BioClima

# D3.1 Hierarchical Statistical Framework for Dynamic Ecological Assessment

Funded by
the European Union

| Document Identification | | | |
|---|---|---|---|
| Related WP | WP3, WP2 | Document reference | D3.1 |
| Status | Draft | Submissions date | 30/06/2026 |
| Related deliverables | - | Dissemination Level | PU |
| Lead Author | Philip Stanley Mostert (NTNU) | Version | 1.0 |
| Co-Authors | Bob O'Hara (NTNU)<br>Charalampos Paraskevas (Neuralio)<br>Stelios Kotsopoulos (Neuralio) | Reviewers | Elnaz Neinavaz (ITC)<br>Jiří Sedlák (VSUO) |
| Keywords | hierarchical statistical modelling, integrated species distribution models, biodiversity monitoring, EBVs, ECVs, Earth Observation, AI/ML, early warning, ecological risk assessment | Website | https://bioclima.net/ |

This document is issued within the frame and for the purpose of the BioClima project. This project has received funding from the European Union's Horizon-RIA Framework Programme under Grant Agreement No. 101181408. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the European Commission.

| Document History | | | |
|---|---|---|---|
| Version | Date | Editors | Changes |
| 0.1 | 10/2/2026 | NTNU | Initial Draft |
| 0.2 | 11/06/2026 | NTNU | Second version after internal review |
| 0.3 | 29/06/2026 | NTNU | Third version after review |
| 1.0 | 30/06/2026 | NTNU | Final Version |

# Table of Contents

# List of Tables

# List of Acronyms

| Acronym | Description |
|---|---|
| AGB | Above-Ground Biomass |
| AI | Artificial Intelligence |
| API | Application Programming Interface |
| AWS | Amazon Web Services |
| BA | Burned Area |
| CAMS | Copernicus Atmosphere Monitoring Service |
| CARD-BS | Terrain-Corrected Backscatter |
| CAS | Chinese Academy of Sciences |
| CCI | Climate Change Initiative |
| CEDA | Centre for Environmental Data Analysis |
| CEMS | Copernicus Emergency Management Service |
| CHM | Canopy Height Model |
| CLMS | Copernicus Land Monitoring Service |
| CRS | Coordinate Reference System |
| CSV | Comma-Separated Values |
| DEM | Digital Elevation Model |
| DIAS | Data and Information Access Services |
| DSM | Digital Surface Model |
| DTM | Digital Terrain Model |
| DwC-A | Darwin Core Archive |
| EBV | Essential Biodiversity Variable |
| EC | European Commission |
| ECMWF | European Centre for Medium-Range Weather Forecasts |
| ECV | Essential Climate Variable |
| EML | Ecological Metadata Language |

| Acronym | Description |
|---|---|
| ENVISAT | Environmental Satellite |
| EO | Earth Observation |
| ET | Evapotranspiration |
| ETM+ | Enhanced Thematic Mapper Plus |
| EVI2 | Two-band Enhanced Vegetation Index |
| ESA | European Space Agency |
| FAO | Food and Agriculture Organization |
| FAIR | Findable, Accessible, Interoperable and Reproducible |
| FAPAR | Fraction of Absorbed Photosynthetically Active Radiation |
| FCover | Fraction of Green Vegetation Cover |
| FLAM | Wildfire Climate Impacts and Adaptation Model |
| FSC | Fractional Snow Cover |
| GAM | Global Forest Model |
| GBIF | Global Biodiversity Information Facility |
| GeoBON | Group on Earth Observations Biodiversity Observation Network |
| GeoTIFF | Geographic Tagged Image File Format |
| GNDVI | Green Normalized Difference Vegetation Index |
| HADISD | Hadley Centre Integrated Surface Database |
| HPC | High-Performance Computing |
| HR-VPP | High-Resolution Vegetation Phenology and Productivity |
| HTTPS | Hyper Text Transfer Protocol Secure |
| IAM | Identity and Access Management |
| ICOS | Integrated Carbon Observation System |
| INLA | Integrated Nested Laplace Approximation |
| ISDM | Integrated Species Distribution Model |
| IoT | Internet of Things |
| JWT | JSON Web Token |

| Acronym | Description |
| --- | --- |
| JSON | JavaScript Object Notation |
| LAI | Leaf Area Index |
| LGM | Latent Gaussian Model |
| LGCP | Log Gaussian Cox Process |
| LLL | Low-Light-Level |
| LiDAR | Light Detection and Ranging |
| LST | Land Surface Temperature |
| MARL | Multi-Agent Reinforcement Learning |
| MCMC | Markov Chain Monte Carlo |
| MERIS | Medium Resolution Imaging Spectrometer |
| MODIS | Moderate Resolution Imaging Spectroradiometer |
| MII | Multispectral Imager |
| ML | Machine Learning |
| MSI | Multispectral Instrument |
| NDSI | Normalized Difference Snow Index |
| NDRE | Normalized Difference Red-Edge Index |
| NDVI | Normalized Difference Vegetation Index |
| NDWI | Normalized Difference Water Index |
| NetCDF | Network Common Data Form |
| NFI | National Forest Inventory |
| NFS | Network File System |
| NPP | Net Primary Production |
| OGC | Open Geospatial Consortium |
| OIDC | OpenID Connect |
| PC | Penalizing Complexity |
| PNG | Portable Network Graphic |
| PPI | Plant Phenology Index |

| Acronym | Description |
| --- | --- |
| RGB | Red-Green-Blue |
| RTC | Radiometrically Terrain Corrected |
| SAR | Synthetic Aperture Radar |
| S2GLC | Sentinel-2 Global Land Cover |
| S3 | Simple Storage Device |
| SDG | Sustainable Development Goals |
| SDGSAT-1 | Sustainable Development Goals Satellite-1 |
| SDM | Species Distribution Model |
| SeBMS | Swedish Butterfly Monitoring Scheme |
| SITES | Swedish Infrastructure for Ecosystem Science |
| SLSTR | Sea and Land Surface Temperature |
| SMOS | Soil Moisture and Ocean Salinity |
| ST | Seasonal Trajectories |
| STAC | Spatio-Temporal Asset Catalog |
| TIS | Thermal Infrared Spectrometer |
| TM | Thematic Mapper |
| UAV | Unmanned Aerial Vehicle |
| URL | Uniform Resource Locator |
| VHR | Very High Resolution |
| VI | Vegetation Index |
| WP | Work Package |

# Executive Summary

This deliverable defines the hierarchical statistical framework that underpins dynamic ecological assessment in the BioClima project. The framework is designed to evaluate biodiversity distributions, ecosystem condition, and ecological dynamics under changing climate, land use, and disturbance regimes, while preserving explicit treatment of uncertainty, heterogeneous data sources, and downstream operational use. In the Description of Action, D3.1 is positioned as a joint methodological report for WP2 and WP3 and as a key input to Tasks 3.2 and 3.3, as well as to Tasks 2.1–2.4; accordingly, the present document treats D3.1 as a cross-work-package methodological backbone rather than as a narrow modelling note.

The methodological core of D3.1 is a hierarchical, integrated, and uncertainty-aware statistical framework. This choice is dictated by the project's observational architecture. D1.1 (Analysis of Data for Integration) shows that the project combines species occurrence records from large biodiversity infrastructures and citizen science, structured ground observations, earth observation (EO)-derived covariates, unmanned aerial vehicle (UAV) products, interoperable metadata services, and derived essential biodiversity variable (EBV) and essential climate variable (ECV) layers. A statistically robust framework must therefore distinguish the latent ecological process from the observation process and must preserve differences in data quality, bias, detection structure, and spatial or temporal support, rather than forcing all source types into a single simplified format (Prassou et al., 2025; Isaac et al., 2020; Simmonds et al., 2020).

At the ecological core, the deliverable adopts a spatio-temporal latent-process model in which biodiversity distributions, ecological intensities, or related response surfaces are explained by environmental covariates, EBVs, ECVs, and structured spatial and temporal dependence. Presence-only data, detection and non-detection data, and count data are linked to this latent process through dataset-specific observation models. This makes the framework suitable for integrated species distribution modelling and closely aligned with current statistical ecology practice for biodiversity inference from heterogeneous sources (Isaac et al., 2020; Koshkina et al., 2017).

The deliverable also explicitly restores and expands the artificial intelligence (AI) and machine learning (ML) sections. The AI/ML layer is not treated as a substitute for ecological-statistical inference, but as a complementary analytical and operational layer for EO and UAV feature extraction, habitat and land-cover mapping, wildlife detection, anomaly detection, advanced EBV computation, and downstream early warning and decision support. This positioning is supported by the Description of Action and by the UAV software and data-fusion architecture already defined in D1.1 (Prassou et al., 2025).

The pilot capability mapping included in the source package confirms that the framework must be modular. The pilot portfolio spans Northern Finland, Austria, Romania, Czechia, the European Arctic, Crete, and northern Belgium, each with different ecological settings, stressors, temporal demands, data streams, and expected outputs.

D3.1 therefore defines a common modelling core combined with pilot-specific adaptations in observation models, covariates, temporal support, and operational outputs. This is essential for preserving methodological coherence while enabling valid implementation across forests, tundra, agricultural landscapes, wetlands, and Mediterranean ecosystems (Prassou et al., 2025).

The principal outputs of D3.1 are methodological and operational. They include a common model specification, a pilot-adaptive covariate logic, a clear definition of uncertainty products, structured interfaces to AI/ML modules, and a pathway toward scenario analysis, early warning, and decision support. These outputs are designed to support later tasks on deep learning, system-level implementation, case-study application, and calibration and validation, while remaining scientifically defensible and consistent with Horizon Europe reporting expectations (Prassou et al., 2025).

# 1. Introduction

## 1.1 Purpose of the document

The purpose of this deliverable is to define the hierarchical statistical framework that the BioClima project will use for dynamic ecological assessment of biodiversity under climate change. More specifically, the document establishes the modelling logic, data-integration principles, uncertainty strategy, and implementation pathway needed to connect heterogeneous biodiversity observations with EO and environmental covariates in a joint analytical framework. In the project architecture, D3.1 provides the methodological baseline for later model enhancement, EBV computation, early warning, and decision-support development.

## 1.2 Relation to other project work

The deliverable is tightly connected to the broader analytical architecture of the project. WP1 defines the data, metadata, harmonisation, interoperability, and platform logic through in situ, citizen-science, EO, UAV, and crowdsourcing resources. WP2 investigated the impact of climate-change related extreme events on ecosystem structure, functioning, health and vulnerability using EBVs and ECVs. WP3 develops the hierarchical statistical framework, AI-enhanced analytics, and system-facing components for monitoring, early warning, and decision support. WP4 applies these outputs in case studies, strategies, and implementation contexts. Within this architecture, D3.1 functions as the methodological bridge between ecological inference and operational analytics, and as a foundational input to Tasks 3.2 and 3.3, as well as to Tasks 2.1–2.4.

## 1.3 Structure of the document

This document is organised into several sections to provide a comprehensive presentation to the proposed hierarchical statistical framework. As an introduction, Chapter 2 provides the ecological, statistical, and AI/ML background for the framework and clarifies the role of species distribution models (SDMs), integrated modelling, EBVs, ECVs, and ecological risk assessment. Chapter 3 then defines the ecological model, the AI/ML model, the supporting software environment, and the interface between the two. Chapter 4 links the framework to the pilots and to the platform capabilities identified for implementation. Chapter 5 translates D1.1 into model-facing requirements. To conclude the document, Chapter 6 defines the main outputs of D3.1, its implementation recommendations, and its principal conclusions.

# 2. Background on the ecological and AI/ML models

## 2.1 Ecological background

Reliable knowledge of biodiversity patterns across space and time is essential for effective conservation, climate adaptation, ecological restoration, and land management. In the present project, this need is particularly strong because the targeted systems span boreal and Arctic regions, forests and wetlands, agricultural mosaics, Mediterranean pine systems, and areas exposed to drought, wildfire, land-use intensification, and ecosystem degradation. Under such conditions, monitoring static biodiversity status is insufficient; what is required is a dynamic framework capable of describing change, dependence, risk, and response under shifting environmental and management conditions.

The central ecological premise is that biodiversity change is driven jointly by species-specific processes, environmental gradients, climatic stress, habitat structure and landscape configurations. This means that the inference problem is unavoidably both biological and statistical: the same ecological state may be recorded differently by structured surveys, opportunistic observations, UAV-based monitoring, or EO-derived proxies. A credible framework must therefore infer ecological structure from incomplete and uneven observational evidence rather than assume that observed records are direct and unbiased representations of the full ecological process (Lavergne et al., 2010, Bellard et al., 2012).

### 2.1.1 Species distribution modelling

SDMs are among the principal model-based tools for linking biodiversity occurrence observations to environmental conditions. Their purpose is to estimate how biodiversity responses (for example, presences, absences and abundance) vary across geographic space and time, and how these responses are associated with covariates such as climate, land cover, habitat configuration, topography, disturbance, or management. Modern SDMs are increasingly expected to support not only descriptive mapping but also temporal extrapolation, scenario analysis, and uncertainty communication for operational and policy contexts. This favours explicit statistical formulations over opaque correlation-based methods, especially in settings with heterogeneous data sources and decision-facing outputs (Elith and Leathwick, 2009; Franklin, 2013; Zurell et al., 2020).

A central challenge in applying SDMs is accounting for spatial and temporal dependence. Biodiversity observations collected close together in space or time tend to be more similar than those collected farther apart, and failing to account for this dependence can lead to biased parameter estimates, residual autocorrelation, and overconfident inference. Therefore, D3.1 treats structured spatial and temporal effects as fundamental components of SDMs rather than optional refinements, ensuring that ecological processes and uncertainty are represented more realistically (Cressie and Wikle, 2015).

### 2.1.2 Species occurrence data

D1.1 shows that the project relies on a rich but heterogeneous biodiversity observation landscape. Open infrastructures such as the Global Biodiversity Information Facility (GBIF), together with citizen-science platforms (for example, iNaturalist) and monitoring networks (for example, the United Kingdom Butterfly Monitoring Scheme), greatly expand the spatial and temporal reach of biodiversity data and make large-scale ecological inference feasible in ways that were not possible in earlier decades. At the same time, these sources differ substantially in taxonomic coverage, observer effort, metadata richness, spatial accessibility bias, and the degree to which absence or effort information is available. This diversity is valuable, but only if handled with an appropriate observation model (GBIF, 2026; iNaturalist, 2026; Dennis et al., 2017; Prassou et al., 2025; Isaac et al., 2020).

D1.1 is especially clear on the need to avoid naive aggregation of heterogeneous occurrence data. Presence-only data dominate many large biodiversity archives and are affected by uneven effort and accessibility. Structured monitoring schemes can counterbalance some of these weaknesses but often lack the same coverage. The statistical implication is direct: the workflow should not reduce all datasets to the lowest common denominator. It should preserve different observation processes and exploit their complementary strengths within a shared inferential structure (Prassou et al., 2025; Isaac et al., 2020; Simmonds et al., 2020).

### 2.1.3 Integrated species distribution models

Integrated species distribution models (ISDMs) respond directly to the challenge of data heterogeneity. In this class of models, multiple biodiversity datasets are treated as different observations of the same underlying ecological process, while each dataset retains its own observation model. This allows presence-only records, structured survey data, detection/non-detection data, and count data to be combined without discarding information about how each source was generated. The resulting models improve the coherence of ecological inference and often reduce uncertainty relative to single-source models when observation biases are modelled explicitly (Isaac et al., 2020; Koshkina et al., 2017; Simmonds et al., 2020).

Integrated modelling is therefore an appropriate choice to implement as a direct response to the project's data architecture. D3.1 documents the coexistence of open occurrence data, structured monitoring streams, EO and UAV derivatives, and crowd-interpreted contextual layers. The most defensible response to that architecture is a hierarchical integrated framework capable of preserving source-specific assumptions while generating unified ecological inference (Prassou et al., 2025).

### 2.1.4 AI/ML models

The AI/ML component in BioClima should be understood as a family of geospatial and ecological learning models embedded within the wider hierarchical statistical framework, rather than as a single standalone algorithm. Its primary role is to increase the amount, resolution, and operational value of information that can be extracted from the project observation system. This includes information from EO, UAV, in-situ, citizen-science, sensor, and derived model products. At the same time, the ecological-statistical framework, based around ISDMs, remains the scientific backbone for inference on species distributions, ecosystem state, and uncertainty. AI/ML therefore complements, extends, and operationalises the statistical framework; it does not replace it.

This distinction is important because the BioClima data architecture is intrinsically multi-source and multi-scale. D3.1 documents a monitoring ecosystem that combines spaceborne data from Sentinel, Landsat, Moderate Resolution Imaging Spectroradiometer (MODIS), ERA5, High-Resolution Vegetation Phenology and Productivity (HR-VPP)-related products, and other Copernicus resources; local and hyperlocal UAV products from Red-Green-Blue (RGB), multispectral, thermal infrared and LiDAR sensors; ground-based biodiversity and climate observations; citizen-science records; internet of things (IoT) and weather-station time series; and derived or crowdsourced layers such as biomass-change drivers and habitat suitability indicators. AI/ML is required to transform this evidence base into model-ready products, especially where raw observations are too high-dimensional, too noisy, or too spatially detailed to be used directly in hierarchical ecological models (Prassou et al., 2025). We can distinguish between four types of AI/ML models for this modelling framework.

A first AI/ML model family consists of interpretable supervised machine-learning methods applied to fused geospatial features. Random Forests, gradient-boosting models, and related classifiers can provide robust baselines for habitat classification, vegetation-condition assessment, forest-health screening, and land-cover change detection. These models are particularly useful in early implementation phases because they can handle mixed predictors, nonlinear responses, and moderate training-data volumes, while still offering practical interpretability through feature-importance diagnostics and class-probability outputs. In D3.1, such models should be treated as baseline and benchmarking components before more complex deep-learning architectures are introduced (Breiman, 2001; Prassou et al., 2025).

A second model family consists of computer-vision and deep-learning models for EO and UAV imagery. Convolutional Neural Networks, U-Net-style segmentation architectures, transformer-based segmentation models, and object-detection networks can support pixel-level habitat and land-cover mapping, detection of vegetation degradation, extraction of canopy and surface-condition patterns, and automated wildlife detection from UAV or camera-trap imagery. In practical terms, these models can generate classified rasters, detection layers, density heatmaps, habitat masks, and confidence scores. These products can then serve either as direct platform outputs or as covariates, validation layers, or local observation supports for the hierarchical statistical model.

A third model family concerns spatio-temporal learning. Many BioClima indicators are not static (for example, species range shifts); they are expressed as trajectories of vegetation productivity, phenology, land surface temperature, soil moisture, snow cover, burned area, canopy condition, or habitat degradation. Sequence-aware models, temporal anomaly detectors, and recurrent or attention-based architectures can be used to identify deviations from expected ecological behaviour, distinguish seasonal dynamics from stress responses, and summarise time-series signals into features that can be consumed by ecological models or early-warning components. This is particularly relevant for drought, heat stress, wildfire exposure, vegetation decline, shrubification, and climate-driven ecosystem transitions.

A fourth model family is graph-based and connectivity-oriented learning. Ecological risk is often mediated by spatial configuration: habitat fragmentation, corridor continuity, distance to refugia, land-use barriers, edge effects, and adjacency between ecological units. Graph Neural Networks and related spatial learning approaches can represent landscape patches, habitat cells, sensor locations, or ecological units as nodes linked by spatial, functional, or hydrological relationships. In BioClima, this logic is relevant for mapping ecological connectivity, identifying fragmentation pressure, supporting corridor analysis, and preparing features for spatial conservation planning.

The AI/ML layer also provides the analytical basis for Task 3.2 (Reinforcement learning in ecological data analysis of essential biodiversity variables). In that context, Multi-Agent Reinforcement Learning (MARL) should be understood as a scenario-exploration and decision-support method rather than as a conventional prediction model. Landscape units, management zones, or ecological planning cells can be represented as agents or decision entities that evaluate possible actions, such as conserving, restoring, monitoring, restricting pressure, or prioritising intervention. Reward functions can integrate biodiversity suitability, species richness proxies, habitat condition, carbon sequestration, wildfire exposure, management cost, and local constraints. This allows the system to explore trade-offs and synergies between biodiversity conservation, climate mitigation, and land-management priorities.

The outputs of the hierarchical statistical model are central to this reinforcement-learning layer. Species probability surfaces, occupancy or intensity estimates, uncertainty maps, habitat suitability layers, EBV-derived products, and scenario-conditioned environmental covariates become the state variables, constraints, or reward components of MARL experiments. Conversely, MARL outputs should feed back into the decision-support system as prioritisation maps, policy scenarios, recommended monitoring actions, or comparative what-if analyses. This two-way relationship ensures that the AI layer remains grounded in statistically defensible ecological inference and does not operate as an opaque optimisation engine detached from uncertainty and ecological meaning.

A further role of AI/ML is anomaly detection and early-warning preparation for Task 3.3 (System Architecture, Implementation Plan and Data Quality Assessment). Early warning systems used in the BioClima project cannot be limited to fixed threshold exceedance, because ecological degradation is often gradual, context-specific, and visible only through the combination of multiple indicators. AI/ML models can support detection of unusual trajectories in land surface temperature, vegetation indices, canopy health, soil moisture, burned-area risk, habitat fragmentation, or species-probability surfaces. However, alert generation should remain confidence-aware. Signals should be accompanied by uncertainty information, provenance, and the indicator family that triggered the warning. They should also clearly indicate whether the signal represents an observed change, a modelled risk, or a scenario-based projection.

This background also implies a strict validation logic. AI/ML outputs must be calibrated and validated against field measurements, structured observations, UAV-derived high-resolution products, expert interpretation, and statistically inferred uncertainty surfaces. The relevant validation metrics will differ by task: classification accuracy and confusion matrices for habitat maps, intersection-over-union or mean pixel accuracy for segmentation, precision and recall for species detection, error metrics for continuous EBV or ECV estimates, and stability tests for scenario and reinforcement-learning outputs. Validation should also include spatial transferability checks between pilots and explicit documentation of model assumptions, training data, preprocessing steps, and known limitations.

Operationally, the AI/ML layer must be interoperable with the BioClima platform architecture. D1.1 already defines a Python-based UAV integration package organised around ingestion, harmonisation, fusion, EBV/ECV derivation, AI analytics, validation, and pipeline orchestration. This same design logic should guide the wider AI/ML implementation in WP3. Models should be deployable as modular services, support geospatial formats such as GeoTIFF, NetCDF, GeoJSON and tabular outputs, preserve metadata and provenance, and remain compatible with Spatio-Temporal Asset Catalog (STAC)-based catalogues, Open Geospatial Consortium (OGC) APIs, Findable, accessible, interoperable and reproducible (FAIR) data principles, and controlled-access platform components (Prassou et al., 2025).

Finally, AI/ML in D3.1 must remain explainable and decision-facing. The objective is not only to improve predictive accuracy, but also to produce outputs that ecologists, pilot teams, policy stakeholders, and platform users can interpret and act upon. For this reason, the AI/ML layer should expose confidence scores, feature-importance indicators where available, validation summaries, model versions, and links to source data. Human-in-the-loop review remains essential, especially for ecological alerts, conservation prioritisation, and scenario-based management recommendations. In this way, Neuralio's WP3 leadership role is reflected in an AI/ML architecture that is technically advanced, operationally reusable, and scientifically aligned with the hierarchical ecological framework of D3.1.

## 2.2 Current applications of the ecological model

Integrated biodiversity modelling has already shown utility in large-scale settings where disparate species-occurrence data and environmental covariates are combined within FAIR-aligned analytical workflows to produce hotspot maps and associated uncertainty surfaces across the heterogeneous landscape of Norway. That experience is relevant because it demonstrates that the model family adopted in D3.1 is mature enough to support operational ecological summaries in real-world settings with heterogeneous inputs and strong reproducibility requirements (Finstad et al. 2025; Wilkinson et al., 2016; Isaac et al., 2020).

The same modelling logic extends naturally to temporal applications, where range shifts, phenology, multi-year trends, and climate-linked changes can be represented through time-indexed covariates and structured temporal random-effects. This is particularly relevant in the present project because several pilots involve long-term warming, vegetation-state shifts, drought pressure, wildfire exposure, and land-use transitions which might not be captured adequately by purely static models.

## 2.3 Perceived challenges

The first major challenge is computational. Spatio-temporal integrated models become increasingly demanding as the number of time periods, latent effects, observation streams, and hyperparameters grows. Even when ecological logic would support high temporal or taxonomic resolution, practical implementation may require aggregation choices, staged workflows, or reduced complexity in specific pilots.

The second challenge concerns metadata and observational interpretation. Many biodiversity data streams do not arrive with sufficiently rich effort metadata or fully reconstructable sampling protocols. This limits the direct specification of observation models and requires closer interaction with data providers, pilot teams, and domain experts to recover plausible bias structures or effort surrogates. Without that step, model sophistication alone cannot guarantee ecological validity (Prassou et al., 2025).

The third challenge is cross-scale integration. The project combines broad-coverage EO variables with local and high resolution UAV measurements, structured field observations, and pilot-specific process-model outputs such as PREBAS- or wildfire-related layers. These inputs do not align naturally in spatial support, time step, uncertainty profile, or analytical role. A central task of D3.1 is therefore to define how these sources are harmonised without damaging the ecological meaning of the underlying data (Prassou et al., 2025).

### 2.3.1 Proposed solutions to the challenge

The first solution is methodological modularity. The framework is built around a common latent-process logic and a common uncertainty policy, but allows flexible observation models, covariate families, and temporal structures. This enables a shared inferential grammar across pilots without forcing identical datasets or identical model complexity in every case.

The second solution is computational pragmatism. The preferred Bayesian route is Integrated nested Laplace approximation (INLA) based estimation for latent Gaussian models, because it offers major computational advantages over conventional Markov chain Monte Carlo (MCMC) methods in the model family relevant to D3.1. Where necessary, staged, sequential, or hybrid fitting strategies may be used to reduce the computational burden, provided that the assumptions are explicit and that uncertainty propagation is preserved as far as possible (Rue et al., 2009; Figueira et al., 2024).

The third solution is a clear division of roles between ecological inference and AI/ML. The hierarchical statistical model remains responsible for ecologically grounded inference and structured uncertainty, while AI/ML modules handle image-derived analytics, higher-dimensional feature extraction, local detection or classification, anomaly detection, and downstream operational tasks. This division is both scientifically sound and practically aligned with the Description of Action and D1.1 (Prassou et al., 2025).

## 2.4 Essential climate variables and Essential biodiversity variables

The project is explicitly organised around EBVs and ECVs. D1.1 shows that the observational system includes variables relevant to ecosystem structure, ecosystem extent, habitat condition, species distribution, abundance, and phenology, alongside climate-related variables such as land surface temperature, soil moisture, precipitation, evapotranspiration, land cover, snow cover, and vegetation indices. The Description of Action positions EBVs and ECVs as methodological connectors between data acquisition, modelling, case-study interpretation, and policy relevance (Prassou et al., 2025).

Within this deliverable, EBVs and ECVs serve three roles. First, they act as structured covariate families or response-linked constructs in statistical modelling. Second, they provide a shared conceptual vocabulary for comparing results across pilots and work packages. Third, they form the bridge to downstream use, because outputs expressed in EBV and ECV terms are more readily transferable to monitoring systems, scenario analysis, early warning, and conservation planning (Prassou et al., 2025).

## 2.5 Estimating species risk and ecological risk assessment

Dynamic modelling of species distributions, occupancy, ecological intensity, or habitat condition can support risk-oriented interpretation. In conservation settings, such information often relates to decline in range, change in occupancy, habitat contraction, or broader ecological vulnerability. Although D3.1 does not attempt to replicate formal Red List workflows (for example, the IUCN Red List of Threatened Species), it does establish the statistical conditions under which risk-relevant indicators can later be derived, such as: rates of change, sensitivity to environmental and climate changes, uncertainty-bounded decline signals, habitat degradation patterns, or range-shift proxies (IUCN Standards and Petitions Committee, 2024).

In the BioClima project, ecological risk assessment extends beyond individual species to include habitat degradation, structural and functional ecosystem stress, wildfire exposure, drought-related vulnerability, and biodiversity shifts under climate or management scenarios. The term “risk assessment” is therefore used in D3.1 in a broad but rigorous sense: the production of uncertainty-aware quantities that can support later classification, prioritisation, or warning functions without overstating what the available data alone can establish.

# 3. Description of the methodological approach

## 3.1 Introduction

The methodological architecture combines a hierarchical ecological-statistical core with an AI/ML extension layer. The ecological-statistical core is responsible for inference on biodiversity distributions, temporal change, and uncertainty. The AI/ML layer is responsible for augmenting this framework through EO and UAV feature extraction, image-based analysis, advanced EBV computation, anomaly detection, and downstream system integration. Together, these components provide a modelling stack consistent with both the project work plan and the practical requirements documented in D1.1 (Prassou et al., 2025).

### 3.1.1 Description of the ecological model

The ecological model follows a state-space logic in which the latent ecological process is separated from the observation process. At the process level, the model represents biodiversity intensity, suitability, occupancy, or a related continuous ecological response as a function of environmental covariates and structured spatial and temporal effects. Firstly, we assume that the observed point pattern of species observations over some geographic region $\Omega \in \mathfrak{R}^2$ and time period $T \in \mathfrak{R}$ is a realization of an inhomogeneous Poisson Point process model. With this assumption, the number of points within this region is assumed to be distributed according to a Poisson distribution, with mean given by:

$$\mu_\Omega = \int_T \int_\Omega \lambda(s,t)dsdt,$$

Where: $\lambda(s,t)$ denotes the expected ecological intensity or related latent response at some location $s$ and time $t$. A generic representation of the core process is:

$$\lambda(s,t) = exp(\eta(s,t))$$

$$\eta(s,t) = \alpha_0 + \alpha_t + \sum_{i=1}^{p} \beta_i X_i(s,t) + \omega(s,t)$$

where, $\alpha_0$ is a term capturing baseline intensity, $\alpha_t$ captures time-period effects, $X_i(s,t)$ are environmental covariates varying across space and time, chosen for their relevance in describing distribution patterns for a given species, $\beta_i$ are associated coefficients to the environmental covariates, and $\omega(s,t)$ is a spatial or spatio-temporal latent random field accounting for residual dependence and unmeasured drivers not included in the model. This follows the logic of log-Gaussian Cox processes (LGCPs) and related latent Gaussian models (LGMs) that are well established in spatial ecology and integrated species distribution modelling. Since this model is represented as a LGM, the Bayesian INLA methodology may be used to provide a fast and accurate approximations to the posterior marginals for the latent effects and hyperparameters (Møller et al., 1998; Illian et al., 2008; Isaac et al., 2020; Rue et al. 2009).

The exact structure of $\alpha_t$ and $\omega(s, t)$ is flexible and may thus vary by application. In simpler settings, time effects may be fixed or independent random effects, and spatial fields may be replicated across time slices. In more complex settings, the model may include persistent and transient spatial components, autoregressive temporal structure, temporally varying covariate effects, or case-specific temporal supports. D3.1 deliberately preserves this flexibility because the pilot portfolio differs significantly in time horizon, observation density, environmental forcing, and ecological process scale.

Furthermore, since these models are estimated in a Bayesian framework, appropriate prior distributions need to be chosen for all the unknown parameters of the model. LGCPs are known to be highly sensitive to the choice of prior distributions, making the selection of priors a critical part of the analysis. Additionally, the chosen priors need to be justifiable, ecologically relevant and interpretable. For the fixed effects of the model, it is typical to assume weakly informative priors to allow the data to dominate the effect. Since the spatial effects typically dominate the models, penalising complexity (PC) priors may be considered. These priors are also weakly informative, and are designed to pull the model towards its simplest realisation to reduce overfitting (Sørbye et al. 2018; Fuglstad et al., 2019).

At the observation level, each dataset is assigned a model compatible with its sampling and detection process. Presence-only data are perhaps the most common source of occurrence data in ecology, and are handled through point-process or thinned point-process formulations, given that they are often collected opportunistically and thus influenced by various sampling biases (for example, collecting observations closer to urban environments). In this model, the data is thought to be generated by two processes: the ecological model described by a Poisson point process model, as described above, and an independent thinning model $p(s, t)$ that describes the likelihood of observing a given species, conditioned on some location $s$ and time period $t$ (Sicacha-Parada et al., 2021; Warton et al., 2013).

The thinning function $p(s, t)$ is typically described by a log-linear function of flexible spatial effects or spatial-covariates $z_j(s, t)$ used to account for observer bias, with associated coefficients $\gamma_j$. Mathematically, this is given by:

$$p(s,t) = \gamma_0 + \sum_{j=1}^{z} \gamma_j Z_j(s,t),$$

Where $\gamma_0$ is an intercept term representing baseline thinning. If the covariates influence both the intensity and the detectability of the species, then it may be difficult to separate out the effect and estimate absolute intensity. Data integration methods may provide a solution: combining presence-only data with structured survey data that allows the estimation of detection probabilities and ensures the identifiability of model parameters (Miller et al., 2019).

Detection/non-detection data are linked through Bernoulli-type models connected to the latent process. In this case, data are often collected with a binary '0' for absence or '1' for at least one presence within some grid cell. These data may be linked to the ecological model described above by modelling the probability that there is at least one individual in location $s$ at time $t$ $p(N(s,t) > 0)$ with a complementary log-log link function. Mathematically this is given by:

$$log(-\ log(1 - p(N(s,t) > 0))) = \eta(s,t),$$

Which as shown by previous applications, links the probability of detection/non-detection to the intensity function of an underlying Poisson process (Kéry and Royle, 2015).

Finally, count or abundance data may be represented naturally through Poisson or negative-binomial (commonly chosen in situations where the data is overdispersed) components where appropriate. Here, data are typically collected at defined sampling sites across area $B_k$, $k = 1, 2,..., K$, and the number of individuals detected are noted. In this case, the number of counted individuals in the $k^{th}$ sampling site follows a Poisson distribution with a mean defined as:

$$\mu_k = p \cdot \int_T \int_B \lambda(s,t) ds dt,$$

Where $p$ is the probability of observing an individual. If the sampling area is small enough, the above equation may be approximated as:

$$\mu_k \approx p \cdot |B| \cdot \bar{\lambda},$$

Where: $|B|$ is the area of $B$ and $\lambda(s,t) = \bar{\lambda}\ \forall s \in B$ . If $|B|$ is available, it is natural to include it as an offset to model densities (species per some unit of area) and to account for varying cell sizes. However, the sample area is often not well defined and thus $|B|$ may be unavailable. Furthermore, $p$ can often only be obtained through auxiliary information such as repeated counts. In these cases, these terms are often assumed to be constant across observations, which leads to an unidentifiable intercept term.

Putting these all together, the likelihood function for the ecological model may be written as a combination of the process model and the individual observation models:

$$L(Y|X, \phi, \theta) = p(\lambda(s,t), X, \phi) \cdot \prod_{l=1}^{M} p(Y_l|\lambda(s,t), \theta_l),$$

Were: $p(\lambda(s,t), X, \phi)$ is the process model, $p(Y_l|\lambda(s,t), \theta_l)$ is the observation model for dataset $l$, $l = 1, 2, ..., M$, $\phi$ are parameters for the process model and $\theta_l$ are parameters associated with the $l^{th}$ observation model.

In later extensions, additional formulations may be used for structured event data or related source types, provided that the ecological interpretation and data-generation mechanism are explicit. This dataset-specific observation logic is one of the most important features of the framework because it preserves information and limits inferential distortion when heterogeneous sources are integrated (Isaac et al., 2020; Koshkina et al., 2017).

Typically, the models described above are tailored for small-scale case studies, and only use a subset of all the available occurrence data. However, in order to gain a holistic understanding on the state of biodiversity across large areas, these models need to be estimated for large groups of species using semi-automated, reproducible, and transparent workflows. These use well-defined pipelines to transform raw data into interpretable and functional estimates of species spatio-temporal patterns and trends, which can be used to influence management decision making (Mostert et al. 2025).

The primary outputs from these models will be spatio-temporal maps of predicted species intensity, which will provide a graphical summary of where biodiversity hotspots are located across an area at a relevant resolution, and how biodiversity has changed across some period of time. The coefficients estimated from the model will provide some insight on how climate- and land use changes can influence these dynamics. Moreover, the predictions can be integrated with future climate and land-use forecasts to illustrate how biodiversity may respond due to a changing world. Furthermore, maps of sampling bias may be obtained from these models, which could highlight where future data collection needs to be prioritised.

In addition, a critical methodological principle is explicit uncertainty propagation. The DoA requires predictive insights together with their associated uncertainties, and this requirement applies across process estimation, observation bias handling, environmental covariates, and scenario-based projection. In practical terms, the framework must generate posterior summaries, uncertainty surfaces, interval estimates where appropriate, and provenance linking model outputs to specific data releases, preprocessing decisions, and modelling configurations.

### 3.1.2 Description of the AI/ML model

The AI/ML model in D3.1 should be understood as a layered geospatial-AI architecture rather than as a single standalone estimator. Its purpose is to augment the hierarchical ecological-statistical framework with scalable learning components that can process high-dimensional EO, UAV, in situ, citizen-science, sensor, and scenario data. The ecological-statistical model remains responsible for estimating latent ecological states and their uncertainty. The AI/ML layer is responsible for transforming complex data streams into model-ready EBV and ECV features, identifying spatial and temporal patterns that are difficult to specify manually, supporting scenario exploration, and providing analytical outputs for the reinforcement learning and warning systems as required in Tasks 3.2 and 3.3 respectively. This division of responsibilities is important for the BioClima project because it preserves ecological interpretability while enabling the operational analytics expected from WP3, including reinforcement learning, early warning, monitoring, and decision support (Prassou et al., 2025).

From a modelling perspective, the AI/ML layer receives five main families of inputs. The first family consists of EO raster and time-series products, including multispectral, radar, thermal infrared data, vegetation-indices, phenology, productivity, land cover, biomass, soil moisture, and climate layers. The second family consists of UAV products, including RGB imagery, multispectral imagery, thermal infrared imagery, LiDAR-derived terrain and canopy products, and sensor-specific derivatives. The third family consists of in situ and monitoring data, including field measurements, habitat inventories, weather stations, Internet-of-Things time series, camera-trap imagery, and structured biodiversity observations. The fourth family consists of citizen-science and participatory data, especially species observations and contextual information that can enrich occurrence data and local interpretation. The fifth family consists of scenario and process-model outputs, such as forest-growth, carbon, hydrological, climate, and wildfire-related layers. The AI/ML model does not treat these inputs as interchangeable. Instead, each input is associated with its spatial support, temporal resolution, provenance, quality information, and intended role in the broader ecological workflow.

The first functional layer is multi-sensor feature extraction. In D1.1 of the BioClima project, the UAV integration package already defines a practical implementation basis for this role through ingestion, harmonisation, fusion, EBV/ECV derivation, AI analytics, validation, and pipeline orchestration modules. At baseline, fused multi-sensor features such as canopy height, NDVI, NDRE, thermal indicators, topographic attributes, and habitat labels can be analysed with interpretable machine-learning models such as Random Forests. This is appropriate for early-stage habitat classification, forest-health mapping, site-level EBV derivation, feature screening, and the production of transparent pilot-specific baselines. The same modular interface should remain open to gradient boosting, convolutional models, transformer-based segmentation, and other advanced methods where the data volume, annotation quality, and pilot objectives justify their use.

The second functional layer is image-based detection and segmentation. Wildlife population monitoring requires automated or semi-automated detection of individuals from UAV imagery, camera traps, or video feeds. For this role, object-detection models such as YOLO-type architectures can be used to generate bounding boxes, confidence scores, species-count tables, and density heatmaps. In the Romanian pilot, this logic is directly relevant to turtle detection from UAV RGB and thermal data, while similar image-detection workflows may support other pilots where camera traps or UAV imagery are available. Habitat and land-cover mapping requires a related but distinct modelling family: pixel-level or object-level classification of EO and UAV imagery using methods such as U-Net-like convolutional segmentation models, SegFormer-like transformer models, or hybrid approaches. The resulting products should include classified habitat maps, class-area statistics, class probabilities, and quality layers, rather than only hard class labels.

The third functional layer is ecological connectivity and landscape-structure learning. Land-cover and habitat maps can be transformed into graph representations in which nodes represent habitat patches, landscape units, monitoring sites, or ecological features, and edges represent spatial adjacency, permeability, distance, corridor strength, or movement constraints. Graph-based methods can then support analysis of habitat fragmentation, corridor prioritisation, and landscape connectivity. This is particularly relevant for pilots where biodiversity risk depends not only on local habitat conditions but also on the arrangement of habitats across the landscape. In this role, graph neural networks and related spatial-learning methods are not replacements for ecological theory; they are tools for representing structured spatial relationships in a form that can be coupled to ecological interpretation and conservation planning.

The fourth functional layer is EBV and ECV computation and enrichment. AI/ML outputs should be explicitly translated into variables that are meaningful for the BioClima monitoring framework. For ecosystem extent, the AI/ML layer can provide classified habitat and land-cover products. For ecosystem structure, it can provide LiDAR-derived canopy height, vertical complexity, vegetation roughness, and habitat heterogeneity. For ecosystem function and condition, it can provide vegetation indices, productivity proxies, phenology indicators, forest-health masks, drought-stress indicators, and thermal-stress proxies. For species and habitat distribution, it can provide detection-assisted occurrence support, microhabitat features, and habitat suitability covariates. For ECV-related interpretation, it can provide local land-surface temperature, microclimate indicators, soil-moisture proxies, and climate-driver representations. These outputs should be treated as derived analytical products with uncertainty and provenance, not as direct ecological truth.

The fifth functional layer is the interface between AI/ML and the hierarchical ecological model. This interface is bidirectional. Upstream of the ecological model, AI/ML products can serve as covariates, bias indicators, habitat masks, observation-support layers, or validation datasets. Examples include segmentation-derived habitat maps, UAV-derived canopy-height products, thermal refugia layers, detection-assisted occurrence points, and EO-derived stress indicators. Downstream of the ecological model, posterior predictive surfaces, uncertainty maps, model residuals, EBV-derived layers, and scenario-conditioned predictions can be used by AI/ML modules for pattern recognition, anomaly detection, prioritisation, and reinforcement learning. The key methodological rule is that uncertainty must not be discarded at the interface. If an image classifier produces uncertain labels, or an object detector has low confidence, that information should be propagated as a confidence, probability, or quality layer. Similarly, posterior uncertainty from the hierarchical model should be passed to downstream decision tools so that decisions can distinguish high-risk/high-confidence cases from high-uncertainty cases that require more monitoring.

The sixth functional layer is scenario-aware learning and MARL for Task 3.2. In this layer, the landscape can be represented as a set of interacting spatial units or agents. Each unit can be described by a state vector containing EBV indicators, ECV indicators, habitat condition, species probability or intensity estimates, posterior uncertainty, carbon or biomass indicators, disturbance exposure, management constraints, and socioeconomic or operational context where available. Candidate actions may include conservation, restoration, monitoring intensification, fuel-load reduction, habitat connectivity enhancement, land-management change, or no intervention, depending on the pilot and decision context. The reward function can then combine biodiversity objectives, carbon sequestration, ecosystem condition, risk reduction, implementation cost, and stakeholder-defined constraints. In this interpretation, MARL is not introduced as an opaque optimisation tool. It is a structured scenario-exploration layer that uses statistically grounded ecological products to compare potential interventions under climate, land-use, and disturbance scenarios.

A generic conceptual representation for the reinforcement-learning interface is therefore: state equals ecological condition plus climate pressure plus uncertainty plus management context; action equals an intervention or monitoring decision; reward equals a weighted combination of biodiversity gain, carbon or ecosystem-function gain, risk reduction, cost, and policy constraints. This formulation is deliberately generic because the pilots differ in ecological process, available data, and decision objective. Northern Finland may emphasise biodiversity-carbon trade-offs and spatial prioritisation, Austria may emphasise wildfire risk and forest impacts, Romania may emphasise turtle detection and habitat degradation, Czechia may emphasise agricultural and post-disturbance landscape change, the European Arctic may emphasise climate-driven vegetation transitions, Crete may emphasise forest health and thermal stress, and Belgium may emphasise grassland biodiversity hotspots and land-management resilience. D3.1 should therefore define the AI-ready data objects and interface logic, while the pilot-specific reward functions and learning configurations are refined in Task 3.2.

The seventh functional layer is anomaly detection, early warning, and decision-support preparation for Task 3.3. AI/ML models can be used to detect departures from expected ecological trajectories, abrupt land-cover or vegetation-condition changes, anomalous thermal patterns, drought or wildfire precursors, habitat degradation signals, and unexpected changes in species-related indicators. These warnings should be derived from combinations of time-series models, spatial anomaly detection, scenario-conditioned thresholds, and residual analysis from the ecological model. The output should not be a single alarm without explanation. It should include the spatial footprint of the signal, the relevant EBV/ECV drivers, the confidence or uncertainty level, the temporal persistence of the anomaly, and the provenance of the data layers that generated it. This is essential for producing warnings that are useful to stakeholders and compatible with the broader BioClima decision-support system.

The AI/ML model must also support explicit validation and uncertainty assessment. For classification and segmentation tasks, appropriate metrics include overall accuracy, class-wise precision and recall, F1 score, confusion matrices, and intersection-over-union-type measures. For object detection, evaluation should include detection precision, recall, localisation quality, species-count error, and performance under different flight heights, sensor types, seasons, and background conditions. For probabilistic outputs, calibration diagnostics are needed so that predicted probabilities can be interpreted consistently. For time-series and early-warning outputs, validation should include temporal hold-out tests, spatial block validation, pilot-level cross-validation, and expert review where ground truth is limited. Validation outputs should be stored together with the model version, training data version, input data release, preprocessing configuration, and performance summary.

Interpretability is a core requirement rather than an optional add-on. Random-Forest-style feature importance, interpretable covariate summaries, partial-dependence-type diagnostics, spatial residual maps, saliency or attention diagnostics for image models, and rule-based post-processing summaries can all help connect AI/ML outputs back to ecological interpretation. This is especially important because the Description of Action recognises the risk that complex AI models may be difficult for stakeholders to understand. The BioClima AI/ML layer should therefore favour transparent reporting, traceable workflows, human-in-the-loop review for sensitive outputs, and clear separation between scientific prediction, model uncertainty, and management recommendation.

Operationally, the AI/ML model should be implemented as a deployable and interoperable set of services rather than as isolated research scripts. The D1.1 software logic already points toward a Python-based geospatial stack using libraries such as rasterio, geopandas, numpy, and scikit-learn, with extensibility toward deep-learning frameworks. The WP3 platform requirements add the need for GPU-enabled training and inference, scalable batch processing of imagery, model versioning, experiment tracking, containerised execution, and reusable APIs. Outputs should be encoded in formats that can be consumed by the rest of the BioClima platform, such as GeoTIFF for rasters, NetCDF or CoverageJSON for gridded or time-series coverages, GeoJSON or OGC API - Features for detections and vector outputs, JSON for model metadata and summaries, and OGC API - Maps or WMS for visual layers.

Finally, the AI/ML model should be governed by a common set of quality and provenance rules across pilots. Each AI/ML output should document the input data sources, spatial and temporal support, preprocessing steps, model family, model version, training and validation data, performance metrics, uncertainty or confidence representation, and intended ecological interpretation. This common reporting logic will allow the AI/ML layer to remain flexible across pilots while preserving comparability and scientific traceability. In summary, the AI/ML component of D3.1 is a modular operational extension of the hierarchical ecological framework: it converts complex observations into EBV/ECV-relevant features, supports image analytics and connectivity modelling, prepares scenario-aware reinforcement-learning inputs, and supplies early-warning and decision-support products without weakening the uncertainty-aware ecological foundation of the deliverable.

## 3.2 Description of available software

The software environment must support both rigorous ecological inference and scalable AI/ML implementation. The project documentation indicates that this can be achieved through a dual-language architecture centred on R for the ecological-statistical core and Python for AI/ML, EO and UAV processing, and modular orchestration (Prassou et al., 2025).

### 3.2.1 Description of available software for the ecological model

The ecological model is most naturally implemented in R, which remains the principal language of statistical ecology and offers a mature ecosystem for hierarchical and spatial modelling. The recommended modelling stack includes PointedSDMs for user-facing integrated species distribution modelling, R-INLA for latent Gaussian approximation, inlabru for spatial model construction, sf for vector geospatial handling, terra for raster processing, and ggplot2 for model-aware visualisation. This combination is coherent, reproducible, and directly aligned, as required with our statistical family (Mostert and O'Hara, 2023; Rue et al., 2009; Bachl et al., 2019; Wickham, 2016).

This software layer is especially appropriate because it supports explicit specification of latent fields, dataset-specific likelihoods, structured random effects, and posterior uncertainty. It also aligns well with pilot-specific workflows in which raster extraction, model fitting, diagnostics, and map production must remain linked and reproducible rather than dispersed across unrelated scripts or environments.

### 3.2.2 Description of available software for the AI/ML model

The AI/ML layer is more naturally implemented in Python, as already demonstrated in D1.1 of the BioClima project. The immediately relevant stack includes rasterio, geopandas, numpy, and scikit-learn for geospatial processing, feature engineering, and baseline machine learning. For pilots that require deep-learning-based image analytics or sequence learning, the same modular architecture can be extended with modern deep-learning libraries and deployment practices, while preserving compatibility with the project's metadata, validation, and orchestration logic (Prassou et al., 2025).

The pilot capability requirements add further operational detail. Wildlife monitoring requires GPU-accelerated training and inference, model versioning, and scalable image processing. Habitat mapping requires direct support for multiband EO and UAV data, evaluation metrics suited to segmentation and classification, and a container-friendly deployment pathway. Forest projection and wildfire-risk workflows require scenario-ready services and reusable APIs. These needs imply that the AI/ML software environment in D3.1 should be conceived from the outset as deployable, container-aware, and interoperable with platform services rather than merely as a research prototype.

## 3.3 Linking the ecological model to the AI/ML model

The link between the ecological model and the AI/ML layer is one of the most consequential design decisions in this deliverable. The ecological model provides a statistically principled representation of biodiversity patterns and their uncertainty. The AI/ML layer provides additional analytical power for image-derived features, advanced EBV computation, simulation support, higher-dimensional pattern recognition, anomaly detection, and operational scenario analysis. The two must therefore be linked through explicit interfaces that preserve semantic meaning, provenance, and quality indicators rather than through opaque hand-offs. An interface structure is summarised in Table 1.

Table 1: Interface structure.

| Data Type | Description | Linked EBVs/ECVs | Data Type |
|---|---|---|---|
| Species probability or intensity surfaces | Ecological model | AI/ML modules, pilots, system layer | Advanced EBV computation, hotspot analysis, prioritisation |
| Uncertainty maps and posterior summaries | Ecological model | AI/ML modules, decision-support layer | Risk-aware filtering, anomaly interpretation, confidence-aware outputs |
| EO and UAV feature rasters | AI/ML and data-proces sing layer | Ecological model | Covariates, habitat proxies, micro-scale explanatory variables |
| Detection or segmentation products | AI/ML image models | Ecological model, pilot tools | Local occurrence support, habitat degradation layers, validation inputs |
| Scenario-conditio ned covariate layers | Scenario-pr eparation layer | Ecological model and decision tools | Forecasting, stress testing, scenario exploration |

| Data Type | Description | Linked EBVs/ECVs | Data Type |
|---|---|---|---|
| Prioritisation-ready biodiversity layers | Ecological model and AI/ML synthesis | Conservation-planning modules | Spatial ranking, protection or management support |

# 4. Linking to the case studies

## 4.1 Description of the case studies

The pilot portfolio confirms that D3.1 must remain modular, scalable, and ecologically sensitive to context. Across the pilots, the framework must support biodiversity inference in boreal and Arctic environments, wildfire-prone forests, steppe and riparian habitats, agricultural mosaics, Mediterranean pine systems, and grassland landscapes under intensive land-use pressure. These differences are substantial, but they do not weaken the case for a common framework. On the contrary, they make a common framework essential, provided that it supports pilot-specific configuration of observation models, covariates, time structures, and outputs. In this sense, the pilots should be understood not only as examples of application, but also as practical tests of whether the framework can operate across different ecosystem types, spatial scales, data availability levels, and management or decision-support needs (Prassou et al., 2025).

The integrated pilot table below synthesises the pilot use case matrix supplied for this revision and translates it into implementation logic for D3.1. The table below is used here as a project-facing design instrument rather than as a bibliographic source.

Table 2: Pilot use case matrix related to the deliverable.

| Pilot area | Main objective | Main ecosystems and land use | Principal stressors | Main inputs | Relevant EBVs and ECVs | Main implications for D3.1 |
|---|---|---|---|---|---|---|
| Northern Finland | Estimate current and future carbon stocks and fluxes; identify areas with high biodiversity and carbon-sequestration potential | Boreal forests, mountain birch, heath, wetlands, protected areas | Rapid warming, moth outbreaks, reindeer overgrazing | Habitat inventories, Landsat, Sentinel-1/2, LiDAR, PREBAS- and YASSO-related outputs | GPP, NPP, phenology, vegetation height, NEP, soil carbon, biomass carbon, snow cover | Requires hybrid coupling between biodiversity inference, carbon-related process-model outputs, and spatial prioritisation using conservation-planning tools; strong emphasis on scenario-ready outputs and cross-scale integration |

| Pilot area | Main objective | Main ecosystems and land use | Principal stressors | Main inputs | Relevant EBVs and ECVs | Main implications for D3.1 |
|---|---|---|---|---|---|---|
| Austria | Identify future wildfire hotspots and evaluate impacts on ecosystem services and biodiversity | Alpine forests, grasslands, croplands, protected areas | Wildfire and climate change | Historical burned areas, climate drivers, forest structure, soil properties | Forest structure, burned area, deadwood, litter, soil carbon, temperature, humidity, wind, precipitation | Requires linkage between ecological risk modelling, wildfire scenarios, and forest-growth modelling, with re-usable spatial outputs for scenario assessment |
| Romania | Monitor biodiversity and climate variables at local scale; detect species and habitat degradation from UAV data | Dry grasslands, steppe, riparian systems, turtle habitats | Drought, grazing pressure, tourism, erosion, instability | Turtle surveys, vegetation plots, geomorphology, UAV RGB, multispectral, thermal and LiDAR, Sentinel-2, HR-VPP | Species presence and abundance, habitat composition and structure, land surface temperature, soil moisture, vegetation indices | Requires hyperlocal multi-sensor AI embedded within the broader statistical framework, especially for wildlife detection and habitat-state covariates |

| Pilot area | Main objective | Main ecosystems and land use | Principal stressors | Main inputs | Relevant EBVs and ECVs | Main implications for D3.1 |
|---|---|---|---|---|---|---|
| Czechia, Vysočina | Monitor agricultural and biodiversity responses to climate change and management shifts | Agricultural landscapes, post-calamity forests, orchards, grasslands, ponds, hedgerows | Drought, heatwaves, extreme weather, bark beetle outbreaks, invasive species | Permanent plots, weather stations, IoT sensors, Sentinel-1/2/3, HR-VPP, UAV data, citizen-science inputs, process models | Ecosystem function, structure, composition, land surface temperature, precipitation, snow cover | Requires multi-source integration across management, hydrology, ecology, and change detection, with flexible temporal structure and local validation |
| European Arctic | Quantify biodiversity change under accelerated warming and link it to climate and land use | Tundra, boreal forest, wetlands, heathland, ecotones | Accelerated warming, drought, late frost, grazing, borealisation, shrubification | Field plots, surveys, sensor networks, Sentinel-2/3, UAVs, Copernicus products, LPJ-GUESS and ORCHIDEE outputs | Structure, function, composition, surface temperature, $CO_2$ flux, snow cover, soil temperature, soil moisture, shrub cover | Requires strongly spatio-temporal models tied to earth-system and ecosystem-change variables and capable of representing large-scale climate-driven transitions |

| Pilot area | Main objective | Main ecosystems and land use | Principal stressors | Main inputs | Relevant EBVs and ECVs | Main implications for D3.1 |
|---|---|---|---|---|---|---|
| Greece, Crete | Assess climate-extreme impacts on Mediterranean pine forest health, structure, and function | Mediterranean pine forest in protected areas | Drought, thermal stress, pest pressure | Canopy condition, mortality and health data, UAV RGB, multispectral, thermal and LiDAR, Sentinel data | NPP, LAI, vegetation height, land surface temperature, precipitation, snow cover | Requires close integration of EO, UAV, and field observations for forest-health EBVs, thermal-stress proxies, and biodiversity-condition monitoring |
| Northern Belgium | Identify biodiversity hotspots and carbon-sequestration potential in grasslands and support resilient land management | Agricultural grasslands from intensive to semi-natural systems | Intensive agriculture, nitrogen deposition, urbanisation, altered precipitation, drought | Field plots, sensors, Sentinel-1/2/3, HR-VPP, Copernicus products, citizen-science inputs, hydrological and soil data | Ecosystem function, species composition, land cover, productivity, phenology, land surface temperature, evapotranspiration, soil moisture, soil organic carbon, CO2 flux | Requires habitat suitability modelling, Bayesian inference, geo-AI, and spatial optimisation to support hotspot identification and management strategies |

The purpose of this chapter is not to claim full empirical completion of all pilots. Its purpose is methodological. The pilots define the range of observation types, environmental drivers, stressors, temporal needs, and output requirements that the common framework must support. They therefore function as design constraints and validation contexts for D3.1 rather than as a set of completed results.

The pilot capability mapping also identifies the main functional capabilities that the platform and modelling architecture must support. These capabilities translate pilot-specific ecological questions into operational analytical requirements.

**Table 3: Pilot use case function capabilities matrix related to the deliverable.**

| Functional capability | Purpose in the D3.1 context | Representative models or algorithms | Principal outputs | Most relevant pilots |
|---|---|---|---|---|
| Wildlife population monitoring | Detect and count animals from imagery and support local biodiversity inference | Object detection and image classification workflows | Detections, density heatmaps, count tables | Romania |
| Habitat and land-cover mapping | Derive habitat, vegetation, and land-cover products used as inputs or outputs | Segmentation and EO/UAV classification workflows | Classified maps, class statistics, habitat layers | Romania, Czechia, northern Belgium, European Arctic, Crete, northern Finland |
| Habitat suitability modelling | Estimate spatial biodiversity suitability or probability surfaces | Bayesian SDMs, hierarchical regressions, INLA-based models | Probability surfaces, uncertainty maps, coefficients | Northern Belgium and, more generally, all pilots requiring species-distribution inference |

| Functional capability | Purpose in the D3.1 context | Representative models or algorithms | Principal outputs | Most relevant pilots |
| --- | --- | --- | --- | --- |
| Forest ecosystem projection | Simulate growth, carbon balance, and future forest conditions | PREBAS-type suites and related ecosystem models | Time series of GPP, biomass, carbon stocks, stand projections | Northern Finland |
| Wildfire risk modelling | Estimate hotspot risk and biodiversity impacts under scenarios | Fire-risk models and linked forest-growth workflows | Fire hotspots, burned-area projections, impact summaries | Austria |
| Spatial conservation planning | Rank areas for conservation or management action | Zonation-type optimisation and multi-criteria planning workflows | Priority maps, performance curves | Northern Finland, with potential extension to Austria and northern Belgium |

This capability structure has direct implications for D3.1. It confirms that the ecological core must generate outputs that are both scientifically meaningful and technically reusable, and that some downstream uses rely on image analytics, containerised services, or optimisation modules beyond the narrow bounds of conventional SDMs. D3.1 must therefore define not only a model, but a reusable inferential layer within a wider analytical system.

# 5. Data and resource requirements

## 5.1 Earth Observation Data

EO data form the main environmental predictor base of D3.1 and are indispensable for both spatial generalisation and temporal analysis. D1.1 identified a broad EO landscape including Sentinel-1, Sentinel-2, Sentinel-3, Landsat, MODIS, derived biomass products, ERA5, HR-VPP-related methodologies, and additional Copernicus datasets. Together, these products support modelling of land cover, vegetation condition, hydrology, phenology, thermal stress, biomass, and broader ecosystem structure and function (Prassou et al., 2025).

Sentinel-1 contributes radar-based information relevant to vegetation structure, water content, and soil moisture, while Sentinel-2 provides multispectral optical data for habitat mapping, vegetation indices, and ecosystem-condition estimation. Sentinel-3 contributes land-surface-temperature information of high value for drought and heatwave studies. Landsat data adds historical depth, and MODIS products contribute phenological and recurrent observational context. The Bioclima project also makes use of processed products such as biomass maps, land-cover layers, anomaly metrics, and atmospheric or hydrological climate datasets. This diversity should be treated as a structured covariate family within D3.1 rather than as an undifferentiated stack of raster layers. For robust pilot implementation, the use and interpretation of individual EO layers should explicitly account for their spatial resolution, temporal frequency, cloud-cover sensitivity, harmonisation level, and the availability of suitable ground validation data, especially when translating EO products into EBV- and ECV-relevant indicators.

D1.1 further specified that EO access and harmonisation are organised through a standards-based platform environment using object storage, STAC principles, APIs, and controlled access. For D3.1, this has direct methodological relevance. EO covariates must be spatially and temporally aligned, described with consistent metadata, and retrievable in a reproducible way. Their provenance and processing status must remain attached to model-ready layers so that the resulting statistical outputs remain traceable and reusable downstream (Prassou et al., 2025; Wilkinson et al., 2016).

## 5.2 Species occurrence data and related resources

Species occurrence data will be central to D3.1 implementations and will often come from open biodiversity infrastructures, especially GBIF, as well as structured monitoring schemes and citizen-science streams. D1.1 is explicit that these sources are rich but methodologically uneven and that structured data with interpretable effort remain essential for understanding observation bias and informing observation models. The relevant data basis therefore includes occurrence records, ground observations, event-based monitoring data, and ecological metadata where available (Prassou et al., 2025; GBIF, 2026).

The project also depends on in situ and monitoring infrastructures beyond strictly biological occurrence records. D1.1 identifies ground-based climate and ecosystem resources such as ICOS and other site-level measurement networks that can support calibration, validation, and ecological interpretation. In other pilots, permanent plots, weather stations, IoT sensor networks, mortality records, and stakeholder observations contribute additional context. These resources are needed not only for validation but also for ecological model specification and interpretation (Prassou et al., 2025).

UAV and airborne products form a further major resource class. In Romania and Crete, D1.1 shows that LiDAR, thermal, RGB, and multispectral UAV products can contribute to habitat mapping, condition assessment, canopy metrics, local thermal proxies, and wildlife monitoring. The UAV integration software further demonstrates that these products can be transformed into EBV and ECV objects, uncertainty-aware classification outputs, and pilot-ready summary layers. Within D3.1, such UAV-derived products can function as covariates, local ecological evidence, or validation layers for EO- and occurrence-based inference (Prassou et al., 2025).

Crowdsourcing resources are also relevant. D1.1 identifies a Geo-Wiki-based tool for attributing forest biomass gains and losses to drivers. From the perspective of D3.1, this means that crowd-interpreted change drivers can become explanatory layers, validation resources, or interpretive context linked to biodiversity and ecosystem change. These products are especially useful where raw EO change signals alone do not provide a defensible account of ecological drivers (Fritz et al., 2012; Prassou et al., 2025).

A summary view of the data basis relevant to D3.1 is provided in Table 4.

Table 4: Data basis relevant to the deliverable.

| Data family | Main examples | Principal role in D3.1 | Linked EBVs and ECVs |
|---|---|---|---|
| EO data | Sentinel-1/-2/-3, Landsat, MODIS, biomass products, ERA5, HR-VPP-related layers | Covariates, climate forcing, habitat proxies, scenario inputs | Ecosystem structure, ecosystem extent, phenology, land cover, land surface temperature, soil moisture, snow cover |

| Data family | Main examples | Principal role in D3.1 | Linked EBVs and ECVs |
|---|---|---|---|
| Airborne and UAV data | LiDAR, multispectral, RGB, thermal | Local ecological evidence, habitat and health layers, structural metrics, validation support | Canopy structure, habitat condition, surface-temperature proxies, species and habitat distribution proxies |
| In situ and occurrence data | GBIF, professional surveys, monitoring schemes, plots, field observations | Model response data, observation-process definition, calibration and validation | Species distribution, abundance, composition, phenology |
| Sensor and IoT data | Weather stations, soil-moisture sensors, local environmental monitoring | Pilot-scale climate context and validation | Soil moisture, temperature, precipitation, habitat condition |
| Derived and crowdsourced data | Geo-Wiki outputs, anomaly metrics, habitat suitability layers, ecosystem products | Explanatory context, driver layers, validation, downstream interpretation | Ecosystem structure, species distribution, land cover, biomass |

In addition to data, D3.1 requires substantial resource infrastructure. Large raster stacks, high-resolution UAV products, spatio-temporal Bayesian estimation, and

GPU-based AI workflows imply a need for scalable compute, storage, containerisation, controlled access, metadata stewardship, and reproducible code management. These are not merely technical conveniences; they directly influence what models can realistically be estimated, validated, and reused across pilots (Prassou et al., 2025).

# 6. Conclusion

The central conclusion of this deliverable is that the project's analytical needs are best served by a hierarchical, integrated, and uncertainty-aware statistical framework capable of combining heterogeneous biodiversity observations with climate, land-use, and ecosystem variables across space and time. This conclusion follows directly from the data architecture documented in D1.1, the task logic and deliverable requirements defined in the Description of Action, and the ecological and operational diversity of the pilots (Prassou et al., 2025).

The framework is strong precisely because it does not oversimplify the project's evidence base. It treats observational heterogeneity, spatial and temporal dependence, and uncertainty as structural features of the problem. It also treats AI/ML as an integrated but distinct analytical layer, preserving ecological interpretability while enabling higher-order image analytics, advanced EBV computation, and later early-warning functions. This combination makes D3.1 both scientifically rigorous and operationally relevant. By combining pilot-specific model configuration with harmonised EBV/ECV-oriented outputs, the framework provides a practical basis for generating interpretable information that pilot teams and stakeholders can use for biodiversity monitoring, spatial prioritisation, early-warning preparation, and decision-support processes.

# 7. Bibliography

[1] Bachl, F. E., Lindgren, F., Borchers, D. L., & Illian, J. B. (2019). inlabru: An R package for Bayesian spatial modelling from ecological survey data. Methods in Ecology and Evolution, 10(6), 760-766.

[2] Bellard, C., Bertelsmeier, C., Leadley, P., Thuiller, W., & Courchamp, F. (2012). Impacts of climate change on the future of biodiversity. Ecology letters, 15(4), 365-377.

[3] Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.

[4] Cressie, N., & Wikle, C. K. (2015). Statistics for Spatio-Temporal Data. Wiley.

[5] Dennis, E. B., Morgan, B. J., Brereton, T. M., Roy, D. B., & Fox, R. (2017). Using citizen science butterfly counts to predict species population trends. Conservation biology, 31(6), 1350-1361.

[6] Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics, 40, 677–697.

[7] Figueira, M., Conesa, D., López-Quílez, A., & Paradinas, I. (2025). A computationally efficient procedure for combining ecological datasets by means of sequential consensus inference. Environmental and Ecological Statistics, 32(2), 495-521

[8] Finstad, A.G., Perrin, S.W., Adjei, K.P., Mostert, P.S., Togunov, R.R., Herfindal, I., Topper, J.P., Grytnes, J., Chipperfield, J. & O'Hara, R.B. (2025). Arbeidsflyt for modellering av heildekkande utbreiingskart for atsmangfald. Published online: https://www.miljodirektoratet.no/publikasjoner/2025/mai-2025/arbeidsflyt-for-modellering-av-heildekkande-utbreiingskart-for-artsmangfald/

[9] Franklin, J. (2013). Species distribution models in conservation biogeography: developments and challenges. Diversity and distributions, 19(10), 1217–1223.

[10] Fritz, S., McCallum, I., Schill, C., Perger, C., See, L., Schepaschenko, D., van der Velde, M., Kraxner, F., & Obersteiner, M. (2012). Geo-Wiki: An online platform for improving global land cover. Environmental Modelling & Software, 31, 110–123.

[11] Fuglstad, G. A., Simpson, D., Lindgren, F., & Rue, H. (2019). Constructing priors that penalize the complexity of Gaussian random fields. Journal of the American Statistical Association, 114(525), 445-452.

[12] GBIF. (2026). Global Biodiversity Information Facility data portal. Available online: https://www.gbif.org/

[13] Illian, J., Penttinen, A., Stoyan, H., & Stoyan, D. (2008). Statistical Analysis and Modelling of Spatial Point Patterns. Wiley.

[14] iNaturalist (2026). iNaturalist. Retrieved on June 30 2026 from https://www.inaturalist.org.

[15] Isaac, N. J. B., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., Freeman, S. N., Golding, N., Guillera-Arroita, G., Henrys, P. A., Jarvis, S., Lahoz-Monfort, J., Pagel, J., Pescott, O. L., Schmucki, R., Simmonds, E. G., & O'Hara, R. B. (2020). Data integration for large-scale models of species distributions. Trends in Ecology & Evolution, 35(1), 56–67.

[16] IUCN Standards and Petitions Committee. (2024). Guidelines for Using the IUCN Red List Categories and Criteria.

[17] Kéry, M., & Royle, J. A. (2015). Applied Hierarchical Modeling in Ecology: Analysis of distribution, abundance and species richness in R and BUGS: Volume 1: Prelude and Static Models. Academic press.

[18] Koshkina, V., Wang, Y., Gordon, A., Dorazio, R. M., White, M., & Stone, L. (2017). Integrated species distribution models: Combining presence-background data and site-occupancy data with imperfect detection. Methods in Ecology and Evolution, 8(4), 420–430.

[19] Lavergne, S., Mouquet, N., Thuiller, W., & Ronce, O. (2010). Biodiversity and climate change: integrating evolutionary and ecological responses of species and communities. Annual review of ecology, evolution, and systematics, 41(1), 321-350.

[20] Miller, D. A., Pacifici, K., Sanderlin, J. S., & Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. Methods in Ecology and Evolution, 10(1), 22-37.

[21] Mostert, P. S., & O'Hara, R. B. (2023). PointedSDMs: An R package to help facilitate the construction of integrated species distribution models. Methods in Ecology and Evolution, 14(5), 1200–1207.

[22] Mostert, P. S., Bjørkås, R., Bruls, A. J., Koch, W., Martin, E. C., & Perrin, S. W. (2025). intSDM: an R package for building a reproducible workflow for the field of integrated species distribution models. Ecology and Evolution, 15(3), e71029.

[23] Møller, J., Syversveen, A. R., & Waagepetersen, R. P. (1998). Log Gaussian Cox processes. Scandinavian Journal of Statistics, 25(3), 451–482.

[24] Prassou, E., Tsilimanis, G., Sandric, I., Mostert, P. S., & Shchepashchenko, D. (2025). D1.1 Analysis of Data for Integration.

[25] Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B, 71(2), 319–392.

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

[27] Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J. B., & O'Hara, R. B. (2020). Is more data always better? A simulation study of benefits and limitations of integrated distribution models. Ecography, 43(10), 1413–1422.

[28] Sørbye, S. H., Illian, J. B., Simpson, D. P., Burslem, D., & Rue, H. (2019). Careful prior specification avoids incautious inference for log-Gaussian Cox point processes. Journal of the Royal Statistical Society Series C: Applied Statistics, 68(3), 543-564.

[29] Warton, D. I., Renner, I. W., & Ramp, D. (2013). Model-based control of observer bias for the analysis of presence-only data in ecology. PloS one, 8(11), e79168.

[30] Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer.

[31] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018.

[32 ] Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J. J., Leitão, P. J., Park, D. S., Peterson, A. T., Rapacciuolo, G., Schmatz, D. R., Schröder, B., Serra-Diaz, J. M., Thuiller, W., Yates, K. L., Zimmermann, N. E. & Merow, C. (2020). A standard protocol for reporting species distribution models. Ecography, 43(9), 1261–1277.