ARPHA Preprints

PENSOFT

Forum Paper

Author-formatted document posted on 05/04/2024

Published in a RIO article collection by decision of the collection editors.

DOI: https://doi.org/10.3897/arphapreprints.e124640

Prototype Biodiversity Digital Twin: Real-time bird monitoring with citizen science data

Julian Lopez Gordillo, Patrik Lauha, Ari Lehtiö, Ossi Nokelainen, Anis Rahman, Allan Souza, Jussi Talaskivi, Gleb Tikhonov, Aurélie Vancraeyenest, Otso Ovaskainen

ARPHA Preprints *Author-formatted document posted on 05/04/2024*. DOI: https://doi.org/10.3897/arphapreprints.e124640

Prototype Biodiversity Digital Twin: Real-time bird monitoring with citizen science data

Authors

Julian Lopez Gordillo[1], Patrik Lauha[2], Ari Lehtiö[3], Ossi Nokelainen[4,5], Anis U. Rahman[5], Allan T. Souza[6], Jussi Talaskivi[3], Gleb Tikhonov[2], Aurélie Vancraeyenest[7] and Otso Ovaskainen[5,2*]

1 Naturalis Biodiversity Center, Leiden, The Netherlands

2 Organismal and Evolutionary Biology Research Programme, Faculty of Biological and Environmental Sciences, P.O. Box 65, FI-00014 University of Helsinki, Helsinki, Finland

3 Digital Services, University of Jyväskylä, P.O. Box 35, 40014 Jyväskylä, Finland

4 Open Science Centre, University of Jyväskylä, P.O. Box 35, 40014 Jyväskylä, Finland

5 Department of Biological and Environmental Science, P.O. Box 35, FI-40014 University of Jyväskylä, Jyväskylä, Finland.

6 Institute for Atmospheric and Earth System Research INAR, Forest Sciences, Faculty of Agriculture and Forestry, P.O. Box 27, 00014 University of Helsinki, Helsinki, Finland

7 CSC – IT Center for Science Ltd., P.O. Box 405, 02101 Espoo, Finland

*Corresponding author: email otso.t.ovaskainen@jyu.fi

Abstract

Bird populations respond rapidly to environmental change making them excellent ecological indicators. Climate shifts advance migration, causing mismatches in breeding and resources. Understanding these changes is crucial to monitor the state of environment. Citizen science offers vast potential to collect biodiversity data. We outline a project that combines citizen science with AI-based bird sound classification. The mobile app records bird vocalizations that are classified by AI and stored for re-analysis. Also, it shows a shared observation board that visualizes collective classifications. By merging long-term monitoring and modern citizen science, this project harnesses both approaches' strengths for comprehensive bird population monitoring.

Keywords

Citizen science, bird monitoring, acoustic monitoring, artificial intelligence, species distribution modelling

Introduction

Bird populations are showing rapid and alarming responses to environmental change. One highly conspicuous phenomenon is that of bird migration, in particular the arrival of migratory birds to Europe during spring. Due to climate change, these migratory events are rapidly shifting to earlier, creating ecological mismatches e.g. between the timing of breeding and resource availability. The ongoing rapid changes in bird populations make it increasingly relevant to better understand the mechanisms driving such changes, and to continuously monitor the fate of bird populations.

A great number of people are interested in birds, and citizen science has a long history in bird research. While citizen science projects have provided huge amounts of valuable biodiversity data, a high proportion of the data provided by citizen science projects suffers from common fundamental limitations. One such limitation is variation in the skills of the observers in species identification, leading to high rates of both false positives and false negatives. Another such limitation is spatiotemporal bias in observation effort, as citizen science projects are typically not based on systematic or randomized sampling schemes but rather on opportunistic sampling. As variation in observer skills and bias in sampling effort can be difficult to quantify and report in the metadata, their effects are often difficult to correct for while using the data to scientific inference, potentially leading to biased inference. Despite these limitations, citizen science has great potential, as it can produce much larger datasets than data acquired by professional researchers (Vohland et al. 2021).

This project aims to combine long-term bird monitoring programs with citizen science to make the best out of the two worlds. To avoid some of the common pitfalls of citizen science projects, the data are not based on the identifications made by the citizens, but by a new mobile phone application MK (acronym of the Finnish name of the application ”Muuttolintujen kevät”, meaning Spring of Migratory Birds) that we developed for the purpose of this project. The birds' vocalizations in the audio recorded by the phone app are identified and classified by an AI-based backend (Lauha et al. 2022), removing variation in observer skills in species identification. As the application submits not only the classifications but also the raw audio files to the server where it is stored, the data can be reanalyzed

with progressively improving classification models. The application was launched in spring 2023, attracting 140,000 users who submitted 3 million recordings during 2023. While data from 2023 was acquired with an opportunistic recording scheme, in spring 2024 we published a new version that enables citizen scientists to submit also standardized point counts in preselected locations.

The phone application implements a common observation board where the classifications obtained collectively by all users can be visualized. A key aim of the project, which is still to be implemented, is to use the citizen science observations to generate continuously updating predictions of bird spatiotemporal distributions and singing activity.

Objectives

The objective of this Biodiversity Digital Twin prototype is to investigate if and how citizen science can be employed to real-time bird monitoring, in a way that produces robust data also for scientific analyses. To achieve this, we aim to make the data compatible with existing long-term data on birds by implementing a point count module and generating calibration data by conducting point counts simultaneously by bird experts and by the phone application. We aim to develop an internet portal that shows data and predictions with minimal delay compared to the real-world system, delivering a proof-of-concept of a real-time digital twin of biodiversity. A further important objective of this project is to increase the public awareness of science on nature and the ongoing environmental change.

Workflow

The overall workflow of this prototype digital twin is illustrated in Fig 1. Citizen scientists record bird vocalizations with a mobile phone application. The audio is sent to a centralized server that runs a CNN model to classify the birds and returns the classifications that are shown in the mobile application. The classifications are compared to prior predictions based on long-term bird monitoring, the results of which comparison is used to update predictions on bird spatiotemporal distributions and singing activity. The workflow of the overall modelling approach is illustrated in Fig. 2, and the workflow for generating prior distribution is illustrated in Fig. 3.

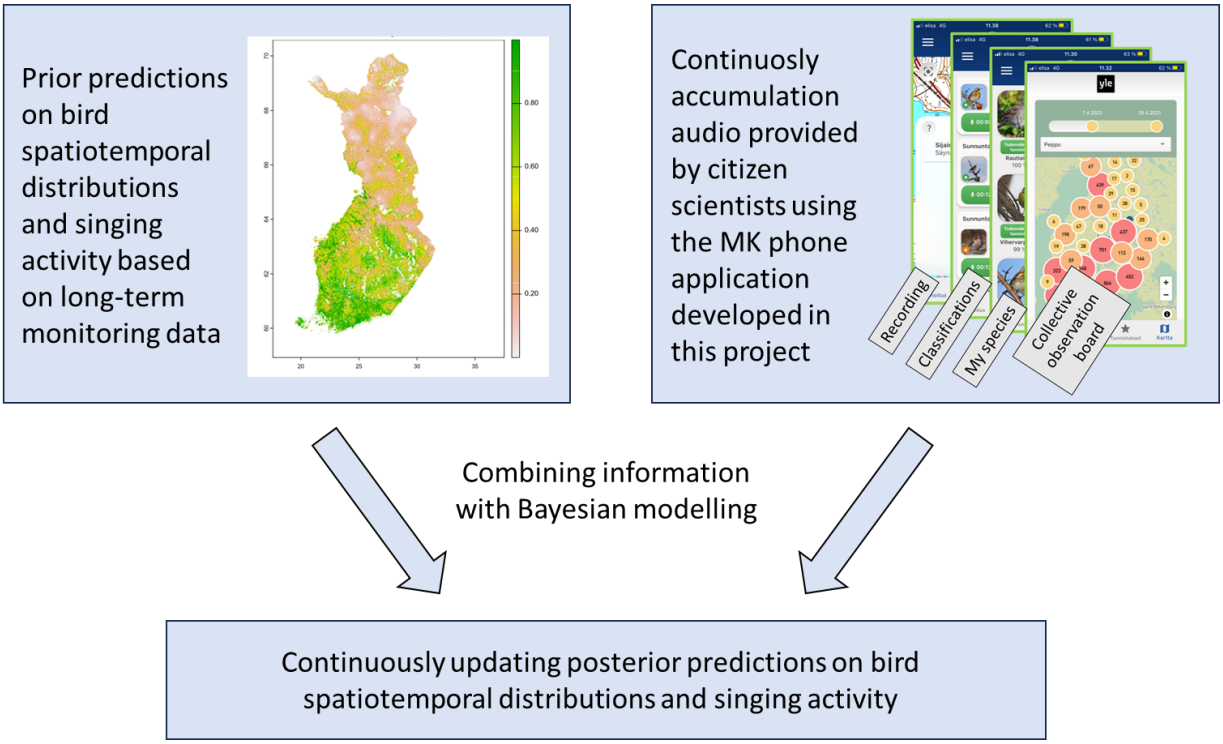

Figure 1. A conceptual diagram of the digital twin prototype. The core aim of this project is to test the feasibility of generating essentially real-time updating predictions on bird spatiotemporal distributions and singing activity by combining prior information based on long-term monitoring data with continuously accumulating new information provided by citizen scientists.

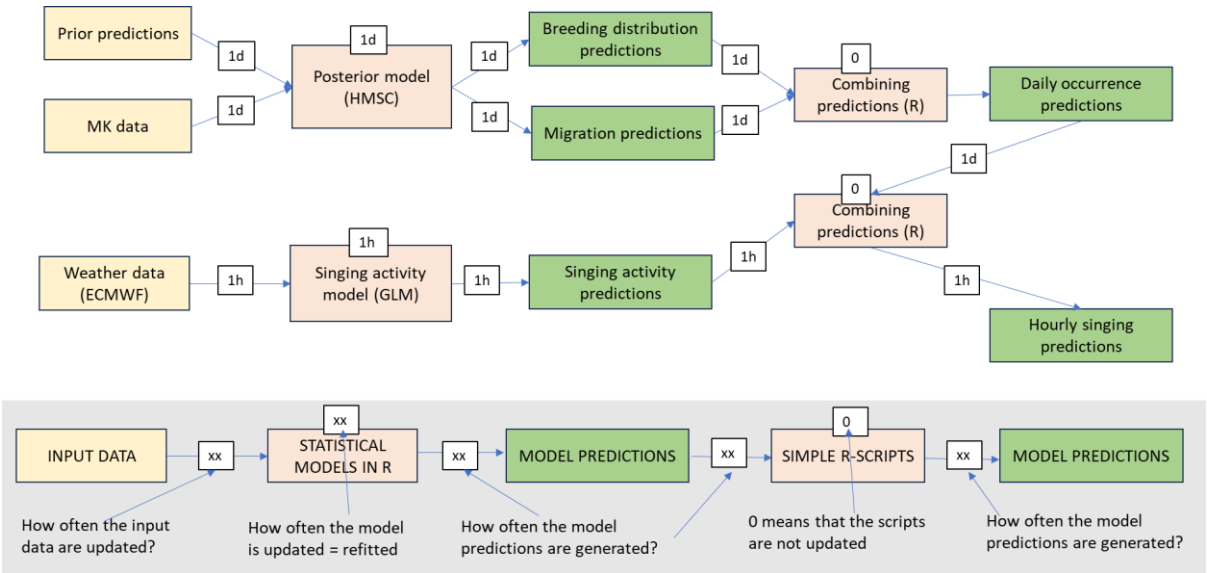

Figure 2. An overview of the modelling strategy for combining prior predictions with the MK phone application data to provide continuous updating predictions of bird distributions and their singing activity. The graphical legend on the bottom of the figure explains the colors and symbols used.

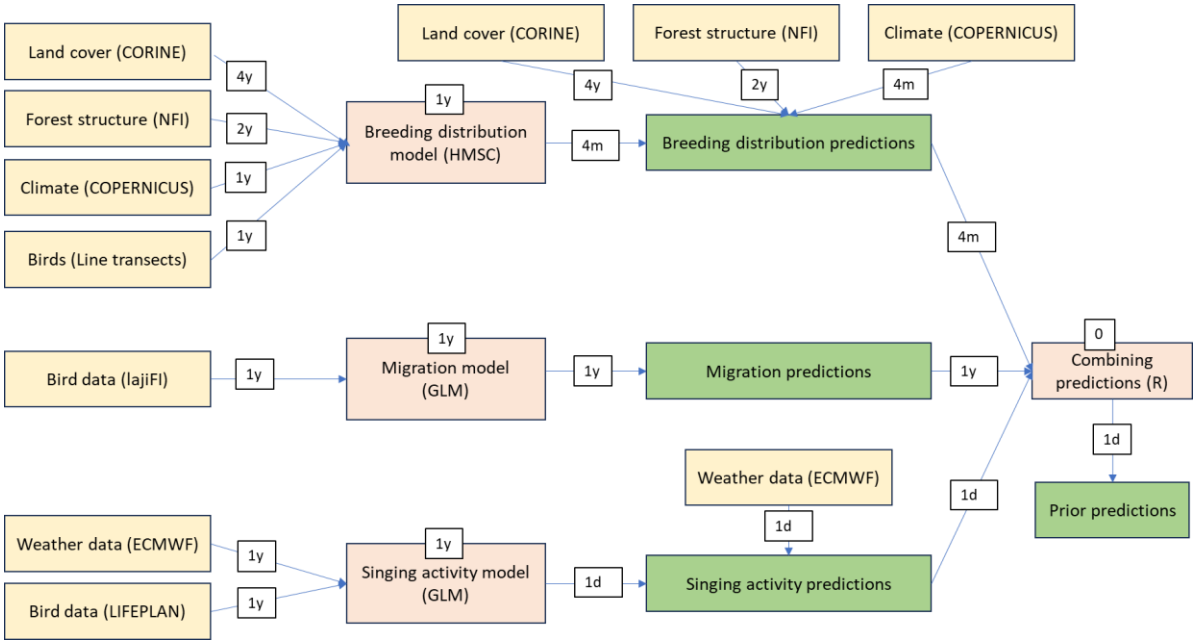

Figure 3. A more detailed description of modelling strategy used to generate the prior predictions based on the long-term data. The colors and symbols used follow the graphical legend of Fig. 2.

Data

As illustrated by the yellow boxes in Figs. 2 and 3, the project combines the following types of data:

- Data recorded by citizen scientists by the mobile phone application MK. The raw audio data consist of .wav files, and the metadata contain information about the user (anonymized), date, time, duration, latitude, and longitude. The classifications made by AI methods describe for each recording the species classified from the recordings, and the reliability of the classifications in units of probability.
- Weather and climatic data derived from Copernicus / ECMWF.
- Land cover data derived from CORINE.
- Transect line counts of birds obtained through collaboration with Finnish bird monitoring program.
- Earlier citizen science data on Finnish birds derived from laji.FI.
- Systematic audio recordings of birds made by the ERC-synergy project LIFEPLAN.

Model

The overall modelling strategy for combining prior predictions with the MK phone application data and weather predictions is illustrated in Fig. 2. This involves fitting the joint species distribution model HMSC (hierarchical model of species communities) at daily intervals to the continuously accumulating audio data submitted from citizen scientists through the mobile phone application MK. The HMSC model considers the species classifications from the phone application data as the response vector and incorporates prior predictions as an offset, estimating spatial and temporal latent factors. These factors signify locations and times where bird occurrences deviate from predictions by the prior model. The latent factors are then used to update prior predictions into posterior predictions of current spatiotemporal distributions of birds, which are further multiplied by the vocalization activity predictions informed by weather forecasts to generate predictions of singing activity.

Additionally, the modelling strategy for constructing the prior predictions is illustrated in Fig. 3. It involves a workflow implemented as a combination of R- and Python/TensorFlow scripts. The prior predictions are obtained as a product of three probabilities, which model (1) how common the species is in the spatial location in question during its summer breeding distribution; (2) whether the species is currently in the summer breeding distribution or its overwintering area (relevant only for migratory species); (3) what is the vocalization activity of the species, given the season, time of the day, and weather conditions. The HMSC model is used in modelling line transect bird count data as a function of environmental (land-use, climate and forest structure) and spatial (latent factors) predictors, used to predict the distribution of Finnish birds at 1-ha resolution, covering over 30 million grid cells.

FAIRness

To facilitate the reusability of the data used in this pDT, we will follow the FAIR principles (Wilkinson et al., 2016) by releasing data to relevant open repositories with assigned persistent identifiers (PIDs) and descriptive metadata.

PID systems, such as the widely used Digital Object Identifier (DOI), ensure reliable referencing, boosting discoverability and facilitating proper citation. Metadata provides comprehensive dataset characterization and captures contextual information that would otherwise be difficult or impossible to retrieve. Adhering to established community standards and vocabularies enables consistency and interoperability and fosters collaboration. This applies to both the choice of metadata structure, such as Research Object Crate (RO-Crate) format (Soiland-Reyes et al., 2022), as well as the data it describes.

The project provides open access to non-sensitive data in a designated repository. Most datasets will be openly available, while sensitive ones may be restricted. Reasons for restrictions will be stated (e.g., GPDR), and access requests will be considered ethically and legally. As a concrete example, the raw audio data is not intended to be publicly available but can be provided on request. Furthermore, the associated metadata (which includes detailed information about the dataset as well as the derived classification) will be openly available.

Researchers can access data following repository guidelines, ensuring they can locate, retrieve, and reuse data. Embargoes may restrict access, with clear terms. Open-access licensing encourages reuse with defined permissions. Quality assurance procedures maintain reliability and include validation, verification, and detailed documentation throughout the data lifecycle, ensuring confident reuse.

Whenever possible, adoption of the FAIR principles will extend to other components of the pDT beyond data (e.g. models, workflows), as established by the FAIR Digital Objects (FDO) interoperability framework (De Smedt et al., 2020). The code will be made publicly available via the BioDT GitHub organization (https://github.com/BioDT) or similar relevant repositories, such as on the space for BioDT on the WorkflowHub registry (https://workflowhub.eu/programmes/22) (Goble et al., 2021).

Ultimately, these efforts seek to align with the wider European strategy on the front of FAIRness and open science, as laid out on the EOSC interoperability framework (European Commission. Directorate General for Research and Innovation. and EOSC Executive Board, 2021).

Performance

The HMSC model is pivotal in our modelling strategy despite its computational intensity. It is used for generating prior predictions and analyzing the continuously accumulating audio data from citizen scientists via the MK mobile phone application. The latter consists of model fitting using MCMC approaches and predicting species occurrences at a 1-ha resolution over Finland. Given the daily frequency of these operations, achieving sufficient computational performance is critically important. To address the computational bottlenecks of the R-package Hmsc (Tikhonov et al. 2020), we developed Hmsc-HPC (Rahman et al. 2024), a high-performance computational module implemented in Python/TensorFlow. Leveraging GPU computation, Hmsc-HPC accelerates model fitting up to 1000 times faster than the R-version (Rahman et al. 2024). Integration of model fitting using Hmsc-HPC on LUMI has been implemented. We have successfully executed the model fitting on LUMI for experimental scenarios, paving the way for making predictions for diverse ecological applications. Furthermore, efforts are underway to leverage the accelerated HPC approach provided by Hmsc-HPC for utilizing fitted models in making predictions, thereby enhancing the performance and scalability of our methodology.

Interface and outputs

The RTBM pDT web application is conceived to facilitate interactive engagement, enabling users to interact with the pDT, running simulations and displaying the predictions on a web browser. By selecting specific bird species and spatial and temporal ranges, users can configure the model runs to suit their needs. The interface is currently in the design phase (Fig. 4). The model outputs will be displayed through updated graphs, maps, and tables, providing information on the bird breeding distribution, migration and singing activity, offering insights into ecological and behavioral patterns and trends.

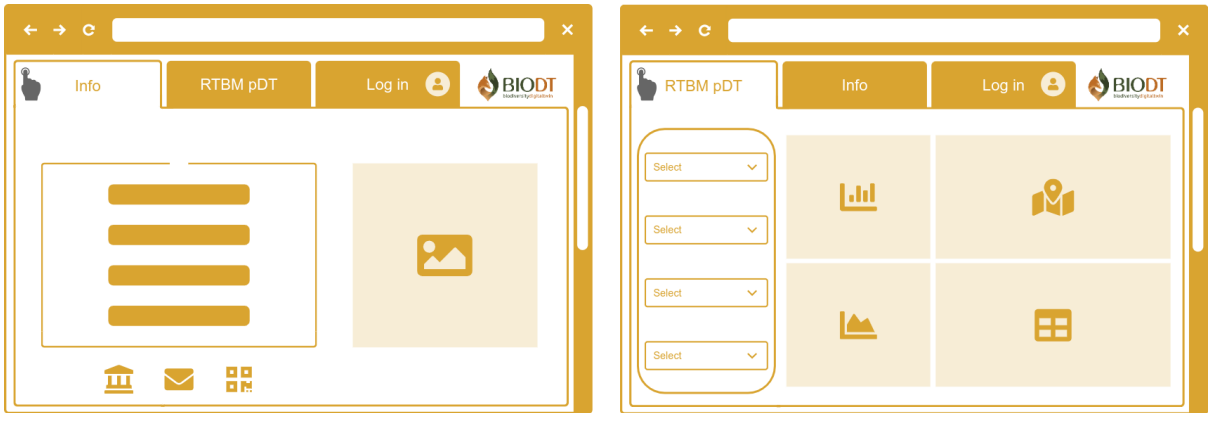

Figure 4. Design of the web application where the users can interact with the RTBM pDT. The figure displays the envisioned features of the web application, including the tabs containing the information on the RTBM, pDT simulation results, and user authentication. There will be a selection of inputs on the RTBM pDT tab (on the left-hand side) and a dashboard on the right hand side of the page displaying the dynamically updated maps, graphs and tables.

Integration and sustainability

The maintenance of the project after the BioDT funding cycle is facilitated by the establishment of the Digital Citizen Science Center that will operate at least until the end of the year 2028 thanks to funding granted by the Jane and Aatos Erkko Foundation. We aim to integrate the project to DestineE and EOSC once doing so is technically feasible, but the details on how to achieve this are still unclear.

Application and impact

Digital twin technologies (DT) have potential to revolutionize biodiversity research, impacting policy frameworks and even market systems. Increasing public awareness of science can inspire masses on environmental initiatives for a common cause: monitoring the state of our environment. Being able to monitor ecological communities in real time through digital technologies can transform biodiversity research. Also, it makes possible to scale data from local to global levels, which can facilitate information-based conservation acts faster than before. Noteworthily, this may include implementing the technology across taxa; a premise, which requires rigorous testing before large-scale reliability could be achieved. Nevertheless, as the information of environmental impact becomes faster and easier through integrating ecological data from various databases, a new era of automated monitoring systems can hasten the green deal and facilitate more sustainable decisions in land use. Further, policies promoting innovation fuel technological advancements, shaping future biodiversity research (and market dynamics), because they promote the development of solutions aligned with Biodiversity Strategy priorities. Policymakers should enact medium and long-term strategies to integrate available DT technology effectively. Collaboration is key to leveraging DT for biodiversity research. Collaboration with stakeholders is vital to maximize benefits, ensure policy alignment and societal impact. Stakeholders include biodiversity Ris, data providers, researchers, policymakers, and industrial actors. Conclusively, policy interventions must align with legislative priorities to drive innovation and achieve sustainable outcomes in future.

Acknowledgements

We thank executive producer Ville Alijoki (Yle Science, Environment and History) for fruitful collaboration: the phone application fast received a broad user community largely thanks to the co-operation with Yle Nature and its promotion of the application in TV, radio, news articles, and social media. The project was funded by the European Union: the HORIZON-INFRA-2021-TECH-01 project 101057437 (Biodiversity Digital Twin for Advanced Modelling, Simulation and Prediction Capabilities), and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 856506: ERC-synergy project LIFEPLAN; and grant agreement No 101123091: ERC-PoC project Breaking the wall between professional science and citizen science by hyperautomation), and the Jane and Aatos Erkko Foundation (grant to establish the Digital Citizen Science Centre for 2024-2028).

References

De Smedt K, Koureas D, Wittenburg P (2020) FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units. Publications 8: 21. https://doi.org/10.3390/publications8020021

European Commission, Directorate-General for Research and Innovation, Corcho, O., Eriksson, M., Kurowski, K. et al., EOSC interoperability framework – Report from the EOSC Executive Board Working Groups FAIR and Architecture, Publications Office, 2021, https://data.europa.eu/doi/10.2777/620649

Goble C, Soiland-Reyes S, Bacall F, Owen S, Williams A, Eguinoa I, Droesbeke B, Leo S, Pireddu L, Rodríguez-Navas L, Fernández JM, Capella-Gutierrez S, Ménager H, Grüning B, Serrano-Solano B, Ewels P, Coppens F (2021) Implementing FAIR Digital Objects in the EOSC-Life Workflow Collaboratory. Zenodo https://doi.org/10.5281/zenodo.4605654

Lauha, P., Somervuo, P., Lehikoinen, P., Geres, L., Richter, T., Seibold, S. and Ovaskainen, O. 2022. Domain-specific neural networks improve automated bird sound recognition already with small amount of local data. Methods in Ecology and Evolution 13, 2799-2810. https://doi.org/10.1111/2041-210X.14003

Rahman, A. U., Tikhonov, G., Oksanen, J., Rossi, T. and Ovaskainen, O. 2024. Accelerating joint species distribution modeling with Hmsc-HPC: A 1000x faster GPU deployment. bioRxiv, 2024.02. 13.580046

Soiland-Reyes S, Sefton P, Crosas M, Castro LJ, Coppens F, Fernández JM, Garijo D, Grüning B, La Rosa M, Leo S, Ó Carragáin E, Portier M, Trisovic A, RO-Crate Community, Groth P, Goble C (2022) Packaging research artefacts with RO-Crate. Data Science 5: 97–138. https://doi.org/10.3233/DS-210053

Tikhonov, G., Opedal, Ø. H., Abrego, N., Lehikoinen, A., de Jonge, M. M., Oksanen, J. and Ovaskainen, O. 2020. Joint species distribution modelling with the R-package Hmsc. Methods in Ecology and Evolution 11, 442-447.

Vohland, K., Land-Zandstra, A., Ceccaroni, L., Lemmens, R., Perelló, J., Ponti, M., Samson, R. and Wagenknecht, K., 2021. Editorial: The Science of Citizen Science Evolves. In: The Science of Citizen Science. Cham: Springer International Publishing. pp. 1–12. DOI: https://doi.org/10.1007/978-3-030-58278-4_1.

Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3: 160018. https://doi.org/10.1038/sdata.2016.18