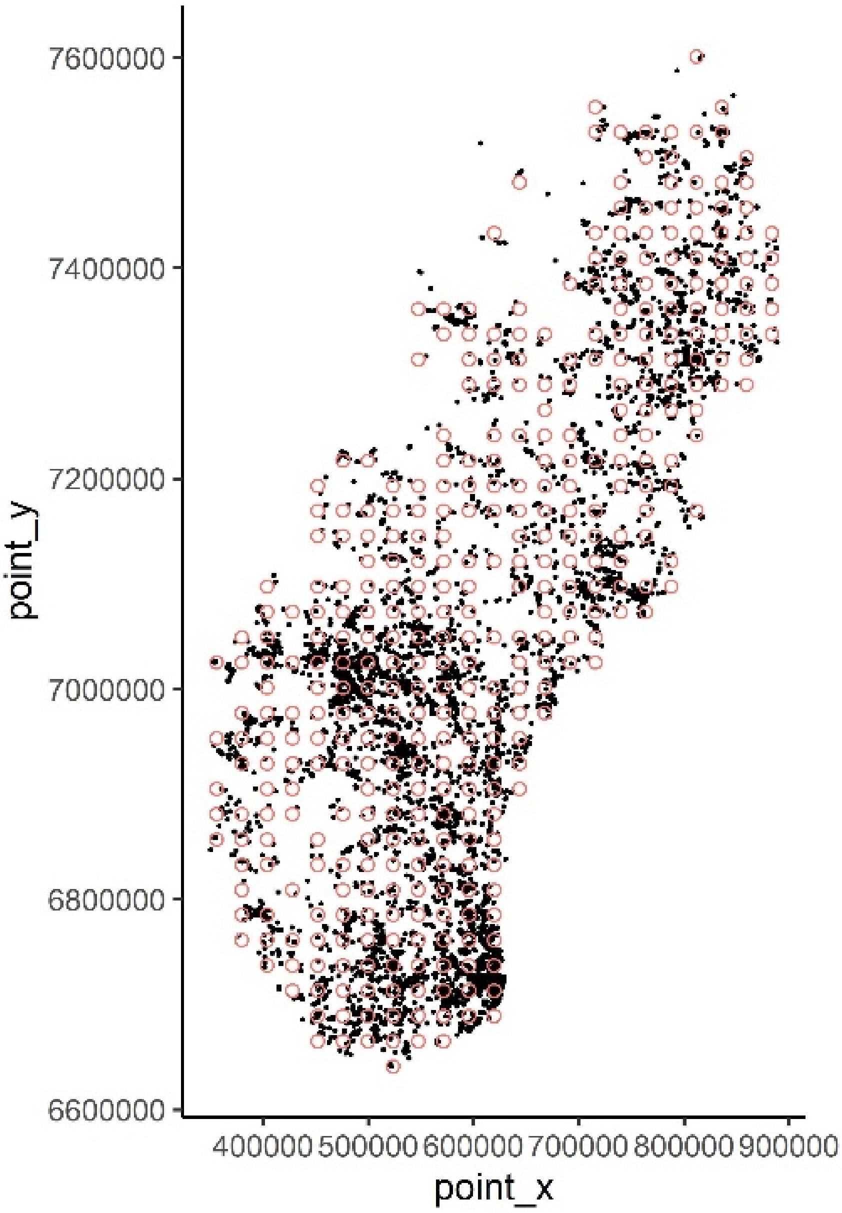
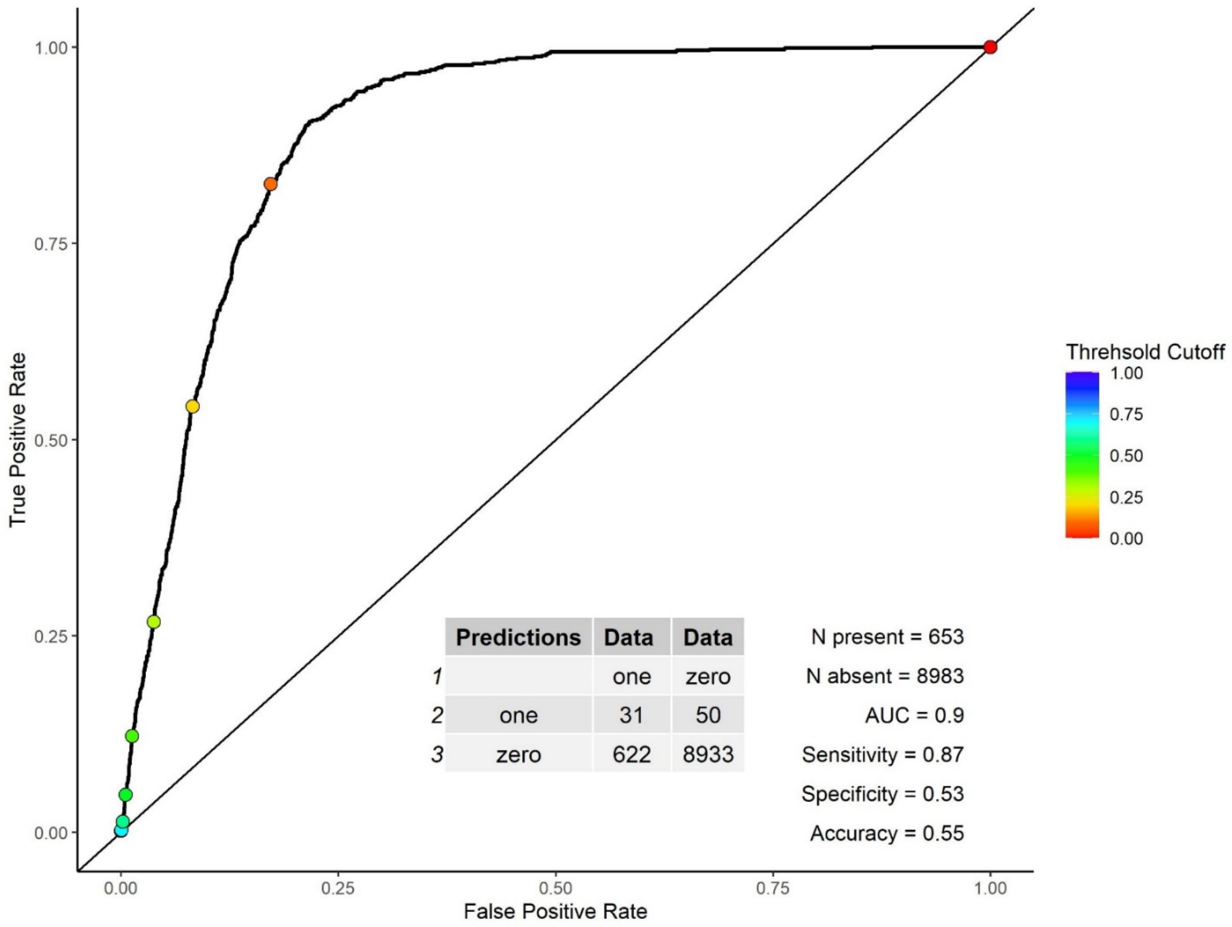


Fig. S3: Locations of used quadrats (black) and 383 knot locations (red) used in Gaussian Predictive Process (GPP).



	Predictions	Data	Data
1		one	zero
2	one	31	50
3	zero	622	8933

N present = 653
N absent = 8983
AUC = 0.9
Sensitivity = 0.87
Specificity = 0.53
Accuracy = 0.55

Fig. S4: ROC courve and model performance estimations for the model in relation to the orgiginal data. Within the plot the confusion matrix shows the frequency of Occurrences (one/present) and Absences (zero/absent) in data to build the model and in predictions over these data. In 8933 cases absences and in 31 cases presences are correctly predicted. In 622 cases the species was present, but the model would predict an absence. In 50 cases the species was absent, but the model would predict a presence.

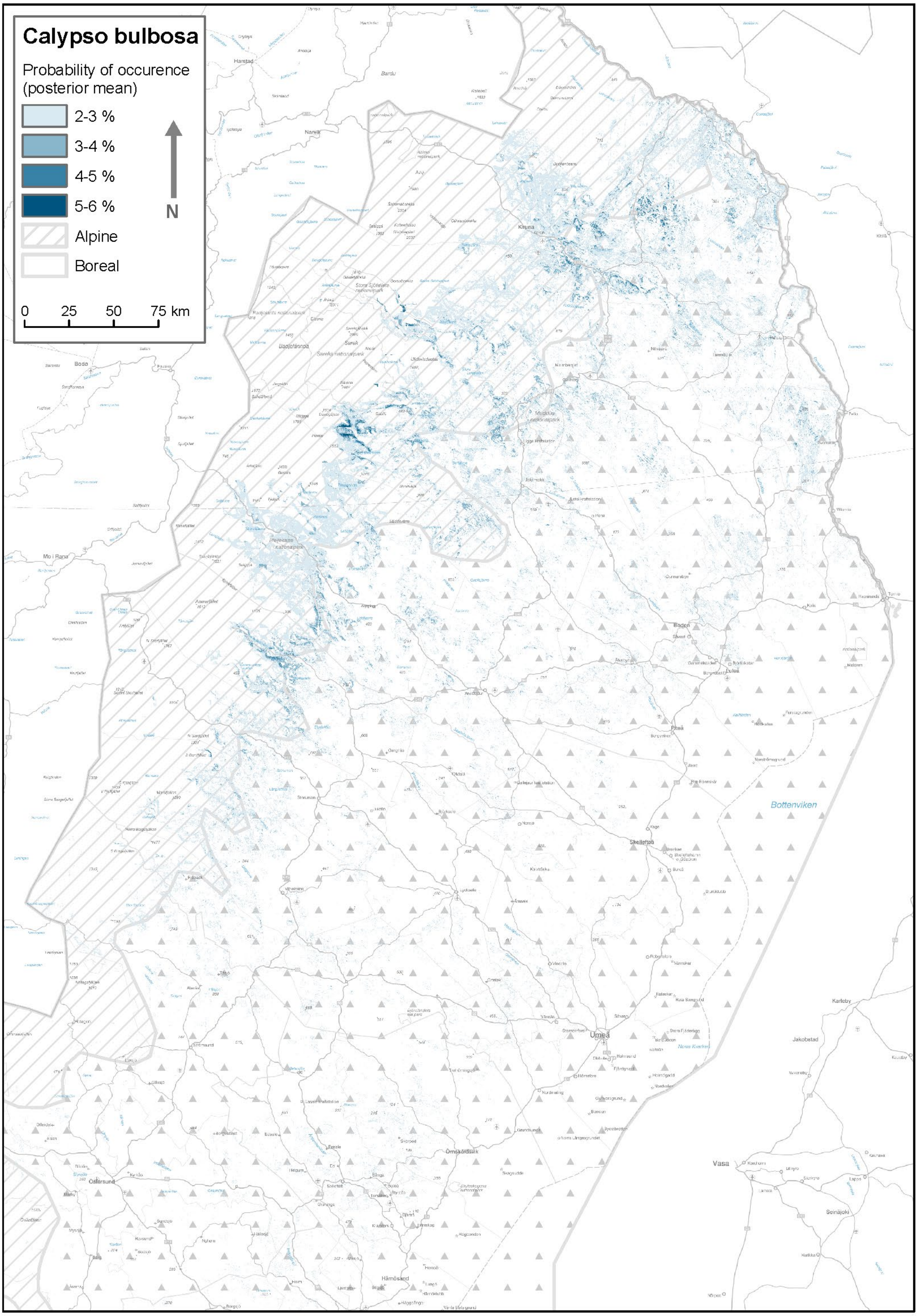


Fig. S5: Predicted probability of occurrence of *Calypso bulbosa* in the alpine and boreal regions of northern Sweden. Background map: © Lantmäteriet.

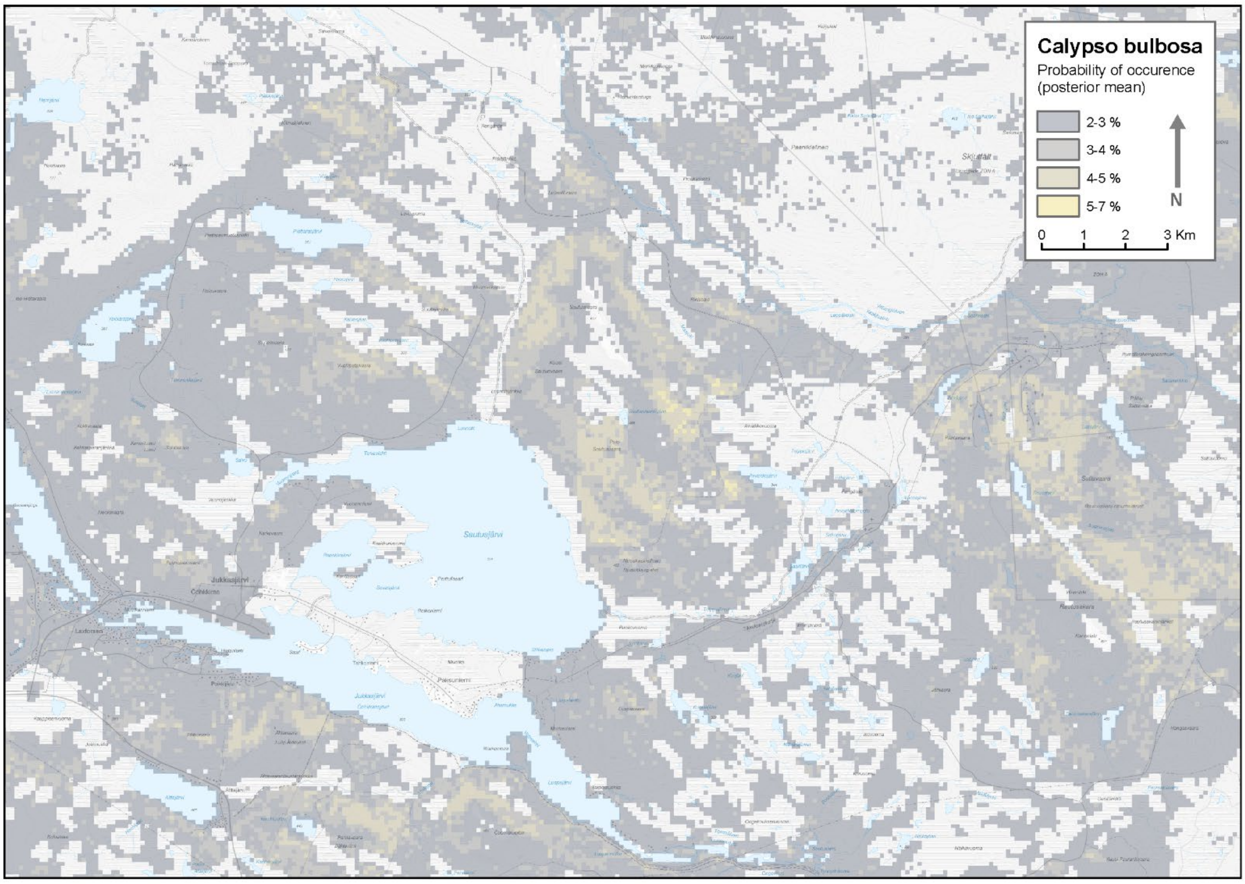


Fig. S6: Predicted probability of occurrence of *Calypso bulbosa* in the area northeast of Jukkasjärvi. Neither presences nor pseudo-absences have been recorded in this area. Background map: © Lantmäteriet.

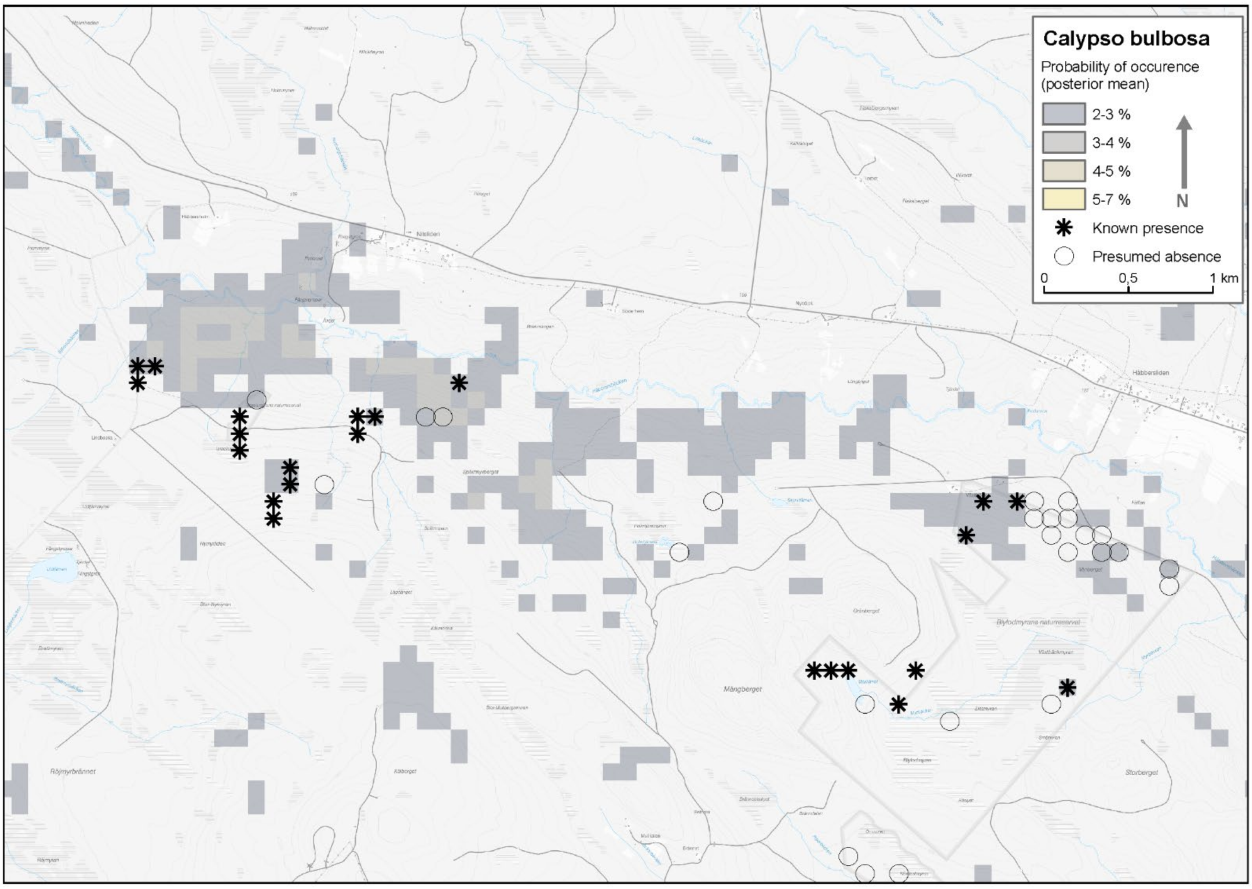
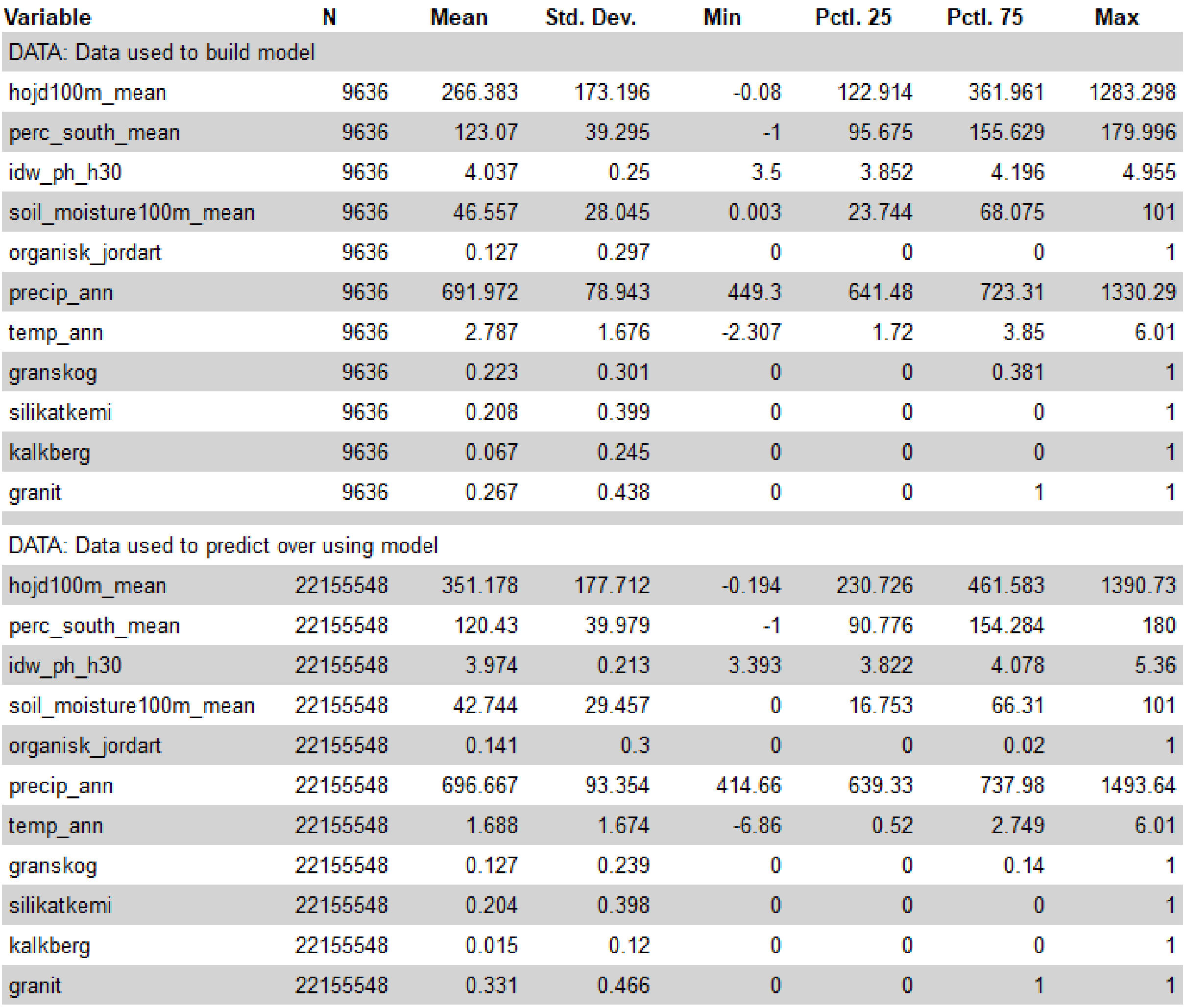


Fig. S7: Predicted probability of occurrence of *Calypso bulbosa* in an area between Skellefteå and Jörn. Recorded presences and pseudo-absences are shown. Background map: © Lantmäteriet.

Table S1: Summary of untransformed data used to build model and untransformed new data used to predict occurrence with developed model.



Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
DATA: Data used to build model							
hojd100m_mean	9636	266.383	173.196	-0.08	122.914	361.961	1283.298
perc_south_mean	9636	123.07	39.295	-1	95.675	155.629	179.996
idw_ph_h30	9636	4.037	0.25	3.5	3.852	4.196	4.955
soil_moisture100m_mean	9636	46.557	28.045	0.003	23.744	68.075	101
organisk_jordart	9636	0.127	0.297	0	0	0	1
precip_ann	9636	691.972	78.943	449.3	641.48	723.31	1330.29
temp_ann	9636	2.787	1.676	-2.307	1.72	3.85	6.01
granskog	9636	0.223	0.301	0	0	0.381	1
silikatkemi	9636	0.208	0.399	0	0	0	1
kalkberg	9636	0.067	0.245	0	0	0	1
granit	9636	0.267	0.438	0	0	1	1
DATA: Data used to predict over using model							
hojd100m_mean	22155548	351.178	177.712	-0.194	230.726	461.583	1390.73
perc_south_mean	22155548	120.43	39.979	-1	90.776	154.284	180
idw_ph_h30	22155548	3.974	0.213	3.393	3.822	4.078	5.36
soil_moisture100m_mean	22155548	42.744	29.457	0	16.753	66.31	101
organisk_jordart	22155548	0.141	0.3	0	0	0.02	1
precip_ann	22155548	696.667	93.354	414.66	639.33	737.98	1493.64
temp_ann	22155548	1.688	1.674	-6.86	0.52	2.749	6.01
granskog	22155548	0.127	0.239	0	0	0.14	1
silikatkemi	22155548	0.204	0.398	0	0	0	1
kalkberg	22155548	0.015	0.12	0	0	0	1
granit	22155548	0.331	0.466	0	0	1	1

Electronic appendix

GIS layer for the probability of finding *Calypso bulbosa* at the hectare level in northern Sweden is available upon request or may be viewed in a specific project in Artportalen.