Data-Centric Paradigm in AI for Earth Science

Sijie Zhao
Nanjing University Shanghai Artificial Intelligence Laboratory
zsj@smail.nju.edu.cn

Abstract

The proliferation of sensors across Earth's spheres has led to an unprecedented accumulation of geoscientific data, now reaching petabyte scales with terabytes added daily. While these data hold immense value for applications such as surface dynamics monitoring, climate forecasting, and natural disaster early warning, their full potential remains constrained by existing research paradigms. The dominant **model-centric** approach in "AI for Earth Science" focuses on refining complex deep learning architectures but largely overlooks the fundamental challenges associated with the data itself. This proposal argues that the model-centric paradigm is hampered by inefficiencies in data management, coarse-grained data utilization, and high-cost data production. These limitations manifest as prohibitive storage and computational costs, significant data security risks, high information redundancy, severe feature entanglement, and unsustainable data acquisition and preprocessing expenses.

To address these foundational issues, this proposal advocates for a paradigm shift towards a **Data-Centric** AI for Earth Science. The central thesis is the development and application of Autoencoder-based frameworks to transform raw, high-dimensional Earth Science data into a compact, complete, and computationally efficient latent manifold. This learned representation will serve as the cornerstone for a new ecosystem of geoscientific analysis. Specifically, this research will focus on three core objectives: (1) achieving **high-efficiency data management** by developing a latent space paradigm that drastically reduces storage and computation costs, and enables secure, privacy-preserving computation; (2) enabling **fine-grained data utilization** by constructing information-dense and feature-disentangled latent spaces that improve model training efficiency and final performance; and (3) facilitating **low-cost data production** by creating unified generative and preprocessing models that operate within the latent space to generate observational data and streamline complex workflows.

1 Introduction

The contemporary era of Earth Science is characterized by a data deluge (Bergen et al., 2019). A vast network of terrestrial, oceanic, atmospheric, and space-based sensors continuously acquires immense volumes of raw data, chronicling the state of our planet's complex systems in real-time. Over decades of accumulation, the corpus of Earth Science data has burgeoned to the petabyte (PB) scale and continues to expand at a rate of terabytes (TB) per day (Rolnick et al., 2022). As a veridical record of Earth's spheres, this data possesses profound scientific and societal value, underpinning critical applications in surface dynamic monitoring, meteorological prediction, and natural hazard pre-warning systems (Reichstein et al., 2019; Huntingford et al., 2019).

The concurrent rise of deep learning has catalyzed a new research direction known as "AI for Earth Science" (Karniadakis et al., 2021). An increasing body of research demonstrates the power of this approach, wherein bespoke deep learning models are trained on raw geoscientific data to achieve state-of-the-art performance on a wide array of tasks (Bi et al., 2023; Lam et al., 2023). This methodology, which prioritizes the intricate design of model architectures to enhance generalizability and task-specific efficacy, can be defined as a **model-centric paradigm**.

However, as the era of large-scale foundation models (Bommasani et al., 2021) arrives, the data appetite of deep learning architectures is growing insatiably. The model-centric paradigm, with its singular focus on architectural innovation, increasingly neglects the foundational role of the data itself (Zha et al., 2025). This oversight has created significant impediments to the progress of AI for Earth Science, which can be categorized into three primary areas: inefficient data management, coarse-grained data utilization, and high-cost data production.

1.1 Problem 1: Inefficient Data Management

- *Prohibitive Storage and Computational Costs:* The sheer volume of Earth Science data imposes substantial financial and logistical burdens (Ballé et al., 2018). The model-centric paradigm, which operates directly on raw or near-raw data formats (e.g., NetCDF, HDF5), incurs exorbitant costs at every stage of the data lifecycle, including storage, distribution, and computation. Processing PB-scale datasets demands massive, often specialized, high-performance computing infrastructure, limiting accessibility for many researchers.
- *Low Data Security:* Geoscientific data, particularly that collected by national agencies over sovereign territories, is often confidential. The direct use of raw data in the model-centric paradigm creates significant security vulnerabilities. Any breach in the storage or processing pipeline can lead to the exposure of sensitive geospatial and environmental information, resulting in serious security incidents and jeopardizing data-sharing agreements (Shokri et al., 2017).

1.2 Problem 2: Coarse-Grained Data Utilization

- *High Information Redundancy:* Earth systems exhibit strong spatial-temporal continuity. Physical properties of matter across the planet's spheres change in a relatively continuous fashion. Consequently, data acquired by sensors often contains substantial redundancy across spatial, temporal, and spectral dimensions, as well as across different data samples (Bengio et al., 2013). In the model-centric paradigm, this vast amount of redundant and often irrelevant information can interfere with the learning process, degrading both the training efficiency and the final performance of deep learning models.
- *High Feature Entanglement:* Earth Science data are observational records of coupled physical systems. As such, the data features representing different physical phenomena (e.g., cloud cover, soil moisture, atmospheric pressure) are often highly entangled. A model-centric approach forces the learning algorithm to disentangle these complex, interacting features from scratch, often in the presence of confounding signals (Locatello et al., 2019). This significantly hampers the model's ability to learn meaningful representations, reducing both training efficiency and ultimate predictive power (Bengio et al., 2013).

1.3 Problem 3: High-Cost Data Production

- *High Data Acquisition Cost:* Different sensor platforms capture complementary aspects of Earth's systems. A comprehensive understanding requires integrating data from multiple sources. However, these sensors often have disparate spatial locations and temporal revisit rates. Achieving full observational coverage necessitates the deployment and maintenance of a multitude of costly sensor systems, from satellites to in-situ networks. The model-centric paradigm treats data as a given commodity, failing to account for the high cost of its production (Tuia et al., 2022). This constrains research to readily available datasets, impeding progress in data-sparse domains.
- *High Data Preprocessing Cost:* Raw data from sensors is rarely suitable for direct use in scientific models. It typically requires extensive preprocessing—including calibration, atmospheric correction, georeferencing, and quality control—to become analysis-ready (Zhu et al., 2017b). These preprocessing workflows are often complex, computationally intensive, and sensor-specific. The model-centric paradigm relies on this costly preprocessed data, ignoring the bottleneck it creates and limiting research to the subset of data for which these resources have already been expended.

1.4 Objective: Data-Centric Paradigm

Against this backdrop, and informed by my research experience and an extensive literature review, I propose to pioneer a **data-centric paradigm** within the AI for Earth Science domain, as shown in Figure 1. This paradigm aims to leverage techniques from representation learning (Bengio et al., 2013; LeCun et al., 2015), specifically those related to Autoencoders (Kingma et al., 2013; Hinton and Salakhutdinov, 2006), to transform voluminous Earth Science data from its original, raw data space into a compact and complete latent manifold. The objective is to create a highly efficient scientific data representation that directly enables high-efficiency data management, fine-grained data utilization, and low-cost data production. By shifting the focus from the model to the data, this research will catalyze progress across the entire field of AI for Earth Science (Donoho, 2017; Zha et al., 2025).

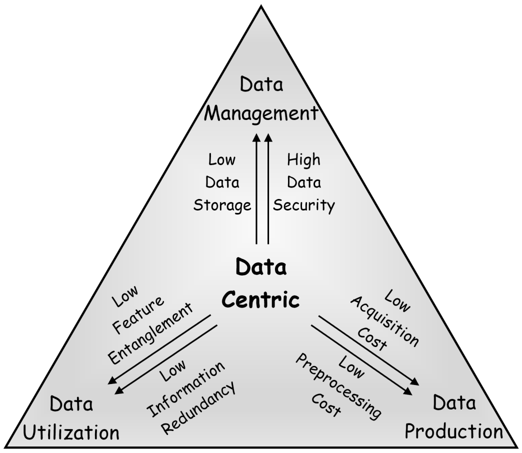

Figure 1: Illustration of the data-centric paradigm.

2 Research Design

2.1 High-Efficiency Data Management

To surmount the challenges of prohibitive storage costs and computational overhead associated with PB-scale geoscientific data, existing research has primarily explored two avenues: data compression and semantic data cubes. Data compression techniques (Ballé et al., 2018), while effective at reducing storage footprints, introduce additional data computation costs. To be utilized, the data must be decompressed, which introduces significant computational latency and can negate the storage benefits during active analysis. Semantic data cubes, on the other hand, leverage foundation models to map high-dimensional raw data to low-dimensional semantic features (Lacoste et al., 2023). While this reduces both storage and usage costs, the semantic features represent a lossy abstraction of the original data. This process discards a vast amount of low-level information, rendering the representation unsuitable for tasks requiring fine-grained detail, for generative modeling (Dhariwal and Nichol, 2021), and for any analysis that necessitates recovery of the original data.

2.1.1 The Latent Space Research Paradigm

Given these limitations, I propose the development of a **Latent Space Research Paradigm**. The core concept is to employ Autoencoder-based models (Kingma et al., 2013) to map high-dimensional geoscientific data into a low-dimensional latent space. This transformation will produce a low-storage latent representation that preserves nearly all the information content of the original data (Hinton and Salakhutdinov, 2006). Critically, subsequent data usage and model training will be conducted directly within this compact latent space, thereby drastically reducing both storage requirements and computational costs throughout the entire analytical workflow.

I have completed a successful proof-of-concept for this paradigm in the domain of meteorology (Zhao et al., 2025). Using an Autoencoder framework, I transformed the 244.34 TB ERA5 reanalysis dataset into a mere 0.43 TB latent representation. This achieved a remarkable compression ratio of 566 while maintaining an exceptionally high reconstruction fidelity, with a mean squared error of 4×10^{-4}. Furthermore, meteorological forecasting models trained directly on this latent representation achieved performance comparable to models trained on the original raw data, and even demonstrated superior performance in the prediction of extreme weather events.

Building on this success, my future work will involve extending the Latent Space Research Paradigm to other domains of Earth Science, including remote sensing, oceanography, and solid earth geophysics. The goal is to establish a generalized framework for creating efficient latent representations for all major spheres of Earth science data, thereby systematically reducing storage and computational costs across the discipline.

2.1.2 Secure Latent Space Representation

Operating within the Latent Space Research Paradigm provides a unique opportunity to enhance data security. I propose to construct a **Secure Latent Space Representation** designed specifically to defend against data reconstruction attacks (Hitaj et al., 2017). This will ensure that even if the latent representation is compromised, recovering the sensitive information within the original data remains computationally infeasible.

To achieve this, I will explore methods from the field of Adversarial Privacy (Edwards and Storkey, 2015). The proposed framework involves augmenting the standard Autoencoder with an additional data reconstruction attacker model. The system will be trained using a generative adversarial network a-like setup (Goodfellow et al., 2014). The Autoencoder's objective will be twofold: to accurately compress and reconstruct the original Earth Science data, while simultaneously maximizing the error of the attacker model that attempts to reconstruct the data from the latent representation. Through this adversarial training process (Ganin et al., 2016), the Autoencoder will learn to generate a latent space that is optimized for utility in authorized downstream tasks but is inherently obfuscated and resilient to unauthorized reconstruction, thus providing robust privacy guarantees (Abadi et al., 2016).

2.2 Fine-Grained Data Utilization

The coarse-grained utilization of data, characterized by high redundancy and feature entanglement, is a major impediment to the efficiency and performance of current AI models in Earth Science (Locatello et al., 2019). My research will address this by designing latent spaces that are explicitly structured for fine-grained analysis.

2.2.1 High Information-Density Latent Space

To combat the high degree of information redundancy in geoscientific data, I propose to construct a **High Information-Density Latent Space**. This will be achieved by reducing redundancy along both the spatial-temporal-channel dimensions and the sample dimension.

- *Spatio-Temporal-Channel Dimension:* Geoscientific data exhibits strong correlations due to spatial proximity, temporal continuity, and spectral similarity. I will design a novel Autoencoder architecture that disentangles these redundancies while preserving the independence of the temporal and channel dimensions for targeted analysis. Specifically, after the encoder maps the input data to a latent representation, this representation will be decomposed into two components: a single *common feature vector* that captures the shared information across time and channels, and multiple *specific feature vectors* that capture the differential characteristics unique to each time step and channel. During decoding, the model will combine the common feature vector with a selected specific feature vector to reconstruct the data for the corresponding time and channel. This architecture will drastically reduce information redundancy while maintaining a structured and dimension independent latent space that allows for the selective use of data from specific times and channels.
- *Sample Dimension:* During the model training process, not all data samples contribute equally to learning. Many samples may be redundant or of low value. I will develop a dynamic sample weighting scheme to reduce redundancy along the sample dimension. This will involve evaluating the contribution of each sample to the model's capability, using metrics such as sample loss or prediction uncertainty. Based on these metrics, the training process will dynamically down-weight or even discard redundant, low-value samples. This approach, related to curriculum learning (Bengio et al., 2009) and active learning (Settles, 2009), will focus the model's learning capacity on the most informative data, significantly improving training efficiency while maintaining or even enhancing final model performance.

2.2.2 Feature-Disentangled Latent Space

To address the challenge of highly coupled features in Earth Science data, I propose the construction of a **Feature-Disentangled Latent Space**. This space will be designed to decouple features at different levels of abstraction, leading to more efficient and effective model training. This will be accomplished through two primary strategies:

- *High-Level and Low-Level Feature Disentanglement:* I will design a composite loss function for the Autoencoder that encourages the separation of high-level and low-level features. In addition to the standard data reconstruction loss which preserves low-level fidelity, I will incorporate a semantic loss by leveraging the representational power of pre-trained vision foundation models like DINOv2 (Oquab et al., 2023) or vision-language models like CLIP (Radford et al., 2021). By regularizing the latent space to align with the semantic space of these powerful foundation models, the Autoencoder will be forced to learn a representation where high-level concepts (Chen et al., 2016) are disentangled from pixel-level details. This will enable the model to perform a complete range of downstream tasks, from fine-grained analysis to high-level classification, with greatly improved efficiency and performance (Yao et al., 2025; Xu et al., 2025; Ma et al., 2025).

- *Preservation of Data Priors:* Raw scientific data often possesses inherent prior properties, such as the equivariance of image semantics to spatial transformations (Cohen and Welling, 2016). To ensure these priors are preserved in the latent space, I will employ a self-supervised training objective. For a given transformation, such as image rotation, the transformation will be applied to both the original data and its latent representation (Chen et al., 2020). The model will then be trained to ensure that decoding the transformed latent representation yields the transformed original data. By enforcing consistency under such transformations (He et al., 2020), the model is compelled to learn a latent space that respects the intrinsic symmetries of the data, leading to more robust and generalizable representations (Kouzelis et al., 2025).

2.3 Low-Cost Data Production

The high costs associated with data acquisition and preprocessing represent a fundamental bottleneck in Earth Science research. My proposed data-centric paradigm offers a path to mitigate these costs by leveraging the efficiency of the latent space.

2.3.1 Low-Cost Data Generation

To address the high cost of data acquisition from multiple sensor platforms, I propose to build a **Unified Cross-Modal Generative Model**. This model will leverage the inherent correlations between different types of observational data (e.g., optical, thermal, radar) and their efficient representations in the latent space.

The framework will involve constructing a unified transformation model that operates entirely within the shared latent space learned from multiple data modalities. This model will be trained to translate the latent representation of one data type into the latent representation of another, similar to unsupervised image-to-image translation techniques (Zhu et al., 2017a; Isola et al., 2017). Consequently, from a small set of observations—or even a single observation type—at a specific time and location, the model will be able to generate a complete suite of high-quality, physically consistent, and related observational data. This will enable low-cost data acquisition and a dramatic expansion of spatial-temporal data coverage, effectively "filling in the gaps" in our observational record using modern generative architectures (Ho et al., 2020; Ramesh et al., 2021).

2.3.2 Efficient Data Preprocessing

To tackle the high cost and complexity of traditional data preprocessing workflows, I propose to construct a **Unified Preprocessing Model** for raw sensor data. This model will leverage the efficient representations of both raw and preprocessed data within the latent space.

The core idea is to frame data preprocessing as a translation task. A unified model will be trained to map the latent representation of raw sensor data directly to the latent representation of its corresponding high-quality, analysis-ready counterpart. By learning this transformation in the compact and information-rich latent space, the model can bypass the complex, multi-step, and often heuristic-driven procedures of traditional preprocessing pipelines (Rombach et al., 2022). This approach will enable the rapid and efficient preprocessing of a wide range of Earth Science data, dramatically lowering the barrier to entry for utilizing raw sensor data and accelerating the pace of scientific discovery (Allen et al., 2025).

3 Expected Results

The proposed data-centric paradigm is expected to optimize Earth Science data across multiple dimensions, culminating in the establishment of a holistic foundational data platform. This platform will be composed of three integrated components: (1) a **Data Management Platform** characterized by low storage overhead and high data security; (2) a **Data Utilization Platform** defined by low information redundancy and minimal feature entanglement; and (3) a **Data Production Platform** that enables low-cost data generation and preprocessing, as shown in Figure 1. The management platform will empower researchers to acquire, store, and utilize geoscientific data cost-effectively and securely. The utilization platform will enhance the training efficiency and ultimate performance of deep learning models. The production platform will facilitate scientific inquiry across vast spatio-temporal domains covered by both real and generated data. Collectively, these components will establish a complete, closed-loop ecosystem for AI for Earth Science, encompassing data production, management, and utilization.

3.1 High-Efficiency Earth Science Data Management Platform

The high acquisition cost of public data and the secure handling of private data are two persistent challenges in the Earth Sciences (Guo, 2017). The proposed Latent Space Research Paradigm and Secure Latent Space Representation, developed under the data-centric framework, are expected to enable the construction of a **High-Efficiency Earth Science Data Management Platform**. This platform will allow large-scale, data-intensive deep learning research to be conducted entirely within a compact and secure latent space. This will drastically reduce data storage and computational costs while mitigating the security risks associated with data leakage, thereby attracting a broad community of researchers to conduct novel Earth Science research using latent representations and accelerating the advancement of the AI for Earth Science field.

Specifically, large-scale public datasets and proprietary datasets (made available through agreements) from various Earth Science domains will be transformed into their latent space counterparts and hosted on this platform. This will dramatically lower the costs of data acquisition, storage, and utilization while enhancing data security and reliability. For instance, it is projected that the 20 PB ERA5 dataset (Hersbach et al., 2020) in meteorology could be converted into a 40 TB latent representation; the 50 PB Sentinel-2 dataset (Drusch et al., 2012) in remote sensing could be reduced to a 70 TB representation; and the 50 TB Ocean Reanalysis System 5 dataset (Copernicus Climate Change Service, 2021) in oceanography could be compressed to a 250 GB representation. As all subsequent data analysis will be performed directly in the latent space, this approach will not only achieve significant cost reductions but also inherently enhance data security.

3.2 Fine-Grained Earth Science Data Utilization Platform

Information redundancy and feature entanglement are pervasive issues that impede effective data utilization in Earth Science (Runge et al., 2019). The specialized Autoencoder models developed within the data-centric paradigm, which are designed to significantly reduce these inefficiencies, will form the basis of a **Fine-Grained Earth Science Data Utilization Platform**. This platform will provide the research community with access to these domain-specific Autoencoders and the highly efficient latent representations derived from their application to large-scale datasets. Consequently, when training deep learning models with this optimized data, researchers are expected to achieve substantial improvements in both training efficiency and final model performance. This will facilitate the deployment and practical application of advanced models across various domains, unlocking the scientific and applied value of deep learning within the AI for Earth Science field.

3.3 Low-Cost Earth Science Data Production Platform

The high costs of data acquisition and preprocessing have long been primary constraints on data availability, limiting the scope of scientific research in Earth Science (Bergen et al., 2019). The Unified Cross-Modal Generative Model and the Unified Preprocessing Model, integral to the proposed data-centric paradigm, will enable the creation of a **Low-Cost Earth Science Data Production Platform**. This platform will offer the research community access to these powerful models and the vast quantities of generated data they produce. This synthetic data, when integrated with observational data, will achieve unprecedented spatio-temporal coverage of Earth's various spheres. As a result, for any given time and location, the possession of even a single modality of observational data will be sufficient to generate a complete, multi-modal suite of observations at low cost. This will profoundly enhance the availability and coverage of Earth Science data, which in turn will support a deeper understanding of extreme events and complex physical structures. By fundamentally advancing the field at the data level, this platform is expected to drive significant scientific discoveries and applied progress in AI for Earth Science.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

Anna Allen, Stratis Markou, Will Tebbutt, James Requeima, Wessel P Bruinsma, Tom R Andersson, Michael Herzog, Nicholas D Lane, Matthew Chantry, J Scott Hosking, et al. End-to-end data-driven weather prediction. *Nature*, pages 1–8, 2025.

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*, 2018.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Karianne J Bergen, Paul A Johnson, Maarten V de Hoop, and Gregory C Beroza. Machine learning for data-driven discovery in solid earth geoscience. *Science*, 363(6433):eaau0323, 2019.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016.

Climate Data Store Copernicus Climate Change Service. Oras5 global ocean reanalysis monthly data from 1958 to present. *Copernic. Clim. Change Serv.(C3S) Clim. Data Store (CDS)*, 2021.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017.

Matthias Drusch, Umberto Del Bello, Sébastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote sensing of Environment*, 120:25–36, 2012.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Huadong Guo. Big earth data: A new frontier in earth and information sciences. *Big Earth Data*, 1(1-2):4–20, 2017.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049, 2020.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

Briland Hitaj, Giuseppe Ateniese, and Fernando Perez-Cruz. Deep models under the gan: information leakage from collaborative deep learning. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pages 603–618, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Chris Huntingford, Elizabeth S Jeffers, Michael B Bonsall, Hannah M Christensen, Thomas Lees, and Hui Yang. Machine learning and artificial intelligence to aid climate change research and preparedness. *Environmental Research Letters*, 14(12):124007, 2019.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

Theodoros Kouzelis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. Eq-vae: Equivariance regularized latent space for improved generative image modeling. *arXiv preprint arXiv:2502.09509*, 2025.

Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al. Geo-bench: Toward foundation models for earth monitoring. *Advances in Neural Information Processing Systems*, 36:51080–51093, 2023.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding. *arXiv preprint arXiv:2502.20321*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and F Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743): 195–204, 2019.

David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)*, 55(2):1–96, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):2553, 2019.

Burr Settles. Active learning literature survey. 2009.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W Mathis, Frank Van Langevelde, Tilo Burghardt, et al. Perspectives in machine learning for wildlife conservation. *Nature communications*, 13(1):792, 2022.

Wanghan Xu, Xiaoyu Yue, Zidong Wang, Yao Teng, Wenlong Zhang, Xihui Liu, Luping Zhou, Wanli Ouyang, and Lei Bai. Exploring representation-aligned latent space for better generation. *arXiv preprint arXiv:2502.00359*, 2025.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15703–15712, 2025.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey. *ACM Computing Surveys*, 57(5):1–42, 2025.

Sijie Zhao, Feng Liu, Xueliang Zhang, Hao Chen, Tao Han, Junchao Gong, Ran Tao, Pengfeng Xiao, Lei Bai, and Wanli Ouyang. Transforming weather data from pixel to latent space. *arXiv preprint arXiv:2503.06623*, 2025.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017a.

Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience and remote sensing magazine*, 5(4):8–36, 2017b.