NASA'S AI FOUNDATION MODELS FOR SCIENCE: CURRENT INITIATIVES, WORKFLOW, ROADMAP AND LESSONS LEARNED

Rahul Ramachandran
Science Research and Projects Division (ST10)
NASA Marshall Space Flight Center
rahul.ramachandran@nasa.gov

Tsengdar Lee
Earth Science Division
NASA Headquarters
tsengdar.j.lee@nasa.gov

Kevin Murphy
Office of Chief Science Data Officer
NASA Headquarters
kevin.j.murphy@nasa.gov

***Abstract*—NASA's scientific data archives have grown rapidly, surpassing 150 petabytes and expected to exceed 500 petabytes by 2029. This expansion presents considerable challenges in data management and analysis. To address these, the Office of the Chief Science Data Officer (OCSDO) has implemented a strategy centered on AI foundation models (FMs) to enhance workflows across NASA's five science divisions. These models, trained with self-supervised pretraining, create versatile, application-agnostic representations that can be adapted efficiently to specific tasks. The OCSDO's "5+1" strategy integrates tailored FMs for each science division with a large language model for cross-domain tasks. Notable initiatives include the INDUS language model suite covering all science divisions, the Prithvi-HLS model for optical remote sensing, and the Prithvi-WxC model for atmospheric analysis. These models reduce computational demands and data labeling needs while performing well on existing benchmarks. Current efforts focus on new models for heliophysics (Surya-SDO), lunar studies, and biological research, undertaken with new partnerships. The FM development process follows science standards prioritizing transparency, accuracy, and relevance. Key workflows encompass pretraining, adaptation, and inference. Lessons learned from this work emphasize the value of balancing costs and benefits, fostering interdisciplinary collaboration, designing science use case-driven models, and maintaining rigorous validation. This paper details the design, implementation, and roadmap of NASA's FMs, illustrating their potential role in advancing scientific discovery through AI-powered methodologies.**

***Index Terms*—foundation models, artificial intelligence, open science, large language model.**

I. INTRODUCTION

NASA's scientific data archives are large and expanding at a rapid pace. Currently, the archive exceeds 150 petabytes and is projected to surpass 500 petabytes by fiscal year 2029. It holds over 1.5 billion files, with an annual data ingestion rate of 50 petabytes. This large scientific repository will continue to grow even after the instruments that originally collected the data cease operations. Managing and analyzing this huge volume of data presents considerable challenges.

To address these challenges, NASA is adopting new strategies to manage its datasets efficiently and to extract valuable scientific insights. The Office of the Chief Science Data Officer (OCSDO) is leading this effort by improving research methodologies and reducing barriers for users to effectively utilize complex scientific data. Specifically, the OCSDO has implemented a comprehensive strategy to leverage artificial intelligence (AI) for advancing scientific research. This strategy has four components. First, establishing a scalable infrastructure to support the development and deployment of AI models within NASA's operational framework. Second, promoting interdisciplinary collaboration by integrating and analyzing data from diverse sources and missions, while also fostering AI literacy across NASA's Science Mission Directorate. Third, enhancing the efficiency of data-related tasks such as search, discovery, analysis, and visualization. And finally, building predictive models and simulations to address varied scientific challenges and support researchers in drawing meaningful insights.

The OCSDO is actively exploring the potential of AI foundation models [1] to augment current scientific analysis workflows. Unlike traditional AI models, which are constrained by labeled data availability and lack of model generalizability, Foundation models use self-supervised learning during pretraining. This approach allows them to be fine-tuned for a wide range of applications, significantly reducing the effort required for downstream tasks.

To demonstrate their utility, OCSDO has developed a "5+1" strategy. This involves collaboration with NASA's five science divisions to create tailored foundation models for each division. Additionally, a large language model will serve as a cross-disciplinary tool applicable across all divisions. OCSDO works with science teams from different divisions while providing expertise in AI, systems engineering, infrastructure, and training to ensure the success of these initiatives.

The development of these foundation models follows the same rigorous standards applied to all NASA data products and algorithms. OCSDO's design and development process emphasizes transparency, accuracy, and reliability. This paper outlines our approach used in designing and implementing the foundation models, highlights current initiatives and released models, and presents a roadmap for future development. It also examines the obstacles encountered during this endeavor and the strategies employed to overcome them.

II. Background on AI Foundation Models

AI foundation models (FMs) are large-scale, pre-trained models designed to learn versatile, general-purpose representations of data [2]. These representations enable FMs to perform a wide range of tasks across domains with minimal fine-tuning. Through pretraining on vast datasets and pretext tasks, FMs develop an understanding of the relationships, structures, and latent features embedded in the data.

Foundation models map input data (x) to a feature representation (z) learning a function F. This representation captures essential patterns and relationships within the data, which can then be fine-tuned to a task-specific function $G(z)$ to produce outputs such as classifications or predictions. An ideal FM can generalize to unseen tasks without additional adjustments as zero-shot generalization [2]. Most real-world FMs still require some degree of fine-tuning to adapt to specific downstream applications.

Pretext tasks play a critical role in FM training. These artificial, unlabeled tasks teach the model to extract meaningful features from raw data. Examples include masked language modeling (predicting missing words in text), image inpainting (reconstructing missing parts of an image), and contrastive learning (identifying similarities and differences between data points). This pretraining process allows FMs to develop generalized representations that are applicable across diverse domains [2].

One of the strengths of FMs lies in their ability to perform effective representation learning. They automatically identify and extract meaningful features—often called latent factors—from data, which reduces reliance on manual feature engineering [3]. These representations are smooth (changing gradually with input variations), hierarchical (breaking down complex concepts into simpler ones), and sparse (focusing only on the most relevant features) [3]. Furthermore, FMs often use manifold learning to simplify high-dimensional data by mapping it onto lower-dimensional spaces, preserving key relationships while reducing complexity [3]. For instance, a model might map images of faces onto a manifold where dimensions represent features like expression, lighting, and angle.

AI FMs offer significant advantages in their ability to support multimodal learning, domain transfer, and task-specific adaptability. They align and integrate relationships across different data types, such as linking radar data (e.g., Sentinel-1) to optical data (e.g., Sentinel-2) using shared feature representations [2]. Additionally, FMs enable domain transfer by applying structural insights from one domain to another; for instance, a model trained on optical satellite imagery from one instrument can assist in analyzing imagery from another instrument as long as their underlying instrument characteristics are similar.

FMs are adaptable, allowing their learned representations to be fine-tuned and reshaped for specific tasks, optimizing their performance for diverse applications. From a mathematical perspective, foundation models transform raw input data into meaningful representations through a function $F(x) = z$. These representations can then be fine-tuned for downstream tasks using a task-specific function $G(z) = y$. This is a two-step process—general representation learning followed by task-specific adaptation. By learning general-purpose representations from diverse datasets, they simplify workflows, enable cross-domain applications, and support a wide range of downstream science tasks.

III. Workflow

The workflow for developing and applying AI FMs for science involves three key phases: pretraining, adaptation, and inference. Each phase has distinct objectives, characteristics, and challenges, requiring careful planning and collaboration to achieve optimal results.

In the pretraining phase, the goal is to create a general-purpose AI model for a science domain by training it on a massive dataset of scientific data using self-supervised learning techniques. This phase is computationally intensive, requiring access to high-performance computing clusters and significant technical expertise in AI/ML, the scientific domain, and data engineering. Pretraining is typically a collaborative effort, involving multiple stakeholders to ensure that the model addresses the diverse needs of the scientific community. Success in this phase relies on access to large, high-quality scientific datasets and the ability to preprocess and utilize them effectively. Additionally, designing the model requires careful consideration of architecture, hyperparameters, and training objectives to ensure generalizability and value for downstream tasks. This phase also involves balancing the cost and effort of FM development against its potential downstream benefits. Open collaboration among different researchers, institutions, and organizations is crucial to maximizing the effectiveness of these models. Additionally, rigorous validation and evaluation procedures are necessary to ensure the quality, reliability, and relevance of the FMs for their intended applications.

The adaptation phase focuses on tailoring the pretrained FM to specific scientific research or application tasks. Compared to pretraining, this phase requires less computational power and small labeled datasets. It is user-focused, targeting individual researchers or data practitioners and using different approaches such as fine-tuning, few-shot learning, or zero-shot learning.

Finally, in the inference phase, the adapted FMs are deployed for real-world scientific research applications. This phase emphasizes the importance of providing user-friendly tools and services, enabling researchers or decision makers to interact with and benefit from the models.

All three phases must be considered if FMs for science are to be developed and deployed to advance scientific research effectively and responsibly.

IV. Current Initiatives

NASA is actively developing advanced AI models tailored to its scientific needs, with notable initiatives including INDUS, Prithvi-HLS, and Prithvi-WxC.

INDUS is a suite of encoder language models developed in collaboration with IBM Research, specifically designed for NASA's scientific domains, covering Earth science, biological and physical science, heliophysics, planetary science, and astrophysics [4]. Designed for NASA's specialized tasks, INDUS addresses the shortcomings of general-purpose language models, such as RoBERTa, and domain-specific models like SCIBERT. The suite includes several models: INDUSBASE, a masked language model built on the RoBERTa BASE architecture and trained on the NASA Science Mission Directorate (SMD) corpus. INDUSSMALL, an efficient version of INDUSBASE created using knowledge distillation. INDUS-RETRIEVERBASE,is fine-tuned for dense retrieval with contrastive learning and INDUS-RETRIEVERSMALL, a distilled retriever model optimized for speed and latency. To evaluate these models, NASA also developed benchmarks for named entity recognition (NER), question answering (QA), and information retrieval (IR), including a climate-focused NER task (CLIMATE-CHANGE NER), NASA-QA for extractive QA in Earth sciences, and NASA-IR, an IR dataset spanning multiple domains. Our evaluations show that INDUS models outperformed general-purpose models on these benchmarks, with knowledge-distilled versions offering faster inference without compromising accuracy. Applications of INDUS include its integration into NASA's Science Discovery Engine (SDE) for improved document retrieval and search and automated tagging and classification of scientific data using fine-tuned models.

Prithvi-HLS was the first FM built using NASA science data. With two iterations to date, Prithvi-HLS aims at optimizing optical remote sensing tasks related to satellite imagery. Prithvi-HLS v1.0 is NASA's first foundation model built on the Harmonized Landsat Sentinel (HLS) dataset, a 100-million-parameter model trained on one year of HLS data over the continental United States (CONUS) [5]. It demonstrated the advantages of a FM in reducing computational demands and labeled data requirements for downstream tasks. Prithvi-HLS v2.0, an improved version developed in response to user feedback, retained the masked autoencoders (MAE) architecture with a vision transformer (ViT) backbone while incorporating temporal and location metadata encoded as weighted sums [6]. Trained on NASA's HLS dataset (2015-2024) with an improved sampling strategy, this model was pretrained on a Jülich Supercomputing Centre system, requiring approximately 23,000 GPU hours for the 300-million-parameter version and 58,000 GPU hours for the 600-million-parameter version. Prithvi-HLS v2.0 has been benchmarked on the GEO-Bench evaluation framework and has outperformed six leading models and its predecessor, Prithvi-EO-1.0, particularly for medium-resolution tasks [6].

Prithvi-WxC is a 2.3-billion-parameter model trained on NASA's MERRA-2 reanalysis dataset. It has been designed to capture atmospheric states across multiple variables and scales. The model is pretrained for forecasting and masked reconstruction tasks [7]. Prithvi-WxC has been used for applications such as downscaling, parameterization, and zero-shot prediction, making it a useful research tool for advancing atmospheric research [8].

Through these initiatives, NASA is seeking to increase the use of AI to enhance scientific research by addressing the challenges in data analysis, interpretation, and application across various domains.

V. Roadmap and Lessons Learned

Our roadmap for FM development for science has several key initiatives and milestones across diverse scientific domains.

The new version of Prithvi WxC V2 will focus on data assimilation to enhance its capabilities to support atmospheric modeling workflows. The updated version is set for release in June 2025, incorporating improvements to better support weather forecasting and scientific research needs. In addition, to encourage NASA's research and application communities to use these FMs in their work and to support NASA's new Earth Science to Action strategy, the Earth Science Division has developed additional research and application solicitations to further enhance these FMs and to build applications and tools leveraging these FMs. These announcements are available in NASA's Research Opportunities in Space and Earth Science (ROSES 2025).

Surya-SDO, a heliophysics foundation model, uses the Solar Dynamics Observatory (SDO) data to support solar corona and magnetic field analysis [9]. This model targets applications such as active region detection, solar wind forecasting, and coronal hole predictions. Surya-SDO will scale to 7–21 billion parameters using computational resources provided by the NSF NAIRR. The first version (V1) is scheduled for release in June 2025, with a community workshop planned for later in the year to engage researchers to evaluate the model and gather feedback.

For planetary sciences, a Lunar FM development effort will be launched in early 2025, with a V1 release targeted for October 2025. This initiative aims to support specialized applications related to lunar exploration and research.

For biological and physical sciences, the development of a Biological and Physical Sciences FM is tentatively planned to begin in fall 2025. This model will focus on addressing the unique challenges and opportunities in this field.

NASA is also strengthening its collaboration with the European Space Agency (ESA) on the joint development of AI FMs. The NASA science team is supporting an ongoing ESA-funded multimodal model which will be tentatively released in September 2025. Additionally, a joint NASA-ESA AI workshop for Earth observation is scheduled from May 5–7, 2025, in Italy. This workshop will emphasize fostering international cooperation in developing open FM models to drive AI innovation in Earth sciences.

The development of these AI FMs over the last two years have unearthed some lessons for maximizing their scientific value and usefulness. It is important to balance cost with scientific impact. Efforts should focus on FMs that offer the greatest potential to drive meaningful advances in research and discovery.

Collaborations grounded in open science principles have emerged as the critical ingredient for success. Cross-institutional partnerships expand resources and expertise, enabling broader and more impactful applications of FMs. Equally important is building interdisciplinary teams that integrate AI specialists, data scientists, and domain experts. This diverse expertise ensures that models address both technical and scientific challenges effectively.

Science-driven design is another essential principle. Grounding model development in a wide variety of science use cases ensures that the resulting tools are practical and address important research needs. Alongside this, a deep understanding of complex and nuanced scientific data as well as underlying scientific principles is crucial. Having science domain expertise enables the team to handle data and domain intricacies effectively—from preprocessing to model design and training.

Finally, rigorous validation is critical for ensuring model trust worthiness. Robust evaluation frameworks help build confidence in the models, ensuring they meet the standards necessary for scientific research.

These lessons should form the basis for any future FM initiatives, emphasizing strategic focus and collaboration to maximize their impact in advancing science.

REFERENCES

[1] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang, "On the Opportunities and Risks of Foundation Models," 2021. [Online]. Available: https://arxiv.org/abs/2108.07258

[2] Y. Yuan, "On the Power of Foundation Models," 2022. [Online]. Available: https://arxiv.org/abs/2211.16327

[3] Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," 2012. [Online]. Available: https://arxiv.org/abs/1206.5538

[4] B. Bhattacharjee, A. Trivedi, M. Muraoka, M. Ramasubramanian, T. Udagawa, I. Gurung, N. Pantha, R. Zhang, B. Dandala, R. Ramachandran, M. Maskey, K. Bugbee, M. Little, E. Fancher, I. Gerasimov, A. Mehrabian, L. Sanders, S. Costes, S. Blanco-Cuaresma, K. Lockhart, T. Allen, F. Grezes, M. Ansdell, A. Accomazzi, Y. El-Kurdi, D. Wertheimer, B. Pfitzmann, C. B. Ramis, M. Dolfi, R. T. de Lima, P. Vagenas, S. K. Mukkavilli, P. Staar, S. Vahidinia, R. McGranaghan, and T. Lee, "INDUS: Effective and Efficient Language Models for Scientific Applications," 2024. [Online]. Available: https://arxiv.org/abs/2405.10725

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

[6] D. Szwarcman, S. Roy, P. Fraccaro, E. Gíslason, B. Blumenstiel, R. Ghosal, P. H. de Oliveira, J. L. d. S. Almeida, R. Sedona, Y. Kang, S. Chakraborty, S. Wang, A. Kumar, M. Truong, D. Godwin, H. Lee, C.-Y. Hsu, A. A. Asanjan, B. Mujeci, T. Keenan, P. Arevalo, W. Li, H. Alemohammad, P. Olofsson, C. Hain, R. Kennedy, B. Zadrozny, G. Cavallaro, C. Watson, M. Maskey, R. Ramachandran, and J. B. Moreno, "Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications," 2024. [Online]. Available: https://arxiv.org/abs/2412.02732

[7] J. Schmude, S. Roy, W. Trojak, J. Jakubik, D. S. Civitarese, S. Singh, J. Kuehnert, K. Ankur, A. Gupta, C. E. Phillips, R. Kienzler, D. Szwarcman, V. Gaur, R. Shinde, R. Lal, A. Da Silva, J. L. G. Diaz, A. Jones, S. Pfreundschuh, A. Lin, A. Sheshadri, U. Nair, V. Anantharaj, H. Hamann, C. Watson, M. Maskey, T. J. Lee, J. B. Moreno, and R. Ramachandran, "Prithvi WxC: Foundation Model for Weather and Climate," 2024. [Online]. Available: https://arxiv.org/abs/2409.13598

[8] S. K. Mukkavilli, D. S. Civitarese, J. Schmude, J. Jakubik, A. Jones, N. Nguyen, C. Phillips, S. Roy, S. Singh, C. Watson, R. Ganti, H. Hamann, U. Nair, R. Ramachandran, and K. Weldemariam, "AI Foundation Models for Weather and Climate: Applications, Design, and Implementation," 2023. [Online]. Available: https://arxiv.org/abs/2309.10808

[9] S. Roy, T. Singh, M. Freitag, J. Schmude, R. Lal, D. Hegde, S. Ranjan, A. Lin, V. Gaur, E. E. Vos, R. Ghosal, B. N. Patro, B. Aydin, N. Pogorelov, J. B. Moreno, M. Maskey, and R. Ramachandran, "AI Foundation Model for Heliophysics: Applications, Design, and Implementation," 2024. [Online]. Available: https://arxiv.org/abs/2410.10841