$_{int}$SDM: A reproducible framework for integrated species distribution models

Philip Mostert[1,2*] Ragnhild Bjørkås[2] Angeline J.H.M. Bruls[2]
Wouter Koch[2,3] Ellen C. Martin[2]

1. Department of Mathematical sciences, Norwegian University of Science and Technology, 7491, Trondheim, Norway; 2. Center for Biodiversity Dynamics, Norwegian University of Science and Technology, 7491, Trondheim, Norway; 3. Norwegian Biodiversity Information Centre, Trondheim, Norway

* *philip.s.mostert@ntnu.no*

Abstract

Integration of data is needed to address many of the problems currently threatening biodiversity. There has been an exponential increase in quantity and type of biodiversity data in recent years, including presence-absence, counts, and presence-only citizen science data. Species Distribution Models (SDMs) are frequently used in ecology to predict current and future ranges of species, and are a common tool used when making conservation prioritization decisions. Current SDM practice typically underutilizes the large amount of publically available biodiversity data and does not follow a set of standard best practices. Integrating data types with open-source tools and reproducible workflows saves time, increases collaboration opportunities, and increases the power of data inference in SDMs. Here, we address the discipline-wide call for open science and standards in SDMs by (1) proposing methods and (2) generating a reproducible workflow to integrate different available data types to increase the power of SDMs. We provide an R package, $_{int}$SDM, which reduces the learning curve for the use of integrated SDMs and makes them an accessible tool for use by non-programmers. We provide code and guidance on how to accommodate users' diverse needs and ecological questions with different data types available on the Global Biodiversity Information Facility (GBIF), the largest biodiversity data aggregator in the world. Finally, we provide a case study of the application of our reproducible workflow by creating SDMs for vascular plants in Norway, integrating presence-only and presence-absence species occurrence data, climate, and habitat data.

Keywords: Citizen science; Data integration; Integrated species distribution model; R package; Reproducible workflow

Introduction

Open science has been put forward as a transparent and reliable way to cope with the large scale of the questions the world needs answered to address the state of climate- and biodiversity crises [1, 2]. The culmination of open science practices is an open and reproducible workflow [3, 4]. Ecological data are typically based on observation, making repeatable data uncommon [5]. A reproducible workflow is a continuation of the reproducible scientific method, and a necessary step when working with different complex data sources, when working collaboratively, or when producing an analytical tool which will be made publicly available. A reproducible workflow allows users to create, share, or replicate others' analyses in a standardized way. Reproducible workflows have been shown to save time, increase sharing, and increase the quality of data [6–8].

Data integration techniques allow researchers to combine data from multiple sources under the same modeling framework, increasing the scope or scale of interest [9, 10]. This integration of data often increases the precision and accuracy of models, and further increases inference about, for example, an organism's distribution in time and space. It also allows data use to extend beyond original intent at the time of collection, blending different sampling methods and dataset properties [11]. One increasingly important source of data is citizen science, but this has some limitations (e.g. taxonomic bias [12, 13]). Additionally, citizen science data generally consists of presence-only data, meaning that it does not provide any information on where a species does not occur [14].

Species Distribution Models (SDMs) are a class of biodiversity models which delineate or forecast a species' range (i.e. distribution) based on occurrence data and environmental data [15]. SDMs are one of the most frequently used and widely reviewed topics in ecological data analysis [16], and they commonly assist managers to make consequential conservation decisions and risk assessments [17]. Best-practice standards for SDMs provide reliability, transparency, consistency, and call for open data and methods used for constructing the SDMs [16]. SDMs are more accurate with presence-absence data than with presence-only data [9]. When both data types are available, point-process models provide an opportunity to incorporate both datasets and create more precise SDMs than if only one data type was used [9]. Point processes provide a flexible framework for developing integrated SDMs, combining data representing the locations of individual organisms and occupancy.

In this paper, we use presence-absence and presence-only data available on GBIF (Global Biodiversity Information Facility, the largest biodiversity data aggregator in the world) and produce a reproducible workflow for integrated SDMs. We provide an R [18] package $_{int}$SDM, aiming to reduce the learning curve for the use of integrated species distribution models (ISDMs) and make them an accessible tool to use by non-programmers. Within our package, we provide code to download data and produce SDMs, along with guidance on how the script and/or data could simply be altered to accommodate the users' diverse needs and ecological questions with variable data types available on GBIF. Herein we highlight a case study of the application of our reproducible

workflow to create species distribution models of vascular plants in Norway by integrating a variety of presence-only and presence-absence species occurrence, climate, and habitat data. We present the step-by-step approach to integrate datasets into a single species distribution model, via an easy-to-use R package ($_{int}$SDM).

Methods

Reproducible Workflow

The steps to create a reproducible workflow for analyzing ecological data are four-fold (Figure 1):

1. Identify and locate sources of data which will be integrated in the model.
2. Clean and filter data based on specific questions and ecological knowledge.
3. Integrate and standardize different data types.
4. Run data through analysis of choice.

In this case study, we specifically create a workflow to create Species Distribution Models (SDMs).

Identify and locate sources of data

Data discovery needs to be documented. This becomes a critical step as data are combined for use in analysis and terms of use and quality control vary. Adequate data documentation is a key component of methods reproducibility [19], and metadata standards have been proposed which encourage tracking information on data authorship, preparation (i.e., cleaning, changing, or validating), and model fitting [20]. Given the wide range of available species' observational data and environmental variable data available for inclusion in SDMs, properly documenting the sources, metadata, and scales are necessary for reproducibility. Documenting data discovery and inclusion is also necessary to assess the comparative value of SDMs when different datasets are used to model the same species' distribution.

In the last decade, numerous large-scale databases have been developed with the purpose of storing, hosting, and making available large numbers of datasets (e.g., Global Biodiversity Information Facility (GBIF), GenBank, and Global Earth Observation System of Systems (GEOSS)). These data collections are collated geographically or by subject and data from these sources are increasingly used to answer large-scale biological (i.e., macroecological) questions. Open data tend to have similar characteristics, with the majority being presence-only data, citizen-science collected, and heterogeneous in data-collection procedures [21].

We here create a tool which aids users in creating SDMs from data available on GBIF. This data repository was chosen because of its current prevalence in the SDM community, its volume of different data types, and its high standard

of assessment (for reviews of GBIF data used in peer reviewed literature, see Heberling *et al.* [10]). Data available on GBIF lend themselves to integration into reproducible workflows for a multitude of reasons. Data and all occurrence downloads are assigned a Digital Object Identifier (DOI), which always resolves to dataset or download pages. GBIF data follow data standards set by Biodiversity Information Standards [22], such as Darwin Core, Ecological Metadata Language, and the Biological Collection Access Service standards [10]. Finally, these data are all open and FAIR (i.e., findable, accessible, interoperable, and reusable) [3].

Biogeographical data are compiled in step 1 of the workflow to address specific research questions about biodiversity. Here, we compile data about species occurrences, abundance, geographical distribution, functional traits of organisms, etc. from GBIF. The data compiled in step 1 are likely provided at varying resolutions, with most tending to be disaggregated [23]. In addition to biogeographical data, this step of the workflow includes compiling environmental data, habitat data, or other variable data to be included in the models. Possible sources of data include WorldClim, LiDAR, etc.

Process data

This step allows users to clean and filter data based on quality and specific questions or ecological knowledge. During this step of the workflow, users process both occurrence data and environmental data obtained in step 1. Resources and R packages exist for data cleaning [24], standardization and validation [25], and formatting metadata (e.g. Merow *et al.* [20]).

Different data sets will often have different scales or resolutions. Spatial scale can vary between several datasets along resolution (i.e., the spatial size of individual sampling locations) and extent (i.e., the total study area [26]), but need to be standardized for use in SDMs. This different scaling influences the environmental variables available for inclusion. During this step, spatial aggregation of data with smaller grain sizes occurs. Finally, during this step a user filters data based on, for example, species name, locality of interest, or ecological parameters of interest.

Integrated Species Distribution Modeling

As an addition to this paper, we created $_{int}$SDM, an R package to help facilitate the modeling of point process integrated SDMs in a reproducible format. Table 1 presents an overview of SDM R packages currently used across the field of ecology; our package differs by allowing modelers to integrate a collection of disparate datasets in a way that is not only reproducible, but also easy to use for those with a minimal background in statistics and R programming.

$_{int}$SDM is built around the R-INLA package [27], which uses the integrated nested Laplace approximation (INLA) methodology [28] to approximate Bayesian latent Gaussian models. INLA is preferred to traditional Bayesian approximation techniques for continuous space models, for its speed and computational efficiency in estimating Gaussian random fields via the stochastic partial differential

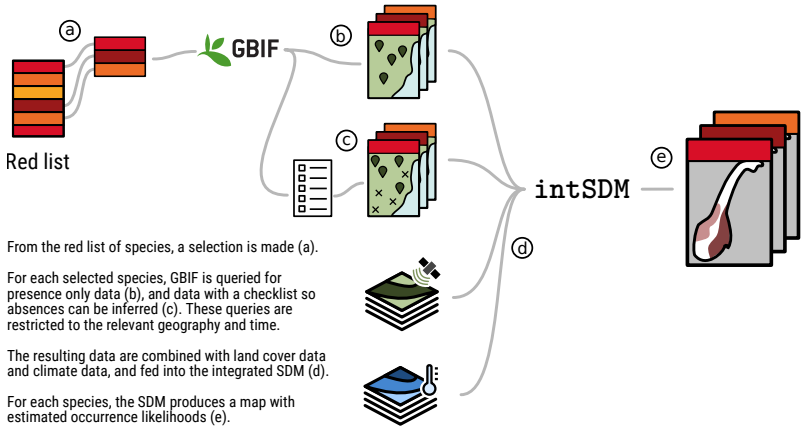

Figure 1. Workflow steps begin with a selection of species (a), a GBIF query of presence-only data (b) and presence-absence data (c), and a combination of landcover data (d). The package $_{int}$SDM is used in the final step to estimate occurrence likelihoods (e).

equation (SPDE) approach [29], which are used to account for unmeasured covariates and potential autocorrelations. The package also uses inlabru [30], an R package which builds wrapper functions around R-INLA to assist in the construction and estimation of spatial point process models.

The $_{int}$SDM package is available on CRAN [31]. Comments and concerns regarding the package may be addressed on the package's GitHub repository: `https://github.com/PhilipMostert/intSDM`.

Framework behind the integrated model

Our integrated model is defined as in Isaac *et al.* [9], where the hierarchical state-space model is the combination of a process model, used to reflect the "true" distribution of the species as well as separate observation models for each of the datasets, where the likelihood for each of these models is dependent on the underlying sampling protocol of the individual dataset. Here, we treat the species' "true" distribution as an inhomogeneous Poisson process, with an intensity function described by spatially varying covariates and a spatially varying Gaussian random field used to account for unmeasured covariates and potential spatial autocorrelation in the model.

The observation processes for our state-space model depend on the type of data at hand, which typically come in the form of: count data, presence-absence or presence-only data. Following Isaac *et al.* [9] we model our count data in this framework as a Poisson point process with expected value given by the integral of the intensity function over the study area, and the presence-absence data with a binomial likelihood with a cloglog link function, which allows us to relate the

probability of occurrence of the species to the log of the intensity function. Given that citizen science data are often collected opportunistically and may contain a multitude of different sampling biases [32], we therefore model the presence-only data from a thinned Poisson process to account for imperfect detection in the collection process.

Output interpretation and use

The outputs of these models give us an understanding of how the underlying environment drives species richness and distribution across an area, and are typically applied to some form of biodiversity assessment (such as selecting areas for restoration and protection).

These results can then be used to make future projections on which species are most sensitive to what environmental conditions, and how the changes in these conditions could force range shifts across time. Such knowledge could be useful e.g., for management practitioners when identifying important areas for protection of biodiversity in the future, or when assessing risks of invasive species [17].

A popular output from these types of models are prediction maps of the intensity function of the model, which gives a reflection of the relative abundance of the species across a map; adding a temporal-component to the model could reflect how this abundance changes over a time period.

However estimates may be heavily skewed when not accounting for biases in low-quality data, such as presence-only data (see for example [33]), and thus these biases need to be accounted for prior to predictions in order to obtain more reliable estimates. For example presence-absence data are often used to account for spatial biases commonly found in opportunistic data. Simmonds *et al.* [34] showed that accounting for biases in presence-only data with a bias-adjusted covariate or, in the absence of such, a second spatial field, provided more reliable predictions than those models which did not account for biases.

Name of package	**Description**	**Reference**
biomod2	A collection of popular tools and algorithms used to construct an ensemble of species distribution models.	`https://cran.microsoft.com/snapshot/2016-05-25/web/packages/biomod2/biomod2.pdf`
dismo	Provides a variety of tools to assist ecologists in the analysis of species distribution models.	`https://cran.r-project.org/web/packages/dismo/dismo.pdf`

Hmsc	Package used to run joint species distribution models, allowing ecologists to combine community data, ecological covariates, trait data and phylogenetic relationships.	https://doi.org/10.1111%2F2041-210X.13345
zoon	Allows ecologists to create and share species distribution models through a reproducible workflow.	https://besjournals.onlinelibrary.wiley.com/doi/pdfdirect/10.1111/2041-210X.12858
Wallace	R-based GUI using species occurrence and environmental data to build SDM's (using dismo) and visualize results in a map.	https://doi.org/10.1111/2041-210X.12945

Table 1. A selection of R packages which are frequently used in the analysis and construction of species distribution models

Case study

As a case study (available as a vignette in the Github repository for $_{int}$SDM), we use our reproducible workflow for integrating data to create a species distribution map for vascular plant species in Norway. Such a map could be used for several purposes, for instance when locating new areas to protect for conservation. According to the Norwegian Red List of 2021, 22% of vascular plants in Norway are considered threatened, i.e., placed in the Red List categories Critically Endangered (CE), Endangered (EN) or Vulnerable (VU) [35]. A major goal when designating new protected areas, both in Norway [36] and globally [37] is to protect species diversity and prevent extinction. It is essential to be able to accurately map the distribution of threatened species to locate areas of importance. To do this, all available data in various standards and qualities must be combined in an effective way.

Here, we utilize GBIF presence-only and presence-absence data collected on vascular plants in Norway. The presence-absence data are collected and provided by NTNU University Museum [38] and University of Oslo [39], using standardized cross-lists containing most vascular plants in Norway which, in addition to providing information about presences, also allows for inference of species absences. In this case study, we have selected the three most abundant species in the data also categorized as being threatened on the Norwegian Red List [35]; *Arnica montana* L. (VU), *Fraxinus excelsior* L. (VU), and *Ulmus glabra* Huds. (EN). We combined the species occurrence data with one environmental variable (annual mean temperature) in the ISDM to produce a map with estimated relative abundance for each species.

Results

We completed our example analysis using the functions developed for our package; and the script used for the analysis is provided in the 'vignettes' folder in our GitHub repository.

The output from this script were distribution maps reflecting the predicted intensity for the three studied species (Figure 2). Since no species records were available for the northern part of Norway, the predicted intensity function for those areas is constant. However the predictions suggest some abundance of our three species towards the south-western part of Norway, especially with regards to *Arnica montana*, where the intensity was the largest.

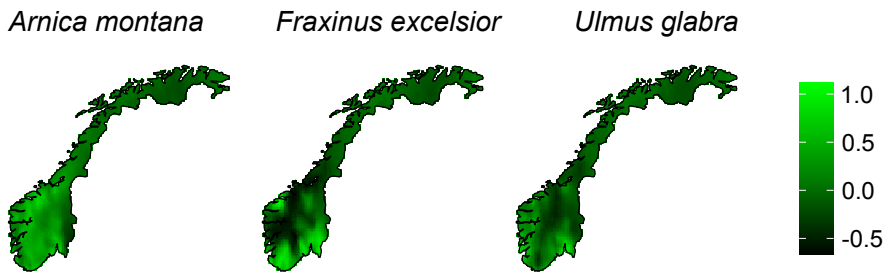

Figure 2. Mean of the predicted intensity for the three species, *A. montana*, *F. excelsior* and *U. glabra* in Norway, from SDMs based on presence absence and presence-only data for each respective species, and annual mean temperature.

Discussion

$_{int}$SDM provides an open, reproducible, and flexible workflow integrating different data types from multiple sources to use integrated SDM's to construct maps of species distributions. Specifically, it facilitates the combination of presence-only and presence-absence data, which has the potential to increase the quality of SDM's. In the presented case study, we used $_{int}$SDM to produce maps of predicted species distributions for three red listed vascular plant species in Norway. Such a map could provide a knowledge basis for biodiversity conservation. Given the flexibility of $_{int}$SDM, it can also be used to predict distributions of other taxa and geographical areas of interest.

Furthermore, reproducible workflows like $_{int}$SDM have the potential to implement and thus promote best practice standards in the model building process of SDMs (such as those proposed by Araújo *et al.* [40] and Zurell *et al.* [19]), which have been sketched out in order to improve the overall transparency of reproducible workflows, raise awareness of model quality and reduce potential misuse of model outputs.

Despite the substantial added value of integrating several sources of data in SDMs, the use of disparate data may still cause some challenges. Presence-only

data are often the most abundant data source, but observations are likely to be biased towards conveniently located, densely populated areas [32] or certain species [12, 13]. Integrating presence-absence and presence-only data can increase the quality of SDMs, but how much depends on the amount available presence-absence data, and the ability to correct for bias in the presence-only data [34]. Datasets of different sizes and quality affect results by producing biased estimates towards the more abundant data sources [41]. Common solutions involve sub-sampling or down-weighting the more abundant datasets - however these methods are often subjective and can lead to different conclusions. For species with fewer occurrences, issues connected to low data availability may arise, such as inflating these species' uncertainty [41]. This causes problems when constructing SDMs for particularly rare species. In our case study, this issue was avoided by choosing species for which we had the most data, although in cases where rare species are used this must be accounted for.

For an open workflow to be truly reproducible, data need to be open too; therefore there is an ongoing effort towards making biodiversity data available. Data from GBIF are FAIR [3] and formatted according to prevailing standards such as Darwin Core [42]. Despite this, there was deficient documentation and metadata for species occurrence data in our case study, requiring consultation with data owners to understand the collection methods. A lack of necessary documentation is not unique to this specific data set, highlighting that there are still challenges to resolve to make biodiversity science truly open and reproducible. Understanding the output of the model properly is essential when using a tool made by others. This is not unique for $_{int}$SDM but rather applies to all available R-packages or tools for data analyzes. Providing properly documented scripts and a vignette illustrating the use of $_{int}$SDM is the first step to mitigate this issue. It is also important to understand the data included in SDMs in order to account for weaknesses and data quality differences between the datasets, and to make sure data are properly prepared before use. Knowledge about the species or system under study is also key to identify and include the necessary data to produce high quality SDMs, and make sure that the produced output is sensible. The package presented here maximizes usefulness of available open source data through integration, and provides a structured framework for other users to follow in order to continue producing open, reproducible workflows.

Acknowledgements

We are grateful to Anders Finstad and Bob O'Hara for their supervision and input on the manuscript, and Rune Sørås for his contributions to the early stages of the methodology development.

Data availability

All code and data used in this manuscript are freely available in the *R* package, $_{int}$SDM, which is available on the Comprehensive R Archive Network: `https://CRAN.R-project.org/package=intSDM` [31]. The *R* code is also available on GitHub: `https://github.com/PhilipMostert/intSDM`. The version of the package used for this manuscript (v1.0.0) is archived on Zenodo; `https://doi.org/10.5281/zenodo.7078852` [43].

Competing interests and funding

The authors declare no competing interest. All authors are affiliated with the Center for Biodiversity Dynamics at the Norwegian University of Science and Technology, funded by the Research Council of Norway (SFF-III 223257). WK was funded by the Research Council of Norway (grant 272947).

Author contributions

All authors: Conceptualization, Methodology, Validation, Writing - Original Draft, Writing - Review & Editing. Additionally, WK: Visualization, PM: Software.

References

1. IPCC Working Group II. *Climate Change 2022: Impacts, Adaptation, and Vulnerability* (eds Pörtner, H.-O. *et al.*) (Cambridge University Press, 2022).
2. IPBES. *Thematic assessment of the sustainable use of wild species of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services* en (eds Fromentin, J.-M. *et al.*) `https://doi.org/10.5281/zenodo.6448567` (2022) (IPBES Secretariat, Bonn, Germany, July 2022).
3. Wilkinson, M. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3** (Mar. 2016).
4. Powers, S. M. & Hampton, S. E. Open science, reproducibility, and transparency in ecology. *Ecological Applications* **29,** e01822 (2019).
5. Schnitzer, S. A. & Carson, W. P. Would ecology fail the repeatability test? *BioScience* **66,** 98–99 (2016).
6. Donoho, D. L. An invitation to reproducible computational research. *Biostatistics* **11,** 385–388 (2010).
7. Lind, E. M. Unified data management for distributed experiments: A model for collaborative grassroots scientific networks. *Ecological informatics* **36,** 231–236 (2016).

8. Fidler, F. *et al.* Metaresearch for evaluating reproducibility in ecology and evolution. *BioScience* **67,** 282–289 (2017).

9. Isaac, N. J. *et al.* Data integration for large-scale models of species distributions. *Trends in ecology & evolution* **35,** 56–67 (2020).

10. Heberling, J. M., Miller, J. T., Noesgaard, D., Weingart, S. B. & Schigel, D. Data integration enables global biodiversity synthesis. *Proceedings of the National Academy of Sciences* **118,** e2018093118 (2021).

11. Miller, D. A., Pacifici, K., Sanderlin, J. S. & Reich, B. J. The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution* **10,** 22–37 (2019).

12. Amano, T., Lamming, J. D. & Sutherland, W. J. Spatial gaps in global biodiversity information and the role of citizen science. *Bioscience* **66,** 393–400 (2016).

13. Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R. & Legendre, F. Taxonomic bias in biodiversity data and societal preferences. *Scientific reports* **7,** 1–14 (2017).

14. Dorazio, R. M. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography* **23,** 1472–1484 (2014).

15. Austin, M. P. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological modelling* **157,** 101–118 (2002).

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

17. Guisan, A. *et al.* Predicting species distributions for conservation decisions. *Ecology letters* **16,** 1424–1435 (2013).

18. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2022). `https://www.R-project.org/`.

19. Zurell, D. *et al.* A standard protocol for reporting species distribution models. *Ecography* **43,** 1261–1277 (June 2020).

20. Merow, C. *et al.* Species' range model metadata standards: RMMS. *Global Ecology and Biogeography* **28,** 1912–1924 (2019).

21. Reichman, O. J., Jones, M. B. & Schildhauer, M. P. Challenges and opportunities of open data in ecology. *Science* **331,** 703–705 (2011).

22. International Working Group on Taxonomic Databases. *TDWG* `www.tdwg.org` (2022).

23. König, C. *et al.* Biodiversity data integration—the significance of data resolution and domain. *PLoS biology* **17,** e3000183 (2019).

24. Ribeiro, B. R. *et al.* bdc: A toolkit for standardizing, integrating and cleaning biodiversity data. *Methods in Ecology and Evolution* (2022).

25. De Lima, R. A., Sánchez-Tapia, A., Mortara, S. R., Ter Steege, H. & de Siqueira, M. F. plantR: An R package and workflow for managing species records from biological collections. *Methods in Ecology and Evolution* (2021).

26. Levin, S. A. The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture. *Ecology* **73,** 1943–1967 (1992).

27. Martins, T. G., Simpson, D., Lindgren, F. & Rue, H. Bayesian computing with INLA: new features. *Computational Statistics & Data Analysis* **67,** 68–83 (2013).

28. Rue, H., Martino, S. & Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)* **71,** 319–392 (2009).

29. Lindgren, F., Rue, H. & Lindström, J. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73,** 423–498 (2011).

30. Bachl, F. E., Lindgren, F., Borchers, D. L. & Illian, J. B. inlabru: an R package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution* **10,** 760–766 (2019).

31. Mostert, P., Bjørkås, R., Bruls, A., Koch, W. & Martin, E. C. *intSDM: Reproducible Integrated Species Distribution Models Across Norway using 'INLA'* version 1.0.0. Sept. 2022. `https://CRAN.R-project.org/package=intSDM`.

32. Sicacha-Parada, J., Steinsland, I., Cretois, B. & Borgelt, J. Accounting for spatial varying sampling effort due to accessibility in Citizen Science data: A case study of moose in Norway. *Spatial Statistics* **42,** 100446 (2021).

33. Gelfand, A. E. & Shirota, S. Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. *Ecological Monographs* **89,** e01372 (2019).

34. Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J. & O'Hara, R. B. Is more data always better? A simulation study of benefits and limitations of integrated distribution models. *Ecography* **43,** 1413–1422 (2020).

35. Norwegian Biodiversity Information Centre. *Norsk rødliste for arter 2021* (Norwegian Biodiversity Information Centre, Trondheim, 2021).

36. Naturmangfoldloven. *Lov om forvaltning av naturens mangfold §33* `https://lovdata.no/lov/2009-06-19-100/%C2%A733` (2022).

37. International Union for Conservation of Nature. *The IUCN Red List of Threatened Species* `https://www.iucnredlist.org` (2022).

38. Norwegian University of Science and Technology. *Vascular plant field notes, NTNU University Museum. Version 1.101.* 2021. `https://doi.org/10.15468/kkb2x0`.

39. University of Oslo. *Vascular Plants, Observations, Oslo (O). Version 1.181.* 2019. https://doi.org/10.15468/tvnjk7.

40. Araújo, M. B. *et al.* Standards for distribution models in biodiversity assessments. *Science Advances* **5,** eaat4858 (2019).

41. Zipkin, E. F. *et al.* Addressing data integration challenges to link ecological processes across scales. *Frontiers in Ecology and the Environment* **19,** 30–38. https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/fee.2290 (2021).

42. Wieczorek, J. *et al.* Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLOS ONE* **7,** 1–8. https://doi.org/10.1371/journal.pone.0029715 (Jan. 2012).

43. Mostert, P., Bjørkås, R., Bruls, A., Koch, W. & Martin, E. C. *intSDM: Reproducible Integrated Species Distribution Models Across Norway using 'INLA'* version 1.0.0. Sept. 2022. https://doi.org/10.5281/zenodo.7078852.