Journal of Environmental & Earth Sciences

https://journals.bilpubgroup.com/index.php/jees

REVIEW

The Digital Frontier: AI Applications in Environmental Monitoring and Earth Science Research

Xuebin Wang, Huanle Zhang *

School of Electronic Information Engineering, Jiangxi Polytechnic Institute, Xinyu 338000, China

ABSTRACT

Emerging as the new frontier of environmental monitoring and environmental science studies, artificial intelligence (AI) promises to allow the process of scalable inferences when new volumes of satellite, airborne, in situ, and model-generated data are taken into account. This review brings together the data ecosystems, methodological principles and application evidence that shape the new frontier of Earth observation, namely the digital frontier. Measurement physics: We explain the effects of measurement physics on problem formulation, label uncertainty, and missingness, and how current machine-learning practices are naively transferred to other domains, despite these domains exhibiting different possibilities that could affect model performance. After this, we discuss principal AI strategies focusing on representation learning and self-supervised pretraining, spatio-temporal deep learning in map and prediction, multi-modal fusion, and generative learning in gap filling, downsizing, and reconstruction. Specific focus is made on physics-guided and hybrid modeling approaches that jointly integrate learned components with mechanistic models to enhance plausibility, extrapolation and uncertainty quantification, calibration, and interpretability needed to gain scientific credibility and operational decision support. In the fields of application that we have considered, land systems, atmosphere and air quality, hydrology and water resources, cryosphere, ocean and coasts, natural hazards and urban environments, we discuss common patterns of success and failure, with operational readiness spanning almost equally evaluation design, data governance, lifecycle maintenance, and architecture choice. Our final contribution is research and community priorities such as Earth system foundation models, resilient extremes and out of distribution beneficial products, decision facing probabilistic products and responsible governance that deal with bias, privacy and dual-use risks. This combination of directions defines a roadmap on the way to

*CORRESPONDING AUTHOR:

Huanle Zhang, School of Electronic Information Engineering, Jiangxi Polytechnic Institute, Xinyu 338000, China; Email: zhl6466@163.com

ARTICLE INFO

Received: 1 April 2026 | Revised: 19 May 2026 | Accepted: 25 May 2026 | Published Online: 12 August 2026

DOI: https://doi.org/10.30564/jees.v8i8.13362

CITATION

Wang, X., Zhang, H., 2026. The Digital Frontier: AI Applications in Environmental Monitoring and Earth Science Research. Journal of Environmental & Earth Sciences. 8(8): 222–250. DOI: https://doi.org/10.30564/jees.v8i8.13362

credible prototypes to reliable and reproducible and beneficial Earth AI systems.

Keywords: Earth Observation; Environmental Monitoring; Spatiotemporal Deep Learning; Physics-Guided AI; Uncertainty Quantification

1. Introduction

Monitoring the environment and Earth science studies are on the brink of fast methodological evolution. Constellations of Earth-observing satellites, flying campaigns, autonomous platforms, and planes and growing in situ networks of sensors are now continuously monitoring the planet [1,2]. Having reanalysis products, high-resolution Earth system simulations, and an ever-increasing amount of human-generated observations (citizen science and social sensing) are now complementing these data streams. They can offer unprecedented spatial coverage, and temporal density, but also a bottleneck of scientific/operational value lies in our capacity to mass, heterogeneous, noisy, and incomplete measurements into timely, reliable data about environmental condition, change, and hazard [3].

The solution to this bottleneck has become one of the focal points of artificial intelligence (AI) as a generalization of machine learning, deep learning, and the new approaches of foundation models. Over the past ten years, AI techniques have shifted out of niche applications, including land-cover classification and object detection in remote-sensing imagery, to general applications throughout the observation-model-decision pipeline. The systems of AI currently support mapping and monitoring of ecosystems, atmospheric composition, hydrologic extremes, cryospheric dynamics, and ocean biogeochemistry; they find widespread use as a component within scientific workflow, be it quality control and data fusion, or as a surrogate model or data assimilation. It is not a simple replacement of instruments. It is already transforming the way the issues of Earth science are crafted, the synthesis of evidence based on the multiplicity of sources, and the expression of uncertainty to the decision-makers [4–6].

Three trends of the digital frontier in Earth observation are mutually reinforcing. The former is the magnitude and heterogeneity of data. In the modern world, remote sensing has been projected over optical and thermal imagers, synthetic aperture radar, LiDAR, and hyperspectral tools with individual dissimilarities to the surface and atmospheric conditions. Meanwhile, sensor webs are used to measure at high frequency meteorology, hydrology, air quality, soil moisture, and ecological variables. This is because these sources provide complementary views, yet their combination is challenging in nature: they vary in terms of spatial resolution, revisit frequency, calibration standards, measurement physics, and error structure. The second is the intensification of the decision timelines. Most environmental uses, such as wildfire detection, flood mapping, harmful algae bloom detection, and extreme heat risk detection, can only be useful when done in a fast and repeated fashion. There is an increased demand for operational agencies and stakeholders for near-real-time products, updated predictions, and early warning signs. The traditional analysis pipelines that are typically built into retrospective studies and well-curated datasets can end up being unable to satisfy the speed and automation demands of such environments. The third movement is the growing ability of Earth system modelling that brings both opportunities and challenges. non-probabilistic High-resolution simulations can offer hypotheses about processes that are rich and counterfactual, whereas they are computationally costly, sensitive to parameterizations, and lingo on areas of data scarcity. Meanwhile, the difference between what is modeled and what is observed, variations in scale, sampling, and definitions of variables, makes it challenging to compare and assimilate them directly. In this scenario, no automatic translation is required between data magnitude and varying model levels to actionable insight [7–10].

AI is at the crossroads of these directions. It provides scalable pattern recognition with large data sets, flexible means to combine heterogeneous sources, and new directions to speed up or scale up process-based models. However, the flexibility that enables AI to perform successfully can also cause risks, including overfitting to spurious correlations. This condition is not robust to changes in domain availability presented as new sensors and unobservable areas, and the interpretability of learned representations can be limited. One of the key reasons for conducting this review is to explain in what ways AI is consistently improving environmental monitoring and Earth science studies, and in what ways it is limited by data, evaluation practices, and governance[11].

The initial excitement in the field was largely because it had been working in the area where convolutional neural networks and segmentation models had distanced dependent variables towards more challenging and increasingly accurate results, such as in land-cover classification, deforestation, and built-environment labeling. Although imagery has continued to be a significant area of focus, it has expanded significantly. Modern AI systems often occur when there is multi-modal fusion, and this information includes satellite data, meteorological fields, elevation models, socio-environmental covariates, and in situ measurements. They also pay more attention to spatiotemporal prediction, which trains to predict the changing aspects of fields like precipitation, air quality, or surface water dynamics as opposed to fixed labels. Monitoring rare or high-impact events, such as methane super-emitters, abrupt glacier calving, and landslides going on, and rapid disturbances in ecosystems centered on the use of change detection and anomaly detection. However, other common applications of AI include super-resolution and downscaling, coarse observation or model output into fine-scale estimates required on the local management scale, gap filling, and quality control in the presence of missingness due to cloud cover, sensor outages, or variable sampling, and so forth. Simultaneously, emulator and surrogate models are currently being created to model costly portions of Earth system models and allow fast scene exploration[12,13].

These jobs are consistent with the various scientific and functional aims, and they put various demands of evaluation and trust. Speed, robustness, and targeted uncertainty may be more important in operational situations than marginal increases in average accuracy. False alarms and undetecteds have asymmetric implications to early warning systems, and working on rare extremes is often the determining factor. Predictive skill is useful, though not enough, in the sciences; a test of credibility is that models reproduce relationships in line with established processes, are stable to plausible perturbations, and can be used to make predictions that can be criticized and reproduced. Consequently, developments in Earth AI should be assessed not by the benchmark scores but by the capacity to handle the realistic spatial and temporal generalization questions and their ability to communicate the ambiguity upon making decisions[14].

Since AI can be loosely defined, it is necessary to fix its intended meaning in the context of environmental monitors and Earth science. It means here computational algorithms that are learning predictive or generative relations with the data, and classical machine learning, such as tree-based, deep learning, such as convolutional networks and transformers, probabilistic or Bayesian, and methods that involve physical constraints, or are coupled with numerical models[3,15]. The foundation models (large models trained on general data sets on self-supervised tasks and subsequently transferred to downstream tasks) are also considered in the review since they are being suggested in general-purpose engines in Earth observation and downstream prediction. Nevertheless, Earth science is unlike most mainstream areas of AI application in a number of aspects that render direct transfer of methods unreliable. There is a high degree of autocorrelation both spatially and temporally of environmental data, and hence naive random splits may inflate estimates of performance. Ground truth can be sparsely distributed, latent, or uncertain, and measurements can frequently be a retrieval algorithm or an algorithmic inference of state variables. And there are several of its central objectives that demand more than prediction, such as the conservation of physical quantities, plausible extrapolation past the training distribution, backing estimates of uncertainty that are sensitive to measurement error as well as to epistemic constraints. However, the pertinent question is not whether AI is more effective than a baseline, but on what conditions it can deliver credible and beneficial inference concerning the Earth system.

There are a number of reasons that can justify the increased rate of adoption of AI in environmental monitoring. Satellite and reanalysis data, as well as cloud-based platforms, have massive open repositories, reducing accessibility barriers to training data-intensive models. Self-supervised learning has minimized the use of curated labels, which is very important in areas where the cost and availability of annotation are scarce. The use of hardware and distributed training allows testing and testing high resolution models of space and time, and pre-training on a large scale. Meanwhile, groundbreaking inventions of the methodology (transformers, diffusion model, graph neural networks, and neural operators) have widened the toolbox to include not just image classification anymore. Lastly, the rising stress of climate extremes and cumulative environmental risks has put additional pressure on the need to have systems with surveillance, prediction, and adaptation capabilities in challenging conditions [16,17].

As well, these drivers contribute to more tension that needs to be resolved in case AI systems are to be scientifically plausible and have operational credibility. Generalization under domain shift is one such problem. The processes and conditions prevailing in the environment change between regions, seasons, and extremes; an environmental-specific model cannot work across environmental variations in the management of the land, aerosol regimes, vegetation phenology, topography, or sensor viewing geometry. Sensor improvements and switching constellations bring on more distribution shifts. Close attention to dataset design, validation splits that mirror intended use (e.g., holdout-by-region/holdout-by-time), and implicit attention to out-of-distribution conditions are, therefore, the factors that render robustness. The second difficulty is the quantification of uncertainty and the relevance of decisions. Most environmental choices are threshold and risk-sensitive, such as issuing a smoke warning, defining the extent of a flood, declaring a drought, or deciding how to invest in infrastructure, and point estimates are often inadequate. The stakeholders require probabilistic products or confidence limits that help to trade off costs against losses and present model boundaries, particularly in the rare extremities. However, most AI systems are not calibrated well, and the estimates of uncertainty are often optimistic, where the model predicts failure most. The third challenge is related to scientific validity and interpretability. AI is able to fit complicated patterns regardless of underlying mechanisms, which increases the possibility of spurious relationships and incorrect attributions, especially when the model builds upon artifacts in labels, sampling, or preprocessing. Such things as interpretability tools can be useful, but should be used with care, as intuitive explanations are not always faithful. These reasons encourage the growing attention paid to physics-guided learning and hybrid modeling, which impose constraints (e.g., conservation law or tying learned components to mechanical models) to enhance plausibility and extrapolation [17–19].

Despite the rapid proliferation of AI applications in environmental monitoring and Earth science, a critical gap remains: existing reviews typically focus on isolated methodological families (e.g., deep learning for image classification) or single application domains (e.g., hydrology or atmospheric composition). Few provide a cross-domain, end-to-end synthesis that explicitly links the unique characteristics of Earth observation data, missingness, sampling bias, measurement physics, and non-stationary domain shifts to appropriate methodological choices, validation practices, and operational governance. From the standpoint of the environmental and earth sciences, this gap is problematic because environmental decision-making (e.g., hazard response, resource allocation, regulatory compliance) demands AI systems that are not merely accurate but also scientifically credible, robust to changing conditions, and ethically governed [20,21].

As environmental monitoring increasingly relies on heterogeneous, noisy, and incomplete data, there is an urgent need to critically assess under what conditions AI-based inferences are reliable. Without a systematic framework, the field risks deploying models that fail under domain shift, produce overconfident or uncalibrated uncertainty, or perpetuate biases—undermining both scientific discovery and environmental management.

This review provides a structured synthesis that (i) characterises Earth observation data ecosystems and their core challenges, (ii) surveys principal AI methodologies (representation learning, spatiotemporal deep learning, generative models, physics-guided hybrid modelling, uncertainty quantification) with explicit attention to their strengths and failure modes, (iii) maps these methods to key application domains (land, atmosphere, water, cryosphere, oceans, hazards, urban, carbon), and (iv) proposes a validation-and-governance framework for operational trust. By integrating these dimensions, the review moves beyond fragmented accounts to offer a coherent roadmap for developing AI systems that are robust, calibrated, interpretable, and responsibly governed, directly aligning with JEES's mission to advance environmental and Earth systems science [5,22,23]. **Figure 1** summarizes this end-to-end view by situating common Earth AI components within an observation–model–decision pipeline, emphasizing where uncertainty, validation, and governance must enter the workflow.

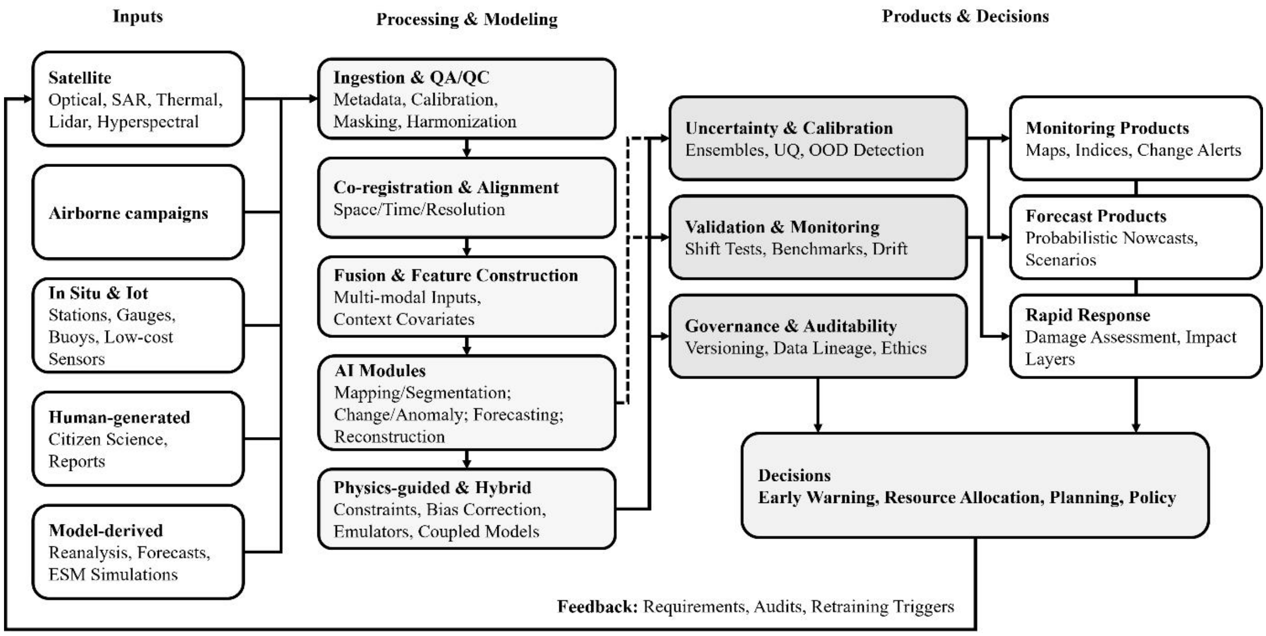

Figure 1. Observation-model-decision pipeline for Earth AI.

2. Data Ecosystems and Problem Formulation

AI is not always constrained to environmental monitoring and Earth science by the capacity of the model [6,20]. These determinants are the nature of the data ecosystem structure and the translation of scientific questions into learning problems, which are the most dominant. Such datasets on the environment have a heterogeneous form in terms of modality, scale, and error; they are determined by the physics of instruments, retrieval algorithms, sampling patterns, and socio-technical constraints on labeling and access. Therefore, one modeling methodology might seem to be very successful when applied to a single data design and fail to work when applied to a different domain of use. The main review of the key sources of data and their properties, the most frequent types of data limitations and shifts, the most prevalent approaches to label construction and supervision, the formulations of the learning problem that can be applied across domains, and the practices of evaluation are more illustrative of the real-world generalization.

Table 1 summarizes the primary data source classes used in environmental monitoring and Earth science, the variables they typically support, and the dominant limitations that shape model design and evaluation.

Table 1. Core data sources, typical variables, and limitations.

Data Source Class	Representative Modalities/Products	Typical Variables Used in AI	Common Preprocessing Steps	Dominant Limitations/Biases	Typical AI Uses
Satellite remote sensing	Optical multispectral, thermal IR, SAR, LiDAR, hyperspectral	Land cover, vegetation indices, surface temperature, inundation, deformation proxies, water color	Atmospheric correction, cloud/shadow masking, speckle filtering (SAR), co-registration, radiometric normalization	Cloud gaps (optical), view/illumination effects, speckle (SAR), calibration drift, resolution mismatch	Mapping/segmentation, change detection, anomaly detection, downscaling, gap filling
Airborne campaigns	Airborne LiDAR, hyperspectral, aerial imagery	Canopy structure, fine-scale land surface features, coastal morphology	Georeferencing, mosaicking, strip adjustment, calibration	Limited temporal coverage, campaign-specific artifacts, small spatial footprint	High-quality labels, calibration/validation, transfer learning
In situ monitoring	Stations, gauges, buoys, flux towers, low-cost sensors/IoT	Meteorology, discharge, PM2.5, soil moisture, water quality, fluxes	QA/QC, drift correction, harmonization, outlier handling	Sparse & biased placement, maintenance outages, instrument drift, representativeness errors	Data fusion, calibration targets, forecasting anchors
Model-derived products	Reanalysis, numerical forecasts, ESM simulations	Gridded atmospheric fields, soil moisture, radiation, wind, precipitation, climate scenarios	Regridding, bias correction, temporal alignment	Model bias/structural error, assimilation artifacts, not ground truth	Covariates, pretraining data, downscaling targets, emulators
Human-generated observations	Citizen science, reports, social sensing	Event occurrence, local impacts, species sightings	De-duplication, geocoding, privacy filtering	Strong behavioral sampling bias, privacy/sensitivity, uneven reporting	Event detection support, validation, exposure estimation

2.1. Observation Sources and Their Measurement Characteristics

In general, four major types of data are used in environmental AI: remote sensing, in situ sensing, model-generated products, and human-observed data. Remote sensing has continued to focus due to its spatial resolution and repeatability [24]. The optical multispectral imagery can offer abundant information about the reflectance of the surface, which allows making inferences related to the land cover, vegetation growth and water characteristics, but it is susceptible to cloud contaminations and illumination effects. Thermal sensors, surface temperature proxies to urban heat, evapotranspiration and wildfire action, have the disadvantage of the spatial resolution and sensitivity to atmospheric disturbances, limiting their interpretation. Synthetic aperture radar (SAR) is provided with the privilege of cloud-penetrating and sensitivity to surface roughness and moisture, which is invaluable when it comes to flood mapping, soil moisture proxies, and deformation monitoring, yet the nontrivial preprocessing and learning problems raised by speckle noise and viewing-geometry effects limit its application. LiDAR provides geometric structure and canopy height data that is hard to access via passive imagery, but which is usually spatially sparse or is done periodically instead of continuously. Hyperspectral instruments provide finer spectral signatures that can be used to identify the material and retrieve biogeochemical values, but have higher dimensionality and more sophisticated calibration and less revisit than multispectral instruments.

In situ sensors are more sensitive to a wide range of variables, but have lower spatial coverage. Temporally dense records such as meteorological stations, river gauges, air-quality monitors, soil probes, and ocean buoys can be used to anchor remote-sensing products, model validation, and provide local early warning. However, they capture a high degree of spatial sampling bias in accessible or populated areas, and their measurements can be biased in site conditions, maintenance cycles, and instrument drift. However, increasingly more cheap sensors and IoT networks are larger-scale, and a trade-off between size and accuracy is common, so calibration, quality control, and uncertainty models come into the spotlight of any AI pipeline that consumes such inputs [25,26].

Products derived through models fill in the observational gaps, and they give physically compatible fields; they do not become ground truth. Reanalysis datasets incorporate various observations into numerical frameworks to generate complete spatiotemporal estimates of at least some variables in the atmosphere and on the surface (global). Much like the Earth system model simulations, the simulations produce scenarios and long-term projections with pre-determined forcings. Such sources may be invaluable to pretraining, data augmentation, and doing identity tasks like downscaling and emulating them, but they encode assumptions and biases of parameterizations and assimilation schemes. They can lead models to memorize model-specific artifacts as opposed to observational reality when used as training targets, when the object of training is in regions where direct measurements are limited [10,22].

The human-generated observations provide a unique data stream whose strengths and weaknesses are different [27]. By supplementing event detection and local validation, especially in times of disaster, citizen science, crowdsourced images, incident reports, and social sensing can be utilized. Nonuniform sampling, privacy issues, and difficult methods of measuring are, however, introduced by them and rely on human behavior and infrastructure. Combined with physical observations, such data can make them more practical but may also enhance bias when not treated openly.

The most important scientific fact that emerges in all these sources is that the data of the environment are measurements of processes by instruments and not realizations of abstract variables. Physics and retrieval algorithms mediate the mapping between measurement and state to form labels, which tend to be proxies and have structural uncertainty. Good AI systems thus must have clear consideration of what is being measured, at what scale, with what error structure, and model design is not even contemplated [15].

2.2. Data Challenges: Missingness, Noise, Bias, and Shift

Environmental datasets have systematic missingness, which is neither randomized nor inconsequential [28]. Cloud cover eliminates optical measurements in a pattern that is seasonally, regionally, and extreme event related, which may confound learning unless the patterns of missingness are incorporated. Sparse revisits and non-random sampling may cause aliasing in quickly changing phenomena like convective storms or flash floods. In situ networks have outages, maintenance discontinuities, and relocation of stations, and generate discontinuities, which can be minor but disastrous to time-series learning.

The structure of noise and uncertainty is similar. Sensor noise, atmospheric effects, viewing-angle effects, reregistration error, in situ sensor drift, and site-specific microclimates are all present in remote sensing data, and products derived using it have both observational error and model bias. Notably, the error statistics tend to be heteroskedastic and depend on illumination, surface type, aerosol load, or even the precipitation regime. When these errors are viewed as homogeneous, they will result in overconfident predictions and worse performance when reliable monitoring is just the required setting [29].

The problem of sampling bias is everywhere and has scientific and ethical consequences [30]. Land cover, biodiversity, or pollution ground truth is often concentrated around areas where field campaigns and monitoring stations are available, which in many cases are in wealthier or more accessible areas. The labelling of disasters might be more comprehensive in locations with good reporting infrastructure. Such biases may produce the second category of models that may seem correct when tested on standard validation data, but that actually perform worse in underserved areas, generating inequities when resources are allocated using products and misleading information are being communicated.

The most perennial operational issue is domain shift [31]. Geographic and climatic regime changes, season and phenological changes, sensor changes and processing lines, and event changes all occur. The AI model that was trained on a collection of sensors can also experience various calibration or spectral reactions following a platform update; a wildfire smoke model may fail on the extreme levels of the compounds that have never been observed before. Moreover, the nonstationarity of climate change is itself added, so the conditions distribution in the future can be different than the historical evidence on which the training is done. Resilience to such changes needs both explicit design decisions in dataset building and evaluation and mechanisms that allow them to realize they are not functioning within their competency. **Figure 2** conceptualizes the major sources of uncertainty and shift in environmental data ecosystems and illustrates how they propagate from measurement and labeling into model training, evaluation, and decision products.

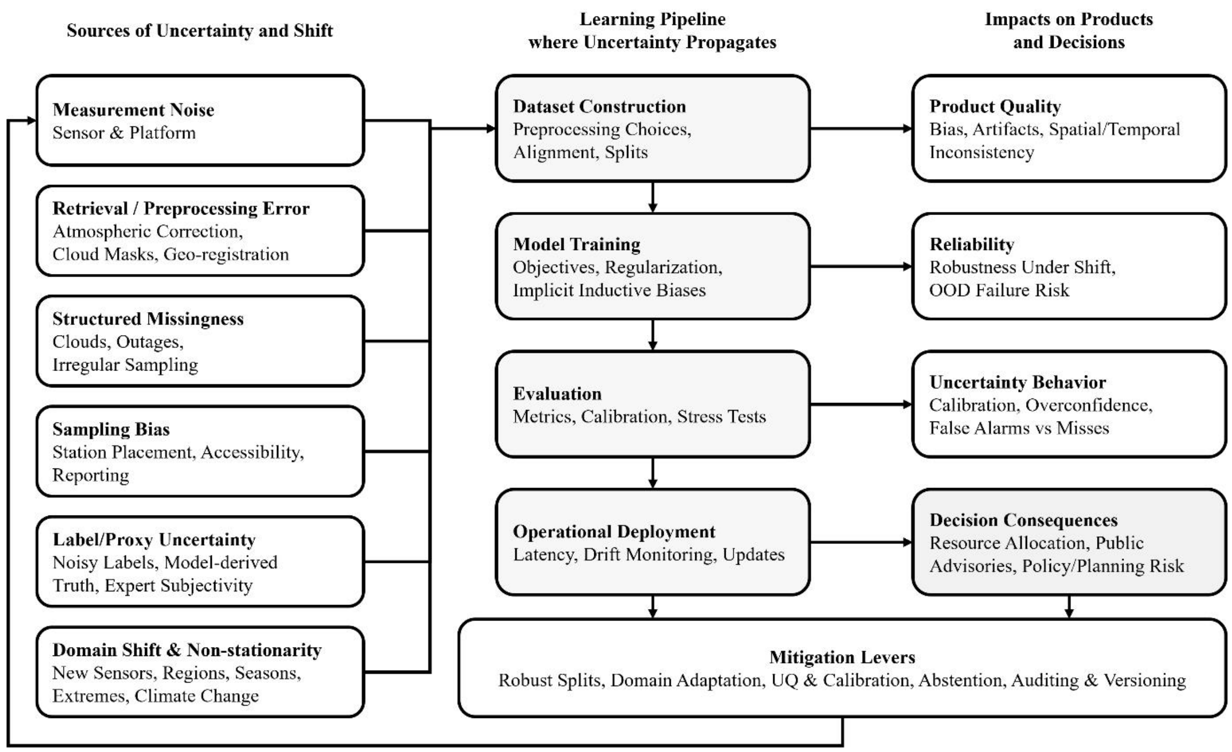

Figure 2. Sources and propagation of uncertainty in environmental monitoring.

2.3. Supervision, Labels, and the Construction of Ground Truth

The Earth science AI is characterized by the lack of and ambiguity of labels [6,21]. Most objects of interest are costly to observe directly, and even ground truth products usually depend on retrieval algorithms or human interpretation as opposed to actual observation. Examples include biomass maps being adjusted to small field plots, inundation labels being based on manual updating, and an inventory of emissions as a mixture of measurements, assumptions, and reporting behavior. Consequently, the labels may be sparse and uncertain, and errors in the labels can be strongly correlated to the covariates that are being used to perform prediction, making learning and evaluation difficult.

In order to solve these limitations, the range of supervision paradigms has materialized. Labels maintained by experts are needed when benchmarking or doing anything high-stakes, but they are not easily scalable. Weak supervision uses heuristics and/or noisy proxies or uses legacy products to create training signals on a large scale, which accepts that labels are inaccurate but informative. Large unlabeled absence of supervision of pretraining has been of interest, especially since it enables models to acquire general spatiotemporal representations that can be tuned to downstream tasks with relatively few labels. The concept of semi-supervised and pseudo-labeling is an extension of the concept of labeled sets, but they grow annotations produced by a model, which increases the risk of enforcing extraneous bias when confidence estimations are uncalibrated [32,33].

The human-in-the-loop approach is becoming a more common practice in the case of operational products that need scaling and responsibility [34]. Active learning may be able to focus on informative or uncertain samples, where it can allocate scarce expert attention where it can produce the greatest impact. The domain experts know the process, and this can be added to the interactive labeling systems, enhancing the quality of labels and their interpretability. Scientifically, those workflows can be correlated with the hypothesis-driven inquiry, where the labelling effort is focused on reangle of models testing model hypotheses instead of merely increasing the sample size.

An important implication of the methodology is that the target variable should be defined carefully. Latent states are often represented by many proxies, and the correlation between the proxy and the state may depend on the part of the country or the time of the year. The model that is trained to forecast a proxy can prove to be very successful yet scientifically misleading in the event that the relationship between the proxy-state is not stationary. As such, explicit label provenance, uncertainty, and known failure modes are a benefit of environmental AI because it allows more meaningful interpretation of model performance [35].

2.4. Problem Formulations in Environmental Monitoring and Earth Science

The scope of environmental usage is varied in the objectives and assumptions of learning formulations. Tasks involving static mapping, e.g., land cover classification or infrastructure extraction, are often formulated as supervised classification or segmentation, the main objective of which is to delineate the space at a given time. Change detection builds on this frame by focusing on differences over time, usually with the constraints of misregistration, seasonal variability, and unequal quality of observations. In such environments, the modeling challenge may often not involve so much learning a class boundary as it involves decontaminating actual change and nuisance variability [22,36].

Most high-value applications are spatiotemporal in nature and thus require formulations of forecasting [37]. This can include the prediction of future conditions of spatial fields, like precipitation, smoke concentration, or surface water extent, or the prediction of the occurrence of events, like the risk of wildfire ignition or a flood. Forecasting settings change the focus to the fitting of patterns in historical data in favor of learning dynamics that are not perturbed. They also ask keener questions of causality and confounding as predictors and targets can be coupled by feedback and common drivers.

The other significant category is that of inverse problems and reconstruction. Super-resolution, downscaling, gap filling, and sensor fusion may be viewed as regression or conditional generation, in which the objective is to predict high-resolution or missing data under partial information. These activities are particularly applicable in remote sensing when clouds, restrictions of revisit, and varying sensor resolution generate systematic gaps. But reconstruction models have the ability to give an illusion of a plausible-looking structure that is not real, and it is uncertainty quantification and physical constraints that can be used in science [38,39].

Graph-based formulations are inherent when the data is observed on non-regular networks like river basins, road networks or sparse grids of sensors. In this situation, the underlying domain is not a regular image lattice, and relationships are based on connectivity, flow, and transport processes. Likewise, multi-task learning can be applied to numerous problems in which the information on related targets is learned together, e.g., soil moisture and evapotranspiration, vegetation indices, and biomass, to take advantage of the joint information and enhance generalization [40].

Lastly, an increasing number of works strive to go beyond being predictive in formulations to causal and decision-based. In climate change and environmental policy related to the scientific question, the issue may be how interventions will affect it, how to attribute changes to the driving factors, or how to assess counterfactual situations. Prediction is not the answer to these questions, as predictive accuracy does not imply proper causal thinking. Although the observational constraints often restrict the full causal identification, the clear identification of the difference between correlational prediction and causal inference can prevent overclaiming and encourage the hybrid methods that combine the knowledge about the processes and the formal assumptions of causality [41,42].

2.5. Evaluation Design and Metrics Aligned to Scientific and Operational Goals

Earth science AI is often misguided by the optimism that evaluation imparts due to its tendency to break the properties of environmental data. Random block cuts within the pixels or timesteps may be information leaking they carry strong spatial and temporal autocorrelation, and give scores on validation that are not indicative of actual deployment. More informative designs of evaluation differentiate training and testing on a regional, time, or event-based basis, based on the intended purpose. An example is that, when the objective is to deploy in new geographies, one should use a holdout-by-region design; when the objective is future forecasting, one needs a purely forward-chaining temporal split. Event-based assessment that challenges the accuracy on specific disasters is usually more significant than the mean accuracy in the usual circumstances when the objective is hazard response [43,44].

Selection of metrics should also be a consequence. A common performance in the monitoring case is asymmetric, with false alarm and missed detection costs varying. The detection of rare-events can also necessitate measures that are sensitive to recall when prevalence is low, yet the false positive, which would saturate the capacity to provide response, are reduced. To evaluate probabilistic products, it requires correct scoring rules and calibration diagnostics to forecast. In mapping, the area-based accuracy can hide the failure of minority classes that are either ecologically or socially significant. In addition, the performance of the uncertainty estimation must be considered in an explicit manner since inappropriately calibrated models may be operationally damaging despite the point estimates appearing to be well-functioning [21].

Its reproducibility and comparability are also hard since datasets have varied preprocessing, definitions of labels, spatial resolutions, and split strategy. Benchmarking activities are good, but they have to be supported with clear records of preprocessing and evaluation procedures. More scientifically informative comparisons are in many instances no longer between two neural architectures, but rather between an AI approach and a physically motivated baseline, such as a statistical model, revision algorithms, and model outputs in form of processes underscoring its physical character. These comparisons can help to understand what information the model is taking advantage of and whether the gains can be viewed as a significant advancement or artificially enhanced improvement [45,46].

2.6. Implications for Method Choice and Scientific Interpretation

The meaning of success and the approaches to use are determined at the end of the data ecosystem and problem formulation. In cases where the labels are sparse and nonstationary, representation learning and adequate validation may be of more significance than architectural novelty. In cases where the missingness is organized, observation processes have to be considered in the models instead of considering gaps to be random. In situations where products are to be used in making decisions, calibration of uncertainty and strength to shift are often of primary importance, and may even override incremental improvements in average accuracy [47].

Scientifically, interpretation also comes as a result of these considerations. Even a model that does a good reconstruction of missing imagery can be weak when used to identify subtle trends in the case of systematic bias. A model that forecasts the proxy variable correctly can still fail to show the underlying mechanism in case the relationship between the proxy-state and its state variable is different across regimes. Hence, environmental AI should be assessed as not just a predictive exercise, but as a measurement and inference pipeline that may be integrated into the intricate physical and socio-technical systems [48,49]. This framing provides the impulse to the methodological review, which then follows, where model classes are considered concerning how they meet the constraints outlined in this section (or not).

3. Core AI Methodologies for Environmental and Earth System Applications

The variety of environmental questions and data modalities has resulted in a variety of methodological landscapes [14,50]. However, the majority of AI systems applied in environmental monitoring and Earth science can be structured around a few recurrent concepts: learning transferable representations in large non-labeled archives; modeling spatial structure and temporal dynamics jointly; using generative models to reconstruct, enhance, and simulate; imbuing physical knowledge through constraints of hybridizing with numerical models; and generating uncertainty estimates and explanations that can be interpreted scientifically to make operational decisions. This part is a survey of these families of methodologies and highlights their assumptions, the kind of problems they are well adapted to, and commonly occurring failure modes under practical situations.

3.1. Representation Learning and Transfer across Sensors, Regions, and Tasks

The main limitation in Earth science AI is that the data are labeled and few, and disproportionate, whereas the number of unlabeled observations is many [6,51]. Representation learning gets around this mismatch by training models to learn general-purpose features on large data sets and apply them to particular tasks of monitoring. The driving force, where the aim is not to minimize label requirements but to increase resistance to distribution shift, in the environmental case, can be the desire to learn representations that are less dependent upon a particular region or type of event.

The aspect of self-supervised learning has gained significant power due to the fact that it does not rely on explicit labels and builds training goals based on the data itself[52]. Predictive sample options that mask parts of visualizing or time sequence to learn multi-view invariance or can utilize variable views, facilitate representations of texture, structure, phenology, and context to be context-dependent. These transformations tend to carry over to lower-level application tasks, e.g., land-cover mapping, change detection, or anomaly detection, particularly with small relaxed Johnny-betray sets. But this process is not automatic in Earth science since nuisance variation usually has scientific implications. As an example, the difference in illumination or seasonal spectral variations can be ignored during generic computer vision, but can be critical information for vegetation health or surface dampness. Representation learning effectiveness is then determined by the extent to which pretraining objectives maintain environmentally significant variability and avoid spurious shortcuts [53].

Another common issue is the transfer learning of sensors and resolutions [21]. The spectral response functions, radiometric calibration, geolocation quality, or noise behavior of a model trained on one satellite can differ on a different platform. Domain adaptation algorithms seek to close these divides by matching feature mass, building sensor-invariant representations, or literally modeling sensor-specific conditions. Practically, the transfer is also commonly most effectively based on algorithmic correspondence in conjunction with prudent preprocessing decisions, due to the fact that errors like misregistration, cloud masks, and atmosphere-adaptation pipelines tend to bridge the apparent gap in the domains. Multi-sensor pretraining, to avoid dependence on a single platform, and the encoding of shared environmental signals across measurement variability, is an increasingly popular usage of multi-sensor pretraining.

3.2. Spatiotemporal Deep Learning for Mapping, Change Detection, and Forecasting

The environmental phenomena are spatial structures that change over time, which is why spatiotemporal modeling is the essence of the methodological requirement of many applications. Convolutional neural networks (CNNs) are still popular in spatial mapping, as they take advantage of the local spatial coherence and scale large images efficiently. Segmentation models and encoder-decoder models are particularly widespread in defining characteristics of surface water, burned areas, urban footprints, and the type of vegetation. The success is in part practical: they are the best at working with high-resolution raster data and can be end-to-end trained on pixel-wise supervision. However, there are also weaknesses of spatial mapping in Earth observation, such as sensitivity to sensor artifact, imbalance of classes, and that smooth boundaries or rare classes are easily misclassified, unless loss functions and sampling schemes are carefully structured [54,55].

Change detection is an extension of mapping, which demands that models make the difference between actual environmental change and confounders like seasonal variations, variable illumination, and suboptimal co-registration [56,57]. Paired or multi-temporal encoders are used to acquire different representations in many of these methods, and change is modeled as a temporal segmentation task in others, which finds transitions between a time series. The methodological challenge is that nothing can be considered as no change, but it is a distribution of possible variability. Models with good results on training sets may not cope with operations where the conditions of observations are more diverse than during training.

Forecasting tasks extend further by the fact that they need models to learn dynamics, not patterns only [58]. Temporal modeling has gone beyond recurrent structures, temporal convolutional networks, to transformer-based sequence models, which are capable of building long-range dependencies. In the case of gridded fields, the models apply the spatial encoders to the temporal modules or apply architectures to process space and time together. The main challenge in precipitation nowcasting, air quality prediction, or flood evolution modeling is that the dynamics can be exogenously driven, and multi-scale interactions, and rare extremes may not be adequately covered by training data. This tends to make predictive models better to use physical covariates, put known drivers into the model explicitly, or be linked to numerical forecasts, as opposed to solely basing the model on past observations.

Graph-based deep learning is especially applicable in cases where the observation space is not a regular grid [59]. Natural graphs: River networks. The edges of a river network represent the flow in the river network. Transportation corridors: The edges of a transportation network represent the flow of a transportation network. Power infrastructure: The edges of a power infrastructure represent the flow of power. Sparse sensor networks: A sparse sensor network can be represented as a graph. Graph neural networks are capable of passing information through such structures, which allow learning to respect connectivity and directionality. Graph formulations are more realistic in hydrologic forecasting, e.g., the upstream-downstream relationship, than the rasterized method. The construction of graphs, however, makes assumptions concerning connectivity and pathways of processes; when the graph does not match the actual physical connection, the model can be simply systematically wrong.

Transformer architectures are given special attention as this has redefined spatiotemporal learning due to their ability to focus on long contexts and very large spatial scales [60,61]. Transformers are applied in Earth observation for imagery and global field purposes, where it can be useful to capture nonlocal interactions and context. However, their flexibility also brings new dangers, such as the propensity to memories region-specific patterns during training splits that are not well-planned and the computational cost of attention systems in high spatial resolution. Real-world uses usually make use of approximations, hierarchical attention, or patch-based processing, creating the opportunity to restore locality, which, unless the model design is sensitive to large-scale teleconnections, can be restrictive to sensitivity to large-scale teleconnections.

3.3. Multi-Modal Fusion and Learning with Heterogeneous Inputs

One of the characteristic features of environmental AI is that heterogeneous data streams are to be integrated, with various coverage, resolution, uncertainty, and semantics [62,63]. Multi-modal fusion techniques strive to integrate these sources into a consistent inference mechanism, whether by mapping, prediction, or attribution. Fusion can be done either earlier, involving the input fusion and learning a shared representation, or later, involving the fusion of modality-specific encoders, which are concatenated by attention, gating, or probabilistic aggregation.

The modalities used in early fusion must be closely aligned (have a shared spatial and time grid), e.g., multispectral bands or derived indices of the same platform. It is harder when the inputs have different resolutions or times, as with satellite imagery with station measurements, or station measurements with a meteorological reanalysis. In this situation, alignment has been incorporated into the learning problem, and models are required to handle implicit interpolation and aggregation. Late fusion can give the option of more flexibility since each modality can be operated to its natural scale, yet may fail to use fine-grained cross-modal interactions when the fusion mechanism is overly coarse [64,65].

The enduring methodological issue is that fusion models can be biased towards the modality that is most readily available to be utilized instead of balancing out the sources [66]. By way of example, socio-environmental covariates can be very strong predictors of a particular outcome and mask the physical signal, resulting in an accurate, but scientifically uninformative or ethically objectionable model. The design of fusion systems thus needs to have clarity concerning what it is supposed to be used in: to provide operational prediction, physical understanding, or both. The need to separate features of physical evidence and features of context is frequently a desirable property in high-stakes monitoring, and quantifying the contributions made by each type is often of interest.

3.4. Generative Modeling for Reconstruction, Enhancement, and Simulation

The importance of generative models is growing due to the fact that most Earth observation tasks are ill-posed or inverse. The existence of cloud gaps and sensor noise, coarse resolution, and incomplete coverage are all incentives to approaches that may recover missing data or change between scales of observation. Examples of these include super-resolution and downscaling, both that make the model learn fine-scale structure by taking coarse inputs. Gap filling in optical imagery, denoising SAR, harmonizing products across sensors are also done using similar techniques [67,68].

The generative models that have been developed recently are of interest to us due to the methodological appeal, i.e., the realism of the results they generate which is similar to the data distribution. Diffusion models, variational autoencoders, and adversarial methods are able to train complicated conditional distributions and produce realistic high-resolution fields. But plausible does not mean true, and the danger of hallucination is especially high in the case of science. A model has the ability to produce aesthetically plausible patterns of a fine scale that cannot be backed by the underlying measurements and hence bias trend analysis, distort event boundaries or mislead downstream decision-making. It is due to this that scientific applications tend to necessitate constraints that fix reconstructions to the evidence that is seen, and uncertainties that are reflective of ambiguity instead of sweeping that ambiguity under the carpet by a single sharp output [69].

Data augmentation and so-called synthetic data are also completed through generative modeling, occasionally to make more representation of rare events or under-sampled areas [70]. Although there is a benefit of augmentation to enhance robustness, the approach introduces the threat of introducing unrealistic artifacts when the generative model is not capturing the actual variability of extremes. The importance of synthetic augmentation then lies in the fact that it should be validated cautiously with held-out real events and the fact that augmented samples should not leak information on test regions or periods.

Another associated direction is generative surrogates, where one tries to model parts of an Earth system model and generate fast approximations of a complex process. In this case, it is not image fidelity but statistical fidelity and computational speed. Emulators are capable of speeding up scenario exploration, sensitivity analysis, and quantifying uncertainty, especially when conventional simulation is too costly to execute at a high resolution. However, emulators need to be considered rough approximations whose inaccuracies could increase as they are combined into larger pipelines. The scientific risk is that the results of the scenarios in fast surrogates can easily induce overconfidence unless the constraints are specified quantitatively [71,72].

3.5. Physics-Guided Learning and Hybrid Coupling with Numerical Models

Since Earth system processes can only be studied by conservation laws and known dynamics, the data-driven learning approach alone can also be vulnerable to extrapolation[58]. Physics-guided methods overcome this by providing explicit structure to the learning system by having physical structure, a differentiable approximation of the governing equations, or a linkage between learned components and numerical models.

One is by putting restrictions on outputs in the form of loss terms or architectural structure and rewarding outputs by conservation of mass or energy, nonnegativity, monotonicity, or other domain-specific properties. Such restrictions have the capability of enhancing plausibility and minimizing catastrophic failures, especially in extrapolative regimes. The other method is based on physics-informed neural networks, which introduce the idea of using the residue of the solution of a differential equation in the training process and can, in principle, with sparse data, learn solutions to known equations. They can be sensitive to the quality and boundary conditions, measurement noise, and the complexity of the system, and have a hard time with stiff systems or high-dimensional nonlinear processes, without being specially designed to do so[73,74].

Neural operators and other methods strive to learn functions between functional spaces, providing one method to estimate operators of PDE solutions under different inputs[75]. These approaches are appealing to the imitation of physical models and to boost components of data assimilation and forecasting. They can promise to learn reusable operators that can make inferences that are generalized to conditions, but their performance is determined by whether the conditions in the training data can cover the relevant regime. In case the learned operator is trained in typical conditions, it will become unstable at the extremes where nonlinearities prevail.

The most practical way to impact an operation can be hybrid modeling[76]. Instead of substituting the numerical models, AI components are placed within a particular section of the pipeline where they enhance the output, e.g., parameterization of subgrid processes, bias corrections, downscaling, or observation operators which convert model state to measurement space. Hybrid systems have the potential to alleviate the rigidity and understandability of physical models with data-driven components. Nevertheless, they also provide new interfaces where errors may propagate and they need to be carefully adjusted to prevent cases of double counting of information where both model-based and data-driven signals are correlated.

Another hybrid application that has proven especially useful is that of data assimilation and inverse modeling, in which the application is to deduce latent state variables or emissions using a combination of observations and dynamical models[77]. AI is able to speed up assimilation by learning the surrogates to the costly forward models or approximating the update step. The methods have the potential to make computational activities more tractable, and also allow higher frequency updates, but need to be proven to be stable and biased, particularly in cases where the observations are uncommon or sparse.

3.6. Uncertainty Quantification and Reliability in Environmental Decision Pipelines

The attribute of uncertainty is never an auxiliary feature in the environmental monitoring system; it is part of scientific plausibility and safety of operation[78]. Combination of measurement error and retrieval uncertainty, sampling bias and model-form uncertainty Environmental datasets include all three types of uncertainties, and these uncertainties tend to be greatest when the extremes are rare and during domain shift. This may therefore imply that AI systems with good point estimates but imprecise uncertainty may be not as useful as those with a little worse mean performance but with predictable probabilistic performance.

Uncertainty may methodologically be dealt with either as aleatoric, which is due to irreducible noise or epistemic, which is due to limited knowledge and model uncertainty[79,80]. Aleatoric uncertainty can be estimated by many systems predicting conditional variance or distributional parameters and this is applicable when the noise of measurement simultaneously varies with condition. Ensembles, Bayesian neural approximations, stochastic regularization methods are popular methods of approximating epistemic uncertainty, where the aim is to maximize uncertainty where the model is not on the training distribution. Practically, the ensembles are still a strong foundation of the baseline since they are not that complicated and tend to enhance the accuracy and calibration but raise the cost of calculation.

The calibration is necessary since the uncertainty outputs should be associated with actual error frequencies[81]. Usage Reliability diagrams Reliability diagrams are of particular use, especially in Earth science, where coordination across the globe can conceal systematic errors in particular climate or land kind. Proper scoring rules Region- or regime-specific calibration tests are thus relevant here. In threshold-dependent decisions, e.g., the issuance of warnings of hazards, uncertainty is preferably considered in terms of the decision, in terms of how much probabilistic outputs enhance expected utility, or in terms of the reduction of false alarms at a given miss rate.

Another reliability issue is that models may become complacent in the presence of a dataset shift [82]. Detection and abstention methods in out-of-distribution are thus becoming more and more relevant, especially in operational products that are required to fail-safely. The right action in times of severe uncertainty is not a good guess, but a warning signal that a person needs more information or he or she should look at the situation.

3.7. Interpretability, Scientific Insight, and the Risk of Spurious Explanations

Interpretability in environmental AI has two different functions, namely scientific knowledge and functional responsibility [83]. Scientific interpretation is concerned with relating model behavior to plausible mechanisms, whereas operational accountability is concerned with explaining predictions to the stakeholders as well as diagnosing failures. The goals are not identical, and they relate to each other; techniques that yield useful narratives do not always follow the internal reasoning of the model.

Saliency maps and attribution methods are popular to highlight influential pixels or regions in the case of imagery-based models, whereas feature attribution can be used in the case of time-series models to highlight influential predictors or lags. The concept-based interpretability methods are trying to regulate the correspondence between the internal representation and human-interpretable concepts, including vegetation type, moisture regime, or atmospheric stability. In earth science, interpretability is most useful as it is applied to establish that the model employs physically significant cues and that it does not employ cloud edges, scanline noise, or administrative boundaries in labels [84].

One of the recurrent dangers is that the means of explanation can be misleading if it has not been proven faithful. Since environmental data are very structured, descriptions may appear natural as a model takes advantage of confounders. Additionally, the interpretability is also complex when there are proxy targets and retrieval products; a model can well explain how it predicts a retrieval artifact, but not how it can deduce the underlying state. Viable interpretability would then contain (strict) interpretability assays: perturbation experiments, originality checks, tests across regime changes, in cooperation with domain specialists, who will be able to assess whether learned patterns can be interpreted as possible processes [85].

3.8. Efficiency, Scalability, and Operational Deployment Engineering

The environmental monitoring systems have a limitation that is likely to be different than the normal research environment. The high-resolution images, world coverage, and continuously updating images leave computational and storage requirements that may consume the model selection. The optimization of latency and throughput in near real-time applications can also override minor improvements in accuracy and unfavourable deployment conditions can consist of edge devices, constrained bandwidth or periodic connectivity [86,87].

Compression techniques like model compression, quantization, distillation and tiling algorithms are thus typical in functional AI on Earth. Streaming inference pipelines are required to process continuous data input, quality management, and product creation, which bring about the challenges of sound engineering methods. Model drift is a phenomenon that should be monitored especially since sensor calibration varies, land cover is not stationary, and the climate system is nonstationary. The operational systems are then required to contain retraining, recalibration and auditing mechanisms with version control and traceable metadata [88].

Scientific issues that touch on deployment realities are such as reproducibility. Extreme sensitivity of results in Earth AI to preprocessing choices, cloud masking, resampling scheme and split design. As a result, operationally credible systems have the advantage of clear data provenance, preprocessing pipelines, and assessment procedures as well as shared benchmarks where possible. Reproducibility can also be promoted by standardized reporting of data sources, training regimes, compute budgets, and uncertainty calibration procedures, even in cases where models cannot be completely open because of licensing or security issues[34].

These methodological themes put together suggest that Earth AI can hardly be a success in architecture. It arises in the correspondence of the representation learning with spatiotemporal modelling, fusion design, physical constraints, uncertainty estimation, interpretability, and deployment engineering with the facts of environmental data and objectives of scientific and operational application[89]. **Table 2** shows the key Earth AI approach categories to best-fit task types, typical of strengths, common failure behavior when the task is shifted, and deployment practical concerns. This correspondence is most evident in the application domain where there is a need to contend with the physics of the domain, label ambiguity, and constraints on decision, analyzed in the next section.

Table 2. Methodology map: AI families, best-fit tasks, strengths, and failure modes.

Method Family	Typical Inputs	Best-Fit Task Types	Strengths in Earth Applications	Frequent Failure Modes	Practical Notes (Compute, Data, Deployment)
Self-supervised/representation learning	Large unlabeled archives (imagery, time series)	Transfer to mapping, change detection, retrieval-like regression	Reduces label needs; improves transfer across regions/seasons	Learns shortcuts if augmentations remove meaningful variability; sensor-specific artifacts	Pretraining cost can be high; fine-tuning usually light
CNN/UNet segmentation	Raster imagery (single or multi-temporal)	Mapping, event delineation (flood, burn, urban)	Efficient and strong spatial inductive bias	Over-smoothing; sensitivity to class imbalance and label noise	Often easiest to operationalize; benefits from tiling + QA
Transformers (spatial/spatiotemporal)	Large-scale rasters/fields + context	Long-context mapping, forecasting, global field prediction	Captures nonlocal context and long-range dependencies	High compute; can memorize geography if splits leak; brittle OOD	Efficient attention variants often needed for high-res
Graph neural networks	Sensor networks, river basins, connectivity graphs	Network forecasting, transport/flow-related inference	Respects topology and directional dependencies	Graph misspecification; transfer issues across basins	Graph construction is a scientific assumption
Generative models (diffusion/VAE/GAN)	Partial/noisy observations + conditioning	Gap filling, super-resolution, harmonization	Handles ill-posed inverse problems; produces realistic samples	Hallucination; bias in long-term trends; uncertainty underreported	Requires strict validation; consider conservative post-processing
Physics-guided/hybrid models	Observations + constraints + numerical model outputs	Emulation, bias correction, downscaling, constrained inference	Better plausibility/extrapolation; leverages known structure	Constraint mismatch; interface bias propagation	Governance needed for model updates and coupling stability
Ensemble/Bayesian approximations	Any of the above	Calibrated probabilistic products	Improves calibration; supports risk-based decisions	Cost increases; still overconfident under severe shift	Often best "default" for operational uncertainty

4. Application Domains and Case Study Taxonomy

The different applications of AI in environmental monitoring and earth science can be best perceived with the help of a task-based taxonomy, which is used to correlate scientific questions with the type of things to observe, modeling options, and the limitations of deployment[90]. The most effective systems, no matter what domain they are applied to, always seem to be based on multiple sources of data or multi-modal observations, as well as multi-spatiotemporal inference, and uncertainty-aware decision criteria. Simultaneously, the notion of ground truth coupled with the incumbency of domain-specific measurement physics and process complexity heavily influences the types of generalization challenges to be practical, what sort of generalization failure is most impactful, and what failure in terms of non-admission. Argue This compartmentalizes the key areas of use and classifies representative pattern case studies in the office concerning their common problem formulation, data ecosystems and operational maturation.

Table 3 provides a cross-domain taxonomy of typical tasks, primary data streams, common modeling patterns, and the validation emphasis most relevant to operational use.

Table 3. Application taxonomy and operational maturity across domains.

Domain	Core Tasks	Primary Data Streams	Typical Model Patterns	Validation Emphasis	Operational Maturity (Typical)
Land & ecosystems	Land cover/use, deforestation, disturbance, biomass proxies	Optical/SAR, LiDAR, climate covariates, field plots	Segmentation + multi-temporal change models; retrieval surrogates	Holdout-by-region/biome; disturbance regime tests	Medium–High for mapping; Medium for biophysical retrievals
Atmosphere & air quality	PM2.5 estimation, plume tracking, emissions screening	Satellite aerosol/trace gases, monitors, reanalysis	Fusion regressors; spatiotemporal forecasting; hybrid bias correction	Season/region stratification; extremes evaluation; calibration	Medium–High for mapping; Medium for forecasting

Table 3. *Cont.*

Domain	Core Tasks	Primary Data Streams	Typical Model Patterns	Validation Emphasis	Operational Maturity (Typical)
Hydrology & water	Flood mapping/forecasting, drought indices, water quality	SAR/optical, gauges, precipitation/reanalysis	SAR segmentation; graph/basin models; hybrid routing	Event-based tests; forward-chaining; basin transfer	High for flood mapping; Medium for forecasting
Cryosphere	Snow/ice mapping, sea-ice motion, glacier change	SAR/optical, altimetry proxies, limited in situ	Spatiotemporal tracking; segmentation under harsh conditions	Sensor/season transfer; uncertainty in trends	Medium
Oceans & coasts	HAB indicators, SST anomalies, shoreline change	Ocean color, SST, in situ, tides/waves	Reconstruction + forecasting; region-adaptive models	Water-type stratification; coastal mixed pixels	Medium
Hazards & response	Wildfire, floods, landslides, damage assessment	Multi-sensor imagery, forecasts, human reports	Rapid segmentation + anomaly detection; human-in-loop	Event holdout; timeliness; false-alarm costs	High for rapid mapping; Medium for prediction
Urban & exposure	Built form, heat islands, exposure mapping	High-res imagery, thermal, census proxies	Segmentation + fusion; careful governance	Bias audits; cross-city transfer; privacy constraints	Medium

4.1. Land Systems and Terrestrial Ecosystems

One of the first and most commonly deployed applications of AI in Earth observation has been land use, mostly due to the comparative ease of sensing the surface of terrestrial environments with optical sensors and the ability to label large numbers of targets with existing maps and human analysis [21,91]. The functions are usually land cover, land use, vegetation condition, loss, and degradation of forests, and the recording of disturbances such as fires, storms, pests, and drought. These issues are commonly framed as supervised segmentation or change detection, where more and more use is made of multi-temporal sequences to see which changes are real and which are seasonal.

Ecosystem surveillance is no longer categorical mapping, but instead of categorical mapping, it is a shift to quantification of biophysical properties of biomass, canopy height proxy, leaf area dynamics, or phenological timing. In this case, AI models are often used as retrieval proxies, which are trained to predict statistical connections between spectral or structural cues and field measurements. Field plots are sparse and biased, meaning that the performance of field plots depends on the overlap between training data and the target domain of inference, with models being able to be highly apparent but poorly performing in ecologically dissimilar areas. The hybrid strategy, which involves the use of climate and topographic covariates, can enhance the stability; however, the interpretability is also problematic, as the physical signal might be overwhelmed by contextual predictors. In turn, the evaluation of ecosystem applications is being demanded more and more, which conditions isolation of generalization between biome types and disturbance regimes, and not on arbitrary spatial divisions that seep into the local context [92].

4.2. Atmosphere, Air Quality, and Composition Monitoring

Spatiotemporal inference in non-homogeneous observation conditions with heavy reliance on meteorology and transport is commonly sought in atmospheric applications [93]. Issues that are important are estimation of near-surface particulate matter concentrations, the tracking of smoke and dust plumes, estimation of emission hotspots, and producing high-resolution air quality fields using sparse monitoring stations. The combination of satellite-based aerosol data, meteorological reanalysis, land-surface features, and ground monitors is a common form of AI model, where the problem is perceived as regression with complicated missingness and nonstationarity.

Due to the fact that atmospheric composition is formed with the help of transport and chemistry, exclusively statistical models can have the tendency of extrapolating poorly during unusual meteorological conditions or extreme events [94,95]. Forecasting and early warning are consequently commonly based on matching learned models with the numerical weather forecast and chemical transport models, either by bias correction, downscaling, or learned emulators that are faster at certain parts. The issues of uncertainty and calibration have a significant role since the performance of decisions is often associated with threshold violations and health warnings. One recurring issue, though, is that the accuracy of models at different locations and times of year may fluctuate widely because of differences in aerosol types, vertical mixing, and viewing geometry, so regime-consistent assessment and out-of-distribution detection is a key aspect of responsible implementation.

4.3. Hydrology, Floods, Droughts, and Water Quality

Hydrologic applications have a high level of spatial connectivity, multi-scale processes, and a high degree of label ambiguity[96]. Remote sensing flood mapping is widely considered a segmentation problem that frequently incorporates SAR to handle cloud cover and flood forecasting needs, spatiotemporal modeling that is conditioned on precipitation, soil moisture proxy, and upstream conditions. The presence of both flash floods and compound events underscores the introduction of one of the fundamental constraints of data-driven approaches: the regimes of the greatest concern are uncommon and can be poorly modeled by historical data, whereas the physical processes external to runoff and routing may vary based on land use and antecedent moisture.

Drought monitoring combines both long-term variations and short-term anomalies, which may involve multi-sensor time series designed to estimate the index or to determine the state of soil moisture[97]. In this case, artificial intelligence can enhance spatial and temporal consistency, yet the most important scientific threat here is confusing correlation and mechanism, in particular when the objects to be trained are products of models themselves. The situation is further complicated by the fact that water quality indicators such as turbidity, chlorophyll-related proxies, or indicators of harmful algal blooms can be monitored optically, which means that optical indicators are subject to atmospheric correction error and locally-specific optical waters. Models in such environments have the advantage of uncertainty-conscious outputs, as well as an explicit set of constraints that do not allow hallucinated structure in gap-filled or downscaled products to be confused with actual biogeochemical change.

Graph representations and basin sensitivity architectures are becoming more popular in hydrology since river networks and catchments provide structure that is not naturally represented in raster models[98]. The effectiveness of these methods, however, is subject to the faithfulness on the supposed connectivity, or to the observance of mass balance and routing timescales on the learning formulation. Flood extent mapping and rapid assessment are operational when at their finest, operational readiness, whereas predictive hydrology and drought attribution are subject to other specifics such as data constraints, nonstationarity, and the existing match of models and observations.

4.4. Cryosphere and High-Latitude Monitoring

Some of the cryosphere uses are to map the snow cover, estimate snow water equivalent proxies, to track changes in glaciers and ice-sheets, and to characterize sea-ice concentration and motion[99]. The environments impose unique methodological problems as the conditions of observation are extreme, and labels are sparse, and dynamics may be both rapid and nonlinear. Polar darkness and clouds are setbacks of optical measurements, whereas we desire SAR due to its speckling signatures and its geometry dependence; and a physical validation of in situ methods is scarce due to access.

The AI helps to enhance segmentation and tracking of ice and snow features, maintain the temporal consistency of products, and allow changes to be monitored more frequently[100]. In the case of sea ice, motion estimation and prediction point towards the value of the spatiotemporal models capable of supporting a representation of advection-like dynamics. In the case of glaciers and ice sheets, calving fronts, crevasses/melt features, and speeding up analysis of large image archives are delimited with the help of AI. But these systems need to be strong to sensor variations and seasonal variation in surface characteristics, and they frequently need to treat uncertainty with care as errors can build when tracking minor trends over extended periods of time. The conflict between visual appeal reconstructions and scientific fact can also be seen in cryosphere applications, and independent measurement validation and conservative uncertainty reporting are especially significant in this context.

4.5. Oceans and Coastal Systems

The sparse in situ sampling and observation of physics at the ocean has limitations on what can be observed on land. Satellite products may imply surface properties, and the quality of retrieval depends upon the atmospheric conditions, sun glint, and water optical properties. Applications of AI include the detection and prediction of harmful algal blooms, coastal water quality estimation, coastal heatwaves identified by surface temperature variations, and coastal erosion and shoreline change. Several of these activities involve critical disaggregation of signal and confounders that are brought about by atmospheric correction, mixed pixels around the shoreline, and variable sediment loads [101,102].

Spatial resolution is of particular importance to coastal applications, as well as the combination of tide, wave, and meteorological context. Effective systems, therefore, tend to blend remote measurements with hydrodynamic models and local measurements. The need to predict blooms or hypoxia-related proxies requires spatiotemporal frameworks that are able to process delayed responses and transport. Its operational value is large, and uncertainty is also consequential, as false alarms may be economically costly, whereas missed incidents may be deadly to ecosystems and human health. Strength of work in water types and seasons, thus it is one of the main evaluation criteria, and models originally successful in a particular area might not be successfully transferred [103].

4.6. Natural Hazards and Disaster Response

Defense response: Data on disaster response has been observed to be a significant force in scrutinizing Earth AI owing to its time pressure inference and scalability [43]. Wildfire detection and burned area mapping, smoke plume tracking, flood extent delineation, landslide susceptibility and detection, storm damage assessment, and quick mapping of infrastructure damage are all examples of its use. The methodological focus is traditionally speed and strength, and the pipelines are designed in a way that consumes new imagery in a short span of time, filters what can be used, and generates products that can be used in maintaining situational awareness.

Such environments are also manifestations of the shortage of fixed standards. The occurrence of disasters is irregular, non-uniform, and non-stationary, and sometimes the most useful performance can be made on the tail of extreme events. Models might fail unpredictably in new circumstances, i.e., when it comes to unusual burn patterns, floodwater contaminated with debris, or complicated patterns of urban damage. Decision rules that are uncertain and event-based validation are therefore very important. Combined methods involving AI detection and physical prognostication, and human analysts in their analysis are usually the ones that offer the best operation pattern especially when resources are utilized based on the use of products. The number of ethical concerns also increases due to the fact that disaster products might also affect who is provided with aid and how promptly, thus discrimination, openness, and control become the key elements of the system design [104].

4.7. Urban Environments, Infrastructure, and Socio-Environmental Exposure

Urban uses include mapping of constructed infrastructures, assessment of impervious surface and vegetation cover, description of urban heat islands and exposure and vulnerability to hazard and pollution [105]. AI-based techniques are particularly adapted to processing high-resolution imagery in accessing urban structure, yet such items can be difficult due to heterogeneity in building materials, roofing, shadows, and informal settlements that may be erratically represented in labelled data.

Predictive performance may be improved as well as interpretability and fairness issues, are raised when the socio-demographic covariates are included in models to estimate risk or exposure [106]. Under these circumstances, it is necessary to make it clear whether the aim is to map physical hazard, estimate human exposure or it is possible to predict the results determined by both environmental and social determinants. Both objectives have varying acceptable input, assessment plans, and governance quality. In applications such as urban heat, thermal remote sensing can be combined with land cover and meteorology; however, inference about the health risk also relies on population vulnerability and adaptive capacity. Consequently, the system-on-a-chip and system-on-a-chip deployable systems are now giving a higher weight to clearly distinguishing between environmental cues and socio-economic correlates, low uncertainty feedback, and deployment strategies that are modeled to capture the diversity of urban morphologies and urban climates.

4.8. Carbon Cycle, Greenhouse Gases, and Climate-Relevant Monitoring

Relevant monitoring of climate can be associated with mapping of carbon stocks and carbon flux proxies, identifying sources of methane and other greenhouse gas emissions, and monitoring land and ocean indicators that are related to climate feedbacks [107]. A lot of these activities are dependent on proxy observations and model-generated products, thereby resulting in a stratified uncertainty structure. As an illustration, the use of biomass proxies to estimate carbon stocks can rely on sparse field plots with carbon stock estimates, and the identification of methane can be influenced by instrument noise, atmospheric effects, and retrieval effects.

The AI can add value by increasing detection sensitivity, developing spatial resolutions of products, and screening large archives of observations at very large scales [108]. Anomaly detection and source localization in emissions monitoring have been identified as suggesting the importance of models that can differentiate between real plumes and confounders, and models that can measure confidence in a detection. In the case of carbon monitoring, spatial and time coherence is paramount due to the fact that the policy and reporting settings of this field typically prioritize the trends and the changes instead of one-time estimates. Such environments are thus very demanding to uncertainty quantification, bias auditing, and reproducibility. Also, their emphasis is made on the role of governance: monitoring products can be regulatory, economical as well as geopolitical, thus methodological transparency and cautious validation are just as significant as raw predictive performance.

4.9. Cross-Domain Synthesis: Recurring Case-Study Patterns and Maturity Levels

In the spheres, we can single out several application patterns. Rapid mapping activities (e.g., delineation of flood extent or burned areas) may become operationally mature earlier since they can be described as segmentation with fairly straightforward visual information and because they can be tested against event-specific ground truth or expert judgement [109]. Gap-filling and downscaling reconstruction activities are common but must be interpreted with a degree of caution since they produce visually plausible results that can conceal uncertainty and bring subtle biases into the analysis of subsequent results. Task forecasts are also the most problematic to predict in a reliable manner, not merely due to the complex dynamics, but also because the evaluation should entail the process of real forward prediction in the nonstationary and infrequent extremes.

The other cross-domain pattern is that hybrid systems often work better than either the data-driven or purely mechanistic approaches in the operational environment [58]. Improved robustness and trust can be achieved by bias-correcting numerical forecasts, combining sparse in situ data and satellite data, and enforcing learned reconstructions using physical consistency. However, hybridization also makes it difficult to assign errors, and it is important to validate the interface closely to ensure that systematic bias is not propagated.

Lastly, although model architecture is sometimes used to determine the maturity of an application, the ecosystem surrounding it strongly influences it: the latency and continuity of data, the availability of ground truth, standardized evaluation procedures, the practice of uncertainty communication, and arrangements that govern the usage and auditing of products [48]. These considerations are the stimulus in the following section that deals directly with validation, trust, ethics, and governance as the conditions of responsible operational deployment, as well as of credible scientific inference.

5. Validation, Trust, Ethics, and Governance for Operational Use

Environmental monitoring AI systems are gradually becoming more significant in hazard response decisions, resource allocation decisions, compliance with regulatory requirements, and long-term planning decisions [110]. In such situations, performance claims should be backed by evaluation schemes that represent actual deployment conditions, and model deliverables should be credible in aspects that go beyond average precision. Trust can be created by a sequence of evidence that includes data provenance, calibration of the benchmarking design, resilience to shift, interpretability, reproducibility, and ethical governance. **Figure 3** summarizes these dependencies as a stepwise validation-and-governance framework that connects technical evaluation to deployment safeguards and lifecycle management. This section is a synthesis of the practices and open problems, telling whether Earth AI systems will prove to be scientifically plausible and operationally dependable.

5.1. Scientific Validity and Benchmarking beyond Convenience Splits

Evaluation, which works to inflate performance unintentionally, is a continuing barrier to reliable Earth AI [111]. The data of the environment are highly correlated with space and time, and most of the data sets contain repetitive measurements of the same places, times, or events. A model can find success in memorizing local context instead of generalizing relationships by randomly splitting training and test sets over pixels, patches, or even timesteps. The resultant scores can appear spectacular, but they will fail when models are used in other regions, other sensors, or other climate regimes.

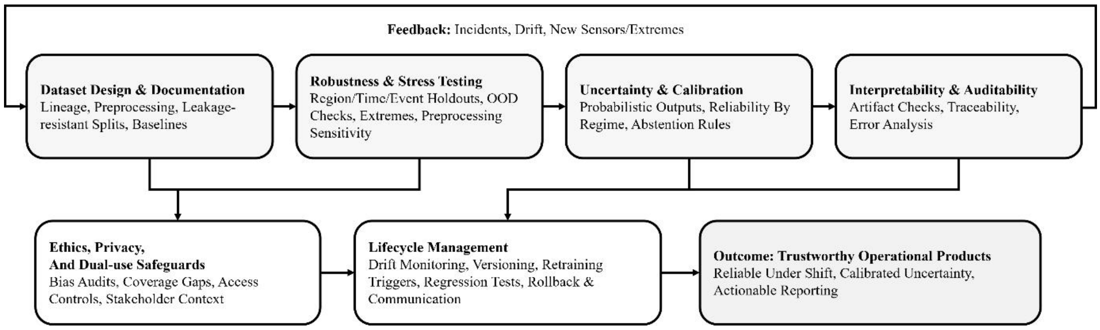

Figure 3. Framework for trustworthy operational Earth AI.

Scientific validity thus relies on benchmarking designs that align with intended use. Holdout-by-region evaluation is necessary when the target of the geographic transfer is to be achieved. In the case of operational forecasting, one needs strictly forward-chaining temporal validation. In situations where the objective is disaster response, event-based testing, in which an entire event is held back, is more accurate about the real use of the models in new situations. In both instances, the assessment should be clear on the type of generalization being asserted, as in Earth science, generalization can be transferred across geography, seasonally, across extremes, across sensors, or within a given societal context [69].

The selection of baselines is also very important [112]. Only problems related to neural architecture can be compared, and it can be unclear whether the improvement is a scientific advancement or a manipulation of the dataset artifacts. More informative reference points are given by physically motivated baselines, classical statistical methods, retrieval algorithms, and numerical model outputs. When AI surpasses them, it will be more reasonable to attribute the increase as actual added value instead of methodological trendiness. On the other hand, in cases where a simple base is almost as good, this can either be a pointer that the task is dominated by readily obtained correlations or that the task is not worthy of the extra complexity that would be needed to be used operationally.

5.2. Robustness and Generalization under Domain Shift and Nonstationarity

The environmental monitoring systems do not work in conditions similar to training distributions [36]. Domain shifts are due to sensor variations, variations in processing pipelines, land use variations, and variations in interannual climate. They also occur due to extremities where the physical system goes into regimes that are exceptional or absent in historical records. Climate change introduces another kind of nonstationarity: the future distribution of conditions cannot be a simple extrapolation of the past, and is especially true of the compound events and tipping-point-like behaviors.

The concept of robustness in this context does not refer to a property, but a sequence of behaviors. An effective model must be able to gracefully degrade with a shift instead of breaking down. It is supposed to have predictable error behaviors and give uncertainty estimates that swell when it gets into unfamiliar regimes. It must be robust to missingness schemes as well as to realistic perturbations of preprocessing options, e.g., cloud masking or atmospheric correction. Stress testing, such as controlled perturbation testing, cross-sensor testing, cross-biome or cross-climate testing, and tests specifically aimed at high-impact and rare subsets of events, is therefore the best measure of robustness [113].

Domain adaptation and transfer learning have the potential to decrease the differences in performance, although it comes with more governance questions. Unlabeled target-domain data is commonly used in adaptation, and this can be operationally useful, but also has the risk of obscuring failure modes when continually updated models are not kept at constant evaluation reference levels. This renders versioning, audit records, and constant monitoring very essential. In the absence of these restrictions, it is hard to know whether a model is improving, or whether it is adapting, or whether the model is overfitting itself to the recent conditions in a manner that decreases the future dependability [6].

5.3. Uncertainty, Calibration, and Decision-Centered Evaluation

The thresholds, the resource constraints, as well as the asymmetric costs, are usually the determinants of environmental decisions. Indeed, uncertainty is not a scientific nicety in such environments, and it is a fundamental necessity for safe and interpretable utilisation. However, on Earth it is difficult to have uncertainty in AI due to a combination of measurement error, label ambiguity, sampling bias, and epistemic uncertainty due to a lack of global coverage of data. One confidence score is hardly able to represent these layers, and naive estimates of uncertainty are most misleading around extremes and when the domain changes [114].

Credible uncertainty starts with calibration [114]. With a high-confidence prediction, a model should have an error rate that measures the confidence of the model within the regimes of interest. Calibration should be assessed on a conditional basis, since a model may be well calibrated on average, and systematically overconfident in particular climates, seasons, or land-cover types. In the case of operation products, uncertainty is also to be analyzed in terms of decision-making. The probabilistic flood map would be useful in the event that it allows reasonable tradeoffs between false alarms and misses due to the limited resources of the responders. On the same note, the probabilistic air quality forecast is valuable when it enhances advisory-making as compared to the current baselines. Decision-centered analysis correlates the outputs of the model to the results of the actions, which is more informative than generic measures of accuracy.

The other vital dimension is the derivation of products in the form of uncertainty [81]. A large number of monitoring pipes reduces model outputs to indices, aggregates, or category alerts. Uncertainty can be crookedly propagated by each transformation. As an illustration, a little bias in a reconstructed field can be transformed into huge variations in threshold exceedance maps. Operational trust thus relies on the design of pipelines that maintain uncertainty by way of post-processing and reporting in ways that the stakeholders can understand in a non-ambiguous manner.

5.4. Interpretability, Auditability, and the Boundary between Prediction and Explanation

The need for interpretability has two intersecting purposes: scientific knowledge and operational accountability [6]. The users of science would be interested in understanding whether or not models are responsive to plausible physical cues and whether the learned associations are consistent with process knowledge. Operational users would like to troubleshoot failures, defend decisions, and learn in which situations they should trust or discount outputs. These requirements put a strain on the interpretability techniques that are intuitive as well as accurate to model behavior.

Earth science uses proxy targets and retrieval artifacts to make interpretability more complex [85]. A model can also be trained to predict a proxy with high accuracy by taking advantage of systematic artifacts in the retrieval algorithm or label construction procedure, and generate explanations that seem to be plausible but irrelevant in science. On the same note, confounders related to the target (e.g., geography, infrastructure, or observational conditions) can be used by models without modelling the underlying process. Verifiable interpretability therefore, demands verification, both through perturbation tests, counterfactual criticisms where possible and on domain shifts that violate spurious correlations.

The interpretability is stretched to governance by auditing. When a high stakes monitoring is required, one should be able to trace the manner in which a product was created, which data were utilized, what model version created it, and what preprocessing steps were taken. This is not simply a matter of operational hygiene, but a requirement of scientific reproducibility, and that there be responsibility in the cases where products affect policy or emergency reaction. Auditability also promotes iterative improvement of procedures through systematic error analysis, in place of erratic troubleshooting following the symptom of failure [51].

5.5. Responsible AI, Bias, and Environmental Justice

Environmental monitoring products have the potential to shift the focus and resources. Provided that models historically underperform in areas with sparse monitoring or poorly represented locations on land-cover types, they can strengthen disparities by simply giving lower-quality data in the areas of possible greatest vulnerability. There are numerous sources of bias, which include: unequal distribution of sensors and labels, differences in quality of retrievals between regions, socio-technical variation in reporting and documentation, and historical trends in investment in monitoring infrastructure [115].

Responsible Earth AI thus needs to be the subject of technical and governance action. Any bias can be quantified technically by means of stratified assessment of regions, climates, and demographic situations where suitable, and reduced by means of selective data gathering, active learning, domain adjustment with safeguards, and uncertainty-conscious abstinence. These are governance responsiveness through disclosure of known coverage gaps, effective communication of constraints, and involvement of stakeholders who receive the affected results of monitoring. In the contexts in which monitoring overlaps with privacy, e.g., high-resolution urban mapping or examining human-generated observations, responsible practice also demands mechanisms that restrict abuse and preserve sensitive data [116].

Another ethical issue is that environmental monitoring is dual-use [34]. Conservation or hazard response instruments could be used to spy, enforce or gain strategic advantages. Although lots of Earth observation data are accessible in the open, the introduction of automatic detection and attribution may alter the risk profile by reducing the effort that may be needed to identify targets or make inferences about sensitive activity. This renders access control, the prudent use of releases, and context-sensitive governance to be valuable ingredients in responsible deployment.

5.6. Governance, Documentation, Reproducibility, and Lifecycle Management

Reliable maintenance of Earth AI is maintained as time goes by due to the governance structures outlining standards, documentation, and maintenance practices [117]. Documentation is of a basic nature since it gives the context in which the performance claims can be interpreted. The documentation of data should have provenance, preprocessing, coverage limits, and known failure modes, whereas model documentation should have training regimes, evaluation splits, intended use and uncertainty behavior. Particular attention should be paid to the process of documentation of the observation process and label construction in the environmental contexts, since most of the targets are proxies and even most datasets contain non-obvious biases.

Reproducibility is not just an academic virtue, but operationally, a stable monitoring system would rely on reproducibility [118]. Minor alterations in preprocess, sensor calibration, or model setup can actually modify outputs to the extent that it affects trend analysis and policy interpretation. Reproducible pipelines thus make a strict versioning of both data, code, and models, as well as not yet moving reference benchmarks that can detect regressions. Changelogs and impact assessment assist users in understanding the difference when models are updated, whether the difference reflects a real change in the environment or a drift in the methods.

As Earth AI transitions into sustained services, lifecycle management becomes more centralized [119]. The operational systems have to track drift, handle retraining, handle backward compatibility where necessary and have human supervision channels when the confidence is low. They should also state the protocols of escalating the cases of failure, especially on the respond to hazard. Governance structures to define roles, responsibilities, and validation requirements have the ability to stop the typical situation of models being implemented without a maintenance, audit, and responsible retirement strategy.

Considered collectively, these reasons suggest that environmental AI trust is built as a system property and not a model property. Strong architectures and big datasets are not enough when evaluation is not well aligned with deployment, uncertainty is not calibrated, and there is no governance. On the other hand, moderately complex models can be reliably monitored as long as they are integrated into transparent pipelines, assessed in realistic shift conditions, coupled with calibrated uncertainties, and controlled using definitive standards and ethical protection. Towards these requirements, **Table 4** outlines a reporting and validation checklist of reliable Earth AI systems, including data lineage, split design, robustness, uncertainty calibration, interpretability and governing the lifecycle. This view preconditions the last section that takes into account new directions and community-level developments that are required to develop the field responsibly [120].

Table 4. Trust and governance checklist for operational Earth AI.

Dimension	What Should Be Reported	Recommended Tests	Typical Red Flags	Minimum Operational Safeguards
Data lineage & preprocessing	Source, versions, QA/QC, masking, regridding, alignment	Sensitivity to preprocessing choices	Unreported cloud masks, unknown label provenance	Versioned pipelines; reproducible preprocessing
Split strategy & leakage control	Spatial/temporal/event splits aligned to use case	Holdout-by-region/time; event-based testing	Random pixel splits; autocorrelation leakage	Published split protocol; independent test sets
Baselines & comparators	Physical/statistical baselines; numerical model outputs	Head-to-head by regime and extreme subsets	Only DL-vs-DL comparisons	Include mechanistic baselines; ablations
Uncertainty & calibration	Probabilistic outputs, calibration metrics, reliability by regime	Reliability curves; proper scoring; OOD detection	Overconfident extremes; no regime-wise calibration	Calibrated uncertainty; abstention rules
Robustness & shift	Sensor/season/region transfer performance	Cross-sensor transfer; stress tests	Performance collapses outside training domain	Drift monitoring; retraining policy
Interpretability & auditing	Evidence model uses physical cues; traceable outputs	Perturbation tests; artifact checks	Explanations that track clouds/boundaries	Audit logs; analyst review pathways
Ethics & dual-use	Bias analysis, privacy handling, access controls	Stratified equity checks; privacy threat modeling	Underperformance in low-monitoring regions	Disclosure of coverage gaps; governance board/process
Lifecycle management	Update cadence, model cards, changelogs	Regression tests on fixed benchmarks	Silent model updates; no rollback plan	Versioning, rollback, and user communication

6. Future Directions and Conclusions

AI has ceased being a supporting technology and has been put at the center of both methodology and Earth science studies. On the ground, in the air, water, ice, seas, risks, and cities, learning-based algorithms now commonly speed up mapping, enhance information fusion, allow collections of data to be updated more frequently, and increase the breadth of inferences that can be drawn about non-homogeneous observations. Longest lived progress has been made in which AI is considered as an end-to-end inference system, not an independent predictor, paying attention to observation physics, label provenance, spatiotemporal structure, and contexts of product use.

Simultaneously, as mentioned in the review, the main impediments to impact are more and more methodological and institutional than merely computational. Most of the documented advances are easily shaky in the face of realistic changes in domains such as geographic relocation, sensor reconfigurations and the uncommon extremes. Targets created through modeling and proxy labels have the potential to inflate face validity but hide scientific validity. Generative reconstruction and downscaling may give visually appealing yet scientifically dangerous results in case of no explicit expression of uncertainty. The tools of interpretability are capable of giving convincing stories and not showing the faithful causal or mechanistic grounding. This does not disregard the worth of AI, but rather points to the fact that performance indicators cannot build trust in Earth AI systems that drive high-stakes decisions by themselves.

The formation of more widely applicable spatiotemporal representations and representation systems (foundation models) specific to the Earth system are a central future direction. The promise is massive: models trained on cross-sensor, cross-season, cross-modal precursors might lessen label reliance and offer a mutual depiction interface to various downstream chores. But foundation models cannot be transformative without stringent evaluation procedures that toxic ate generalization in the modes that Earth science needs, such as holdout-by-region, holdout-by-time, stress testing on extremes, calibrated uncertainty in form of distribution shift. The scale can only increase the failure modes and not the safeguards without such safeguards.

The second direction is that of physics-guided and hybrid modeling. The most plausible way of achieving a healthy extrapolation is probably not the data-only forecasting, but the integration that incorporates the advantages of both AI and the mechanistic models. This contains trained emulators that run costly parts, parameterizations based on AI to enhance subgrade resolutions, bias correction and downscaling which maintain physical harmonies, and differentiable or surrogate-based assimilation that can run at higher frequency. The advances in this area requires thoughtful layout of interfaces, strict management of conservation and its constraints, and validation which quantifies stability and bias propagation across coupled pipelines.

The third trend is the uncertainty-conscious, decision-focused products maturation. Environmental monitoring is becoming effective and operational application requires communicated calibration uncertainty that is in harmonization with action thresholds and risk tolerance. Further development of this need's better ways of handling epistemic uncertainty under change, effective abstention and out-of-distribution identification and systematic propagation of uncertainty using post-processing and aggregation stages. It also involves the assessment systems that do not treat uncertainty as a secondary diagnosis but they associate probabilistic performance with the quality of the decision.

The fourth direction is enhanced community standards of benchmarking, documentation and reproducibility. The review highlights that to a great extent, the seeming inconsistency in reported results can be attributed to variations in preprocessing, definition of labels and validation splits. It would be useful to develop common protocols (particularly those of spatial and time cross-validation, event-based testing, and regime stratified reporting) to enhance comparability and minimize the danger of false generalization. Extent to which data lineage, label uncertainty, model intent, and known limitations can be represented are also critical practice of documentation There is absolutely no greater scientific credibility or operational accountability than striking these practices into a standardized form.

Lastly, the responsible governance needs to be developed together with technical capability. The implementation of environmental AI systems can support inequities in the case of uneven coverage monitoring, and new threats in the case of high-resolution detection capabilities being used in malicious applications. To resolve these problems, fairness metrics are not enough but participatory design where justified, transparency regarding coverage gaps, privacy and sensitive site protections and explicit deployment, auditing, and maintenance rules. Trust is eventually gained not just by correctness, but also by exhibited reliability, openness and adherence to social values.

To sum up, AI is widening the digital boundary of environmental surveillance because it is now possible to transform large volumes of heterogeneous data into information that is timely and actionable. This area is now ready to move from the phase of demonstrations of feasibility to one of the standards of reliability. To reach that change, machine learning innovation will need to be more closely related to Earth science principles: physical manner of treatment and uncertainty, assessment that is more realistic and captures rare extremes, hybrid methods that respect physical constraints, and governance systems that facilitate reproducibility and ethical usage. When all these factors proceed simultaneously, AI could assist in the creation of more efficient monitoring systems that are more credible, in that they assist in the process of scientific discovery, enhance the capability of early warning, and make decisions in a world where the pace of environmental change is increasing.

Funding

This work received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data is all presented in the article.

Conflicts of Interest

The authors declare no conflict of interest.

AI Use Statement

During the preparation of this work, the authors used AI tools for grammar checking. The authors subsequently reviewed and edited the content as necessary and take full responsibility for the final content of the published article.

References

[1] Freeman, R.H., 2020. Overview: Satellite Constellations. Journal of Space Operations & Communicator. 17(2).

[2] Ustin, S.L., Middleton, E.M., 2024. Current and near-term earth-observing environmental satellites, their missions, characteristics, instruments, and applications. Sensors. 24(11), 3488.

[3] Binetti, M.S., Massarelli, C., Uricchio, V.F., 2024. Machine learning in geosciences: A review of complex environmental monitoring applications. Machine Learning and Knowledge Extraction. 6(2), 1263–1280.

[4] Gu, Z., Zeng, M., 2024. The use of artificial intelligence and satellite remote sensing in land cover change detection: Review and perspectives. Sustainability. 16(1), 274.

[5] Kazanskiy, N., Khabibullin, R., Nikonorov, A., et al., 2025. A comprehensive review of remote sensing and artificial intelligence integration: Advances, applications, and challenges. Sensors. 25(19), 5965.

[6] Zhao, T., Wang, S., Ouyang, C., et al., 2024. Artificial intelligence for geoscience: Progress, challenges, and perspectives. The Innovation. 5(5), 100691.

[7] Zhao, Q., Yu, L., Du, Z., et al., 2022. An overview of the applications of earth observation satellite data: Impacts and future trends. Remote Sensing. 14(8), 1863.

[8] Panda, S.S., Rao, M.N., Thenkabail, P.S., et al., 2024. Remote sensing systems—Platforms and sensors: Aerial, satellite, UAV, optical, radar, and LiDAR. In Remotely Sensed Data Characterization, Classification, and Accuracies. CRC Press: Boca Raton, FL, USA. pp. 3–57.

[9] Srivastava, A., Jain, S., 2025. Remote Sensing for Environment Assessment: Multispectral, Hyperspectral, and Thermal Imaging Applications. In Remote Sensing for Environmental Monitoring. Springer: Singapore.

[10] Tuia, D., Schindler, K., Demir, B., et al., 2024. Artificial intelligence to advance earth observation: A review of models, recent trends, and pathways forward. IEEE Geoscience and Remote Sensing Magazine. 13(4), 119–141.

[11] Rane, N.L., Mallick, S.K., Rane, J., 2025. Artificial intelligence and machine learning for enhancing resilience: Concepts, applications, and future directions. Deep Science. DOI: https://doi.org/10.70593/978-93-7185-143-5

[12] Bhalani, V., 2025. Multi-Modal AI for Renewable Energy Forecasting: Combining Satellite Imagery, Weather Data, and Historical Patterns. Journal of Engineering and Computer Sciences. 4(7), 66–73.

[13] Deep, G., Verma, J., 2024. Deep learning models for fine-scale climate change prediction: enhancing spatial and temporal resolution using AI. In Big Data, Artificial Intelligence, and Data Analytics in Climate Change Research: For Sustainable Development Goals. Springer: Singapore. pp. 81–100.

[14] Akhtar, Z., Rawol, A., 2025. Harnessing artificial intelligence (AI) towards the landscape of big earth data: Methods, challenges, opportunities, future directions. Journal of Geography and Cartography. 8(1), 10224.

[15] Chauhan, B.V.S., Vedrtnam, A., Wyche, K.P., et al., 2025. Review on artificial intelligence in the environmental monitoring. In Prospects of Artificial Intelligence in the Environment. Springer: Singapore. pp. 29–60.

[16] Liu, X., Gu, D., Chen, Z., et al., 2023. Rise of distributed deep learning training in the big model era: From a software engineering perspective. ACM Transactions on Software Engineering and Methodology. 32(6), 1–26.

[17] Wang, L., Yan, J., Ma, Y., et al., 2024. Cloud computing in remote sensing: A comprehensive assessment of state of the arts. In Remote Sensing Handbook, Volume I. CRC Press: Boca Raton, FL, USA. pp. 399–438.

[18] Pandey, A., 2025. Harnessing AI for Environmental Hazard Mitigation and Resilience: Pathways to a Safer Future. In Geostatistical Insights on Mapping Flood Hazards and Wetland Dynamics. IGI Global Scientific Publishing: Hershey, PA, USA. pp. 151–178.

[19] Roy, P., 2024. Enhancing real-world robustness in AI: challenges and solutions. Journal of Recent Trends in Computer Science and Engineering. 12(1), 34–49.

[20] Alotaibi, E., Nassif, N., 2024. Artificial intelligence in environmental monitoring: In-depth analysis. Discover Artificial Intelligence. 4(1), 84.

[21] Janga, B., Asamani, G.P., Sun, Z., et al., 2023. A review of practical AI for remote sensing in earth sciences. Remote Sensing. 15(16), 4112.

[22] Zhu, J.-J., Yang, M., Ren, Z.J., 2023. Machine learning in environmental research: Common pitfalls and best practices. Environmental Science & Technology. 57(46), 17671–17689.

[23] Li, X., Feng, M., Ran, Y., et al., 2023. Big Data in Earth system science and progress towards a digital twin. Nature Reviews Earth & Environment. 4(5), 319–332.

[24] Zhang, L., Zhang, L., 2022. Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities. IEEE Geoscience and Remote Sensing Magazine. 10(2), 270–294.

[25] Spiridonov, V., Ćurić, M., Novkovski, N., 2025. The New Era of Environmental Monitoring: Insights into Weather, Water, and Climate. In Atmospheric Perspectives: Unveiling Earth's Environmental Challenges. Springer: Cham, Switzerland. pp. 401–443.

[26] Chan, K., Schillereff, D.N., Baas, A.C.W., et al., 2021. Low-cost electronic sensors for environmental research: Pitfalls and opportunities. Progress in Physical Geography: Earth and Environment. 45(3), 305–338.
[27] Huang, X., Wang, S., Yang, D., et al., 2024. Crowdsourcing geospatial data for earth and human observations: A review. Journal of Remote Sensing. 4, 0105.
[28] Zhang, Y., Thorburn, P.J., 2022. Handling missing data in near real-time environmental monitoring: A system and a review of selected methods. Future Generation Computer Systems. 128, 63–72.
[29] Elmes, A., Alemohammad, H., Avery, R., et al., 2020. Accounting for training data error in machine learning applied to earth observations. Remote Sensing. 12(6), 1034.
[30] Hughes, A.C., Orr, M.C., Ma, K., et al., 2021. Sampling biases shape our view of the natural world. Ecography. 44(9), 1259–1269.
[31] Lyu, S., Zhao, Q., Zhou, Z., et al., 2025. Deep learning based domain adaptation methods in remote sensing: A comprehensive survey. arXiv preprint. arXiv:2510.15615.
[32] Li, W., Hsu, C.-Y., Hu, M., 2021. Tobler's First Law in GeoAI: A spatially explicit deep learning model for terrain feature detection under weak supervision. Annals of the American Association of Geographers. 111(7), 1887–1905.
[33] Kondylatos, S., Bountos, N.I., Prapas, I., et al., 2025. Probabilistic machine learning for noisy labels in Earth observation. Scientific Reports. 15(1), 35890.
[34] Stall, S., Cervone, G., Coward, C., et al., 2023. Ethical and responsible use of AI/ML in the earth, space, and environmental sciences. ESS Open Archive. DOI: https://doi.org/10.22541/essoar.168132856.66485758/v1
[35] Ghamisi, P., Yu, W., Marinoni, A., et al., 2024. Responsible AI for earth observation. arXiv preprint. arXiv:2405.20868.
[36] Hino, M., Benami, E., Brooks, N., 2018. Machine learning for environmental monitoring. Nature Sustainability. 1(10), 583–588.
[37] Amato, F., Guignard, F., Robert, S., et al., 2020. A novel framework for spatio-temporal prediction of environmental data using deep learning. Scientific Reports. 10(1), 22243.
[38] Sdraka, M., Papoutsis, I., Psomas, B., et al., 2022. Deep learning for downscaling remote sensing images: Fusion and super-resolution. IEEE Geoscience and Remote Sensing Magazine. 10(3), 202–255.
[39] Vu-Duc, A., Nguyen-Vi, K., Bui-Quoc, B., et al., 2025. A comprehensive survey of super-resolution remote sensing image datasets: Evolution, challenges, and future directions. IEEE Access. 13, 145350–145372.
[40] Pasinetti, A., 2023. Multi-Task Learning for Drought Impact Monitoring across Europe [Master's Thesis]. Andrés Pasinetti, Politecnico di Milano: Milan, Italy.
[41] Bostrom, A., O'Connor, R.E., Böhm, G., et al., 2012. Causal thinking and support for climate change policies: International survey findings. Global Environmental Change. 22(1), 210–222.
[42] Ferraro, P.J., 2009. Counterfactual thinking and impact evaluation in environmental policy. New Directions for Evaluation. 2009(122), 75–84.
[43] Rehan, H., 2022. Enhancing disaster response systems: Predicting and mitigating the impact of natural disasters using AI. Journal of Artificial Intelligence Research. 2(1), 501.
[44] Qin, F., Huang, C., Lin, Z., 2024. Big data and artificial intelligence-driven natural disaster prediction and prevention: Technological advances and prospects. Geographical Research Bulletin. 3, 381–398.
[45] Tuia, D., Roscher, R., Wegner, J.D., et al., 2021. Toward a collective agenda on AI for earth science data analysis. IEEE Geoscience and Remote Sensing Magazine. 9(2), 88–104.
[46] Sun, Z., Cristea, N., Rivas, P., 2023. Artificial Intelligence in Earth Science: Best Practices and Fundamental Challenges. Elsevier: Amsterdam, The Netherlands.
[47] Sharma, S., Layes, J., Esposito, M., et al., 2025. Strategic calibration AI Framework: Adaptive imbalance in dynamic environments. SSRN. DOI: http://dx.doi.org/10.2139/ssrn.5130456
[48] Miller, T., Durlik, I., Kostecka, E., et al., 2025. Integrating artificial intelligence agents with the internet of things for enhanced environmental monitoring: Applications in water quality and climate data. Electronics. 14(4), 696.
[49] Ravi, K., 2025. Socio-technical system challenges in the era of artificial intelligence: A comprehensive analysis. International Journal of Business & Management Studies. 6(9), 75–91.
[50] Rawat, K.S., Yadav, M., 2025. Decoding the intellectual landscape and evolving research paradigm of artificial intelligence in environmental sustainability. IEEE Internet of Things Journal. 12(21), 44343–44366.
[51] Gevaert, C.M., 2022. Explainable AI for earth observation: A review including societal and regulatory perspectives. International Journal of Applied Earth Observation and Geoinformation. 112, 102869.
[52] Gui, J., Chen, T., Zhang, J., et al., 2024. A survey on self-supervised learning: Algorithms, applications, and future trends. IEEE Transactions on Pattern Analysis and Machine Intelligence. 46(12), 9052–9071.
[53] Zhang, H., Xu, J.J., Cui, H.W., et al., 2024. When geoscience meets foundation models: Toward a general geoscience artificial intelligence system. IEEE Geoscience and Remote Sensing Magazine. 13(4),

79–118.

[54] Kumar, R., Bhanu, M., Mendes-Moreira, J., et al., 2024. Spatio-temporal predictive modeling techniques for different domains: A survey. ACM Computing Surveys. 57(2), 38.

[55] Huang, B., Zhao, B., Song, Y., 2018. Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery. Remote Sensing of Environment. 214, 73–86.

[56] Panuju, D.R., Paull, D.J., Griffin, A.L., 2020. Change detection techniques based on multispectral images for investigating land cover dynamics. Remote Sensing. 12(11), 1781.

[57] Kristollari, V., Karathanassi, V., 2022. Change detection in VHR imagery with severe co-registration errors using deep learning: A comparative study. IEEE Access. 10, 33723–33741.

[58] Reichstein, M., Camps-Valls, G., Stevens, B. et al., 2019. Deep learning and process understanding for data-driven Earth system science. Nature. 566(7743), 195–204.

[59] Zhao, S., Chen, Z., Xiong, Z., et al., 2024. Beyond Grid Data: Exploring graph neural networks for Earth observation. IEEE Geoscience and Remote Sensing Magazine. 13(1), 175–208.

[60] Mao, Y., Zhou, H., Chen, L., et al., 2025. A survey on spatio-temporal prediction: From transformers to foundation models. ACM Computing Surveys. 58(4), 85.

[61] Pereira, G.A., Hussain, M., 2024. A review of transformer-based models for computer vision tasks: Capturing global context and spatial relationships. arXiv preprint. arXiv:2408.15178.

[62] Akanbi, A., Masinde, M., 2020. A distributed stream processing middleware framework for real-time analysis of heterogeneous data on big data platform: Case of environmental monitoring. Sensors. 20(11), 3166.

[63] Sun, A.Y., Scanlon, B.R., 2019. How can Big Data and machine learning benefit environment and water management: A survey of methods, applications, and future directions. Environmental Research Letters. 14(7), 073001.

[64] Gomez-Chova, L., Tuia, D., Moser, G., et al., 2015. Multimodal classification of remote sensing images: A review and future directions. Proceedings of the IEEE. 103(9), 1560–1584.

[65] Jiao, T., Guo, C., Feng, X., et al., 2024. A comprehensive survey on deep learning multi-modal fusion: Methods, technologies and applications. Computers, Materials & Continua. 80(1), 1–35.

[66] Dalla Mura, M., Prasad, S., Pacifici, F., et al., 2015. Challenges and opportunities of multimodality and data fusion in remote sensing. Proceedings of the IEEE. 103(9), 1585–1601.

[67] Georgakis, A., 2025. Deep Learning Techniques for Satellite Image Downscaling: A Cross-Sensor Super-Resolution Approach for Cloud Mask Products [Master's Thesis]. National And Kapodistrian University Of Athens: Athens, Greece.

[68] Lasko, K., 2022. Gap filling cloudy Sentinel-2 NDVI and NDWI pixels with multi-frequency denoised C-band and L-band synthetic aperture radar (SAR), texture, and shallow learning techniques. Remote Sensing. 14(17), 4221.

[69] Camps-Valls, G., Tuia, D., Zhu, X.X., et al., 2021. Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science and Geosciences. John Wiley & Sons: Hoboken, NJ, USA.

[70] Bauer, A., Trapp, S., Stenger, M., et al., 2024. Comprehensive exploration of synthetic data generation: A survey. arXiv preprint. arXiv:2401.02524.

[71] Lu, D., Ricciuto, D., 2019. Efficient surrogate modeling methods for large-scale Earth system models based on machine-learning techniques. Geoscientific Model Development. 12(5), 1791–1807.

[72] Bocquet, M., 2023. Surrogate modeling for the climate sciences dynamics with machine learning and data assimilation. Frontiers in Applied Mathematics and Statistics. 9, 1133226.

[73] Willard, J., Jia, X., Xu, S., et al., 2020. Integrating physics-based modeling with machine learning: A survey. arXiv preprint. arXiv:2003.04919.

[74] Cuomo, S., Di Cola, V.S., Giampaolo, F., et al., 2022. Scientific machine learning through physics–informed neural networks: Where we are and what's next. Journal of Scientific Computing. 92(3), 88.

[75] Kovachki, N., Li, Z., Liu, B., et al., 2023. Neural operator: Learning maps between function spaces with applications to PDEs. Journal of Machine Learning Research. 24(89), 1–97.

[76] See, S., Adie, J., 2021. Challenges and opportunities for a hybrid modelling approach to earth system science. CCF Transactions on High Performance Computing. 3(3), 320–329.

[77] Lahoz, W.A., Schneider, P., 2014. Data assimilation: Making sense of Earth Observation. Frontiers in Environmental Science. 2, 16.

[78] Van der Sluijs, J.P., 2007. Uncertainty and precaution in environmental management: insights from the UPEM conference. Environmental Modelling & Software. 22(5), 590–598.

[79] Hüllermeier, E., Waegeman, W., 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine Learning. 110(3), 457–506.

[80] Hofer, E., Kloos, M., Krzykacz-Hausmann, B., et al., 2002. An approximate epistemic uncertainty analysis approach in the presence of epistemic and aleatory un-

certainties. Reliability Engineering & System Safety. 77(3), 229–238.
[81] Caers, J., 2011. Modeling Uncertainty in the Earth Sciences. John Wiley & Sons: Hoboken, NJ, USA.
[82] Turhan, B., 2012. On the dataset shift problem in software engineering prediction models. Empirical Software Engineering. 17(1), 62–74.
[83] Xu, H., Yang, X., Hu, Y., et al., 2024. Trusted artificial intelligence for environmental assessments: an explainable high-precision model with multi-source big data. Environmental Science and Ecotechnology. 22, 100479.
[84] Camps-Valls, G., Reichstein, M., Zhu, X., et al., 2020. Advancing deep learning for earth sciences: From hybrid modeling to interpretability. In Proceedings of the IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020.
[85] Jiang, S., Sweet, L., Blougouras, G., et al., 2024. How interpretable machine learning can benefit process understanding in the geosciences. Earth's Future. 12(7), e2024EF004540.
[86] Kim, J., Zeng, H., Ghadiyaram, D., et al., 2017. Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment. IEEE Signal Processing Magazine. 34(6), 130–141.
[87] Bablu, T.A., Rashid, M.T., 2025. Edge computing and its impact on real-time data processing for IoT-driven applications. Journal of Advanced Computing Systems. 5(1), 26–43.
[88] Tella, A., Molaudi, M., 2025. Ensuring reliable data: Innovative approaches to metadata research and practices. Library Journal (1029-7200). 2025(4), 132–145.
[89] Yu, M., Huang, Q., Li, Z., 2024. Deep learning for spatiotemporal forecasting in Earth system science: a review. International Journal of Digital Earth. 17(1), 2391952.
[90] Persello, C., Wegner, J.D., Hänsch, R., et al., 2022. Deep learning and earth observation to support the sustainable development goals: Current approaches, open challenges, and future opportunities. IEEE Geoscience and Remote Sensing Magazine. 10(2), 172–200.
[91] Lai, J., Kang, X., Lu, X., et al., 2022. A review of land observation satellite remote sensing application technology with new generation artificial intelligence. National Remote Sensing Bulletin. 26(8), 1530–1546.
[92] Turner, M.G., Seidl, R., 2023. Novel disturbance regimes and ecological responses. Annual Review of Ecology, Evolution, and Systematics. 54(1), 63–83.
[93] Wei, G., 2025. Statistical Spatio-Temporal Models with Applications to Natural Processes [PhD Thesis]. Georgia Institute of Technology: Atlanta, GA, USA.
[94] Brasseur, G.P., Jacob, D.J., 2017. Modeling of Atmospheric Chemistry. Cambridge University Press: Cambridge, UK.
[95] Lu, X., Zhang, L., Shen, L., 2019. Meteorology and climate influences on tropospheric ozone: a review of natural sources, chemistry, and transport patterns. Current Pollution Reports. 5(4), 238–260.
[96] Ichiba, A., Gires, A., Tchiguirinskaia, I., et al., 2018. Scale effect challenges in urban hydrology highlighted with a distributed hydrological model. Hydrology and Earth System Sciences. 22(1), 331–350.
[97] Jiao, W., Tian, C., Chang, Q., et al., 2019. A new multi-sensor integrated index for drought monitoring. Agricultural and Forest Meteorology. 268, 74–85.
[98] Lehner, B., Grill, G., 2013. Global river hydrography and network routing: Baseline data and new approaches to study the world's large river systems. Hydrological Processes. 27(15), 2171–2186.
[99] Ryan, J.C., 2024. Cryospheric applications of remote sensing: Snow water equivalent. In Remote Sensing for Characterization of Geohazards and Natural Resources. Springer International Publishing: Cham, Switzerland. pp. 603–621.
[100] Li, W., Hsu, C.-Y., Tedesco, M., 2024. Advancing Arctic sea ice remote sensing with AI and deep learning: Opportunities and challenges. Remote Sensing. 16(20), 3764.
[101] Masria, A., Abouelsaad, O., 2025. Artificial intelligence applications in coastal engineering and its challenges–A review. Continental Shelf Research. 286, 105425.
[102] Muppala, M., 2025. Artificial intelligence, IoT, and sensor technologies for Marine monitoring and climate resilience. In Digital Oceans: Artificial Intelligence, IoT, and Sensor Technologies for Marine Monitoring and Climate Resilience. Deep Science Publishing: New York, NY, USA.
[103] Moriasi, D.N., Gitau, M.W., Pai, N., et al., 2015. Hydrologic and water quality models: Performance measures and evaluation criteria. Transactions of the ASABE. 58(6), 1763–1785.
[104] Khan, S.M., Shafi, I., Butt, W.H., et al., 2023. A systematic review of disaster management systems: Approaches, challenges, and future directions. Land. 12(8), 1514.
[105] Sidiqui, P., Roös, P.B., Herron, M., et al., 2022. Urban Heat Island vulnerability mapping using advanced GIS data and tools. Journal of Earth System Science. 131(4), 266.
[106] Romano, D., 2025. Explainable Artificial Intelligence for Studying the Impact of Environmental Factors on Health [PhD Thesis]. University of Bari and Polytechnic University of Bari: Bari, Italy.
[107] Campbell, A.D., Fatoyinbo, T., Charles, S.P., et al., 2022. A review of carbon monitoring in wet carbon systems using remote sensing. Environmental Research Letters. 17(2), 025009.

[108] Li, W., Hsu, C.-Y., 2022. GeoAI for large-scale image analysis and machine vision: Recent progress of artificial intelligence in geography. ISPRS International Journal of Geo-Information. 11(7), 385.

[109] Campbell, A.D., Fatoyinbo, T., Charles, S.P., et al., 2017. Live Monitoring of Earth Surface (LiMES): A framework for monitoring environmental changes from Earth Observations. Remote Sensing of Environment. 202, 222–233.

[110] Wani, A.K., Rahayu, F., Amor, I.B., et al., 2024. Environmental resilience through artificial intelligence: innovations in monitoring and management. Environmental Science and Pollution Research. 31(12), 18379–18395.

[111] Ghamisi, P., Yu, W., Marinoni, A., et al., 2025. Responsible Artificial Intelligence for Earth Observation: Achievable and realistic paths to serve the collective good. IEEE Geoscience and Remote Sensing Magazine. 13(3), 72–96.

[112] Mamalakis, A., Barnes, E.A., Ebert-Uphoff, I., 2023. Carefully choose the baseline: Lessons learned from applying XAI attribution methods for regression tasks in geoscience. Artificial Intelligence for the Earth Systems. 2(1), e220058.

[113] Van Buren, K., Reilly, J., Neal, K., et al., 2017. Guaranteeing robustness of structural condition monitoring to environmental variability. Journal of Sound and Vibration. 386, 134–148.

[114] Tomsett, R., Preece, A., Braines, D., et al., 2020. Rapid trust calibration through interpretable and uncertainty-aware AI. Patterns. 1(4), 100049.

[115] Sarker, S., Jahan, F., 2025. AI-driven MIS applications in environmental risk monitoring: A systematic review of predictive geographic information systems. ASRC Procedia: Global Perspectives in Science and Scholarship. 1(1), 81–97.

[116] Sentürk, O., 2026. AI-driven climate adaptation: Technical applications, ethical governance, and social inclusion. Acadlore Transactions on Machine Learning. 5(1), 20–31.

[117] Foresti, R., Rossi, S., Magnani, M., et al., 2020. Smart society and artificial intelligence: Big data scheduling and the global standard method applied to smart maintenance. Engineering. 6(7), 835–846.

[118] Thota, M.R., 2022. Next-Generation Observability: AI Techniques for Predictive Performance and Reliability in Data-Intensive Systems. Journal of Scientific and Engineering Research. 9(3), 360–374.

[119] Sakao, T., Funk, P., Matschewsky, J., et al., 2021. AI-LCE: Adaptive and Intelligent Life Cycle Engineering by applying digitalization and AI methods—An emerging paradigm shift in Life Cycle Engineering. Procedia CIRP. 98, 571–576.

[120] Zancan, V., Decuzzi, F., Furano, G., et al., 2024. Dependable Systems and AI in Critical Infrastructures: A Case Study in European Earth Observation Missions. In Proceedings of the 2024 IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT), Didcot, UK, 8–10 October 2024.