Funded by

AETHER-xAI: SELF-EXPLAINABLE AI FOR EARTH OBSERVATION

Requirements and Activity Baseline Report

[Ondertitel van document]

Authors:

M.J.R. (Rob) Knapen, Domantas Giržadas, Arun K. Pratihast, Manuel Winograd, and Maciej J. Soja	Wageningen Environmental Research, the Netherlands
Mirjam Fredriksen and Stefan Jetschny	NILU, Norway
Thijs L. van der Plas	Wageningen University, the Netherlands
Martin Karlson	Linköping University, Sweden
Piotr Wężyk	University of Agriculture Kraków, Poland

Version: 1.0

Date: 15 Dec. 2025

Citation guide:

M.J.R. Knapen, D. Giržadas, A.K. Pratihast, M. Winograd, M.J. Soja, M. Fredriksen, S. Jetschny, T.L. van der Plas, M. Karlson, and P. Wężyk (2025), *AETHER-xAI: Requirements and Activity Baseline Report*, ESA Contract No 4000149494, Version 1.0, December 2025

Acknowledgement:

This work has been funded by European Space Agency project “AETHER: AI for Earth Transparency using Human Explainable Reasoning” (ESA Contract No 4000149494). Dr Peter Naylor (ESA technical officer) and the entire AETHER team are acknowledged for their contribution to this document and the entire project.

Abstract

The Self-Explainable Artificial Intelligence (S-xAI) for Earth Observation (EO) project AETHER develops and demonstrates a transparent EO-AI modelling approach that unites deep learning with explainable reasoning and knowledge grounding. Its architecture integrates spatiotemporal EO embeddings, semantically grounded concept base representations, and retrieval-augmented generation modules to translate satellite data into physically meaningful variables and stakeholder-oriented explanations. Concept annotations are automatically derived from auxiliary spatial data via rule-based templates, enabling large-scale, weakly supervised training while preserving scientific traceability.

AETHER will design, implement, and evaluate a proof-of-concept system across three use cases: (i) detection of urban heat islands and their evolution due to global warming, (ii) crop yield prediction and rapid assessment of the effect of floods, droughts, and fires, and (iii) mapping of biodiversity and its loss due to climate change. The three use cases will share a common embedding backbone and explainability framework for consistency and reusability. The system will produce self-interpretable concept layers, accurate predictions, and text-based explanations grounded in both scientific evidence and stakeholder knowledge.

The proof-of-concept will span two or three of the use cases and will employ 10-30 representative concepts per use case, requiring under ten GPU-weeks for training and less than one terabyte of storage. The results will showcase a scalable, efficient, and trustworthy S-xAI framework that bridges EO observation, concept-based interpretation, and human-centric explanation, advancing transparency, reproducibility, and reliability in EO-AI applications for environmental science and decision support.

TABLE OF CONTENTS

1. Introduction *11*
2. Methodology *12*
2.1. Literature review 12
2.2. Gap analysis 13
2.3. Proposed S-xAI Methodology 14
2.4. Requirements 18
2.5. Software & Hardware 19
2.6. EO Data & Knowledge 21
2.7. Ethical & Privacy Aspects 23
3. Use Cases *25*
3.1. Use case 1: Biodiversity 26
3.2. Use case 2: Crop Yield 27
3.1. Use case 3: Urban Heat Islands 29
4. End Users and Impact *32*
4.1. Use case 1: Biodiversity 32
4.2. Use case 2: Crop Yield 32
4.3. Use case 3: Urban Heat Islands 33
5. Summary & Conclusion *34*
References *35*

Acronyms and Abbreviations (alphabetical order)

AETHER	AI for Earth Transparency using Human-Explainable Reasoning
AI	Artificial Intelligence
AI4EO	Artificial Intelligence for Earth Observation
AUC	Area Under the Curve
CL	Contrastive Learning
CV	Computer Vision
DL	Deep learning
ECV	Essential Climate Variables
ESA	European Space Agency
ESP	Earth Systems Predictability
EO	Earth Observation
GenAI	Generative Artificial Intelligence
GUI	Graphical User Interface
IPCC	Intergovernmental Panel on Climate Change
LLM	Large Language Model
ML	Machine Learning
mIoU	Mean Intersection over Union
POC	Proof of Concept
RAG	Retrieval Augmented Generation
RB	Requirements and Activity Baseline
SOTA	State Of The Art
SoW	Statement of Work
S-xAI	Self-eXplainable Artificial Intelligence
TCAV	Testing with Concept Activation Vectors
TRL	Technology Readiness Level

UC	Use Case
VLM	Vision Language Model
xAI	Explainable Artificial Intelligence

Definitions

The following definitions, many like those of the SoW, are used in the context of this RB report:

Artificial Intelligence (AI). The Oxford Dictionary defines Artificial Intelligence as the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages. Artificial Intelligence is a branch of computer engineering, designed to create machines that behave like humans. *More generically AI focusses on the study and construction of agents (things that act) that do the right thing (that behave rationally).*

Within this document, the term "AI" will therefore be used mainly as a generic term to refer to Machine Learning, Natural Language Processing, Computer Vision, and other techniques adapted to work with Earth Observation data. The term "AI4EO" will refer to the use of EO data with AI techniques.

AI technique. It is an AI-based way of achieving a task. A single AI technique can be applied to various EO use-cases.

AUC (Area Under the Curve) is a scalar metric that summarises performance across all thresholds; in xAI it typically refers to the area under deletion or insertion curves, measuring how rapidly a model's prediction confidence decreases or increases as the most important features (per an explanation) are progressively removed or added. Higher AUC indicates more faithful explanations.

Bottleneck architecture is a machine learning concept where the information flow is reduced to a lower dimensional representation. In the field of deep learning, this concept denotes a vector layer where the preceding and following layers are larger, forcing the model to learn a suitable compression and decompression of the (usual) high-dimensional information.

Chunk is a small, usually fixed-size text window extracted for use by an *LLM* in training or retrieval-augmented generation. Chunks are typically defined in *tokens* (e.g., 256 – 1000 tokens) with optional overlap, and they serve as the direct analogue of EO *patches*.

Clip is a cropped subset of an EO *image* or *tile* extracted using a user-defined area of interest (AOI) such as a polygon or bounding box. Clips have variable size and shape and are used to restrict computation to the relevant geographic region.

Concept is a meaningful, human-understandable feature or property used by e.g. a *Concept Based Model* to describe an input. Concepts serve as intermediate variables between raw data and predictions and can be binary, categorical, or continuous. They are typically chosen because they are semantically clear to domain experts and useful for explaining or guiding the model's predictions.

Concept Based Model (**CBM)** is a machine learning model that makes predictions through an intermediate layer of human-interpretable *concepts*. Instead of mapping inputs directly to outputs, it first infers a set of predefined concepts (e.g., "leaf is yellow", "soil moisture is low") and then uses those concept values to produce the final prediction. This design aims to improve interpretability, allow concept-level supervision, and enable users to inspect or intervene in the model's reasoning.

Contrastive Learning is a machine learning approach where a model learns representations by comparing examples. It is encouraged to pull together ("positively pair") representations of similar or related inputs (e.g. two augmented views of the same image, or an Image and a caption) and push apart ("negatively pair") representations of dissimilar inputs. By learning to distinguish what should be close versus far in representation space, the model builds useful, general-purpose features without necessarily needed explicit labels.

Corpus is a collection of text sources assembled for training, retrieval, or evaluation of language models. A corpus is defined by shared scope, preprocessing rules, and metadata conventions, and is the text-processing analogue of EO *imagery*. Corpus is at the top of the typical LLM / Generative AI hierarchy: *Corpus – Document – Shard – Excerpt – Chunk – Token*.

Deep Learning (DL). It is a subfield of machine learning, using deep neural networks. With the depth of the model being represented by the number of layers, it is often considered that more than three layers (including input and output layer) qualifies as "deep" learning.

Document is a single text source within a *corpus*, such as a PDF, web page, report, or chat thread. Documents are the primary ingestion unit in Large Language Model (LLM) pipelines and typically carry metadata like title, author, date, language, and provenance.

Downstream task. A downstream task is a specific application or prediction problem that uses representations learned earlier (often in a pretraining stage). After a model learns general features, such as *embeddings* via *contrastive* or *self-supervised* learning, it is adapted or evaluated on downstream tasks like classification, retrieval, segmentation, forecasting, or recommendation. The downstream task is "downstream" because it comes after and builds on the learned representations.

Embedding is a learned numerical representation of an item (such as a word, image, document, user or sensor record) as a fixed-length vector (usually with many dimensions). The embedding is trained so that important properties and relationships of the item are captured in the vector's values, making it easier for a model to compare items, find patterns, or use them in downstream tasks.

Embedding Space is the geometric vector space formed by embeddings, where each item is a point (vector) in that space. The space is structured so that distance and directions reflect semantic or functional similarity: items that are related or alike are placed close together, and unrelated items are far apart. Operations in the space (e.g., nearest-neighbour search, clustering, vector arithmetic) can therefore be used to reason about relationships between items.

EO Use-Case. It is a specific application in Earth Observation in which a product or service could potentially be used. Some examples of EO use-cases can be found on the ESA website. An EO use-case is agnostic of its potential solutions, and various solutions (in the scope of this document: AI techniques) can be proposed for a single EO use-case.

Excerpt (or segment) is a variable-length subset of a *document* selected because it matches a task, query, or structural boundary. Excerpts reflect "regions of interest" in text, such as a section, paragraph range, or retrieved span.

Explainable AI (xAI) is a set of methods and model designs that make an AI system's decisions understandable to humans. It aims to reveal *why* a model produced a particular output, what

evidence it relied on, and how its internal reasoning connects inputs to predictions, so users can assess trustworthiness, fairness, and correctness.

Foundation models are a type of AI models that are trained on a massive amount of data and can be adapted to a wide range of tasks.

Generative AI (GenAI) is a class of machine-learning systems that learn patterns from existing data and then produce new content, such as text, images, audio, video, code, or structured data, that is statistically and semantically similar to what they were trained on. In practice, generative AI models don't just label or predict; they create, by sampling from a learned probability distribution over possible outputs conditioned on a prompt or context.

Imagery is a collection of Earth Observation images acquired by a sensor and treated as a coherent product for analysis. Imagery often implies shared spatial reference, acquisition context (time, orbit, sensor), processing level, and metadata, and can refer to s scene set or a multi-temporal stack rather than a single file. Imagery is at the top of the typical EO-AI hierarchy: *Imagery – Image – Tile – Clip – Patch – Pixel*.

Image is a single EO raster representing one acquisition over a geographic area. In EO-AI, an image is typically multi-band (e.g., spectral bands beyond RGB) and georeferenced, meaning every pixel corresponds to a real-world location and ground sampling distance.

Input-Level Explainability refers to methods that explain a prediction by pointing to which parts of the input influence it and how. The focus is on linking the model's decision to specific input features, regions, tokens, or time steps (e.g., "these pixels", "these words", "this sensor segment") that were most responsible for the output.

Large Language Model (LLM) is a neural network trained on large-scale text (and sometimes other modalities) to learn statistical patterns of language for generating, transforming, or interpreting text. LLMs operate over *token* sequences and can be adapted to tasks like question answering, summarisation, extraction, and reasoning via prompting or fine-tuning.

Machine Learning (ML) is the study of computer algorithms that learn how to improve automatically through experience. It is seen as a part of *Artificial Intelligence*. Machine Learning algorithms build a model based on sample data, known as "training data", to make predictions or decisions without being explicitly programmed to do so.

mIoU (mean Intersection over Union) is a localisation metric that measures the average overlap between an explanation map (e.g. saliency or attribution) and a reference region or mask (such as a field boundary or land-cover proxy), computed as the intersection divided by the union of the two areas. Higher mIoU indicates more spatially accurate explanations.

Model-Level Explainability refers to understanding the overall behaviour and internal logic of the model, rather than a single prediction. It aims to describe how the model works globally, its learned rules, representations, decision pathways, and typical failure modes, so users can reason about what the model tends to do across many inputs. Examples include interpretable architectures, global surrogate models, rule extraction, and analysis of learned concepts or features.

Output-Level Explainability refers to methods that make the output itself more interpretable and informative. This includes explaining what the output means, how confident the model is, and how different output components relate to each other (e.g., probabilities, uncertainty ranges, concept scores, or textual rationales that clarify the prediction).

Open vocabulary querying is the ability to search an *embedding space* using arbitrary natural-language queries, without being limited to a fixed set of predefined labels or classes. Both the query and the items in the collection are embedded into the same vector space, and retrieval is done by similarity (e.g., nearest neighbours). Because the "vocabulary" is open, users can ask for *concepts* or descriptions the system was not explicitly trained to classify and still find semantically matching items.

Patch is a small, usually fixed-size window cut from an EO *image* or *clip* to serve as input to a *Machine Learning* model. Patches are commonly sampled with overlap / stride and may carry labels (annotations) for supervised learning, making them the core training and inference unit in EO-AI.

Pixel is the smallest addressable element of an EO *image*, storing one value per band for a specific ground area. Pixel size corresponds to spatial resolution, so each pixel represents a real-world footprint (e.g., 10 m x 10 m) and acts as the atomic unit from which *tiles, clips,* and *patches* are composed.

Prompt is the input, as text or other modalities, given to a *generative AI* model to specify the task, context, constraints, or desired output. Prompts guide the model's generation by conditioning what it produces, ranging from a short instruction to a structured template with examples, rules, or data.

Retrieval Augmented Generation (RAG) is an approach where a generative model (such as a *Large Language Model*) produces answers using both its learned parameters and external information retrieved at query time. Given a user query, the system first retrieves relevant documents or records (often via embeddings and similarity search) and then conditions the generator on that retrieved context to produce a grounded response. This helps improve factual accuracy, coverage of up-to-date knowledge, and traceability of outputs.

Scene is a single, sensor-defined EO acquisition covering a contiguous geographic area captured at one time, often corresponding to a satellite overpass or flight line. A scene is the natural "unit of capture" in remote sensing, and may later be processed into *imagery* products, tiled, clipped, or stacked into time series.

Self-eXplainable AI (S-xAI) refers to models that are inherently structured to produce explanations as part of their normal operation. Rather than adding an explanation after the fact, these models generate predictions through interpretable intermediate steps (e.g. *concepts*, rules, rationales, or explicit feature contributions), so the explanation is tightly coupled to the decision process.

Self-Supervised Learning is a training paradigm where a model learns from unlabelled data by creating its own supervision signal. It does this by solving a proxy (pretext) task whose labels are automatically derived from the data itself (e.g., predicting masked words in text, the next frame in

a sequence, or matching two augmented views of the same image). The goal is to learn general representations that can later be used for other tasks.

Shard (or partition / page) is a standardised subdivision of a *corpus* or large documents created for scalable storage or distributed processing. Shards are system-driven units (e.g., index shards, dataset file splits, or PDF pages) and are not necessarily aligned with model input boundaries.

TCAV (Testing with Concept Activation Vectors) is an explainable AI method that quantifies how much a human-defined *concept* (e.g. vegetation, imperviousness, crop stress) influences a model's prediction by measuring the directional sensitivity of internal neural activations to that concept, enabling explanations in terms that are meaningful to domain experts rather than individual features.

Tile is a standardized spatial subdivision of an EO *image* or *imagery* collection based on a fixed tiling grid. Tiles are primarily a data management and distribution unit that makes very large scenes easier to store, index, and process consistently.

Token is the smallest unit of text processing by an *LLM*, usually a subword or symbol produced by a tokenizer. Tokens are the atomic elements that form sequences and *chunks*, and model limits like context size are measured in tokens.

Tokenizer is a pre-processing component that converts raw inputs into a sequence of tokens (and back again). Depending on the modality, it may segment text into subwords or symbols, images into patches or visual codes, audio into frames or discrete units, or other signals into learned token forms. It defines the token vocabulary and mapping to integer IDs a model consumes, shaping how information is represented, how sequences are formed, and how limits like context size are measured.

Vision Language Model (VLM) is a multimodal neural network trained to jointly process visual inputs (images or video) and text so it can relate what is seen to what is said. VLMs learn shared representations across vision and language, enabling tasks such as image captioning, visual question answering, grounding text in images, and reasoning over combined visual-text context.

Weakly-Supervised Learning is a setting where a model is trained using imperfect, incomplete, or noisy labels instead of fully accurate annotations. The supervision may come from coarse labels (e.g., image-level tags instead of pixel labels), heuristic rules, distant supervision, or crowd-sourced annotations with errors. The model learns to make robust predictions despite the lower quality of the training signal.

1. Introduction

Earth Observation (EO) is a cornerstone of global environmental monitoring, providing continuous, large-scale data on the Earth surface and atmosphere. The growing availability of high-resolution satellite imagery, complemented by airborne and ground-based sensors, enables valuable insights into agriculture, urbanisation, water resources, and climate change, supporting evidence-based environmental and disaster management (Kansakar and Hossain, 2016). The EO landscape is rapidly expanding, producing petabytes of heterogeneous data that demand advanced computational tools for effective analysis (Vance et al., 2024). Situated within the broader "big data" paradigm—defined by volume, velocity, variety, and veracity, and often termed *Big Earth Data* (Sudmanns et al., 2020)—EO increasingly depends on scalable and interpretable analytical approaches.

Machine Learning (ML), and particularly Deep Learning (DL), have become central to EO data exploitation, enabling automated image analysis and predictive modelling for applications such as land cover classification and yield estimation (Paudel et al., 2021; Zhao et al., 2023). Advances in transformer-based foundational models have further expanded EO capabilities (Jakubik et al., 2023). However, DL models remain complex and opaque, raising concerns about interpretability, reproducibility, and trust (Hassija et al., 2024; Taskin et al., 2024). Challenges including hallucinations, limited generalisability, and biased training data have constrained broader adoption (Gawlikowski et al., 2023; Reichstein et al., 2019; Zhu et al., 2017).

To address these issues, eXplainable AI (xAI) has emerged to increase transparency and accountability in AI-driven EO systems (Höhl et al., 2024; Reichstein et al., 2019; Roscher et al., 2020; Wang et al., 2023). Self-*explainable AI* (S-xAI) integrates interpretability directly into model architectures, producing intrinsic explanations—such as reasoning traces or concept activations—without relying on post-hoc methods (Hou et al., 2024). Such approaches are especially relevant in EO, where interpretability underpins trust and scientific validation in high-impact applications like land use monitoring and climate resilience planning (Ghamisi et al., 2024; Taskin et al., 2024). In this context, we will implement a novel self-explainable AI framework that combines EO embeddings and concept-based model-level explainability, inspired by Meta's Large Concept Models, with output-level explainability through retrieval-augmented generation (RAG) and Large Language Models (LLMs). This integrated approach enhances transparency and contextual understanding, enabling structured reasoning and domain-grounded natural language explanations across various environmental EO use cases.

This document contains the introduction of the state-of-the-art (SOTA), gaps, and the proposed S-xAI architecture (Section 2), an overview of the SOTA and data for the use cases (Section 3), and a description of the end-users and impact (Section 4).

2. Methodology

This section contains a literature review (Section 2.1), gap analysis (Section 2.2), S-xAI architecture design and description (Section 2.3), requirements (Section 2.4), software and hardware considerations (Section 2.5), data and knowledge considerations (Section 2.6), and the ethical and privacy aspects (Section 2.7)

2.1. Literature review

As artificial intelligence becomes more broadly available and practically applicable, various domains explore the potential benefits of state-of-the-art machine learning and deep learning methods. Remote sensing and Earth observation domains are no exception. Large deep learning models are actively being developed and applied for a broad range of tasks, such as land use classification (Haider et al., 2025), change detection (Liu et al., 2024), flood risk prediction (Ruthra et al., 2025), and many others. Jakubik *et al.* (2023) have recently published their TerraMind model – a multi-modal "any-to-any" generative model, which outperforms several state-of-the-art models within the field of AI for EO on several relevant benchmark tests. The authors hope and expect the TerraMind model to be used as a foundation for numerous different downstream tasks. Foundation models and other pre-trained models are especially important in the field of Earth observation, where large deep learning architectures are often necessary to appropriately capture and represent complex spatial and temporal relationships. An extensive summary and evaluation of foundation models, applicable in the domain of remote sensing and Earth observation has recently been published by Xiao *et al.* (2024).

Despite the impressive performance of deep learning models on various benchmark tasks, the complexity of their decision-making process is proving to be problematic when the models are meant to inform high-impact decisions. For this reason, explainable AI (xAI) research has gained significant traction in the recent years. In fact, the intensive development in this domain has resulted in numerous different perspectives and approaches to improving the transparency of the decision-making process of AI models. A comprehensive systematic overview of xAI approaches within the domain of remote sensing has been made by Höhl *et al.* (2024).

While most of these methods provide insight about the link between model input and output, some argue that quantifying this link alone is not sufficient for a fully transparent and trustworthy decision. In their large-scale survey of explainable AI in environmental sciences, Schiller *et al.* (2025) point out that despite the recent growth of attention on explainable AI, there is a lack of focus on trust. The authors recommend for a more human-centric approach, where evaluating the user needs, gaining stakeholder trust, and defining explanations on a case-specific basis are of utmost importance. O'Loughlin *et al.* (2025) agree with the notion that *post hoc* input-output relationship explanations are not sufficient for a fully trustworthy model and advocate for component-level explainability. They support their recommendations with positive examples of physics-based modelling, where internal model calculations are grounded in well-defined physical processes, and discuss the potential of physics-informed AI modelling.

Meanwhile, the concept of self-interpretable (or self-explainable) AI is gradually gaining prominence in the field of deep learning. Self-interpretable neural networks take explainability

into account by design. There are multiple ways to approach this. Ji *et al.* (2025) offer a comprehensive survey of self-interpretable neural networks and organise the literature into five main categories: *attribution-based*, which explains predictions by highlighting influential inputs or features; *function-based*, which constraints model structure so its internal computations remain transparent; *concept-based*, which aligns decisions with human-understandable concepts; *prototype-based*, which justifies outputs by comparing them to representative examples; and *rule-based*, which expresses reasoning through explicit logical or decision rules. The authors also emphasise that concept-based self-interpretable models can allow direct human intervention, for instance by refining or correcting the concepts that guide the model's predictions.

Considering the benefits of human intervention-compatible models, it is, perhaps, unsurprising that there are numerous recent developments related to concept-based self-interpretable models. Among these developments, variations and improvements of the Concept Bottleneck Model (CBM) (Koh et al., 2020) are some of the most active and prominent. CBMs work by having the model first predict a set of human-interpretable concepts (such as attributes or intermediate properties) and then use those concepts, rather than raw features alone, to produce the final prediction, creating a "bottleneck" that makes the decision process more transparent and easier to inspect or adjust.

The definition of concepts and their interactions is proving to be a significant challenge with many caveats. Shang *et al.* (2024) highlight the difficulty of collecting an adequate and complete set of concepts. They propose using an optimisable vector-based approach to find missing concepts and linking them back to clear meanings with a novel incremental concept discovery module. Vandenhirtz *et al.* (2024) and Xu *et al.* (2024) propose potential improvements to the CBM with regards to the relationships between different concepts, and how they react to human intervention. Researchers at Meta also acknowledge the benefits of concept-based interpretation (LCM Team et al., 2024). They have presented a novel, concept-based perspective on language modelling: large concept models. This new model architecture can represent text in a more efficient, interpretable, and controllable manner, compared to classic token-based language models.

2.2.Gap analysis

Despite the described advances in deep learning and foundation models for EO, the current state-of-the-art approaches leave several critical gaps that limit their interpretability and practical trustworthiness. Most existing models lack domain-aligned concept spaces that connect EO embeddings to human-understandable geospatial knowledge, and their explanations remain primarily *post hoc*-focused on input-output attributions rather than stakeholder-relevant reasoning. Moreover, model-level and output-level explainability are rarely integrated, resulting in fragmented understanding of how predictions are formed. Temporal dynamics, essential for environmental and agricultural applications, are often ignored, as are mechanisms for incremental discovery of new or evolving concepts.

The S-xAI methodology we propose in the next section addresses these gaps through a dual explainability architecture that combines concept-based model transparency with RAG-

enhanced output reasoning. By aligning EO and textual embeddings into a shared embedding space, the approach enables interpretable, concept-level representations of EO data, while a Retrieval-Augmented Generation component grounds predictions and concepts in domain-specific knowledge to produce clear, human-centric explanations. Its modular design ensures adaptability across use cases and provides a scalable foundation for future extensions towards temporally aware and scenario-driven explainable AI, thus bridging the divide between powerful EO prediction models and trustworthy, scientifically grounded decision support.

2.3.Proposed S-xAI Methodology

Methodology

Our approach will make EO-based predictions better understandable by providing clear text-based explanations. This will be done by:

(i) constructing domain-specific, text-aligned representations (or embedding spaces) from EO data and auxiliary geospatial data, through a concept-based alignment model.

(ii) Enhancing the explainability of the output of this model by a Retrieval Augmented Generation (RAG) based agentic AI component, to generate human-understandable and stakeholder-oriented explanations.

In this way, our S-xAI architecture will leverage both model explainability and output explainability. In other words, with our dual approach we will both text-align the intermediary (embedding) stage of the model and "ground" the model predictions using domain specific knowledge documents (Figure 1).

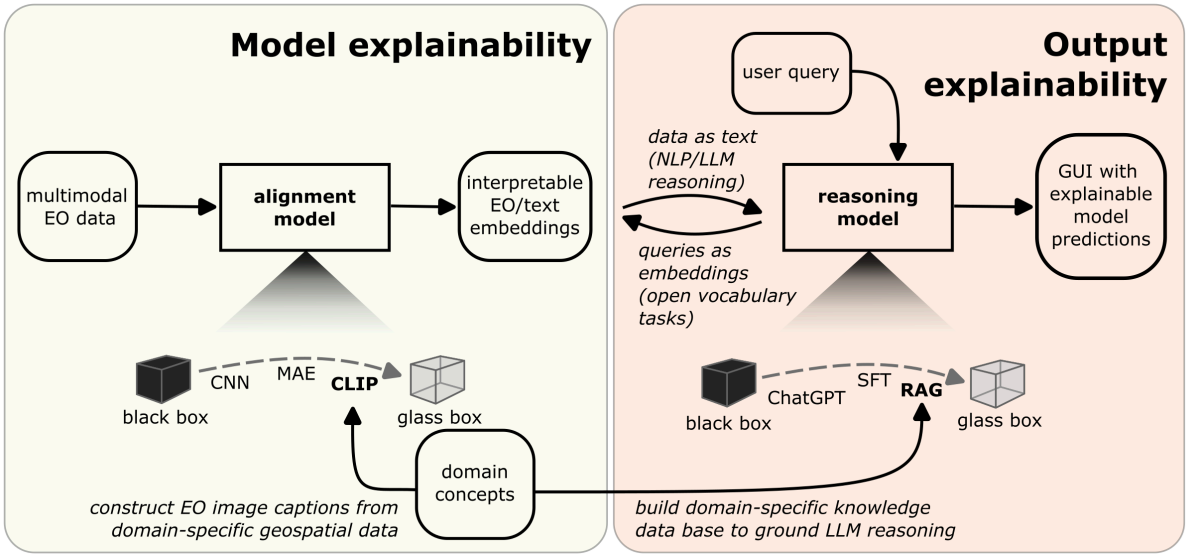

Figure 1: AETHER Self-Explainable EO-AI model concept

The *alignment model* (for concept-based model explainability) in Figure 1 will be developed as follows. First, EO data is encoded to embeddings, either using location encoders (e.g., SatCLIP), EO image encoders, multimodal EO encoders, or pre-trained geospatial foundation models (e.g., AlphaEarth, Terramind). Secondly, text captions are encoded using text encoders (e.g., CLIP text

encoder) to text embeddings. EO embeddings and text embeddings can be aligned by either training both encoders (Radford et al., 2021), or by freezing the EO encoder and only training the text encoder (Zhai et al., 2022), or vice versa. The second method is more flexible, because it works with fixed EO encoders such as large geospatial foundation models, and requires less training data (Zhai et al., 2022), but is constrained by the richness (information content) of the pre-existing EO embeddings. We will try this method first, because fewer data points will be needed, and consider training both encoders if performance is not sufficient compared to benchmarks. Next, concepts will be queried using *open-vocabulary tasks* (i.e., in natural language, not limited to a fixed set of terms or phrases) in the text-aligned EO embedding space, and we will quantify the similarity of embeddings to concept embeddings to quantify how well these concepts are present in the EO data.

The *reasoning model* (for concept-based output explainability) will be developed as follows: we will train shallow classifier or regression heads on the EO embedding space (e.g., a Generalized Linear Model or Random Forest) to predict the target variables. We will try predicting using the embeddings directly and using the similarities with queried concepts. At the same time, we will employ Retrieval Augmented Generation (RAG) to link the task (using the relevant concepts) to relevant parts of domain-relevant knowledge documents and then prompt an LLM-based workflow to explain the model predictions given the concept activations, predictions, and extracted related knowledge. Figure 2 provides a more detailed illustration of the architecture.

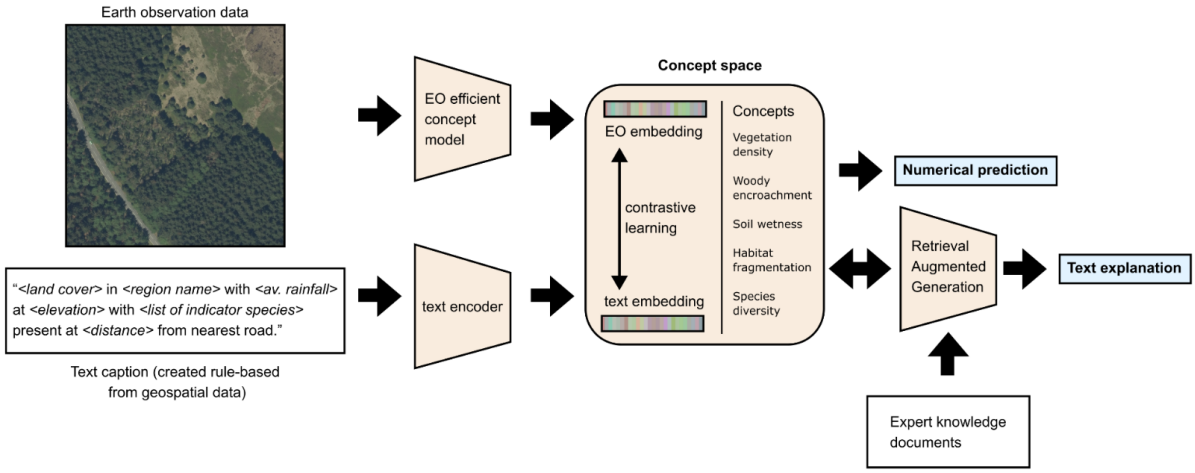

Figure 2: Illustration of the modular architecture of the model, where each component can be developed independently.

RAG module

The RAG approach will be tailored to our system that aligns EO and text embeddings, performs open-vocabulary retrieval, and then generates user-oriented explanations. When tile concepts are produced from auxiliary geospatial datasets and rules on top of e.g. an EO foundation model, retrieval can be powerful, but the resulting concepts are *weak/synthetic labels*. They inherit assumptions and potential drift from the source datasets and rules. The RAG component therefore will serve two core functions: *semantic grounding* (clarifying what a retrieved concept

means, consistently and in user language) and *transparent justification* (explaining why the concept could have been retrieved and how reliable that association is).

The proposed corpus mix (see Table 1) reflects those two functions. A strong backbone of *authoritative definitions, standards, and ontologies* provides stable meanings, synonyms, and hierarchical relations so open-vocabulary queries map to clear, evergreen concepts (ideas or terms whose meaning stays stable over time and across contexts) rather than rule-specific jargon. *EO sensor/product and measurement guides* will ground explanations in what EO can observe, enabling "why" narratives tied to physical signals, scales, and acquisition constraints. To address weak labeling directly, a dedicated slice of *annotation rulebooks and provenance notes* documents data lineage, thresholds, spatial/temporal buffers, and known failure modes, so the generator LLM can state provenance and qualify uncertainty instead of guessing. Complementing this, *confounder/validation references* shall supply default caveats and error modes that help prevent overconfident explanations.

Finally, the corpus must include *short concept explainers, application playbooks, and contextual briefs* to keep outputs useful and audience appropriate. Concept cards enforce consistency and readability for frequent terms; playbooks and context sources enable "so-what" guidance; and case-based or review material anchors responses in realistic magnitudes and common patterns.

Good quality of the documents is critical for RAG to work well. Therefore, a scoring rubric will be provided by WP200 to WP300 to allow easy selection of usable information. This rubric should score on aspects such as: Relevance and scope, authority and stability, concept grounding usefulness, EO explainability value, weak-label provenance value, user-oriented generation value, chunkability and retrievability, and metadata findability. Scores in different categories can be weighted to calculate a final value, that should result in a keep, park, or reject decision for each document.

Table 1: RAG corpus composition overview. Corpus share percentages represent initial values and can be adjusted based on evaluation feedback.

Document Type	Contribution to RAG	Approximate corpus share (%)
Authoritative definitions, standards and glossaries	Canonical meanings for open-vocab concepts; disambiguation; guards against rule drift	25%
Domain taxonomies/ontologies	Structured relationships amongst concepts (broader/narrower/related); improves clustering and explanations	10%
EO sensor/product handbooks and measurement guides	"How EO sees X" and what signals mean; supports "why retrieved?"	15%
Method/application reviews	Consensus workflows and typical assumptions; good for "how reliable / how done?"	10%
Uncertainty, validation and confounder references	Disclaimers for weak labels; explains failure modes	10%

Document Type	Contribution to RAG	Approximate corpus share (%)
Annotation rulebooks and provenance documents	Allows RAG to explain synthetic labels used and their limits	8%
User-oriented concept cards and explainers	Short, concise, friendly answers aligned to audience	15%
Local/sector/context briefs	"So-what" relevance; turns concepts into actionable context	7%

Overall, the chosen mix makes the RAG a *trust and calibration component*: it converts open-vocabulary retrieval into explanations that are semantically correct, physically grounded, provenance-aware, and genuinely usable for decision-making.

Temporal effects

The proposed S-xAI methodology aims to develop concept-based explainable models that operate on embeddings derived from EO data. In its initial version, the methodology focuses on predicting and explaining target variables from single "state" representations, each corresponding to a specific timestamp, allowing for transparent interpretation of relationships between learned EO concepts and model outputs. While this approach provides a clear framework for explainability, it is recognised that many real-world applications, specifically in the environmental domains, inherently depend on temporal dynamics. Therefore, a potential future extension of the methodology will consider incorporating time-series embeddings, enabling the model to capture temporal effects and seasonal trends across, e.g., the crop growing period or longer environmental processes. The use of geospatial foundation models that already encode temporal information, e.g., Presto and AlphaEarth Foundations (Brown et al., 2025; Tseng et al., 2023), is envisaged as a promising direction. Additional extensions, such as the generation of counterfactual or scenario-based concept descriptions to support "what-if" analyses, are identified as valuable follow-up studies. Nevertheless, the initial focus is on achieving concept-level explainability for single-state predictions, while paving the way toward temporally aware and scenario-driven S-xAI approaches.

Unified Architecture

The AI architecture and processing workflow provide a *unified backbone* for all three the EO-based use cases of the project. This backbone will be tailored for each use case based on their goals, data, and concepts. The design is modular and extensible, comprising generic components for data ingestion, EO data encoding, text encoding, concept extraction, downstream task prediction, and RAG-enhanced explanation (see Figure 3). This modularity enables flexible localisation, and adaption to new domains or datasets by varying input data source, trained models, and knowledge corpora, while maintaining a consistent and transparent explainability framework across applications.

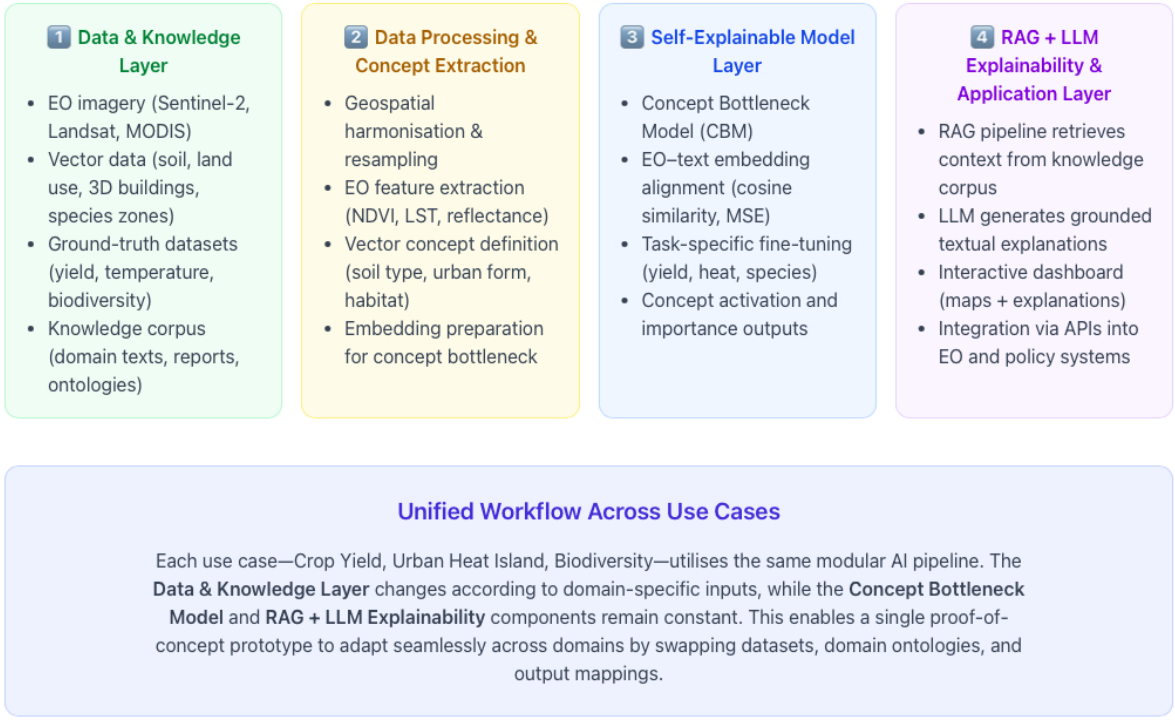

Figure 3: Initial unified Self-Explainable EO-AI architecture for the AETHER project

2.4.Requirements

As described in Section 2.3, the S-xAI architecture will be designed as a modular system, leveraging EO embedding models, a concept-based alignment module, and an advanced RAG component using agentic AI. While this design enhances the explainability, transparency, reusability, and scalability across the use cases, it also introduces specific constraints on model design and data processing:

- **Embedding model dependency**: The semantic representation of EO data relies on a fixed embedding space, which constrains the range of domain-specific fine-tuning applicable per use case.
- **Explainability module integration**: The requirement for interpretability restricts the use of certain black-box models that might be more high-performing.
- **RAG framework consistency**: All use cases must operate with a shared RAG pipeline architecture, limiting the diversity of data access and context retrieval mechanisms.
- **Modularity enforcement**: Each component must remain decoupled, affecting the ability to optimise end-to-end performance for individual use cases.

In the following subsections, the software and hardware requirements, including computational and storage, as well as the data and knowledge requirements for model development and use in the proof of concepts for the use cases, will be further described.

2.5.Software & Hardware

System maturity is progressed through four successive (approximately two-month) prototype iterations (proto-0 to proto-3), with proto-1 defining the S-xAI methodology handoff (WP200 to WP500), proto-2 supporting intermediate review (IR) feedback, and proto-3 achieving TRL-5 at final review (FR).

Table 2: Software and Hardware Requirements

Req ID	Requirement	Metric	Target	Measurement
SW-01	All AI model and training components shall be implemented in Python using open-source packages, validated by PROTO-1. Depends on DATA-EO-01, DATA-AUX-01/02, DATA-ITER-01.	% model code in Python	100%	Repo language scan + environment file review
SW-02	The system shall integrate at least one external LLM API provider for LLM functionality by PROTO-2. Depends on: KNOW-01/02/03, PROMPT-01.	# LLM APIs integrated	>= 1	Integration test with live API calls
SW-03	The RAG/LLM orchestration shall be implemented using a standard OSS framework (LangChain, LangChain4J, or Semantic Kernel) by PROTO-0. Depends on: KNOW-01/02/03, PROMPT-01, DATA-AUX-02.	Framework usage	1 selected framework in use	Codebase inspection
SW-04	The system shall implement operational capabilities (logging, traceability, authentication) using standard libraries by PROTO-2. Depends on: KNOW-02/03, PROMPT-01.	Presence of ops modules	All 3 present	Checklist + integration tests
OUT-01	Model predictions shall be exportable to at least one standard GIS-readable geospatial format by PROTO-2. Depends on: DATA-EO-03, DATA-ITER-01.	# supported GIS formats	>= 1	Export + load test in QGIS
OUT-02	Explanations shall be generated in plain text or Markdown and be exportable/copyable from the GUI by PROTO-2. Depends on: KNOW-03, PROMPT-01.	Explanation format support	Text + Markdown	UI acceptance test
GUI-01	End-user POCs shall be developed in C#/.NET with optional Semantic Kernel integration by PROTO-2. Depends on: PROMPT-01, KNOW-01/03.	Runtime / stack compliance	100%	Build pipeline + code review

Req ID	Requirement	Metric	Target	Measurement
GUI-02	The GUI shall be a modular web application with separate frontend and backend API by PROTO-2. Depends on: DATA-EO-03.	Architecture compliance	Frontend/backend separation implemented	Repo structure + deployment review
GUI-03	The backend API shall be usable both by the GUI and by external clients/automated workflows by PROTO-2. Depends on: DATA-EO-02, KNOW-01/02.	External API usability	>= 3 external endpoints documented and tested	Postman / Swagger tests
API-01	Data and model results shall be exchanged via Web services or WebSockets using JSON and GeoJSON by PROTO-2. Depends on: DATA-AUX-01, OUT-01.	Protocol + format compliance	JSON + GeoJSON supported	Integration tests
DEV-01	Backend services and GUI shall be containerized with Docker and deployable via Kubernetes on NILU's cluster by PROTO-2. Depends on: KNOW-02/03, DATA-EO-02.	K8S deployment success	100% success on NILU cluster	CI deploy + smoke tests
OPS-01	All code shall be publicly hosted on GitHub with CI/CD enabling parallel partner development by PROTO-0. Depends on: DATA-EO-01 / KNOW-02, PROMPT-01.	CI/CD availability	CI + CD pipelines available	Check GitHub Actions / pipelines

The AI model(s) and all model training-related components will be developed in Python, using open-source packages. The machine learning components will be implemented in PyTorch. For LLMs we will use existing LLM API services (e.g., Mistral AI, EuroLLM, OpenAI, Anthropic). The RAG component and related operational functionality of the POCs will be developed using standard frameworks such as LangChain (Python), LangChain4J (Java), or Microsoft's Semantic Kernel (supporting multiple programming languages), leveraging standard ecosystem components (e.g. JVM or CLR based) and libraries for logging, traceability, user authentication, etc.

The AI model(s) will partly use pre-trained model components and train new model components as well. For this, we will use the Wageningen University & Research HPC infrastructure Anunna[1]. As such, data will be stored on local data servers that can be accessed from the HPC. In case more capacity is needed we will temporarily use Microsoft Azure based cloud resources.

Output predictions from the AI model will be exported to geospatial data formats so they can be loaded into GIS software by end-users, if required (depending on the GUI). Explanations will be generated in plain text or Markdown format, and simple copy functionality will be provided in the GUI. The RAG/LLM module potentially can also be used to drive GUI functionality that supports or enhances open vocabulary tasks that allow an end-user to access the latent embedding space

[1] https://wiki.anunna.wur.nl/

using natural language queries (NLQ), i.e. to interact with the explanations and explore the reasons and concepts, e.g., to find areas or time periods with similar or opposite conditions. Integrating such functionality could be part of the proof-of-concept development, driven by feedback from our end-user evaluations.

The end-user application(s) (i.e., the POCs) will be developed using C# and .NET with possible integration of Microsoft's Semantic Kernel for AI-driven functionalities such as interaction with LLMs. The Graphical User Interface (GUI) will be designed as a modular web application, separating frontend and backend API. This makes sure that the service can be used by the GUI but also by external applications or in automated workflows as an AI agent. The GUI will connect to multiple APIs to combine data, e.g., to display geospatial data combined with explanations and background information. Data and model results will be exchanged through standard Web services or WebSocket protocols, using JSON and GeoJSON for structured communication. The backend services and GUI components will be containerized using Docker/Kubernetes within NILU's cluster during development and evaluation, for the duration of the project.

All code will be developed publicly on GitHub with appropriate CI/CD infrastructure to allow parallel development from different partners of the project.

2.6.EO Data & Knowledge

Table 3: EO Data and Knowledge Requirements

Req ID	Requirement	Metric	Target	Measurement
DATA-EO-01	All EO and auxiliary datasets used in the project shall be publicly available (openly licensed) for all use cases, validate by PROTO-1.	Share of datasets with open/public licence	100%	Dataset inventory review + license verification checklist
DATA-EO-02	EO input data shall be sourced (according to the selected EO encoder specification) achieving complete AOI coverage by PROTO-1.	EO data coverage of AOIs	>= 95% spatial coverage per UC AOI	coverage report against AOI / time windows
DATA-EO-03	After the EO encoder is fixed, the project shall deliver end-user guidelines for acquiring required EO data (when rasters are needed) by PROTO-0.	Guideline completeness	1 guideline package per UC where EO rasters required	Documentation review + UC lead sign-of
DATA-AUX-01	For each UC auxiliary geospatial data will be provided as a locations x features table (rows = locations, columns = numerical features), including labelled and unlabelled locations, by PROTO-0.	Auxiliary table availability per UC	1 table per UC	Data handover check + schema validation
DATA-AUX-02	The data loader shall support on-the-fly selection of auxiliary feature columns for caption generation using predefined rules, implemented by PROTO-1.	Dynamic feature selection success rate	>= 99% runs without selection errors	Automated caption-generation runs over auxiliary tables

Req ID	Requirement	Metric	Target	Measurement
KNOW-01	UC leads shall provide a curated concept list for retrieval/explanation by PROTO-0, updated iteratively thereafter.	Concepts per UC	10-30 concepts	Concept registry review + count per UC
KNOW-02	UC leads shall provide domain-relevant knowledge documents per UC for RAG, with first full set by PROTO-0 and iterative curation with WP200 thereafter.	Knowledge documents per UC	>= 50	Document inventory + de-duplication report
KNOW-03	Knowledge documents shall cover all required explanation languages, be machine-readable (plain text / Markdown preferred) and be freely accessible (no DRM / password / encryption), validated by PROTO-0.	Language coverage; machine-readability; access constraints	100% language coverage; >= 95% machine-readable; 0 restricted-access docs	Format scan + access check _ language metadata audit
PROMPT-01	Prompts for explanation tuning, evaluation datasets, and guardrails shall be co-developed with UC leads and version-controlled, with an initial validated prompt set per UC by PROTO-1.	Prompt set completeness	>= 1 explanation prompt + 1 evaluation / guardrail prompt per UC	Prompt repository review + UC lead approval
DATA-ITER-01	Data hyperparameters (e.g. EO patch size, pretraining sample counts, target sample counts, number of concepts, labelling rules) shall be defined and updated iteratively based on model performance, with documented finalised values before each release (PROTO-1 and later).	Hyperparameter documentation coverage	100% of releases have documented data settings	Release checklist + config snapshot stored per version

The project will be based on publicly available data for all use cases, thus limiting issues related to licensing, data protection regulations, or exposure of personal and sensitive information.

EO data will be dependent on the EO encoder. For example, for location encoders or pre-computed geospatial foundation models we will only require latitude/longitude coordinates, while for other encoders we will require EO input data (e.g., Sentinel-2). Therefore, EO data will be sourced by WP200 (who develop the EO encoders) rather than the UC owners in WP300. However, after the model is finished and the EO encoder is fixed, we will include guidelines for end-users how to acquire relevant EO data, if applicable.

Other auxiliary geospatial data (to generate text captions with) are UC specific and will be provided by UC leads. They will be provided in the form of a table, where rows list different locations, and columns list different numerical features (e.g., mean temperature). These data can be provided both for locations with target numerical values (e.g., crop yield or urban temperature) and without. In the latter case, these data can be used for contrastive learning. This auxiliary data can be extensive, because the relevant data columns will be selected later, during on-the-fly

generation of captions using pre-established rules in the data loader. Auxiliary data will be sourced from publicly available datasets.

UC leads will then list the concepts of interest (approx. 10-30) which will be queried, as well as the knowledge documents (100 or more, domain and use case goal relevant, e.g. selected from scientific or grey literature) for the RAG component. These data sets will be curated in an iterative process as the model continues to develop, in collaboration with WP200. Knowledge documents need to be provided in all languages in which explanations need to be generated. Documents preferably are in plain text or Markdown format or at least need to be in a machine-readable format. Besides that, they need to be freely accessible, without copy protection, passwords, or encryption. To tune the explanations to specific end-user needs, prompts will be developed in collaboration with the use cases. Similar for any other prompts that might be required, e.g., to be used as evaluation datasets or guardrails to prevent undesired outputs.

Other specific data requirements, such as patch size, number of pre-training data points, number of target data points, number of concepts, etc., will be established iteratively as the model, architecture, and workflows are developed, guided by model performance and data need priorities.

2.7.Ethical & Privacy Aspects

The AETHER project adheres to the highest standards of ethical conduct and data protection, ensuring that the design and deployment of self-explainable Earth Observation (S-xAI) models remain transparent, responsible, and compliant with European and international regulations. All activities will align with the EU General Data Protection Regulation (GDPR), the EU AI Act, and the European Code of Conduct for Research Integrity (ALLEA - All European Academies, 2023; European Parliament and Council of the European Union, 2024, 2016).

AETHER's ethical framework addresses three core dimensions:

- Data ethics and privacy: The project uses only publicly available or properly licensed Earth Observation and ancillary datasets. No personally identifiable information (PII) is collected or processed. Where socio-economic or location-specific indicators are used (e.g., in the Urban Heat Island case), data are anonymised and aggregated to ensure individuals and communities cannot be re-identified.
- Algorithmic transparency and accountability: The S-xAI architecture incorporates explainability by design, allowing model reasoning and outputs to be interpretable, traceable, and auditable. This approach supports fairness and mitigates risks of bias in training data, ensuring that automated insights can be verified by domain experts and stakeholders.
- Responsible use and societal impact: AETHER promotes equal access to AI-enabled Earth Observation insights and safeguards against misuse. Human oversight remains a key requirement in all decision-support scenarios, especially when outcomes could affect communities or environmental management. The integration of retrieval-augmented generation and large language models follows strict guidelines for factual grounding, security, and bias control.

Table 4: Ethical and Privacy Requirements

Req ID	Requirement	Metric	Target	Measurement
ETH-01	All project activities shall comply with GDPR, the EU AI Act, and the ALLEA European Code of Conduct for Research Integrity, including a strict prohibition on the collection, storage, or processing of any personally identifiable information (PII). Continuously enforced from PROTO-0 onward.	Compliance audit pass rate, PII detection incidents	100% at each review; 0 PII incidents	Formal compliance checklist + internal audit sign-off; automated PII scanning of all data assets + periodic manual spot checks.
ETH-02	Only publicly available EO and ancillary datasets shall be used, validated by PROTO-0 and rechecked at every major release.	Share of datasets with valid public / license proof	100%	Dataset inventory + license / provenance verification
ETH-03	The system shall include bias / fairness evaluation over training and inference data, with mitigation actions documented, first complete cycle by PROTO-1 and repeated each release.	Bias evaluation completion rate; mitigations logged	100% cycles completed; >= 1 mitigation if bias found	Bias test suite run + mitigation log in DMP / repo
ETH-04	Ethical monitoring shall be continuous throughout the project, with ethical risks and mitigation measures documented in the Data Management Plan (DMP) starting from PROTO-1 and updated periodically. An internal ethics lead (NILU) shall formally review and validate these risks and mitigations at each major project milestone.	DMP update cadence; Ethics risk register completeness; Ethics lead review completeness	Periodic DMP updates from PROTO-1 onward; 100% of identified ethical risks tracked and mitigated; 100% of major milestones formally reviewed	DMP version history and documented change logs; Ethics risk register review; Signed ethics review memos by the internal ethics lead at each milestone
ETH-05	All project outputs (including data, models, and documentation) shall comply with FAIR principles and Responsible AI requirements. Assessment methods shall be applied to all use cases, with the first full assessment completed by PROTO-2, and repeated for Final release.	FAIR score; % of FAIR gaps addressed; Responsible AI assessment coverage	FAIR score >= agreed project threshold; >= 80% of FAIR gaps addressed by Final; 100% of UCs assessed by Final	FAIR self-assessment tool results; FAIR and Responsible AI action-tracking log; Completed FAIR and Responsible AI assessment templates with UC lead sign-off

Ethical monitoring will be an ongoing process throughout the project lifecycle. Partners will document ethical risks and mitigation measures in data management plans, supported by an internal ethics lead contributed by NILU as part of the consortium team, and alignment with ESA's data governance requirements. Outputs will comply with FAIR (Findable, Accessible, Interoperable, Reusable) principles (Wilkinson et al., 2016) and Responsible AI best practices, ensuring transparency, reproducibility, and trust across all use cases. As this an AI driven project,

a special focus will be on AI ethics and a tailored assessment and validation[1] developed during the EU Horizon project FAIRiCUBE[2] will be applied.

3. Use Cases

The overarching theme of this project is "Reliable & fast monitoring in unpredictable times", which reflects some of the main challenges of our times. We have selected three use cases:

1. Mapping of biodiversity and its loss due to climate change.
2. Crop yield prediction and rapid assessment of the effect of floods, droughts, and fires.
3. Detection of urban heat islands and their evolution due to global warming.

These use cases collectively span natural, agricultural, and urban ecosystems, ensuring the broad applicability and relevance of the S-xAI framework. They address high-priority policy domains aligned with the EU Green Deal, Farm to Fork Strategy, and Nature Restoration Law, while providing multi-scale and multi-domain validation of AI interpretability methods across diverse geographies, from temperate Europe to tropical Africa and Latin America. Together, they deliver strong scientific and societal impact, advancing biodiversity conservation, food security, and urban climate resilience. Moreover, they leverage rich and openly available datasets, including SatBird, S2BMS, Sentinel-1/2 archives, Copernicus Land Monitoring Service products, and urban climate and socio-economic databases, which support reproducibility, scalability, and FAIR data compliance, thereby enhancing the scientific robustness and long-term utility of the project outcomes.

In the following, we will provide a motivation for each use case, along with an overview of the available experience, models, and datasets. To structure the description and documentation we will make use of AI canvases, similar in concept to business canvases. An AI Canvas is a structured framework designed to describe and analyse how artificial intelligence can address a specific problem or opportunity. It helps teams clarify a use case by systematically capturing key elements such as the objective, data sources, model approach, stakeholders, success metrics, risks, and ethical considerations. In this project, the AI canvas will support the consolidation of multiple use cases by providing a common structure for comparing needs, data availability, and analytical workflows. This enables alignment within the project, identification of shared components (e.g., datasets, models, or components and infrastructure for the prototypes), and prioritisation of efforts.

This work directly aligns with the European Space Agency's (ESA) strategic objectives for xAI by embedding transparent, self-explainable AI methodologies. Furthermore, it will demonstrate the potential for scalability to address broader climate-related hazards, such as floods and storms, and will promote operational readiness by delivering transparent and actionable tools designed to strengthen community resilience. The project will rigorously adhere to FAIR data principles, ensuring that all

[1] https://hub.fairicube.eu/validation-ai-ethics.html

[2] https://fairicube.eu

spatially explicit results and models are made available to urban planners, decision-makers, and local actors to foster widespread adoption and impact (Kumar et al., 2024; Wilkinson et al., 2016).

Following the agreements with ESA, the use case on urban heat islands will receive highest priority during the project. Due to the higher maturity level of the biodiversity dataset, the S-xAI development will begin with that use case instead. If either the biodiversity or crop yield use case shows significant underperformance compared to the xAI SOTA in the related domain (considering a possible prediction accuracy vs explainability trade-off), it may be dropped from further development into a proof-of-concept.

3.1. Use case 1: Biodiversity

Biodiversity monitoring is crucial for understanding the health of ecosystems, detecting changes over time, and guiding conservation efforts to maintain the balance of natural systems. Climate change disrupts species distributions, e.g., by altering habitats and climatic conditions, and increases the frequency of extreme weather events, all of which can push vulnerable species toward (local) extinction (Bellard et al., 2012).

Species distribution models (SDMs) use geospatial data to predict where species occur and are vital for conservation planning, but traditional models rely on manually selected covariates, e.g., land cover, distance from road, and mean temperature, limiting their capacity to capture complex ecological patterns (Beery et al., 2021; Elith and Leathwick, 2009). Recent deep learning approaches improve predictive power by using raw Earth Observation (EO) data (Cole et al., 2023; Teng et al., 2023; van der Plas et al., 2025b, 2025a), but often lack transparency, making it difficult to interpret ecological relevance (Ryo et al., 2021). To address this, *post hoc* explainable AI methods have been used to improve the transparency of deep SDMs, for example by computing LIME and Shapley values (Ryo et al., 2021; Zbinden et al., 2025).

Here, we will go beyond *post hoc* explainable AI methods and develop a self-explainable AI EO model to predict species distributions directly from raw EO data, while explaining what EO features were used to make these predictions, based on documented habitat preferences of the species. See Table 5 for the AI canvas of this use case. This will allow the model to identify from EO data complex concepts not easily defined by hand, such as habitat fragmentation, density of vegetation, habitat mosaics, thus enabling the monitoring of these biodiversity features at scale.

To that end, we will use two machine learning ready, public data sets of Sentinel-2 EO data coupled with species observations: SatBird and S2BMS. SatBird is a data set of bird observations from EBird in the USA and Kenya (Teng et al., 2023), and S2BMS is a data set of butterfly observations from UKBMS in the UK (van der Plas et al., 2025b). As auxiliary geospatial data we will use public geographic, bioclimatic, land cover and human footprint data. We will compare our model performance against the existing benchmarks for these data sets, and ask the model to explain, per species, what EO data features drove the predictions. We will evaluate these explanations both quantitatively, using standard benchmark metrics, and qualitatively, by consulting experts to judge the explanation quality and validity.

Table 5: AI Canvas for the biodiversity monitoring use case

Use case objective Predict multi-species species encounter likelihood / habitat suitability and deliver self-explainable biodiversity indicators at hotspot scale. Success: Strong SDM generalisation across regions + explanations that are *faithful, stable, and ecologically plausible*.	**Data** 1) **EO**: depends on EO encoder. 2) **Ancillary**: S2BMS (Butterfly observations (UK)), SatBird (bird observations (USA), geography/topography, bioclimatic, land-use, human footprint 3) **Captioning rules:** TBD, based on auxiliary data. 4) **Corpus**: Biodiversity reports; ecological literature, EO specs, concept cards.
Modelling approach Shared EO foundation encoder (EO embeddings per tile); text/semantic encoder (embeddings of habitat/species terms). EO – Semantic alignment trained with weak EO-caption pairs (from ancillary data and rules) to form a joint EO-semantic space. A light SDM head predicts species encounter likelihood from EO embeddings. Explainability via open-vocabulary probing of the aligned space (ranked ecological drivers + similarity maps), supported by EO attributions and RAG-LLM narratives grounded in ecology corpus.	**Initial concepts** land-cover type; greenness/NDVI; canopy density; wetland index; open water; cropland intensity; grassland fraction; shrub land; bare soil/rock; urban / impervious; elevation; slope/aspect; distance-to-water; fragmentation/edge density; burn/scar; phenology phase; drought/heat/rain anomalies; human footprint; protected status.
Study regions UK: Butterflies USA: Birds	**Outputs** Species encounter probabilities per location, including text explanations. Can be aggregated to species richness.
Current xAI SOTA MaskSDM (attribution-based explainability)	**Current xAI benchmarks** shapley values, AUC. Benchmarks not relevant as other data was used.
Value proposition Scalable biodiversity monitoring with transparent habitat / pressure drivers, supporting conservation action and policy trust.	**Stakeholders** Conservation agencies/NP managers; biodiversity NGOs / policy units, ecological researchers, citizen-science platforms.

3.2.Use case 2: Crop Yield

Crop yield prediction and rapid damage assessment is vital for ensuring food security, optimizing resource management, and supporting farmers' decision-making processes. In conventional monocultural agriculture, yields are highly sensitive to changes in climate conditions, including altered rainfall patterns, rising temperatures, and more frequent extreme weather events (Lobell et al., 2011). In contrast, agroforestry systems—which integrate crops, trees, and sometimes pastures—offer greater resilience (Ngaba et al., 2024; Santos et al., 2019) but can still suffer yield losses from natural disasters like droughts, floods, storms, and wildfires, which disrupt ecosystem services and reduce landscape stability, and potentially lead to significant carbon emissions and crop yield loss.

Mapping and predicting crop yield in croplands and agroforestry is a challenging task (Muruganantham et al., 2022), primarily due to the temporal, spatial, structural, phenological, and species diversity of the vegetation species, as well as due to the ever-increasing unpredictability of the weather events. Long time series of Sentinel-1 and Sentinel-2 data combined with machine learning offer promising tools for capturing interactions between tree

cover, crop yield, and environmental factors in agroforestry (Oliveira et al., 2025) as well as conventional croplands (Paudel et al., 2022, 2021). Studying this within our project will support the development of scalable, cost-effective monitoring frameworks essential for managing food security and land use in climate-vulnerable regions like the Sudano-Sahel (Bayala et al., 2014; Burke and Lobell, 2017).

The AI canvas for this use case is shown in Table 6. We will leverage a rich collection of crop yield datasets over 50,000 crop-cut observations from GROW Africa (Geyman et al., 2025) and One Acre Fund[1], covering seven African countries (Kenya, Rwanda, Ethiopia, Malawi, Burundi, Tanzania, and Zambia) between 2012 and 2021. These datasets include detailed crop performance records for cereals such as maize, sorghum, millet, rice, wheat, and teff, as well as legumes and root crops like beans, groundnuts, and sweet potatoes. Data attributes include fresh and dry weights, yield (tons/ha), harvest dates, and field-level classifications. In addition, over 10,000 tree-level cocoa yield records from CocoaSoils[2] and parkland agroforestry data from Burkina Faso (Oliveira et al., 2025) will enrich the analysis. Complementary benchmark datasets, including CY Bench (a global reference dataset for sub-national crop yield forecasting) and FLAME (Field-Level Asset Mapping Dataset for England's Agricultural Sector), will support cross-regional validation (Paudel et al., 2025; Sheikh et al., 2025). High-precision georeferenced data (≤0.03 km) and lower-precision samples (≤5 km) provide broad spatial coverage, facilitating scalable model calibration across diverse agro-ecological zones and farming systems.

Together, these datasets form one of the most comprehensive multi-country resources for developing and validating self-explainable AI models for crop yield prediction and agroecosystem monitoring. The S-xAI architecture on the black box models described in the literature (Oliveira et al., 2025; Paudel et al., 2022, 2021) will be used to generate predictions of crop yield using available in-season EO data based on the crop calendar and well-known yield-related concepts, as well as predict crop yield for the next season subject to additional information on meteorological conditions and natural disasters. his will allow us to study the effect of climate change and natural disasters on crop yield, e.g., by altering the precipitation to simulate dryer and wetter seasons and simulating flood and drought events.

Table 6: AI Canvas for the crop yield prediction use case

Use case objective Provide accurate, scalable, and explainable crop yield estimation models integrating EO, ancillary biophysical layers, and contextual datasets. Deliver actionable insights for policy, climate adaptation, and farm-level decision-making.	**Data** 1) **EO**: Sentinel-1, Sentinel-2 2) **Ancillary**: Biophysical (soils, topography), climate (temperature, rainfall), environmental quality, socio-economic and built-up layers 3) **Captioning rules:** TBD, based on auxiliary data. 4) **Corpus**: Agronomic standards, EO documentation, crop and yield modelling literature, validation guidelines, concept cards, local context briefs.
Modelling approach The model uses a shared EO encoder for	**Initial concepts** Key layers include land-cover type, greenness/NDVI,

[1] https://oneacrefund.org/

[2] https://cocoasoils.org/

spatiotemporal embeddings and a semantic encoder for crop, soil, and climate concepts. Weakly supervised EO-caption alignment creates a joint concept space, with a lightweight head predicting crop yield. Explainability is provided via ranked agronomic/biophysical terms, EO attribution maps, and RAG LLM narratives, with provenance and uncertainty calibration.	canopy density, and drought/heat/rainfall anomalies, which can be integrated with additional biophysical, climate, and socio-economic data
Study regions West Africa (Burkina Faso) East Africa (Kenya, Rwanda, Tanzania, Zambia, Malawi & Burundi)	**Outputs** Crop-specific and aggregated yield maps, hotspot change alerts highlighting significant yield gains or losses, explanatory products such as ranked agronomic and biophysical concepts, spatial similarity layers and attribution maps, and RAG-grounded textual rationales that describe key drivers of yield patterns with provenance and uncertainty.
Current xAI SOTA Multimodal EO-weather-soil features; feature attribution (SHAP/IG) for climate, soil and management drivers; temporal attention and growth-stage attribution; counterfactual what-if analysis on rainfall, temperature and inputs; spatial saliency and prototype fields for high- and low-yield patterns.	**Current xAI benchmarks** Faithfulness via deletion/insertion AUC (AUC approx. 0.90), localisation accuracy of attribution maps against field-level proxies (mIoU up to approx. 0.5), stability of spatial and temporal attributions across seasons and domains, counterfactual validity through plausible and minimal perturbations (e.g. weather and inputs), and human alignment via agronomic expert usefulness assessments.
Value proposition A scalable and explainable crop-yield intelligence system that delivers transparent, data-driven insights on agronomic and environmental drivers, strengthening decision-making for climate adaptation, food security, and policy planning.	**Stakeholders** Senior researchers in Burkina Faso universities; Smallholder farmer(s) in Burkina Faso parklands; NGOs

3.1. Use case 3: Urban Heat Islands

Urban areas are characterised by significantly higher temperatures compared to their surrounding rural regions, a phenomenon driven by dense infrastructure, limited green cover, greater absorption and delayed release of heat by anthropogenic structures (Deilami et al., 2018). These Urban Heat Islands (UHIs) exacerbate the vulnerability of urban socio-ecological systems to climate change, particularly as heat waves intensify and high-temperature days become more frequent, posing critical threats to public health, infrastructure, and ecosystems (Tehrani et al., 2024). The consequences are severe, including increased cooling energy demands, increased air pollution, and heightened heat-related health risks, disproportionately affecting vulnerable populations (Hartinger et al., 2024). Projections indicate that by 2050, fatalities linked to UHIs and associated heat waves could surpass those from infectious diseases globally (Hartinger et al., 2024).

UHI is a well-documented phenomenon, but only since the beginning of 2000s has it gained a significant research interest, largely because of the widespread use of remote sensing data and GIS solutions and the sharp rise in Machine Learning and AI based approaches in UHI studies over the last decade (He et al., 2023). State-of-the-art studies apply AI models to process complex, heterogeneous datasets, including satellite imagery, urban morphology indicators, IoT sensor data and socio-economic variables,

which affect urban climatic zones (Shatnawi et al., 2025; Snaiki and Merabtine, 2025). In UHI monitoring, AI solutions are used to create high resolution spatial and temporal maps of thermal conditions (Yi et al., 2025). Deep Learning models, particularly Convolutional Neural Networks (CNNs) and U-Net architectures, are employed for the semantic segmentation of satellite imagery, providing products which can support UHI mitigation policies (Shaamala et al., 2025). Beyond that, latest AI models are transitioning from simple Land Surface Temperature (LST) estimation to UHI forecasting and predictive warning systems (Hoang and Nguyen, 2025).

Despite undeniable recent advancements in UHI analysis, there are still significant challenges connected with available datasets, AI models application, solutions scalability and transferability (Marey et al., 2025). UHI research relies largely on remote sensing data and specialised in-situ measurements. Yet remote sensing satellites equipped with thermal imaging sensors are mostly limited to Landsat, MODIS and VIIRS missions, which do not provide sufficient temporal (Landsat) or spatial (MODIS, VIIRS) resolution. Ground-based sensor networks are costly to establish and maintain, and have limited spatial coverage (Snaiki and Merabtine, 2025). Conventional AI models often fall short in providing transparent explanations for why specific urban neighbourhoods face greater heat risks. The “black box” nature of many AI algorithms can hinder their adoption by urban planners who require clear, justifiable insights for decision-making (Mallick and Alqadhi, 2025; Shaamala et al., 2025). This inherent lack of transparency and the challenge of integrating nuanced, dynamic local variations limit the utility of many current AI-driven assessments in developing targeted and effective urban adaptation strategies (Kumar and Bassill, 2024; Srivastava and Maity, 2023).

To bridge this gap, this project outlines an innovative urban heat island case study focused on developing and applying S-xAI. The study aims to identify urban zones prone to higher temperatures by integrating data on climate exposure, land use patterns, infrastructure characteristics, and socio-economic vulnerabilities. A key objective is to explain how specific urban features such as building materials, vegetation density and type, and urban morphology affect ventilation and influence local heat stress. This approach will yield interpretable and actionable adaptation strategies, such as optimised NBS initiatives and the strategic installation of water bodies and green area to mitigate identified risks. Guatemala City, the Hague, and Kraków have been selected as pilot cities for this research, leveraging diverse datasets including high-resolution canopy cover, Normalised Difference Vegetation Index (NDVI) as a proxy for Urban Temperature Regulation (UTR), and indicators of Urban Green Infrastructure Quality (Bokwa, 2023; van Eupen et al., 2024; Winograd et al., 2023). Other data inputs for UHI models typically include detailed urban structural data, land surface temperatures, vegetation indices, and meteorological parameters like air temperature, humidity, and wind speed (Kumar et al., 2016; Tehrani et al., 2024).

The AI canvas for this use case is given in Table 7. Specifically, this project will pursue the following objectives: first, to develop S-xAI models that integrate satellite-derived land surface temperature data with comprehensive urban morphological and socio-economic datasets, thereby clearly identifying the primary drivers of localised heat risk. Second, to produce readily understandable visual and textual outputs tailored for city planners, policymakers, and community stakeholders, facilitating informed decision-making. Thirdly, to incorporate future climate change projections into the S-xAI models to enable robust, long-term adaptation planning.

Table 7: AI Canvas for the urban heat island use case

Use case objective Map and forecast urban heat patterns at block scale and deliver self-explainable heat-risk indicators for planning and adaptation. Success: Accurate UHI intensity prediction across cities and seasons, with explanations that are physically consistent and actionable for planners.	**Data** 1) **EO**: Landsat-8/9, Sentinel, Copernicus LST, DEM/slope. 2) **Ancillary**: Municipal land-use/land-cover and forest/green area maps; population density and road network; building height, built-up age, and 3D block models; poverty and vulnerability indicators; heatwave frequency/intensity projections. 3) **Labelling rules**: UHI labels from LST deviation relative to city mean and heat-stress comfort thresholds, grouped into 8–10 ordered LST/UHI stress classes (from cold pixels to very high stress) 4) **Corpus**: Urban heat standards, urban form and vulnerability ontologies, EO documentation for LST and urban indicators, UHI modelling literature, validation guidance, concept cards, city context briefs.
Modelling approach A shared EO encoder generates spatiotemporal embeddings per urban tile, while a tabular/graph encoder processes socio-economic and built-environment features. A multimodal head predicts continuous LST and discrete UHI classes with temporal generalisation. Explainability is provided via feature attributions, concept-based probes (e.g., vegetation, imperviousness), and RAG-style narratives with provenance and uncertainty.	**Initial concepts** Key layers include land cover, vegetation and canopy metrics, building and impervious surface characteristics, climate and heat indicators, and socio-economic factors such as vulnerability and access to green space, integrated to support urban heat and greening analyses.
Study regions Guatemala City: Metropolitan area with LST-based UHI gradients and socio-environmental vulnerability zones. Kraków: City-wide LST maps, hot- and cold-spots, detailed built-up and vegetation structure. Optional: Third Latin American or European city with comparable EO and municipal datasets.	**Outputs** High-resolution LST and UHI class maps; hot-spot/cold-spot layers; block heat-risk indicators; priority greening and cooling intervention maps; dashboards for inequality and vulnerability overlays; explanation products: ranked drivers (e.g. vegetation, imperviousness, height), concept similarity maps, attribution maps, text rationales with provenance and uncertainty.
Current xAI SOTA Integrated tabular EO features; concept activation/TCAV for vegetation and imperviousness concepts; counterfactual what-if analysis on greening and densification; prototype/critic tiles for typical hot and cool patterns.	**Current xAI benchmarks** Faithfulness via deletion/insertion AUC (AUC approx. 0.90), localisation accuracy using mask overlap where available (mIoU approx. 0.5), stability under perturbations and domain shift, counterfactual validity via plausibility and minimality checks, concept quality for TCAV through consistent positive concept scores across space and time, human alignment typically assessed through expert usefulness studies.
Value proposition Scalable, explainable UHI intelligence that links physical drivers, social vulnerability, and planning levers, enabling targeted cooling investments, resilient urban design, and climate-health policy with high public trust.	**Stakeholders** Local citizens and community organisations in vulnerable neighbourhoods; city planners and municipal authorities; climate NGOs and international development partners; and urban researchers.

4. End Users and Impact

The project addresses the urgent need for scalable, self-explainable AI methods to support biodiversity monitoring, crop yield prediction, and urban heat island assessment, supporting global commitments under the Kunming-Montreal Global Biodiversity Framework and the UN Sustainable Development Goals (SDGs 2, 11, 13, and 15). The primary end-users include environmental and agricultural agencies, urban planners, research institutions, and policy makers who require explainable, data-driven insights for conservation, food security, and climate adaptation. By integrating Earth Observation data, deep learning, and self-explainable AI, AETHER aims to deliver interpretable and decision-ready outputs that strengthen trust and uptake across sectors. All datasets, models, and analytical workflows will be developed in line with FAIR (Findable, Accessible, Interoperable, and Reusable) principles (Wilkinson et al., 2016) to ensure long-term accessibility and reproducibility. Moreover, natural language-based interfaces will enable both expert and non-technical users to interact intuitively with model results, promoting transparency and inclusion. Ultimately, the project will enhance evidence-based policy and planning, support capacity development, and enable open data reuse across biodiversity, agriculture, and urban sustainability domains, delivering lasting societal and environmental impact beyond the project's duration.

4.1. Use case 1: Biodiversity

This component enhances the explainability of AI models for biodiversity assessment and conservation. End-users include ecology researchers and practitioners from networks such as LTER-LIFE[1], the Netherlands Institute of Ecology (NIOO)[2], and the UK Centre for Ecology & Hydrology (UKCEH)[3]. Engagement will be conducted through established research partnerships and collaborative activities. Two user groups will be directly involved in the testing of the developed AI system: ecologists from UKCEH representing the scientific community, and a data analyst from the Peak District National Park representing conservation practitioners. The explainable outputs will help these users better interpret ecosystem trends, inform conservation priorities, and improve communication with policy and public audiences.

4.2. Use case 2: Crop Yield

The agricultural component focuses on developing self-explainable AI models for crop yield prediction in West Africa, supporting food and income security and climate resilience. Engagement will target governmental agencies, agricultural researchers, NGOs, farmer organisations, and private-sector partners, ensuring open access to yield predictions. Confirmed user groups include university scholars in Burkina Faso representing the scientific community and project managers from a cocoa company in West Africa. Additionally, individual smallholder farmers in Burkina Faso will be asked to join the end user testing. These users will validate and

[1] https://lter-life.nl/en

[2] https://nioo.knaw.nl/en

[3] https://www.ceh.ac.uk/

apply the models to enhance agricultural planning, optimise resource allocation, and guide sustainable intensification in both staple and cash crop systems.

4.3.Use case 3: Urban Heat Islands

The urban component develops explainable AI models to map and mitigate urban heat island effects, delivering actionable information to planners and local decision-makers. In Guatemala City, end users involved in this project will include municipality employees (two confirmed senior staff members from planning city office) and NGOs active in NBS implementation, urban sustainability and social resilience (two confirmed senior staff members of CALMECAC[1]). In Krakow, we have confirmed the collaboration of one senior staff member from the Department of Environment, Climate, and Air (WS), and the collaboration will build on earlier UHI mapping and the city's participation in the EU "100 Climate-Neutral and Smart Cities by 2030" Mission. Furthermore, a representative of the Sendzmir Foundation[2] promoting sustainable development will be involved in the end-user testing and provide feedback from the level of NGOs. Finally, two private residents of central Krakow will provide citizen-level feedback. The resulting tools and indicators, such as canopy cover and NDVI, will support equitable urban planning, informed policymaking, and improved community adaptation to heat-related risks.

[1] https://www.fundacioncalmecac.org/

[2] https://sendzimir.org.pl/en/

5. Summary & Conclusion

The AETHER project will develop a self-explainable AI (S-xAI) framework that integrates deep learning, concept-based reasoning, and retrieval-augmented generation to produce transparent, scientifically grounded, predictions from Earth Observation data. By aligning EO embeddings with human-interpretable concepts and generating text-based explanations, the system will bridge the gap between powerful AI models and stakeholder needs for trust, traceability, and actionable insight. The architecture is modular, scalable, and applicable across three high-impact use cases—biodiversity monitoring, crop yield prediction, and urban heat island assessment—each supported by rich public datasets and extensive end-user engagement. Overall, AETHER will demonstrate how self-explainable AI can improve environmental monitoring and decision-making by offering interpretable outputs, adhering to FAIR data principles, and an inclusive design that empowers scientists, planners, and communities to better understand and respond to climate-related challenges.

References

ALLEA - All European Academies, 2023. The European Code of Conduct for Research Integrity. ALLEA - All European Academies, DE.

Bayala, J., Sanou, J., Teklehaimanot, Z., Kalinganire, A., Ouédraogo, S., 2014. Parklands for buffering climate risk and sustaining agricultural production in the Sahel of West Africa. Curr. Opin. Environ. Sustain. 6, 28–34. https://doi.org/10.1016/j.cosust.2013.10.004

Beery, S., Cole, E., Parker, J., Perona, P., Winner, K., 2021. Species Distribution Modeling for Machine Learning Practitioners: A Review, in: ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS). Presented at the COMPASS '21: ACM SIGCAS Conference on Computing and Sustainable Societies, ACM, Virtual Event Australia, pp. 329–348. https://doi.org/10.1145/3460112.3471966

Bellard, C., Bertelsmeier, C., Leadley, P., Thuiller, W., Courchamp, F., 2012. Impacts of climate change on the future of biodiversity. Ecol. Lett. 15, 365–377. https://doi.org/10.1111/j.1461-0248.2011.01736.x

Bokwa, A., 2023. Cracow Climate and Urban Heat Island.

Brown, C.F., Kazmierski, M.R., Pasquarella, V.J., Rucklidge, W.J., Samsikova, M., Zhang, C., Shelhamer, E., Lahera, E., Wiles, O., Ilyushchenko, S., others, 2025. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. ArXiv Prepr. ArXiv250722291.

Burke, M., Lobell, D.B., 2017. Satellite-based assessment of yield variation and its determinants in smallholder African systems. Proc. Natl. Acad. Sci. 114, 2189–2194. https://doi.org/10.1073/pnas.1616919114

Cole, E., Van Horn, G., Lange, C., Shepard, A., Leary, P., Perona, P., Loarie, S., Mac Aodha, O., 2023. Spatial implicit neural representations for global-scale species mapping, in: International Conference on Machine Learning. PMLR, pp. 6320–6342.

Deilami, K., Kamruzzaman, Md., Liu, Y., 2018. Urban heat island effect: A systematic review of spatio-temporal factors, data, methods, and mitigation measures. Int. J. Appl. Earth Obs. Geoinformation 67, 30–42. https://doi.org/10.1016/j.jag.2017.12.009

Elith, J., Leathwick, J.R., 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annu. Rev. Ecol. Evol. Syst. 40, 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

European Parliament, Council of the European Union, 2024. Regulation (EU) 2024/1689 on harmonised rules on artificial intelligence (Artificial Intelligence Act).

European Parliament, Council of the European Union, 2016. Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation).

Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., Zhu, X.X., 2023. A survey of uncertainty in deep neural networks. Artif. Intell. Rev. 56, 1513–1589. https://doi.org/10.1007/s10462-023-10562-9

Geyman, E.C., Ferris, A., Sahajpal, R., Anderson, W., Lee, D., Hausmann, N., 2025. An Africa-wide agricultural production database to support policy and satellite-based measurement systems. Sci. Data 12, 1087. https://doi.org/10.1038/s41597-025-05257-5

Ghamisi, P., Yu, W., Marinoni, A., Gevaert, C.M., Persello, C., Selvakumaran, S., Girotto, M., Horton, B.P., Rufin, P., Hostert, P., Pacifici, F., Atkinson, P.M., 2024. Responsible AI for Earth Observation. https://doi.org/10.48550/arXiv.2405.20868

Haider, I., Khan, M.A., Masood, S., Algamdi, S.A., Alasiry, A., Marzougui, M., Nam, Y., 2025. Performance of pre-trained deep learning models for land use land cover classification using remote sensing imaging datasets. Environ. Earth Sci. 84, 298. https://doi.org/10.1007/s12665-025-12317-x

Hartinger, S.M., Palmeiro-Silva, Y.K., Llerena-Cayo, C., Blanco-Villafuerte, L., Escobar, L.E., Diaz, A., Sarmiento, J.H., Lescano, A.G., Melo, O., Rojas-Rueda, D., Takahashi, B., Callaghan, M., Chesini, F., Dasgupta, S., Posse, C.G., Gouveia, N., Martins De Carvalho, A., Miranda-Chacón, Z., Mohajeri, N., Pantoja, C., Robinson, E.J.Z., Salas, M.F., Santiago, R., Sauma, E., Santos-Vega, M., Scamman, D., Sergeeva, M., Souza De Camargo, T., Sorensen, C., Umaña, J.D., Yglesias-González, M., Walawender, M., Buss, D., Romanello, M., 2024. The 2023 Latin America report of the Lancet Countdown on health and climate change: the imperative for health-centred climate-resilient development. Lancet Reg. Health - Am. 33, 100746. https://doi.org/10.1016/j.lana.2024.100746

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., Hussain, A., 2024. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cogn. Comput. 16, 45–74. https://doi.org/10.1007/s12559-023-10179-8

He, B.-J., Wang, W., Sharifi, A., Liu, X., 2023. Progress, knowledge gap and future directions of urban heat mitigation and adaptation research through a bibliometric review of history and evolution. Energy Build. 287, 112976. https://doi.org/10.1016/j.enbuild.2023.112976

Hoang, N.-D., Nguyen, Q.-L., 2025. Geospatial Analysis and Machine Learning Framework for Urban Heat Island Intensity Prediction: Natural Gradient Boosting and Deep Neural Network Regressors with Multisource Remote Sensing Data. Sustainability 17, 4287. https://doi.org/10.3390/su17104287

Höhl, A., Obadic, I., Torres, M.Á.F., Najjar, H., Oliveira, D., Akata, Z., Dengel, A., Zhu, X.X., 2024. Opening the Black-Box: A Systematic Review on Explainable AI in Remote Sensing. IEEE Geosci. Remote Sens. Mag. 12, 261–304. https://doi.org/10.1109/MGRS.2024.3467001

Hou, J., Liu, S., Bie, Y., Wang, H., Tan, A., Luo, L., Chen, H., 2024. Self-eXplainable AI for Medical Image Analysis: A Survey and New Outlooks. https://doi.org/10.48550/arXiv.2410.02331

Jakubik, J., Roy, S., Phillips, C.E., Fraccaro, P., Godwin, D., Zadrozny, B., Szwarcman, D., Gomes, C., Nyirjesy, G., Edwards, B., Kimura, D., Simumba, N., Chu, L., Mukkavilli, S.K., Lambhate, D., Das, K., Bangalore, R., Oliveira, D., Muszynski, M., Ankur, K., Ramasubramanian, M., Gurung, I., Khallaghi, S., Hanxi, Li, Cecil, M., Ahmadi, M., Kordi, F., Alemohammad, H., Maskey, M., Ganti, R., Weldemariam, K., Ramachandran, R., 2023. Foundation Models for Generalist Geospatial Artificial Intelligence. https://doi.org/10.48550/arXiv.2310.18660

Ji, Y., Sun, Y., Zhang, Y., Wang, Z., Zhuang, Y., Gong, Z., Shen, D., Qin, C., Zhu, H., Xiong, H., 2025. A Comprehensive Survey on Self-Interpretable Neural Networks. https://doi.org/10.48550/ARXIV.2501.15638

Kansakar, P., Hossain, F., 2016. A review of applications of satellite earth observation data for global societal benefit and stewardship of planet earth. Space Policy 36, 46–54. https://doi.org/10.1016/j.spacepol.2016.05.005

Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P., 2020. Concept Bottleneck Models. https://doi.org/10.48550/ARXIV.2007.04612

Kumar, D., Bassill, N.P., 2024. Artificial intelligence for sustainable urban climate studies, in: Earth Observation in Urban Monitoring. Elsevier, pp. 291–307. https://doi.org/10.1016/B978-0-323-99164-3.00017-3

Kumar, P., Geneletti, D., Nagendra, H., 2016. Spatial assessment of climate change vulnerability at city scale: A study in Bangalore, India. Land Use Policy 58, 514–532. https://doi.org/10.1016/j.landusepol.2016.08.018

Kumar, P., Hendriks, T., Panoutsopoulos, H., Brewster, C., 2024. Investigating FAIR data principles compliance in horizon 2020 funded Agri-food and rural development multi-actor projects. Agric. Syst. 214, 103822. https://doi.org/10.1016/j.agsy.2023.103822

LCM Team, Barrault, L., Duquenne, P.-A., Elbayad, M., Kozhevnikov, A., Alastruey, B., Andrews, P., Coria, M., Couairon, G., Costa-jussà, M.R., Dale, D., Elsahar, H., Heffernan, K., Janeiro, J.M., Tran, T., Ropers, C., Sánchez, E., Roman, R.S., Mourachko, A., Saleem, S., Schwenk, H., 2024. Large Concept Models: Language Modeling in a Sentence Representation Space. https://doi.org/10.48550/ARXIV.2412.08821

Liu, Z., Li, J., Ashraf, M., Syam, M.S., Asif, M., Awwad, E.M., Al-Razgan, M., Bhatti, U.A., 2024. Remote sensing-enhanced transfer learning approach for agricultural damage and change detection: A deep learning perspective. Big Data Res. 36, 100449. https://doi.org/10.1016/j.bdr.2024.100449

Lobell, D.B., Schlenker, W., Costa-Roberts, J., 2011. Climate Trends and Global Crop Production Since 1980. Science 333, 616–620. https://doi.org/10.1126/science.1204531

Mallick, J., Alqadhi, S., 2025. Explainable artificial intelligence models for proposing mitigation strategies to combat urbanization impact on land surface temperature dynamics in Saudi Arabia. Urban Clim. 59, 102259. https://doi.org/10.1016/j.uclim.2024.102259

Marey, A., Zou, J., Goubran, S., Wang, L.L., Gaur, A., 2025. Urban morphology impacts on urban microclimate using artificial intelligence – a review. City Environ. Interact. 28, 100221. https://doi.org/10.1016/j.cacint.2025.100221

Muruganantham, P., Wibowo, S., Grandhi, S., Samrat, N.H., Islam, N., 2022. A systematic literature review on crop yield prediction with deep learning and remote sensing. Remote Sens. 14, 1990.

Ngaba, M.J.Y., Mgelwa, A.S., Gurmesa, G.A., Uwiragiye, Y., Zhu, F., Qiu, Q., Fang, Y., Hu, B., Rennenberg, H., 2024. Meta-analysis unveils differential effects of agroforestry on soil properties in different zonobiomes. Plant Soil 496, 589–607. https://doi.org/10.1007/s11104-023-06385-w

Oliveira, J., Karlson, M., Ouédraogo, A.S., Bazié, H.R., Ostwald, M., 2025. Towards a framework for monitoring crop productivity in agroforestry parklands of the Sudano-Sahel using Sentinel-1 and 2 time series. Remote Sens. Appl. Soc. Environ. 37, 101494. https://doi.org/10.1016/j.rsase.2025.101494

O'Loughlin, R.J., Li, D., Neale, R., O'Brien, T.A., 2025. Moving beyond post hoc explainable artificial intelligence: a perspective paper on lessons learned from dynamical climate modeling. Geosci. Model Dev. 18, 787–802. https://doi.org/10.5194/gmd-18-787-2025

Paudel, D., Boogaard, H., De Wit, A., Janssen, S., Osinga, S., Pylianidis, C., Athanasiadis, I.N., 2021. Machine learning for large-scale crop yield forecasting. Agric. Syst. 187, 103016. https://doi.org/10.1016/j.agsy.2020.103016

Paudel, D., Boogaard, H., De Wit, A., Van Der Velde, M., Claverie, M., Nisini, L., Janssen, S., Osinga, S., Athanasiadis, I.N., 2022. Machine learning for regional crop yield forecasting in Europe. Field Crops Res. 276, 108377. https://doi.org/10.1016/j.fcr.2021.108377

Paudel, D., Kallenberg, M., Ofori-Ampofo, S., Baja, H., Van Bree, R., Potze, A., Poudel, P., Saleh, A., Anderson, W., Von Bloh, M., Castellano, A., Ennaji, O., Hamed, R., Laudien, R., Lee, D., Luna, I., Meroni, M., Mutuku, J.M., Mkuhlani, S., Richetti, J., Ruane, A.C., Sahajpal, R., Shai, G., Sitokonstantinou, V., De Souza Nóia Júnior, R., Srivastava, A.K., Strong, R., Sweet, L., Vojnovic, P., Athanasiadis, I.N., 2025. CY-Bench: A comprehensive benchmark dataset for sub-national crop yield forecasting. https://doi.org/10.5194/essd-2025-83

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., others, 2021. Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning. PmLR, pp. 8748–8763.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204. https://doi.org/10.1038/s41586-019-0912-1

Roscher, R., Bohn, B., Duarte, M.F., Garcke, J., 2020. Explainable Machine Learning for Scientific Insights and Discoveries. IEEE Access 8, 42200–42216. https://doi.org/10.1109/ACCESS.2020.2976199

Ruthra, R., C, L., P, S.K., Regmi, S., 2025. Integrating Hydrodynamic and Deep Transfer Learning Models for Enhanced Flood Risk Assessment in Urban Areas, in: 2025 8th International Conference on Trends in Electronics and Informatics (ICOEI). Presented at the 2025 8th International Conference on Trends in Electronics and Informatics (ICOEI), IEEE, Tirunelveli, India, pp. 1764–1769. https://doi.org/10.1109/ICOEI65986.2025.11013480

Ryo, M., Angelov, B., Mammola, S., Kass, J.M., Benito, B.M., Hartig, F., 2021. Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. Ecography 44, 199–205. https://doi.org/10.1111/ecog.05360

Santos, P.Z.F., Crouzeilles, R., Sansevero, J.B.B., 2019. Can agroforestry systems enhance biodiversity and ecosystem service provision in agricultural landscapes? A meta-analysis for the Brazilian Atlantic Forest. For. Ecol. Manag. 433, 140–145. https://doi.org/10.1016/j.foreco.2018.10.064

Schiller, J., Stiller, S., Ryo, M., 2025. Artificial intelligence in environmental and Earth system sciences: explainability and trustworthiness. Artif. Intell. Rev. 58, 316. https://doi.org/10.1007/s10462-025-11165-2

Shaamala, A., Tilly, N., Yigitcanlar, T., 2025. Leveraging urban AI for high-resolution urban heat mapping: Towards climate resilient cities. Environ. Plan. B 0, 23998083251337864. https://doi.org/10.1177/23998083251337864

Shang, C., Zhou, S., Zhang, H., Ni, X., Yang, Y., Wang, Y., 2024. Incremental Residual Concept Bottleneck Models. https://doi.org/10.48550/ARXIV.2404.08978

Shatnawi, N., Alqaralleh, R.M., Tarawneh, E.R., 2025. Urban heat island in Amman: AI-based modeling of urban morphology and green infrastructure in mitigating thermal stress. Environ. Earth Sci. 84, 498. https://doi.org/10.1007/s12665-025-12507-7

Sheikh, H.A., Singh, A., Kushwaha, N., Christiaen, C., Tkachenko, N., Sabuco, J., Caldecott, B., 2025. A Field-Level Asset Mapping Dataset for England's Agricultural Sector. Sci. Data 12, 1240. https://doi.org/10.1038/s41597-025-05521-8

Snaiki, R., Merabtine, A., 2025. Recent advances on machine learning techniques for urban heat island applications: a review and new horizons. Sustain. Cities Soc. 134, 106943. https://doi.org/10.1016/j.scs.2025.106943

Srivastava, A., Maity, R., 2023. Unveiling an Environmental Drought Index and its applicability in the perspective of drought recognition amidst climate change. J. Hydrol. 627, 130462. https://doi.org/10.1016/j.jhydrol.2023.130462

Sudmanns, M., Tiede, D., Lang, S., Bergstedt, H., Trost, G., Augustin, H., Baraldi, A., Blaschke, T., 2020. Big Earth data: disruptive changes in Earth observation data management and analysis? Int. J. Digit. Earth 13, 832–850. https://doi.org/10.1080/-17538947.2019.1585976

Taskin, G., Aptoula, E., Ertürk, A., 2024. Explainable AI for Earth observation: current methods, open challenges, and opportunities, in: Advances in Machine Learning and Image Analysis for GeoAI. Elsevier, pp. 115–152. https://doi.org/10.1016/B978-0-44-319077-3.00012-2

Tehrani, A.A., Veisi, O., kia, K., Delavar, Y., Bahrami, S., Sobhaninia, S., Mehan, A., 2024. Predicting urban Heat Island in European cities: A comparative study of GRU, DNN, and ANN models using urban morphological variables. Urban Clim. 56, 102061. https://doi.org/10.1016/j.uclim.2024.102061

Teng, M., Elmustafa, A., Akera, B., Bengio, Y., Abdelwahed, H.R., Larochelle, H., Rolnick, D., 2023. SatBird: Bird Species Distribution Modeling with Remote Sensing and Citizen Science Data. https://doi.org/10.48550/ARXIV.2311.00936

Tseng, G., Cartuyvels, R., Zvonkov, I., Purohit, M., Rolnick, D., Kerner, H., 2023. Lightweight, Pre-trained Transformers for Remote Sensing Timeseries. https://doi.org/-10.48550/ARXIV.2304.14065

van der Plas, T.L., Alexander, D., Pocock, M.J.O., 2025a. Monitoring protected areas by integrating machine learning, remote sensing and citizen science. Ecol. Solut. Evid.

van der Plas, T.L., Law, S., Pocock, M.J.O., 2025b. Predicting butterfly species presence from satellite imagery using soft contrastive regularisation. Presented at the CVPR FGVC12 workshop.

van Eupen, M., Winograd, M., Garcia, E., Barahona, R., 2024. Análisis de vulnerabilidad y riesgo climático para la ciudad de Guatemala.

Vance, T.C., Huang, T., Butler, K.A., 2024. Big data in Earth science: Emerging practice and promise. Science 383, eadh9607. https://doi.org/10.1126/science.adh9607

Vandenhirtz, M., Laguna, S., Marcinkevičs, R., Vogt, J.E., 2024. Stochastic Concept Bottleneck Models. https://doi.org/10.48550/ARXIV.2406.19272

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., Anandkumar, A., Bergen, K., Gomes, C.P., Ho, S., Kohli, P., Lasenby, J., Leskovec, J., Liu, T.-Y., Manrai, A., Marks, D., Ramsundar, B., Song, L., Sun, J., Tang, J., Veličković, P., Welling, M., Zhang, L., Coley, C.W., Bengio, Y., Zitnik, M., 2023. Scientific discovery in the age of artificial intelligence. Nature 620, 47–60. https://doi.org/10.1038/s41586-023-06221-2

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.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., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 'T Hoen, P.A.C.,

Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., Van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., 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, 160018. https://doi.org/10.1038/sdata.2016.18

Winograd, M., van Eupen, M., Garcia Piedrasanta, E., Barahona Fong, R., 2023. Taller sobre vulnerabilidad y riesgo climatico pra la identificacion de puntos criticos y exploracion de opciones para una adaptacion con soluciones basadas en la naturaleza.

Xiao, A., Xuan, W., Wang, J., Huang, J., Tao, D., Lu, S., Yokoya, N., 2024. Foundation Models for Remote Sensing and Earth Observation: A Survey. https://doi.org/10.48550/-ARXIV.2410.16602

Xu, X., Qin, Y., Mi, L., Wang, H., Li, X., 2024. Energy-Based Concept Bottleneck Models: Unifying Prediction, Concept Intervention, and Probabilistic Interpretations. https://doi.org/-10.48550/ARXIV.2401.14142

Yi, S., Li, X., Li, D., Dong, X., Wang, R., Xu, Q., 2025. Hyperlocal heat stress around bus stops in Philadelphia: Insights from spatio-temporal microclimate modeling and explainable AI. Comput. Environ. Urban Syst. 122, 102341. https://doi.org/10.1016/j.compen-vurbsys.2025.102341

Zbinden, R., van Tiel, N., Sumbul, G., Vanalli, C., Kellenberger, B., Tuia, D., 2025. MaskSDM with Shapley values to improve flexibility, robustness, and explainability in species distribution modeling. https://doi.org/10.48550/ARXIV.2503.13057

Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., Beyer, L., 2022. Lit: Zero-shot transfer with locked-image text tuning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18123–18133.

Zhao, S., Tu, K., Ye, S., Tang, H., Hu, Y., Xie, C., 2023. Land Use and Land Cover Classification Meets Deep Learning: A Review. Sensors 23, 8966. https://doi.org/10.3390/s23218966

Zhu, X.X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. IEEE Geosci. Remote Sens. Mag. 5, 8–36. https://doi.org/10.1109/MGRS.2017.2762307