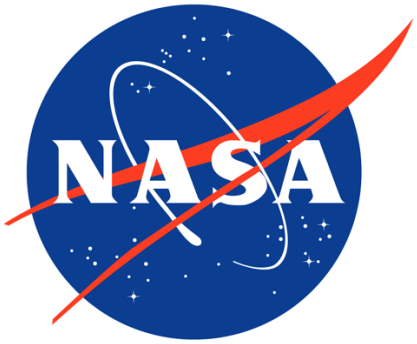

NASA Science Mission Directorate
Data and Computing Architecture Study

Final Report

August 2024

Executive Summary

The Data and Computing Architecture Study aimed to determine whether a coordinated, hybrid cloud-High End Computing (HEC) infrastructure would benefit NASA's Science Mission Directorate (SMD) and its research community in transitioning to Open-Source Science. Use cases and user needs collected via open workshops, a Request for Information (RFI), and Technical Interchange Meetings (TIMs) identified opportunities to fill gaps in High-End Computing (HEC) services, improve efficiency, and develop infrastructure that enables sharing of data and scientific results between SMD-funded collaborators.

This study identified numerous capabilities and services that would be effective in enabling an open scientific process at NASA, including capabilities that would be highly useful across SMD Divisions. In particular, the study identified a strong need for increased cloud computing and HEC services and support. It also identified capabilities that have the potential to reduce duplication within SMD and support Divisions in the implementation of modern scientific collaboration tools and facilitation of interoperability among repositories. A hybrid, centrally managed HEC/cloud architecture that includes common services will benefit all divisions and serve SMD's scientific goals by enabling an open and efficient scientific process.

This study identified programmatic structures including organizational realignments that will support a sustainable, secure, and cost-effective scientific data and computing architecture that fosters innovation and collaboration across the scientific community. These structures included diverse sustainability models (i.e. public-private partnerships, fee-for-service models, and grant funding) to ensure ongoing financial support for this infrastructure.

Findings and recommendations were presented at the November 9, 2023 SMD Science Management Council meeting and received unanimous concurrence from SMD Divisions' leadership.

Study recommendations were organized into five areas that address a Reference Architecture and programmatic aspects:

1. **Open-Source Science Infrastructure**
   Addresses the need to efficiently access/combine data from multiple repositories, leverage modern scientific analysis and collaboration tools, and easily utilize data in cloud environments and high-end computing facilities.

2. **Infrastructure Common Services**
   Addresses specific "baseline" collaboration tools, cloud services, and data services that would be of use to all SMD Divisions.

3. **Improving Efficiency and Access to Computing**
   Addresses need for developers, operators, and users to have flexibility, ease of entry, and surge capacity within data and computing systems.

4. **Managing Cybersecurity Risks (Programmatic)**
   Addresses implementation of cybersecurity across SMD's data and computing resources.

5. **Long-term Sustainability (Programmatic)**
   Addresses organizational framework required to most effectively capitalize on existing and future scientific data and computing resources, including management of cloud resources.

Efficient data and computing capabilities that allow widespread collaboration will be critical to NASA's future groundbreaking discoveries. In response to these findings, it is recommended the Office of the Chief Science Data Officer (OCSDO) should facilitate the development of a core architecture that organizes general Core Services across SMD Divisions, with the goal of preserving the data-management autonomy of each division. This framework will increase researchers' access to computing capabilities, and improve data access, distribution, and sharing.

Background

The collective data repositories across SMD's five science Divisions are expected to expand swiftly due to upcoming missions and the execution of new models. Access to the computational power necessary to analyze these larger data volumes using modern methods will be critical to fulfilling NASA's scientific mission.

SMD Mission and Research Data Repository Storage Estimates

	Mission Data Storage (PB)							Research Data Storage (PB/yr)		
	FY23	FY24	FY25	FY26	FY27	FY28	FY29	Analysis Data*	Model Data**	Total
ESD	37.13	80.19	113.46	147.17	179.38	210.23	243.16	1		
PSD	8.05	10.05	12.05	14.05	16.05	18.05	20.05	0.24	1.6	1.84
BPS (OSDR)	0.24	0.34	0.43	0.52	0.61	0.70	0.79	0.1		
APD	12.81	16.24	21.29	24.90	39.01	39.01	39.01	0.2		
HPD	5.00	15.00	20.00	25.00	27.00	30.00	35.00	.02		
Total	**63.23**	**121.82**	**167.23**	**211.64**	**262.05**	**297.99**	**338.01**	**~2 PB**	**~8-10 PB**	**~12 PB**

***Total analysis data is expected to be ~1 PB.**
- Stored at existing NASA services (NexSci, genelab, etc),
- Community services (Zenodo, OSF, other agency repositories)

****Total model data is expected to be 8-10 PB.**
- Some stored at HEC (e.g. 1 Exobyte tape storage space)

Table 1: Projected data storage estimates for SMD Science Divisions.

The Data and Computing Architecture Study was chartered in June 2022 to answer the question: Can a coordinated hybrid cloud-High End Computing (HEC) infrastructure meet the data and computing needs of SMD, enable efficiencies, and support SMD's transition to Open-Source Science? To answer this driving question, this study had two goals:

1. Identify scientific data and computing capabilities and architectures that enable Open Science, improve access to advanced computing and analytic services, allow open scientific collaboration, manage cybersecurity risks, and balance cost.

2. Identify opportunities that will ensure long-term evolution and sustainability of an SMD- wide data and computing infrastructure to enable Open-Source Science.

The study's scope included the data and computing systems supporting SMD-funded scientific workloads, including scientific modeling and simulation, data processing (L0-L4), data analytics, and Artificial Intelligence/Machine Learning (AI/ML).

Current State

Data and computing systems supporting SMD have traditionally embraced systems tailored for each Division or in many cases, individual missions/projects within Divisions. These uncoupled systems have allowed each division flexibility to support unique science requirements and user needs of their communities. Management of data and computing resources is typically based on the specific needs of each mission or Division, with limited consideration for enabling interdivisional research. Access to data on a particular theme or mission is typically through a specific repository; with 30+ science data repositories, it can be challenging to find data, especially if the user doesn't know specifically what they are looking for.

SMD currently stores over 90 PB of observational data and model results. Over the coming years, the opportunities for utilizing big data and conducting interdisciplinary science will grow, as will the challenges of making these data accessible and usable to all interested users. Each Division is responsible for their data and the repositories it is stored in; the data are currently maintained in a mix of on-premises and cloud storage, with each division maintaining its own pace of migration to the cloud. Demand for HEC services is high across all Divisions with a significant backlog, and demand is growing for the improved infrastructure required to facilitate data-intensive AI/ML work

- Some of today's cutting-edge Artificial Intelligence and Machine Learning (AI/ML) applications, which are likely to be critical to NASA science in the near future, require about five to ten thousand times more computing power than NASA is currently capable of delivering as an agency.
- NASA is able to fund and support small scientific research projects with its existing capabilities, but more computationally intensive AI/ML work on larger amounts of data requires collaboration with other government agencies (e.g. the Department of Energy).
- Data and observations from NASA missions (e.g. the James Webb Space Telescope, a $10 billion investment) will require more computational resources to analyze than are currently available at the agency or from collaborative agreements.

SMD divisions recognize the need to evolve the existing data and computing architecture to support modern computational approaches and methods, as well as make them compliant with the Scientific Information Policy for the Science Mission Directorate (SPD-41a)[1].

Approach

Study activities were guided by an all-civil servant steering team composed of representatives from OCSDO and all five SMD science divisions.

Steering Team Member	**Role/Affiliation**
Kevin Murphy (co-chair)	Chief Science Data Officer
Tsengdar Lee (co-chair)	High-End Computing
Samrawit Gebre	Biological and Physical Sciences Division
Katie Baynes	Earth Science Division
Cerese Albers	Earth Science Division
Roopesh Ojha	Astrophysics Division
Matthew McClure	Heliophysics Division
Megan Ansdell	Planetary Science Division

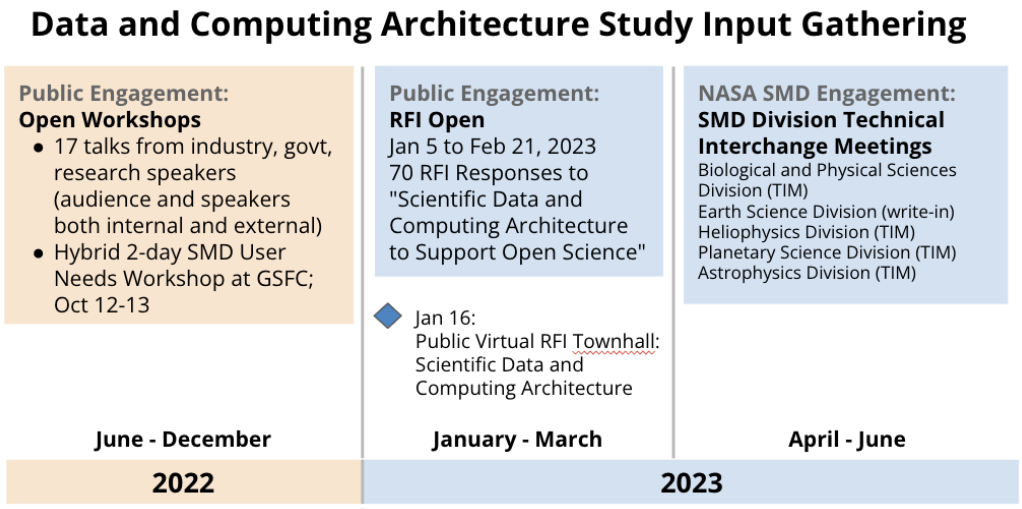

Figure 1: Public (external) and internal (NASA) engagement mechanisms for the Data and Computing Architecture Study.

[1] SPD-41a

Input gathering for the Data and Computing study included 17 open workshops with industry, research institutions, and government agencies, a Request for Information (RFI; Open January through February 2023)[2], and Technical Interchange Meetings (TIMs) with SMD Divisions that covered Cloud and HEC Architecture. Recent SMD studies on HEC[3] and Cloud Architecture were also considered. Study design emphasized a combination of broader outreach (RFI, open workshops) with smaller, highly targeted "deep-dives" with highly expert individuals from SMD Divisions (TIMs).

Publicly solicited input to this study (RFI, open workshops) was broad and sourced from a wide range of individuals and organizations. Input from the RFI (Figure 2) included 70 responses from NASA and a range of outside organizations (including other government agencies, universities, nonprofits, and private companies) that addressed topics related to NASA's scientific data and computing architecture, cloud infrastructure, High-Performance Computing, open-source software and tools, training, and user support. Participants in open workshops included individuals from NASA, other government agencies, and the general public; 279 individuals participated across the 17 open workshops held for the study. Average workshop attendance was 45 participants.

Inputs from the RFI, open workshops, SMD Division, previous NASA-chartered studies of scientific data and computing resources, and Technical Interchange Meetings (TIMs) to identify needs for scientific data and computing services. Aggregated study inputs were used to determine user and Division needs.

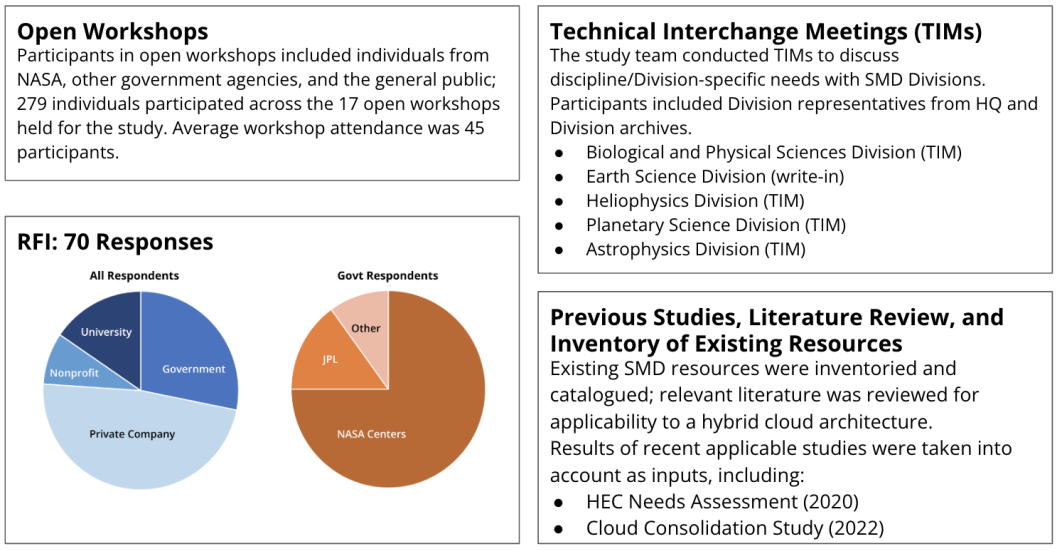

Figure 2: Data and Computing Architecture Study Inputs

These user and Division needs were framed in the context of a reference architecture for SMD scientific data and computing that was developed in response to inputs. Study inputs identified capabilities, needs, and gaps that are addressed by this reference architecture. Use Cases identified by the study provide context for user stories that can then be expanded into actionable implementation plans. Within each layer of the reference architecture (Figure 3), specific required services were identified that satisfied use cases

[2] Request for Information: Scientific Data and Computing Architecture to Support Open Science
[3] NASA High-End Computing Program User Needs Assessment 2020

from study inputs. Findings and recommendations reflect which of these services will be provided by SMD Divisions and which will be provided by OCSDO.

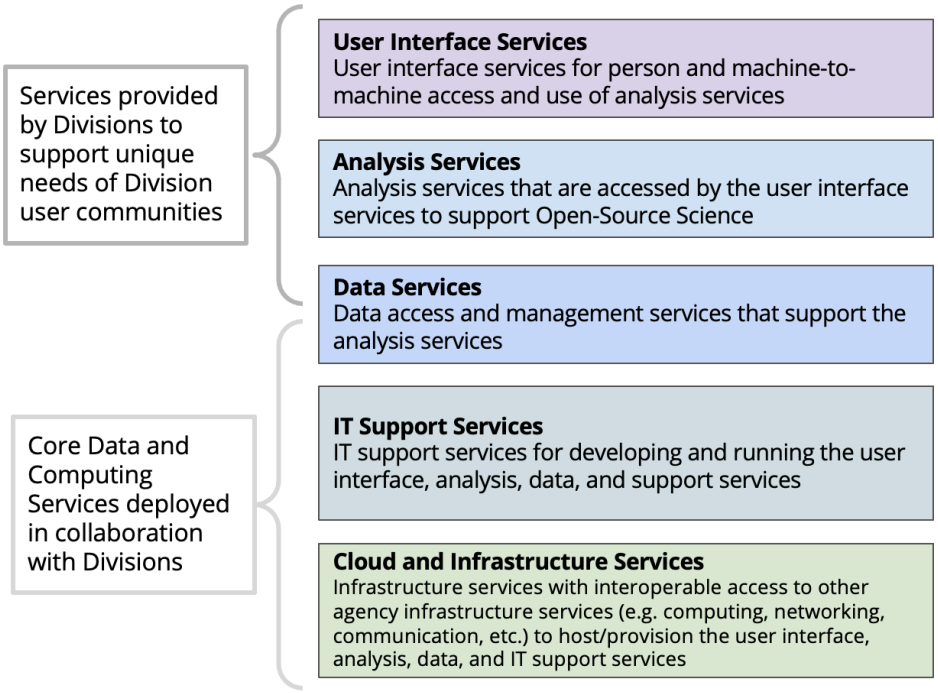

Figure 3: A reference scientific data and computing architecture developed in response to study inputs.

Findings and Recommendations

The Data and Computing Architecture Study resulted in five recommendations that address the Reference Architecture and programmatic aspects:

1. **Open-Source Science Infrastructure**
   Addresses the need to efficiently access/combine data from multiple repositories, leverage modern scientific analysis and collaboration tools, and easily utilize data in cloud environments and high-end computing facilities

2. **Infrastructure Common Services**
   Addresses specific “baseline” collaboration tools, cloud services, and data services that would be of use to all SMD Divisions

3. **Improving Efficiency and Access to Computing**
   Addresses need for developers, operators and users to have flexibility, ease of entry, and surge capacity within data and computing systems.

4. **Managing Cybersecurity Risks (Programmatic)**
   Addresses implementation of cybersecurity across SMD's data and computing resources

5. **Long-term Sustainability (Programmatic)**
   Addresses organizational framework required to most effectively capitalize on existing and future scientific data and computing resources, including management of cloud resources.

Area 1: Open-Source Science Infrastructure

Finding Summary

There is a strong need in SMD's user communities to efficiently access and combine data from multiple repositories (using a uniform method of access), leverage modern scientific analysis and collaboration tools, and easily utilize data in commercial cloud environments and high-end computing facilities. Heterogeneities in data structures, user communities, and scientific data and computing systems create high technical and procedural knowledge barriers to accessing existing SMD scientific data and computing services, which slows the pace of scientific discovery and the adoption of open scientific practices.

Recommendation Summary

Divisions should streamline access to scientific data and computing services to allow a broad array of internal and external users to easily access, use, combine, and share data sets, within and across disciplines. OCSDO should support Divisions in the development of services that enable use of modern scientific collaboration tools and facilitate governance, including adoption and use of standards, for interoperability and accessibility among repositories. When developing these capabilities, OCSDO and Divisions should prioritize support for high-value datasets and services that have the potential to enable interdisciplinary research and/or address high-priority scientific questions.

Recommendation 1.1: User Interface Services

OCSDO and Divisions should provide user interface services for person and machine-to-machine access and use of data analysis services.

Recommendation 1.1.1

Divisions should provide Data Visualization and Collaboration Services that provide a collaborative shared environment where multiple data sciences and analysts can work together with the same data, on the same analysis, to build and integrate data analytics in a collaborative real-time manner. These services should facilitate simple environments for data exploration, visualization, data/analytics sharing and real-time collaboration.

Recommendation 1.1.2

Divisions should provide Interactive Analytics and Visualization Services for the user to interactively perform different analytics in real-time and interactive real-time visualization for large data sets.

Recommendation 1.2: Analysis Services

OCSDO and Divisions should provide analysis services that are accessed by the user interface services to support Open-Source Science.

Recommendation 1.2.1

OCSDO and Divisions should provide Data Analytics and Data Science Platforms that consist of a tool set for mission engineers to absorb, organize, discover, and analyze data to reveal actionable insights that can help improve decision-making and inform business strategy.

Recommendation 1.2.2

OCSDO should provide Operational Machine Learning (MLOps) Platforms that provide data scientists with a specialized collaborative environment that facilitates iterative data exploration, real-time co-capabilities for experiment tracking, feature engineering, and model management, as well as controlled model transitioning, deployment, and monitoring. These platforms should provide support for tracking Machine Learning (ML) and generative AI experiments, reproducible and shareable AI pipelines or workflows, deploying AI for production, and multiple AI tools.

Recommendation 1.2.3

OCSDO should provide AI/ML Tools and Frameworks that include a set of core/common software features to assist in training and developing models.

Recommendation 1.2.4

OCSDO and Divisions should provide an Ontology Service that provides tools and processes for creating and augmenting domain-specific and hierarchical ontologies by different data domains. This service should be capable of visual navigation of ontology, information extraction facilities, collaborative development, support for standard industry domain and core vocabularies, and support for integrating with existing ontologies.

Recommendation 1.3: Data Services

OCSDO and Divisions should provide data access and management services that support the analysis services.

Recommendation 1.3.1

Divisions should provide Data Service Management Services, which encapsulate authorized operations on key enterprise data. These services provide a well-defined way of accessing data thru form and vetted APIs. Data Service Management is used to perform management activities such as creating/exposing data services and discovering and accessing data services.

Recommendation 1.3.2

Divisions should provide Data Access and Delivery Services that make data sets available for enterprise-wide access and use, subject to the defined access policy constraints.

Recommendation 1.3.3

Divisions should provide Enterprise Data Repositories that can consolidate multiple sources of data for ease of management, unified access, and a "one-stop shop" for analytical processing, exploration, and intelligence.

Recommendation 1.3.4

Divisions should provide Data Virtualization and Federation Services that place diverse physical database assets behind technology that provides a single logical interface to multiple data sets. Data virtualization integrates data from disparate sources, locations, and formats, without replicating the data, to create a single, virtual data layer that delivers unified data services to support multiple applications and users.

Recommendation 1.3.5

Divisions should provide Data Acquisition and Conditioning Services that handle data ingest and transformation, in order to populate data store(s) to be discoverable, accessible, and useful for the data consumer.

Recommendation 1.3.6

OCSDO and Divisions should provide a Data Protection Service that includes processes, services and methods used to accomplish privacy, safety, confidentiality, integrity, availability, and recovery of the enterprise datasets. This service secures data from intentional or unintentional unauthorized access and facilitates planning and executing the plan for data backup and disaster recovery.

Recommendation 1.3.7

OCSDO should provide Data Monitoring and Tracking services to track the generation, movement, and use of data throughout its lifecycle. These services facilitate data observability, metric collection; and evaluation.

Recommendation 1.3.8

OCSDO and Divisions should provide Data Discovery Services for authorized users to quickly, easily, and securely discover and access decision-quality data supporting any mission across any domain, in real-time and on-demand. This service provides a self-service metadata repository that allows any user to register, enrich, understand, discover, and consume datasets.

Recommendation 1.3.9

Divisions should provide Data Visualization and Business Intelligence Services. Business intelligence services perform analysis, reporting, data mining, predictive analysis, online analytical processing, and business performance management; data visualization services provide tools for analyzing, extracting data, and presenting the data in graphical representation. These services also include tools for fast interactive data queries from multiple data sources and provide on-the-fly data fusion and interactive visualization for big data use cases.

Area 2: Infrastructure Common Services

Finding Summary

Divisions currently provide a range of tools and services tailored to the specific needs of users in their disciplines. However, there are specific “baseline” collaboration tools, cloud services, and data services that would be of use to all SMD Divisions. The minimal reuse of services and tools across SMD Divisions is a missed opportunity for improving interoperability and efficiency.

Recommendation Summary

OCSDO should support common data services and platforms that support Divisions in developing and deploying discipline-specific infrastructures while reducing or eliminating duplication of effort.

OCSDO should:

- facilitate coordination with the Divisions and other Cooperating Organizations to understand common service needs;
- provide knowledge management of lessons learned and best practices in utilizing common services;
- develop best practices to decrease time-to-setup and improve ease of access for Divisions; and
- improve efficiency by coordinating procurement of commercial cloud resources across Divisions.

Recommendation 2.1: Edge Service Routing Services

OCSDO should provide Edge Service Routing Services that serve as entry points to services and route requests to services as required by the user.

Recommendation 2.2: Identity and Access Management Services

OCSDO should provide Identity and Access Management Services, which are the combination of technical systems, policies and processes that create, define, govern utilization, and safeguard identity information. These services should manage the relationship between an entity and the resources to which access is needed. These services should be capable of managing digital identities, authenticating users (including authorized users without NASA PIV badges), and authorizing access to resources.

Recommendation 2.3: Application and Resource Monitoring Services

OCSDO should provide Application and Resource Monitoring Services, which enable the observation and analysis of application health, performance, and user experience. These services should be capable of observing an application’s complete transactional behavior, automating discovery and mapping of an application and its infrastructure components, monitoring of applications running on various elements of the enterprise, identification and analysis of application performance problems, supporting multi-cloud environments, performance metrics analysis,and integration with application security functionality.

Recommendation 2.4: Workflow Management Services

OCSDO should provide Workflow Management Services, which provide an infrastructure for the set-up, execution, and monitoring of a defined sequence of tasks, arranged as a cloud-native workflow application that is executed in a cloud hosting environment.

Recommendation 2.5: Software Factory Services

OCSDO should provide Software Factory Services, which facilitate an organized approach to software development that provides software design/development teams a repeatable, well-defined pipeline to create, update, and integrate software components. These services result in a robust, compliant, and more resilient process for delivering applications to production. These services should be capable of continuous integration and continuous delivery, as well as providing a structured and modular approach to software design and development, producing software-deployable artifacts, automating activities in the develop, build, test, release, and deliver phases, and integration with cyber security and software quality analysis tools.

Area 3: Improving Efficiency and Access to Computing

Finding Summary

Developers, operators, and users require flexibility, ease of entry, and surge capacity within data and computing systems to meet evolving needs and to support modeling and modern data science workloads such as AI/ML. SMD currently develops and operates over 5 cloud computing environments with duplicative capabilities. SMD is the major user of NASA's on-premises High-End Computing facilities. A lack of coordination and interoperability among these computing capabilities increases user frustration, introduces development complexity, increases cost (uses resources that could otherwise be available for scientific advancement) and impedes scientific collaboration and progress.

Recommendation Summary

OCSDO should develop a single interoperable hybrid cloud/HEC data and computing architecture to reduce duplication and increase efficiency. This architecture should improve the ability of developers, operators, and users to access and process data by streamlining cloud environments, supporting specialized scientific workloads in the most appropriate computing environments, and accelerating access to resources for NASA and external users. Priority should be given to developing an accessible cloud infrastructure that is interoperable with harmonized with other Agency storage and computing capabilities, including HEC capabilities, to meet the needs of generalized and specialized computing requirements including scientific modeling, management and distribution of repository data, standard product production and archival, and specialized computational workflows.

Recommendation 3.1 Hybrid cloud/HEC data and computing architecture

OCSDO should develop a hybrid computing framework that harnesses the benefits of both cloud services and on-premises high-performance computing.

Recommendation 3.1.1

OCSDO should provide a consolidated, managed cloud environment (MCE) and eliminate redundant cloud computing environments (i.e. NGAP, SMCE, MCP, et. al.).

- OCSDO should provide improved access to data and computing resources by refining the processes to estimate (i.e cost modeling/forecasting), request, and begin using cloud resources for researchers and developers;

- OCSDO should provide a Multi-environment Computing Service (Infrastructure as a Service – IaaS) that provides physical as well as virtualized compute capacities in multiple cloud environments and on-premises;
- OCSDO should provide a Storage management solution that facilitates scalable storage (on-premises, co-located, hybrid cloud environment).

Recommendation 3.1.4

OCSDO should provide a Network and connectivity service that facilitates communication and transfer of data and information between various networks and domains.

Recommendation 3.1.5

OCSDO should provide a Communication software that facilitates remote access to systems and the ability to exchange data or browse remote resources.

Recommendation 3.1.6

OCSDO should provide a Cluster management service that aggregates computing and memory resources in the virtual environment.

Recommendation 3.1.7

OCSDO should provide a Load balancing that directs and controls external traffic between the application services and their clients.

Recommendation 3.1.8

OCSDO should provide a Container orchestration service that offers a unified cloud agnostic multi-cluster multi-tenant application management.

Recommendation 3.2: Cybersecurity

OCSDO should provide scientific data and computing architecture that enacts computer security in line with Program and Project requirements and promotes experimentation by allowing users to run a wide range of data analysis in a secure architecture that allows analysis to fail while containing damage.

Recommendation 3.3: Partnerships

OCSDO should establish partnerships/agreements with outside organizations (including government agencies and academic, international, and industry organizations) to increase access to computing resources, increase access to external data sources, reduce risks and promote innovation.

Area 4: Managing Cybersecurity Risks

Finding Summary

Cybersecurity is critical to the successful operation of NASA's scientific data and computing systems; evolving security requirements broadly impact missions, systems, and users. The current heterogeneous structure of NASA's scientific data and computing systems increases the complexity of cybersecurity

implementation, resulting in inconsistency in both the level of support and guidance provided to SMD systems for cybersecurity. In the current state, improper or overly restrictive security categorizations can unintentionally deter and sometimes prevent internal and external users from analyzing data and collaborating together.

Recommendation Summary

OCSDO should provide support in the development of security plans for science data systems to ensure a common interpretation that follows NASA and NIST requirements while allowing for scientific collaboration. Additionally, OCSDO should work collaboratively with the OCIO on security requirements to balance system integrity with open scientific collaboration.

Recommendation 4.1: Cybersecurity Support for Science Data and Computing Systems

OCSDO should provide support to Division repositories during security assessments and in the development of their security plans to ensure they universally interpret and follow NASA and NIST requirements.

Recommendation 4.2 SMD and OCIO Cybersecurity Collaboration

OCSDO should provide strategic collaboration with the OCIO to enhance SMD’s open source science IT capabilities and streamline open science cybersecurity governance.

Area 5: Long-Term Sustainability (Programmatic)

Finding Summary

SMD's scientific missions and models produce NASA's largest pool of publicly available data, requiring critical investment in data and computing infrastructure. High-End Computing (HEC) drives astrophysics theory development, space weather prediction, and climate projections, while cloud computing is advantageous for data processing, analytics, and open science collaborations. As a premier science and technology agency, NASA must maintain cutting-edge cyberinfrastructure to make discoveries and support open science collaborations. The accelerating pace of technological advancement underscores the vital role of data and computing capabilities accessible to both internal and external users.

The current organizational framework overseeing the management of SMD's cloud and HEC infrastructures is presently inadequate to fully capitalize on the substantial existing investments. This limitation, in turn, obstructs the ability to swiftly adapt to changing needs, efficiently allocate resources, and promptly respond to technological advancements. Misalignments within the organizational structure may result in missed opportunities to strategically leverage innovative initiatives and interagency collaboration from major legislation.

Furthermore, the growing need for specialized scientific and data science skills compounds these challenges. The demand for experts in these fields is steadily increasing, and the competitive landscape for recruiting and retaining top-tier talent is becoming more intense. In this evolving landscape, it is

imperative to strategically address the requirement for specialized workforce development and retention in scientific and data science domains.

Recommendation Summary

SMD should decrease friction to onboard external users and internal projects/missions. SMD should ensure users are able to select the correct fit-for-purpose capabilities and quickly deploy new functionality/capabilities. SMD should increase the efficiency of cloud and HEC resources through economies of scale, mitigation of vendor lock-in, cybersecurity enhancements, and access to specialized computing environments. SMD should ensure the discovery and evolution of advanced science data and computing technologies.

The programmatic alignment of HEC and cloud resources should be evaluated to respond to rapidly-changing technology and to match the evolving scientific, technological, and mission goals and needs crucial for long-term viability and success. Strategic scientific data and computing architecture investment decisions should be coordinated to support open science and access to fit-for-purposes data and computing systems.