Prototype Biodiversity Digital Twin: Invasive Alien Species

Taimur Khan[‡], Ahmed El-Gabbas[‡], Marina Golivets[‡], Allan T. Souza[§], Julian Lopez Gordillo[|], Dylan Kierans[¶], Ingolf Kühn[‡,#,¤]

‡ Helmholtz Centre for Environmental Research - UFZ, Halle (Saale), Germany
§ 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
| Naturalis Biodiversity Center, Leiden, Netherlands
¶ KTH Royal Institute of Technology, Division of Computational Science and Technology, Stockholm, Sweden
# Martin-Luther-University Halle-Wittenberg, Halle (Saale), Germany
¤ German Centre for Integrative Biodiversity Research (iDiv), Leipzig, Germany

Corresponding author: Taimur Khan (taimur.khan@ufz.de), Ahmed El-Gabbas (ahmed.el-gabbas@ufz.de)

Academic editor: Sharif Islam

Abstract

Invasive alien species (IAS) threaten biodiversity and human well-being. These threats may increase in the future, necessitating accurate projections of potential locations and the extent of invasions. The main aim of the IAS prototype Digital Twin (IAS pDT) is to dynamically project the level of plant invasion at habitat level across Europe under current and future climates using joint species distribution models. The pDT detects updates in data sources and versions of the datasets and model outputs, implementing the FAIR principles. The pDT's outputs will be available via an interactive dashboard. All input and output data will be freely accessible.

Keywords

Invasive alien species, Digital Twin, climate change, joint species distribution models, Dynamic Data-Driven Application Systems, workflows

Introduction

Invasive alien species (IAS) are a major threat to biodiversity, ecosystem functioning, human well-being and economies worldwide (Pyšek et al. 2020, IPBES 2023). The impacts of IAS will likely increase in the future due to new species introductions (Seebens et al. 2017) and synergies with other drivers of global change, for example, climate and land-use change (Pyšek et al. 2020). It is, therefore, essential to have the capacity to accurately project the potential location and extent of new invasions, relying on the best available data and knowledge of invasion pre-conditions. However, current IAS

modelling involves rather static approaches that do not calibrate with changing real-world conditons, leading to a delay between modelled projections and policy action.

The success of IAS depends on the characteristics of both an invading species and a recipient environment. Incorporating the species' habitat affinity (as part of the environmental preference) and habitat availability into models may substantially enhance the accuracy of resulting predictions and provide more relevant information for policy-making and management. In regional management planning, it is particularly important to know how the overall level of invasion (i.e. the number of IAS) and their spatial extent may vary across habitat types and climate change scenarios.

Previous efforts to model habitat-specific invasions at the European scale undertook an 'assemble first, predict later' approach (*sensu* Ferrier and Guisan (2006)). For plants, this approach encompassed aggregating vegetation data to calculate the number of IAS per habitat and subsequent modelling of the latter as a function of environmental predictors (Chytrý et al. 2009, Chytrý et al. 2012). Alternatively, the level of invasion in habitats can be quantified by modelling individual species and then stacking the responses of those species distribution models (SDMs) (Guisan and Zimmermann 2000). Going one step further, joint species distribution models (jSDMs) allow the modelling of multiple species simultaneously in a single model, accounting for species residual co-occurrence patterns not explained by environmental variation (Tikhonov et al. 2017, Ovaskainen and Abrego 2020, Wilkinson et al. 2020). Compared to single-species or stacked SDMs, jSDMs can better estimate the effects of environmental drivers on species distributions and, by leveraging the residual information, better predict the overall assemblage composition. Further, jSDMs allow rare species (e.g. recently introduced IAS) to borrow information from common species (e.g. widely distributed IAS), facilitating model parameterisation and improving its predictive power (Tikhonov et al. 2017, Ovaskainen and Abrego 2020). jSDMs are hierarchical models that model, in addition to species-specific response to the environment, a shared relationship between the community and environment. Species' responses to the environment are assumed to have a joint structure that depends on phylogenetic relationships between species and their traits; i.e. assuming that species having similar traits respond similarly to the environment (Ovaskainen and Abrego 2020).

Digital Twinning is a dynamic modelling paradigm that models the underlying phyical object or process with updated data to capture the most up-to-date state of the object or process (de Koning et al. 2023). In the Biodiversity Digital Twin project (BioDT; https://biodt.eu/), Digital Twins (DTs) are used to mimic behaviour observed in nature, with the purpose of developing an improved understanding of biodiversity dynamics in response to diverse human pressures, including climate change (Golivets et al. 2024). In the context of IAS modelling, DTs can help improve the accuracy of projections by dynamically integrating environmental variables and species interactions, leading to updated predictions and management strategies, hence bridging the delay between updated projections and policy action. The IAS prototype Digital Twin (pDT) uses jSDMs to predict the level of invasion of vascular plants in Europe at the habitat type level.

Objectives

- **Create a pDT for plant IAS in Europe:** The use of the DT paradigm in ecological research is a burgeoning field (de Koning et al. 2023). Hence creating a prototype DT will help understand how invasion science can benefit from this technology and what are some of the technical requirements for DTs in IAS research.
- **Dynamically project the distribution of plant IAS across Europe:** Generating dynamic projections of the potential future spread of IAS under different global change scenarios using automated workflows will offer a fuller overview of IAS spread over time and the evidence necessary for effective IAS management.
- **Enhance decision-making and operational efficiency**: DTs offer a dynamic and comprehensive virtual representation of IAS systems, enabling real-time monitoring, predictive maintenance and informed decision-making. This approach significantly surpasses traditional static models by providing a detailed lifecycle view of system design, construction, and operation, thus facilitating the detection of issues, enhancing productivity and supporting the validation of results.

Workflow

The IAS pDT follows a layered architecture (Fig. 1) and is composed of the following components:

1) Dynamic Data-Driven Application Systems (DDDAS) based workflows that check for updates in data sources (1.a feedback loops), pull and process the required data into the required format/type (1.b data processing), merge and reconcile the data with the previous version(s) of the data (1.c data assimilation), version the datasets in a way that captures the state of the input data and add metadata to describe the datasets (1.d State + FAIR metadata management) and transfer the updated datasets (1.f data + log files) to the data server (1.e data servicing).

DDDAS is a conceptual framework that synergistically combines models and data to facilitate the analysis and prediction of physical phenomena (Darema 2004). It is a two-layer system design where data assimilation and feedback loops create a bi-directional information flow (Kapteyn and Willcox 2020). The workflows will run periodically at six-month intervals or when new data are detected or available from the sources (e.g. using additional IAS observations or newer versions of climate data).

2) Open-source Project for a Network Data Access Protocol (OPeNDAP) cloud server is where the data is serviced from the previous component. The server is an interface to the twin data (input, output, metadata and log files) and third-party applications to connect to the IAS pDT to request information encapsulated by the DT (Cornillon et al. 2003).

3) jSDMs are the model layer of the IAS pDT, which takes the input data to create detailed model outputs (e.g. level of invasion and species-specific prediction maps).

4) IAS pDT dashboard is the platform where the results of the pDT will be displayed to users and stakeholders. The dashboard aggregates the model results and presents them easily and in a user-friendly manner.

Data

The input data (Table 1) are processed into an equal area projection with a resolution of 10 km. The most recent checklist of naturalised terrestrial alien plant species of non-European origin (n = 1,361; February 2024) was obtained from FloraVeg.EU (Axmanová 2022b). The checklist was standardised against the GBIF taxonomic backbone using the rgbif R package (Chamberlain et al. 2023). Species occurrence records were collated from three data sources (GBIF, EASIN and eLTER; Fig. 2). Further data sources (e.g. DiSSCo, https://www.dissco.eu/ or Biodiversity Data Cubes, https://b-cubed.eu) can be integrated into the workflow in future pDT versions upon availability.

Models were calibrated at the habitat level, i.e. a single model per terrestrial habitat type (see below). CORINE land-cover (CLC) data were converted into the broad habitat classification of Pyšek et al. (2022), based on the habitat descriptions and our ecological knowledge. From CLC data, the percentage coverage of each habitat type per grid cell was calculated to represent habitat availability. Habitat availability maps are used as predictors in the corresponding model. For example, for models done for forest habitat type, forest habitat availability is used as a model predictor. Species-specific habitat affinity data were obtained from SynHab (https://www.synhab.com/), DAISIE (Roy et al. 2020) and FloraVeg.EU (Axmanová 2022a).

CHELSA climatological data (Karger et al. 2017, Karger et al. 2018) were used as the main explanatory variables of the models. For the models, we selected six not highly correlated bioclimatic variables based on their ecological relevance (see Table 1 for more details). These variables were processed into the study area under current and multiple plausible future climate change scenarios. In addition to climate predictors, another two predictors were used in the models:

1. road intensity (total length of roads per grid cell) to represent site accessibility, the level of habitat disturbance and the dispersal of IAS; and
2. railway density (total length of railways per grid cell) as a proxy for IAS dispersal routes.

All data pre- and post-processing steps and model fitting (see “Model” section below) are implemented in the R programming environment (R Core Team 2023). The processed datasets are exported in Network Common Data Form (NetCDF) format for easy access and subsetting when possible.

Model

Models are fitted using the *Hmsc* R package (Tikhonov et al. 2020, Tikhonov et al. 2024). Models are run in a containerised environment on the LUMI supercomputer, Finland (https://www.lumi-supercomputer.eu/). Containerisation technology is designed to encapsulate entire software environments into a single and portable package, including the application, its dependencies and run-time libraries. Singularity is the particular container technology utilised by the models, along with support for Docker containers. Singularity is commonly deployed at High Performance Computing (HPC) sites as it features static portable images, rootless containers and support for HPC schedulers (Kurtzer et al. 2017) along with near bare-metal performance (Torrez et al. 2019). Using Singularity containers allows the management of an independent and self-contained R environment for the model executions, including control over the versions of R and its packages.

Incorporating habitat information into the models provides more robust estimates of the levels of invasion (i.e. sums of predicted individual species presences per grid cell per habitat) that are more informative for management and policy-making. For each habitat type, the main model output is species-specific habitat suitability, which will then be aggregated into the estimates of the level of invasion. The level of invasion under current and projected future climate scenarios is visualised as maps.

Due to the opportunistic nature of the current presence-only data, the total number of vascular plant observations made after 1980 per grid cell in the GBIF database (> 230 million occurrences, March 2024) was used to account for sampling bias, as it is considered a proxy of the sampling effort for vascular plants across Europe. Models are evaluated using spatial block cross-validation to maintain spatial independence between training and testing data.

FAIRness

The IAS pDT aims to move towards higher levels of FAIRness (Wilkinson et al. 2016) by releasing its digital objects on relevant open repositories with a persistent identifier (PID) as well as descriptive metadata. To this end, we are following the Findable, Accessible, Interoperable, Reproducible (FAIR) Digital Objects (FDO) framework for interoperability (De Smedt et al. 2020), implemented through the Research Object Crate (RO-Crate) format (Soiland-Reyes et al. 2022). The IAS pDT references GBIF species occurrence data using a separate Digital Object Identifier (DOI) to access and download the data at each workflow run. For environmental data sources, the pDT provides already existing identifiers from the sources. All data pre- and post-processing steps are implemented using reproducible workflows and made publicly available through the OPeNDAP server. Model outputs are made openly available with persistent identifiers using the same server. All workflow runs are being versioned from execution to execution, along with the

respective input and output data. This adds to the provenance of the pDT by presenting a documented trail that accounts for the origin of the data and its evolution over time.

The model, workflow code and datasets are described through metadata using RO-Crates. Each workflow run is described using RO-Crate, describing all the associated input and output data for that specific workflow run (Fig. 3). The code will also be publicly available via the BioDT GitHub organisation (https://github.com/BioDT), as well as on the space for BioDT on the WorkflowHub registry (https://workflowhub.eu/projects/134) (Goble et al. 2021).

A FAIR Implementation Profile (FIP), based on Magagna et al. (2022) for the IAS pDT, has been created to describe the FAIRness of the data and software, which can inform other communities about the FAIR Enabling Resources used so that they might consider following this approach. The FIP is summarised in Table 2.

Performance

The IAS pDT workflows were tested in steps locally and then moved one by one to LUMI HPC where local running code was the code implemented as Simple Linux Utility for Resource Management (SLURM) jobs using the workflow system described above. Moving the pDT from a local/testing setup to a cloud/HPC environment involved several stages and considerations. Before the migration, the current setup's architecture, performance and requirements were assessed by relevant BioDT project colleagues. This assessment helped to determine the changes and optimisations needed for the transition.

The first step involved replicating the pDT environment in the HPC setup. This included provisioning the necessary infrastructure, such as virtual machines (for the OPeNDAP server), storage and networking components. The architecture may need to be adjusted to efficiently leverage the capabilities and scalability offered by the LUMI platform.

Once the infrastructure was set up, the next phase included migrating the workflows and their data to the new environment (e.g. for Python, R or containers). This involved reconfiguring the application to work optimally on LUMI, using features like batch job submission and parallel computing capabilities. For already-migrated workflows on LUMI, large improvements in run-time were noted due to code parallelisation and the use of a parallel shared file system.

After the migration process, performance metrics are closely monitored to ensure that the pDT operates efficiently in the new setup. Metrics such as response times, throughput, resource utilisation and scalability will be evaluated to identify any bottlenecks or areas for improvement.

Models on a subset of species were tested locally first (along with their evaluation and the preparation of their outputs) before moving to LUMI. On LUMI, the full models were performed in isolated Singularity containers. Resources used by the models (e.g. the total

running time and amount of used memory) are registered. This helps to request sufficient resources for different models in future versions of the pDT and reasonably use the available resources on LUMI.

jSDMs with spatial structures can be highly computationally intensive, if not intractable, for big datasets and large study areas like Europe (Tikhonov et al. 2020a, Rahman et al. 2024). Our models are fitted using the Hmsc-HPC extension (Rahman et al. 2024). The Hmsc-HPC extension is written in Python and uses the TensorFlow library (Abadi et al. 2016). It uses a Graphical Processing Unit (GPU)-compatible implementation of the model fitting algorithm and was shown an up to >1000 speed-up in model fitting (Rahman et al. 2024). Our exploratory models showed an improvement of up to 120 times speed-up (Hmsc-R vs. Hmsc-HPC with GPU on LUMI) for a toy model that implements a Gaussian Predictive Process (GPP; Tikhonov et al. (2020a)) on a subset of the study area and species (90 species, 3581 sampling units (i.e. locations), nine covariates, one spatial random variable). HMSC-R models for these data took ca. 40 hours per chain on parallel (four MCMC chains); while using HMSC-HPC on LUMI-CPU took ca. 100 minutes per chain on average. Using HMSC-HPC on LUMI-GPU for the same data took only ca. 20 minutes per chain. The use of the HMSC-HPC on LUMI-GPU is very promising to speed up our model fitting when upscaling to the full species list at the full European spatial extent.

Interface and outputs

Data Interface

All the processed input datasets and model outputs at each pDT execution will be versioned and stored on the OPeNDAP server for open access to any interested third party. OPeNDAP enables users to access data regardless of its storage format (e.g. NetCDF, Hierarchical Data Format (HDF), General Regularly-distributed Information in Binary (GRIB) etc.). It utilises a client-server architecture, where the client sends data requests to the server and the server responds with the requested data in a format that can be easily used by various analysis tools and software. The server will also serve as a back-end service for the IAS pDT dashboard with aggregated views of the model outputs.

User Interface (UI)

The UI for the IAS pDT is planned as part of the BioDT project-wide web application. It will be a dashboard summarising the results of the model outputs in maps, charts and tables (Fig. 4). Users can interact with the web application and explore the results of the DT.

The dashboard shows maps for the level of invasion under current and projected climate scenarios and uncertainties accompanying model predictions. The results are displayed on a European scale, but users can restrict the visualisation of the results according to the options in the selection box (e.g. country, climate change scenarios, timeframe etc.). In addition to the level of invasion, predicted habitat suitability for each species will be shown.

The web application has different sections (tabs), that display the information related to the IAS pDT and its authors and developing, linking to the relevant sources of information and providing guidance on the usage of the web application. Additionally, the web application displays different levels of details on the “Selection box” and “Dashboard” sections (Fig. 4), based on what the user is interested in exploring in the pDT. General information is displayed in the pDT user section, while more levels of input selection and outputs
will be displayed in the pDT expert section (e.g. model convergence, explanatory/ predictive power, response curves per species and the whole IAS community). User authentication is an IAS pDT web application feature, following the same approach as other BioDT’s pDTs. The contents and design of the UI will be adapted dynamically throughout the development of the IAS pDT to reflect the needs of the pDT team and users. The UI will be developed using R shiny (Chang et al. 2023) following the modularised code environment of the Rhino R package (Żyła et al. 2024).

Integration and sustainability

The sustainability of the project results in all BioDT pDTs is a topic of concern as there is no clear indication whether the project results will be available via the infrastructure currently available in the project or whether pDT teams will seek independent infrastructure. For the IAS pDT, the plan is to keep everything inside the LUMI ecosystem for as long as possible. However, should the need arise, the pDT will be moved to the HPC at the Helmholtz-UFZ (https://www.ufz.de/) called EVE. For this purpose, it has been made sure that all the code in this pDT is self-contained and that the models are containerised to move the pDT to any computational environment in the future.

The input/output datasets in the IAS pDT will be available openly through the OPeNDAP server for anyone to access, along with corresponding metadata and relevant information about versioning. However, no active connection or integration to third-party projects is actively being sought at the time of writing this paper. Additionally, the OPeNDAP server is an independent publicly available software that can be used in use cases beyond this pDT.

Application and impact

Relying on the advanced computing resources and modelling approaches, the IAS pDT will leverage the data from major biodiversity research infrastructures (RIs) to dynamically provide gridded maps of potential plant IAS distributions and the level of invasion in broad terrestrial habitats across Europe under current and future climatic conditions. These projections will allow tracking the invasion potential of several hundreds of plant IAS, including the IAS of European Union concern (i.e. species threatening Europe's biodiversity, human health and the economy; European Commission (2020)), in near-real time, thus providing crucial information for developing adaptive robust IAS policies and management strategies in Europe. In particular, IAS pDT outputs can be incorporated into

species and pathway prioritisation workflows to tackle plant invasions effectively using the most up-to-date information. The pDT workflow's platform-agnostic architecture permits the pDT components' transferability beyond the BioDT environment.

The IAS pDT will also offer valuable support to industries and Small and Medium-sized Enterprises (SMEs). By providing early detection and monitoring capabilities, the pDT will potentially assist industries, such as agriculture, forestry and fisheries, in identifying and mitigating potential threats posed by IAS to their operations and supply chains. This proactive approach helps prevent costly damage to habitats, safeguarding industry interests and enhancing productivity.

The absence of a standardised software design framework for DTs in ecology poses significant hurdles in developing and implementing these systems. Unlike fields with established frameworks facilitating interoperability, the diverse nature of ecosystems and research methodologies in ecology complicates the establishment of standardised approaches. Additionally, many RIs within the ecological community lack implementation of modern methods for data sharing (e.g. Application Programming Interfaces (APIs)), further exacerbating the challenge (Zipkin et al. 2021). This situation results in heterogeneous data landscapes, making it difficult to integrate information seamlessly. The lack of a common framework and modern data access methods not only impedes collaboration and data sharing, but also limits the scalability and effectiveness of DTs in applications of ecological research. Addressing these obstacles presents an opportunity for improvement through initiatives that establish standardised protocols, promote open data practices and provide support for RIS to adopt modern data access methods, ultimately enhancing the capabilities and impact of ecological DT technology.

Adopting DTs in ecological research faces barriers such as technical complexity, lack of standardisation and resistance to change. To overcome these challenges, strategies include providing comprehensive training, securing funding, implementing robust data infrastructure and developing standardised protocols. Additionally, promoting education and outreach, fostering collaborative research initiatives and incentivising innovation can encourage uptake. Demonstrating successful case studies and advocating for supportive policies further facilitate adoption. By addressing these barriers with targeted strategies, DTs can enhance precision and dynamism in IAS modelling and associated conservation efforts.

Acknowledgements

This study has received funding from the European Union's Horizon Europe Research and Innovation Programme under grant agreement No. 101057437 (BioDT project, https://doi.org/10.3030/101057437). Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them. We acknowledge the EuroHPC Joint Undertaking for awarding this project access to the EuroHPC supercomputer LUMI, hosted by CSC (Finland; https://

www.csc.fi) and the LUMI consortium through a EuroHPC Development Access call. This research complies with all relevant regulations and data-sharing protocols outlined in data sources, ensuring adherence to ethical guidelines and legal requirements for the collection, use and dissemination of the data used in this study.

Author contributions

Taimur Khan and Ahmed El-Gabbas contributed equally to this forum paper.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, et al. (2016) TensorFlow: a system for Large-Scale machine learning. 12th USENIX symposium on operating systems design and implementation (OSDI 16).
- Axmanová I (2022a) Broad habitat. www.FloraVeg.EU.
- Axmanová I (2022b) Origin in Europe. www.FloraVeg.EU.
- Chamberlain S, Barve V, Mcglinn D, Oldoni D, Desmet P, Geffert L, Ram K (2023) rgbif: Interface to the Global Biodiversity Information Facility API. R package version 3.7.8. https://CRAN.R-project.org/package=rgbif.
- Chang W, Cheng J,, Allaire J, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B (2023) Shiny: Web Application Framework for R. R package version 1.8.0. https://CRAN.R-project.org/package=shiny.
- Chytrý M, Pyšek P, Wild J, Pino J, Maskell LC, Vilà M (2009) European map of alien plant invasions based on the quantitative assessment across habitats. Diversity and Distributions 15 (1): 98-107. https://doi.org/10.1111/j.1472-4642.2008.00515.x
- Chytrý M, Wild J, Pyšek P, Jarošík V, Dendoncker N, Reginster I, Pino J, Maskell LC, Vilà M, Pergl J (2012) Projecting trends in plant invasions in Europe under different scenarios of future land-use change. Global Ecology and Biogeography 21 (1): 75-87. https://doi.org/10.1111/j.1466-8238.2010.00573.x
- Cornillon P, Gallagher J, Sgouros T (2003) OPeNDAP: Accessing data in a distributed, heterogeneous environment. Data Science Journal 2: 164-174. https://doi.org/10.2481/dsj.2.164
- Darema F (2004) Dynamic Data Driven Applications Systems: A New Paradigm for Application Simulations and Measurements. In: Bubak M, Albada GD, Sloot PA, Dongarra J (Eds) Computational Science - ICCS 2004. 7 pp. [ISBN 978-3-540-22116-6 978-3-540-24688-6]. https://doi.org/10.1007/978-3-540-24688-6_86
- de Koning K, Broekhuijsen J, Kuhn I, Ovaskainen O, Taubert F, Endresen D, Schigel D, Grimm V (2023) Digital twins: dynamic model-data fusion for ecology. Trends Ecol Evol 38 (10): 916-926. https://doi.org/10.1016/j.tree.2023.04.010

- De Smedt K, Koureas D, Wittenburg P (2020) FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units. Publications 8 (2). https://doi.org/10.3390/publications8020021
- European Commission (2020) Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. EU Biodiversity Strategy for 2030. Bringing nature back into our lives 2020, COM/2020/380.
- Ferrier S, Guisan A (2006) Spatial modelling of biodiversity at the community level. Journal of Applied Ecology 43 (3): 393-404. https://doi.org/10.1111/j.1365-2664.2006.01149.x
- 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.4605653
- Golivets M, Sharif I, Wohner C, Grimm V, Schigel D (2024) Building Biodiversity Digital Twins. Rio Special issue/topical collection https://doi.org/10.3897/rio.coll.240
- Guisan A, Zimmermann N (2000) Predictive habitat distribution models in ecology. Ecological Modelling 135 (2-3): 147-186. https://doi.org/10.1016/s0304-3800(00)00354-9
- IPBES (2023) Summary for Policymakers of the Thematic Assessment Report on Invasive Alien Species and their Control of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. In: Roy HE, Pauchard A, Stoett P, Renard Truong T, Bacher S, Galil BS, Hulme PE, Ikeda T, Sankaran KV, McGeoch MA, Meyerson LA, Nuñez MA, Ordonez A, Rahlao SJ, Schwindt E, Seebens H, Vandvik V,SAW (Eds) IPBES secretariat, Bonn, Germany. https://doi.org/10.5281/zenodo.7430692
- Kapteyn M, Willcox K (2020) Predictive Digital Twins: Where Dynamic Data-Driven Learning Meets Physics-Based Modeling. In: Darema F, Blasch E, Ravela S, Aved A (Eds) Dynamic Data Driven Applications Systems. 4 pp. [ISBN 978-3-030-61724-0 978-3-030-61725-7]. https://doi.org/10.1007/978-3-030-61725-7_1
- Karger D, Conrad O, Böhner J, Kawohl T, Kreft H, Zimmermann N, Linder HP, Kessler M (2018) Data from: Climatologies at high resolution for the earth's land surface areas. Dryad https://doi.org/10.5061/dryad.kd1d4
- Karger DN, Conrad O, Bohner J, Kawohl T, Kreft H, Soria-Auza RW, Zimmermann NE, Linder HP, Kessler M (2017) Climatologies at high resolution for the earth's land surface areas. Sci Data 4 https://doi.org/10.1038/sdata.2017.122
- Kurtzer GM, Sochat V, Bauer MW (2017) Singularity: Scientific containers for mobility of compute. PLoS One 12 (5). https://doi.org/10.1371/journal.pone.0177459
- Magagna B, Schultes E, Suchánek M, Kuhn T (2022) FIPs and Practice. Research Ideas and Outcomes 8: 94451. https://doi.org/10.3897/rio.8.e94451
- Meijer J, Huijbregts MJ, Schotten KGJ, Schipper A (2018) Global patterns of current and future road infrastructure. Environmental Research Letters 13 (6). https://doi.org/10.1088/1748-9326/aabd42
- Ovaskainen O, Abrego N (2020) Joint species distribution modelling, with applications in R. Cambridge University Press [ISBN 9781108591720 9781108492461 9781108716789] https://doi.org/10.1017/9781108591720
- Pyšek P, Hulme PE, Simberloff D, Bacher S, Blackburn TM, Carlton JT, Dawson W, Essl F, Foxcroft LC, Genovesi P, Jeschke JM, Kuhn I, Liebhold AM, Mandrak NE, Meyerson

LA, Pauchard A, Pergl J, Roy HE, Seebens H, van Kleunen M, Vilà M, Wingfield MJ, Richardson DM (2020) Scientists' warning on invasive alien species. Biol Rev Camb Philos Soc 95 (6): 1511-1534. https://doi.org/10.1111/brv.12627

- Pyšek P, Sádlo J, Chrtek J, Chytrý M, Kaplan Z, Pergl J, Pokorná A, Axmanová I, Čuda J, Doležal J, Dřevojan P, Hejda M, Kočár P, Kortz A, Lososová Z, Lustyk P, Skálová H, Štajerová K, Večeřa M, Vítková M, Wild J, Danihelka J (2022) Catalogue of alien plants of the Czech Republic (3rd edition). Preslia 94 (4): 447-577. https://doi.org/10.23855/preslia.2022.447
- Rahman AU, Tikhonov G, Oksanen J, Rossi T, Ovaskainen O (2024) Accelerating joint species distribution modeling with Hmsc-HPC: A 1000x faster GPU deployment. bioRxiv https://doi.org/10.1101/2024.02.13.580046
- R Core Team (2023) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
- Roy D, Alderman D, Anastasiu P, Arianoutsou M, Augustin S, Bacher S, Başnou C, Beisel J, Bertolino S, Bonesi L, Bretagnolle F, Chapuis JL, Chauvel B, Chiron F, Clergeau P, Cooper J, Cunha T, Delipetrou P, Desprez-Loustau M, Détaint M, Devin S, Didžiulis V, Essl F, Galil BS, Genovesi P, Gherardi F, Gollasch S, Hejda M, Hulme PE, Josefsson M, Kark S, Kauhala K, Kenis M, Klotz S, Kobelt M, Kühn I, Lambdon PW, Larsson T, Lopez-Vaamonde C, Lorvelec O, Marchante H, Minchin D, Nentwig W, Occhipinti-Ambrogi A, Olenin S, Olenina I, Ovcharenko I, Panov VE, Pascal M, Pergl J, Perglová I, Pino J, Pyšek P, Rabitsch W, Rasplus J, Rathod B, Roques A, Roy H, Sauvard D, Scalera R, Shiganova TA, Shirley S, Shwartz A, Solarz W, Vilà M, Winter M, Yésou P, Zaiko A, Adriaens T, Desmet P, Reyserhove L (2020) DAISIE - Inventory of alien invasive species in Europe. Version 1.7. Research Institute for Nature and Forest (INBO). Checklist dataset accessed via GBIF.org on 2024-03-28. https://doi.org/10.15468/ybwd3x
- Seebens H, Blackburn TM, Dyer EE, Genovesi P, Hulme PE, Jeschke JM, Pagad S, Pysek P, Winter M, Arianoutsou M, Bacher S, Blasius B, Brundu G, Capinha C, Celesti-Grapow L, Dawson W, Dullinger S, Fuentes N, Jager H, Kartesz J, Kenis M, Kreft H, Kuhn I, Lenzner B, Liebhold A, Mosena A, Moser D, Nishino M, Pearman D, Pergl J, Rabitsch W, Rojas-Sandoval J, Roques A, Rorke S, Rossinelli S, Roy HE, Scalera R, Schindler S, Stajerova K, Tokarska-Guzik B, van Kleunen M, Walker K, Weigelt P, Yamanaka T, Essl F (2017) No saturation in the accumulation of alien species worldwide. Nat Commun 8 https://doi.org/10.1038/ncomms14435
- Soiland-Reyes S, Sefton P, Crosas M, Castro LJ, Coppens F, Fernández J, Garijo D, Grüning B, La Rosa M, Leo S, Ó Carragáin E, Portier M, Trisovic A, Community RO, Groth P, Goble C, Peroni S (2022) Packaging research artefacts with RO-Crate. Data Science 5 (2): 97-138. https://doi.org/10.3233/ds-210053
- Tikhonov G, Abrego N, Dunson D, Ovaskainen O (2017) Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. Methods in Ecology and Evolution 8 (4): 443-452. https://doi.org/10.1111/2041-210x.12723
- Tikhonov G, Duan L, Abrego N, Newell G, White M, Dunson D, Ovaskainen O (2020a) Computationally efficient joint species distribution modeling of big spatial data. Ecology 101 (2): e02929. https://doi.org/10.1002/ecy.2929
- Tikhonov G, Opedal OH, Abrego N, Lehikoinen A, de Jonge MMJ, Oksanen J, Ovaskainen O (2020b) Joint species distribution modelling with the r-package Hmsc. Methods Ecol Evol 11 (3): 442-447. https://doi.org/10.1111/2041-210X.13345

- Tikhonov G, Ovaskainen O, Oksanen J, de Jonge M, Opedal O, Dallas T (2024) Hmsc: Hierarchical Model of Species Communities. R package version 3.0-14, https://www.helsinki.fi/en/researchgroups/statistical-ecology/software/hmsc.
- Torrez A, Randles T, Priedhorsky R (2019) HPC Container Runtimes have Minimal or No Performance Impact. IEEE/ACM International Workshop on Containers and New Orchestration Paradigms for Isolated Environments in HPC (CANOPIE-HPC). IEEE https://doi.org/10.1109/CANOPIE-HPC49598.2019.00010
- Wilkinson D, Golding N, Guillera-Arroita G, Tingley R, McCarthy M, Freckleton R (2020) Defining and evaluating predictions of joint species distribution models. Methods in Ecology and Evolution 12 (3): 394-404. https://doi.org/10.1111/2041-210x.13518
- Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, 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 AJ, Groth P, Goble C, Grethe JS, Heringa J, t Hoen PA, 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 SA, 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. Sci Data 3 https://doi.org/10.1038/sdata.2016.18
- Zipkin E, Zylstra E, Wright A, Saunders S, Finley A, Dietze M, Itter M, Tingley M (2021) Addressing data integration challenges to link ecological processes across scales. Frontiers in Ecology and the Environment 19 (1): 30-38. https://doi.org/10.1002/fee.2290
- Żyła K, Nowicki J, Siemiński L, Rogala M, Vibal R, Makowski T, Basa R (2024) rhino: A Framework for Enterprise Shiny Applications. R package version 1.7.0.9000. Enterprise Shiny Applications URL: https://github.com/Appsilon/rhino

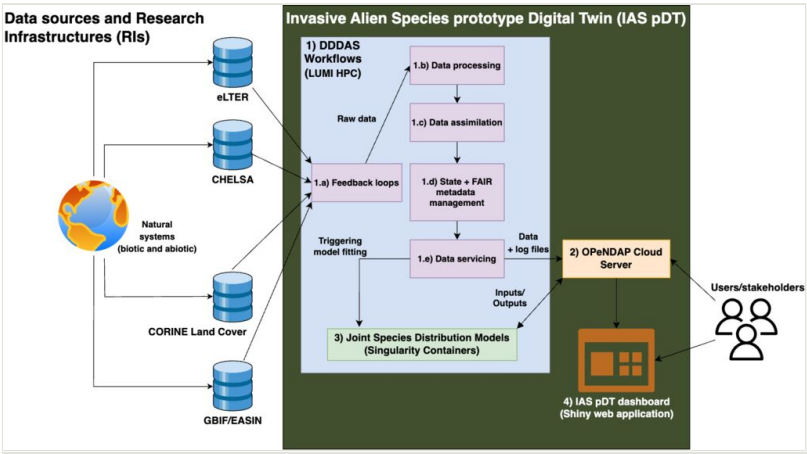

Figure 1.

Figure 1: An overview of the IAS Prototype Digital Twin (IAS pDT) components. Main input data sources include eLTER — the Integrated European Long-Term Ecosystem, critical zone and socio-ecological Research Infrastructure; CHELSA — climatologies at high resolution for the Earth's land surface areas; CORINE — Coordination of Information on the Environment; GBIF — Global Biodiversity Information Facility; and EASIN — European Alien Species Information Network. See Table 1 for more details. Data workflows are based of the Dynamic Data-Driven Application System (DDDAS) paradigm and all data are available under an Open-source Project for a Network Data Access Protocol (OPeNDAP) cloud server.

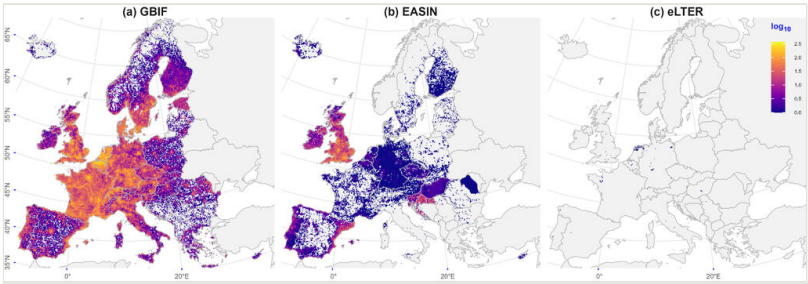

Figure 2.

Figure 2. The $\log_{10}$-transformed number of IAS (invasive alien species) per 10 km × 10 km grid cell in the three data sources (a) GBIF, (b) EASIN and (c) eLTER (updated March 2024). Only observations made after 1980 were considered. For EASIN data, only data from data providers other than GBIF are shown. See Table 1 for more details.

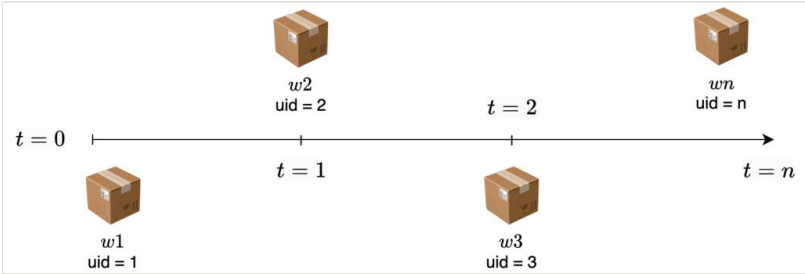

Figure 3.

Figure 3: A visualisation of workflow runs and the associated RO-Crates (represented by boxes), where *t* is the time of the run, *w* is the workflow run and *uid* is the associated unique identifier for the workflow. Each crate represents the metadata representation of all the associated input/output data for a specific workflow run.

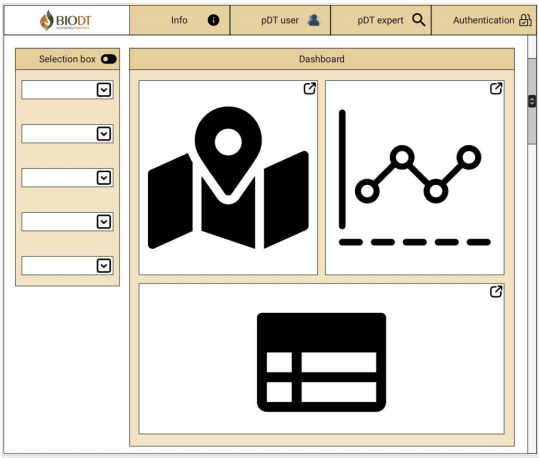

Figure 4.

Figure 4. Wireframe of the IAS pDT dashboard, displaying the envisioned features of the web application (as it was not ready at the time of writing), including the tabs containing the information on the pDT, the user group (pDT user and pDT expert) and user authentication. The web application will have selection boxes (on the left) and a dashboard (on the centre-right) displaying dynamically updated maps, graphs and tables.

Table 1.

Table 1: Input data sources used in the models and their source, spatial and temporal resolution.

Data		Spatial resolution	Temporal resolution	Details	Source
Reference grid		10 km	---	The European Environment Agency's (EEA) reference grid at 10 km resolution at Lambert Azimuthal Equal Area projection (EPSG:3035). All data listed below were processed into this reference grid.	https://www.eea.europa.eu/en/datahub/datahubitem-view/3c362237-daa4-45e2-8c16-aaadfb1a003b
Species observations	Global Biodiversity Information Facility (GBIF)	points	> 1981	The most up-to-date version of occurrence data is dynamically downloaded from GBIF using the rgbif R package (Chamberlain et al. 2023) (> 8 million occurrences, March 2024; Figure 2a). Doubtful occurrences and occurrences with high spatial uncertainty are excluded.	https://www.gbif.org/
	European Alien Species Information Network (EASIN)	points	> 1981	EASIN provides spatial data on 14,000 alien species. Species occurrences were downloaded using EASIN's API. Thirty-four partners shared their data with EASIN (including GBIF). Only non-GBIF data from EASIN were considered in the models (> 692 K observations for 483 IAS; March 2024; Figure 2b).	European Commission - Joint Research Centre - European Alien Species Information Network (EASIN) https://easin.jrc.ec.europa.eu/

	Integrated European Long-Term Ecosystem, Critical Zone and socio-ecological Research (eLTER)	points	> 1981	eLTER is a network of sites collecting ecological data for long-term research within the EU. Vegetation data from 137 eLTER sites were processed and homogenised. The final eLTER dataset comprises 5,265 observations from 46 sites, representing 110 IAS (Figure 2c).	https://elter-ri.eu/
Habitat information	Corine Land Cover (CLC)	100 m	2017-2018	CLC dataset is a pan-European land-cover and land-use inventory with 44 thematic classes, ranging from broad forested areas to individual vineyards. We are currently using V2020_20u1 of CLC data, but the data workflow is flexible to use future versions of CLC data.	https://land.copernicus.eu/en/products/corine-land-cover

Climate data	Climatologies at high resolution for the Earth's land surface areas (CHELSA)	30 arc seconds; ~ 1 km	1981–2010 2011–2040 2041–2070 2071–2100	CHELSA provides global high-resolution data on various environmental variables currently and in different future climate scenarios. Six ecologically meaningful and low-correlated bioclimatic variables are used in the models - temperature seasonality (bio4) - mean daily minimum air temperature of the coldest month (bio6) - mean daily mean air temperatures of the wettest quarter (bio8) - annual precipitation amount (bio12) - precipitation seasonality (bio15) - mean monthly precipitation amount of the warmest quarter (bio18) In addition to current climate conditions, there are nine options for multiple climate CMIP6 models (3 shared socioeconomic pathways [ssp126 - ssp370 - ssp585] × 3 time slots [2011-2040 - 2041-2070 - 2071-2100]).	Karger et al. (2018), Karger et al. (2017) https://chelsa-climate.org/
Road intensity		lines	most recent	The total length of roads per grid cell was computed from the most recent version of the GRIP (Global Roads Inventory Project) global roads database.	Meijer et al. (2018) https://www.globio.info/download-grip-dataset
Railway intensity		lines	most recent	The total length of railways per grid cell was computed from the most recent version of OpenRailwayMap.	https://www.openrailwaymap.org/

Sampling bias	points	> 1981	The total number of vascular plant observations per grid cell in the GBIF database was computed (> 230 million occurrences, March 2024).	https://www.gbif.org/

Table 2.

Table 2. FAIR Implementation Profile (FIP) created using the FIP Wizard: https://fip-wizard.ds-wizard.org/. The "ID" column contains the specific FAIR principle the question addresses (e.g. "A1.1"), as well as whether it refers to data or metadata ("D" or "MD", respectively). The FIP of IAS pDT is accessible on: https://fip-wizard.ds-wizard.org/wizard/projects/20b812be-b4e6-48e6-98c8-5bff3691876c.

		FAIR Enable Resource (FER)	
ID	**Question**	**Name**	**Unique Resource Identifier (URI)**
F1 MD	What globally unique, persistent, resolvable identifier service do you use for metadata records?	UUID | Universally Unique Identifier	http://purl.org/np/RA5ikgqnKqn071dwzXFdiXlnM8hWZRdFKsQjC_e5YRkEw#UUID
F1 D	What globally unique, persistent, resolvable identifier service do you use for datasets?	DOI | Digital Object Identifier	http://purl.org/np/RAnAWGdeI_1GGmDAqv-vZjby5XqbL2ZujNz1vgwK_6cRI#DOI
F2	What metadata schema do you use for findability?	RO-Crate | Research Object Crate	http://purl.org/np/RAcYMfIt1ICpNTg0RCiR0QHfNoSUU-b-5Yw3w06HSL9VA#RO_Crate
F4 MD	Which service do you use to publish your metadata records?	Zenodo | https://zenodo.org/	http://purl.org/np/RAQKRYjUrndhJAbsgnuhr1Z3DecqtWVI1qUTC2cPpyLDY#Zenodo
A1.1 MD	Which standardised communication protocol do you use for metadata records?	OPeNDAP | Open-source Project for a Network Data Access Protocol	http://purl.org/np/RApihvFKR8-JO6eD5nuYMkyEDONbIZZC5uDkjxdqqq0ZQ#OPeNDAP
A1.1 D	Which standardised communication protocol do you use for datasets?	OPeNDAP | Open-source Project for a Network Data Access Protocol	http://purl.org/np/RApihvFKR8-JO6eD5nuYMkyEDONbIZZC5uDkjxdqqq0ZQ#OPeNDAP
A1.2 MD	Which authentication & authorisation service do you use for metadata records?	HTTPS | Hypertext Transfer Protocol Secure	http://purl.org/np/RAF1ANn-BCFop0OBMOC7S8NtG0y_xYhRX4tAu37XZVCo0#HTTPS

A1.2 D	Which authentication & authorisation service do you use for datasets?	HTTPS | Hypertext Transfer Protocol Secure	http://purl.org/np/RAF1ANn-BCFop0OBMOC7S8NtG0y_xYhRX4tAu37XZVCo0#HTTPS
I1 MD	What knowledge representation language (allowing machine interoperation) do you use for metadata records?	JSON-LD | JavaScript Object Notation for Linking Data	http://purl.org/np/RAQKjgd7Ug9xSo4J0REW_AHGOJyaF9-ydj60nunqQ0qVg#JSON-LD
I1 D	What knowledge representation language (allowing machine interoperation) do you use for datasets?	JSON-LD | JavaScript Object Notation for Linking Data	http://purl.org/np/RAQKjgd7Ug9xSo4J0REW_AHGOJyaF9-ydj60nunqQ0qVg#JSON-LD
I2 MD	What structured vocabulary do you use to annotate your metadata records?	RO-Crate | Research Object Crate	http://purl.org/np/RAcYMfIt1ICpNTg0RCiR0QHfNoSUU-b-5Yw3w06HSL9VA#RO_Crate
I2 D	What structured vocabulary do you use to encode your datasets?	RO-Crate | Research Object Crate	http://purl.org/np/RAcYMfIt1ICpNTg0RCiR0QHfNoSUU-b-5Yw3w06HSL9VA#RO_Crate