Unidata: Next-generation Data Services and Workflows to Advance Geoscience Research and Education

Effective use of real-world data is a central component of modern approaches to understanding our physical environment. For more than three decades, Unidata has been a cornerstone data facility in the geosciences, serving a diverse community of education and research institutions with the common goal of sharing data and tools for scientific analysis. Unidata functions at the nexus between data and knowledge, providing its community members with a wide variety of meteorological data and the practical tools, training, and support to maximize the effective use of those data.

There is a widespread recognition among geoscientists that addressing real-world "grand challenge" problems requires a deep and wide integration of knowledge, techniques, and expertise, transcending disciplinary as well as organizational boundaries. As a result, geoscience cyberinfrastructure needs to support open, multi-disciplinary, interdisciplinary, and transdisciplinary teams that bring together diverse expertise and perspectives spanning cultures and practices. In today's data-driven and collaborative research environment, easy and reliable access to compute, storage, and data resources is as necessary as the theoretical foundations and algorithms used for advancing the frontiers of science. It is also evident that the explosion of data from sources as diverse as remote sensing instruments, unmanned vehicles, and ensemble prediction models requires new thinking about the delivery and use of data — that instead of bringing data to scientists, we need to bring their science to the data. This movement towards open, collaborative, data-proximate analysis will bring with it a seismic shift in scientists' workflows; Unidata believes that managing this shift will be key to the success of one of NSF's "10 Big Ideas," the *Harnessing the Data Revolution* initiative (NSF 2016).

Easing the geoscience community's transition to these new data-centered discovery methods is at the core of this proposal. Here, Unidata presents its plan for advancing research and education in the geosciences by working toward a *Science as a Service* paradigm (Madduri and Foster 2016) that brings together data and tools in ways that foster interdisciplinary collaboration, seamless sharing of data, reproducibility of results, and minimal barriers to participation.

Who We Are and What We Do

Unidata is a community data facility for the atmospheric and related sciences, established in 1984 by U.S. universities with sponsorship from the National Science Foundation (NSF). The Unidata Program Center (UPC), the program office for Unidata and the hub for activities related to Unidata's mission, is managed by the University Corporation for Atmospheric Research (UCAR), a consortium currently comprising 117 member universities and academic affiliates providing science in service to society.

Unidata exists to engage and serve researchers and educators who are advancing the frontiers of Earth System science. Our aim is to help transform the conduct of research and education in the atmospheric and related sciences by providing well integrated data services and tools that address the entire scientific data lifecycle, from locating and retrieving useful data, through the process of analyzing and visualizing data either locally or remotely, to curating and sharing the results.

Specifically, some of the UPC's most important activities include:

- Acquiring, distributing, and providing remote access to real-time meteorological data.
- Developing software for accessing, managing, analyzing, visualizing, and effectively using geoscience data.
- Providing comprehensive training and support to users of its products and services.

From the outset, Unidata has been governed by its community. Representatives from universities populate standing committees that help set policies for the program and provide first-hand feedback from users of program software and services. Non-voting representatives from government agencies provide valuable information and advice. Unidata's governance structure ensures that the program stays in tune with the community it serves, and allows it to quickly adjust priorities as the technological landscape and community needs change.

While Unidata's primary mission of serving universities engaged in atmospheric science education and research has remained unchanged through the years, the evolution and broad usefulness of its products and services have greatly enlarged its user base. Today, the Unidata community includes users from all sectors in over 200 top-level Internet domain countries and territories, including more than 1600 academic institutions and more than 60 research labs. Last year, with community input and aid of its governing committees, Unidata developed a strategic plan, creating a roadmap and a vision for the future. The overarching goal embodied in that plan is to work toward a scientific ecosystem in which time to science is shrunk by reducing "data friction" (Edwards, 2010), and data transparency and ease-of-use are significantly increased. This proposal outlines activities we will undertake to accomplish the goals set forth in our strategic plan.

NSF Site Visit Team

In March 2017, the NSF convened a Site Visit Team (SVT) to visit the UPC and conduct a mid-term review of Program activities under NSF 1344155. We are indebted to the SVT panel for the time, energy, and thought they devoted to reviewing the many areas of the Unidata Program during their visit, and we appreciate the comments, findings, and recommendations provided in the "Unidata Site Visit Team Summary Report: 6-7 March 2017." In that summary report, the SVT states:

> *"Overall, the SVT believes that the UPC is performing at an impressive level of efficiency. Unidata works on an impressive array of projects with limited resources. Unidata robustly serves the Atmospheric Sciences Community, provides unique services and works diligently to obtain feedback on its activities and to implement recommendations. UPC has greatly expanded its role in serving a wide variety of geoscience disciplines."*

We are gratified by the SVT's conclusion that Unidata provides effective service to the geoscience community. Plans working toward implementation of many of the SVT's recommendations can be found in the forward-looking sections of this proposal.

Unidata as Open Source Team Leader

Recent years have seen the spread of very robust open source development tools, like GitHub, that dramatically lower the barriers to community contributions to software projects. Unidata software development has long been an open process that encouraged community contributions, but the growth of the open source movement has made it possible to open the doors even wider. Unidata developers have worked to improve the openness and transparency of our software development and planning efforts by making use of the latest open source development and social coding tools.

All Unidata software projects store their source code in GitHub repositories, and many use GitHub's issue tracking system to manage issues and enhancement requests. Using GitHub's "Pull Request" mechanism allows Unidata staff to provide code review and quality control for community contributions, and continuous integration tools like Travis CI and Jenkins ensure that contributed code passes the same rigorous tests that Unidata developers' code is subject to. Various communication channels are used to engage with and build each project's community, encouraging contribution and collaboration while providing guidance concerning how to contribute and what is expected of participants both in terms of code submissions and community conduct.

Unidata Snapshot

The following tables provide a snapshot of the Unidata program in September 2018, with comparisons to metrics from the previous NSF proposal (where available).

Table 1: Data Services	2013	2018
Host machines on the IDD network[1]	520 (232 unique sites)	577 (245 unique sites)
Connections to UCP-hosted IDD clusters		1184 (195 unique sites)
Approximate volume of data ingested into the IDD	315 GB/day	1.47 TB/day
Volume of data pushed to IDD sites	13 TB/day	35 TB/day
Volume of data pulled via remote access protocols	659 GB/Day	1.58 TB/Day
Uptime of UPC data and support infrastructure	99.96%	

Table 2: Software Package Downloads	2013-2018
AWIPS	14,799
GEMPAK	7,207
IDV	33,971
Local Data Manager	10,226
McIDAS	257
MetPy	45,365
netCDF-C Libraries (includes FORTRAN, C++ support)[2]	435,142
netCDF-Java Libraries (Common Data Model)	29,007
Siphon	39,227
THREDDS Data Server	12,775
UDUNITS	24,558

Table 3: Workshops	2013-2018
Training Workshop Participants	146
Users Workshop Participants	152
Regional Workshop Participants	387

Table 4: Miscellaneous	2018
Number of community members registered with Unidata	59,453
Number of countries where Unidata software and services are used	204
Number of academic institutions participating (U.S.)	540
Number of academic institutions participating (worldwide)	1,644
Number of organizations participating worldwide	2,225
Number of community electronic mail lists	40
User support e-mail transactions, 2013-2018	29,975
Number of Community Equipment Awards, 2013-2018	27
Average staff FTEs at the UPC, 2013-2018	21

[1] These metrics are limited to sites of which Unidata is aware. Sites can use the LDM and participate in the IDD without reporting statistics to the UPC; we suspect the number of unreported sites is large.
[2] UPC source code downloads only. This number does not include downloads from repositories at the University of Kyoto and on the Github site, or binary distributions available via package managers on UNIX-like systems.

Unidata Focus Activities

Unidata takes part in a variety of activities on behalf of the geoscience community. While distributing geoscience data in near real-time has always been at the core of Unidata's mission, the Program also develops software for the analysis and visualization of that geoscience data, provides user support and training, and takes on a variety of community service and leadership roles. Because the geoscience community has supported Unidata consistently over the long term, the Program has been able to undertake projects that span years or even decades, supporting thousands of users in the U.S. and around the world. This section briefly describes the interlocking set of activities that form the core of Unidata's work; the Plan of Action that follows describes how these activities will address Unidata's strategic goals for the empowerment of its broad community.

Data Distribution

The UPC coordinates the Internet Data Distribution system (IDD), in which hundreds of universities cooperate to disseminate near real-time Earth observations via the Internet. The data "pushed" to member sites via the IDD include weather radar, satellite imagery, forecast model output, and others — more than 30 distinct data feed types in all. In addition to the "push" data services provided by the IDD, the UPC also provides on-demand "pull" data services via THREDDS, ADDE, AWIPS EDEX, and RAMADDA data servers. The UPC's data servers are not classified as "operational" resources, but they nonetheless have a 99.96% uptime record and are used heavily by organizations that lack the resources to store IDD-provided data locally, or to operate their own data servers (see Table 1 on page 3). UPC's servers are housed in a UCAR co-location computer facility or on a cloud facility for reliability, and are connected to Internet2/National Lambda Rail network, which provides access to ample bandwidth for Unidata's needs.

Software Development

A variety of software packages are developed, maintained, and supported by the UPC. With the exception of the McIDAS-X package, which is supported by Unidata but developed and maintained at the University of Wisconsin, all Unidata software is open source and freely available.

NetCDF

Unidata's netCDF (network Common Data Form) is a collection of data access libraries that provide a machine-independent data format that is self-describing, portable, scalable, appendable, sharable, and archivable — all important qualities for those who wish to create, access, and share geoscience data. NetCDF permits easy access to array-based, multi-dimensional datasets, a task that can be difficult when using other common storage schemes. NetCDF has been adopted widely by the atmospheric sciences community, and is especially popular among climate and ocean modelers. For example, as was the case with the Fifth Assessment Report, model output datasets for the Sixth Assessment Report of the Intergovernmental Panel on Climate Change must be submitted in netCDF format, using the associated Climate and Forecast (CF) metadata conventions. Data from the GOES-R series satellites is also delivered in netCDF format. The resulting large base of netCDF users and data has led to support for the format in more than 80 open source packages and many commercial applications including ArcGIS, MATLAB, and IDL.

MetPy & Siphon

MetPy is a collection of tools written in the Python language for reading, visualizing, and performing calculations with weather data. MetPy aims to mesh well with the rest of the scientific Python ecosystem, adding GEMPAK-like functionality specific to meteorology. MetPy became a supported package at Unidata in 2015; thanks to dedicated work at the UPC and in the MetPy user community it has seen 17 major and minor releases since that time.

Siphon is a collection of Python utilities for accessing data from remote data services such as the THREDDS Data Server. Growing out of work done to support Unidata's Python software training course, like MetPy it has attracted a dedicated base of users and community developers.

AWIPS & GEMPAK

The Advanced Weather Interactive Processing System (AWIPS) is a weather forecasting, display, and analysis package developed by the NWS and NCEP for use in Weather Forecast Offices (WFOs) and National Centers for Environmental Information. Because many university meteorology programs are eager to use the same tools used by NWS forecasters, Unidata community interest in AWIPS is high. As designed for the WFOs, however, AWIPS is a very resource-intensive system; installing and maintaining the NWS version is beyond the capabilities of many universities. As a result, Unidata has reconfigured and repackaged AWIPS for non-operational use, allowing universities access to the system for teaching and research with significantly reduced equipment requirements. By operating cloud-based AWIPS Environmental Data EXchange (EDEX) servers on behalf of the university community, Unidata has lowered barriers to educational use of the system even further.

While NCEP is engaging in little new development of the GEneral Meteorology PAcKage (GEMPAK), Unidata continues to provide support and patch releases as they become available.

The Python-AWIPS library is a collection of utilities for accessing data from the AWIPS EDEX, allowing users to query and render gridded models, satellite and radar imagery, watches and warnings, surface and upper air observations, lightning point data, and static map resources in their Python programs.

Cloud-oriented Software

Unidata's focus on helping community members efficiently use cloud-based computing resources has led to the creation of the CloudStream technique, which uses containerization and application-streaming technologies to facilitate the use of existing desktop software in a cloud-computing environment without the need to re-engineer the application. Using CloudStream, client-side software packages like AWIPS and the Integrated Data Viewer (IDV) can be containerized and uploaded to a cloud-computing resource; from there the application can be displayed and used inside an ordinary web browser. This allows the application to both take advantage of potentially large computing resources available from cloud providers and benefit from the application being located "next to" datasets in cloud-based storage, thus reducing the need to transfer data to the application.

Common Data Model & THREDDS Data Server

Unidata's Common Data Model (CDM) provides an interface for reading and writing files in netCDF and a variety of other scientific data formats. The CDM uses metadata to provide a high-level interface to geoscience-specific features of datasets, including geolocation and data subsetting in coordinate space. Unidata's THREDDS Data Server (TDS) builds on the CDM to allow for browsing and accessing collections of scientific data via electronic networks. Data published on a TDS are

accessible through a variety of remote data access protocols including OPeNDAP, DAP4, OGC Web Map Service (WMS) and Web Coverage Service (WCS), NetCDF Subset Service (NCSS), and HTTP.

The CDM and TDS are widely used in the United States (by NOAA, USGS, NASA, and the Earth System Grid, for example) and internationally, and are part of the deep infrastructure on which next generation capabilities are being built by other organizations. Additionally, many other tools build on the CDM (NOAA's ERDDAP, NASA's Panoply, and CMAS' VERDI, are examples) and on the TDS (NOAA PMEL's LAS and Ferret-TDS, for example).

Local Data Manager

The Unidata Local Data Manager (LDM) system includes network client and server programs designed for event-driven data distribution. It is the fundamental component of the IDD system. The LDM is used by hundreds of sites worldwide, and is integrated into the National Weather Service's AWIPS package.

Integrated Data Viewer

Unidata's IDV is a 3D geoscience visualization and analysis tool that gives users the ability to view and analyze a rich set of geoscience data in an integrated fashion. The IDV brings together the ability to display and analyze satellite imagery, gridded data (such as numerical weather prediction model output), surface observations (METARs), upper air soundings, NWS NEXRAD Level II and Level III RADAR data, NOAA National Profiler Network data, and GIS data, all within a unified interface. The IDV integrates tightly with common scientific data servers (including Unidata's TDS) to provide easy access to many real-time and archive datasets.

Rosetta

Rosetta is a service designed to improve the quality and accessibility of observational data sets collected via datalogging equipment. Rosetta transforms unstructured ASCII data files of the type commonly generated by datalogging equipment into the netCDF and other formats, while minimizing disruption to existing scientific workflows.

UDUNITS

Unidata's UDUNITS package supports conversion of unit specifications between formatted and binary forms, arithmetic manipulation of units, and conversion of values between compatible scales of measurement.

McIDAS

The Man-computer Interactive Data Access System (McIDAS) is a large, research-quality suite of applications used for decoding, analyzing, and displaying meteorological data. The older McIDAS-X system, developed by the University of Wisconsin's Space Science Engineering Center and supported by Unidata, is gradually being replaced by the IDV and by McIDAS-V (which is based on the IDV).

User Support and Training

Users of Unidata software and data rely on the UPC for comprehensive support services. UPC software developers handle user support directly, together responding to an average of more than 350 technical support questions each month. Developers also create product documentation, training materials, and provide hands-on training workshops each year.

The workshops, which have been held at the UPC in Boulder, a variety of host universities, and at AMS annual meetings, have been attended by more than 500 participants from the university, government, and commercial spheres in the past five years. In-person workshops provide UPC developers with an excellent opportunity to interact with software users face-to-face.

In addition to in-person training, Unidata Python developers have created an Online Python Training resource designed to help those new to the language get up to speed quickly. The Python team also creates weekly video tutorials as part of the "MetPy Mondays" program.

Unidata's Data Management Resource Center (DMRC) is an online resource designed to help community members select appropriate data management workflows and navigate the data management requirements of federal agencies. By presenting real-world examples of different types of projects, the DMRC gives researchers ideas about how to combine existing technologies into robust data management processes.

Unidata Science Gateway

The UPC provides cloud-based resources to community members on a demonstration basis through the *Unidata Science Gateway*. The Science Gateways Community Institute (SGCI) defines a science gateway as an online community space for science and engineering research and education, with web-based resources for accessing data, software, computing services, and equipment specific to the needs of a science and engineering discipline. Unidata's Science Gateway takes advantage of resources granted to Unidata on the NSF Jetstream cloud, and provides users with access to a JupyterHub server and a complete IDD feed made accessible via a variety of data services including TDS, ADDE, and RAMADDA servers. Gateway users can experiment with data-proximate computing techniques, or simply use the JupyterHub server as a cloud-based learning environment. Beginning in 2018, Unidata has hosted several workshops and university courses for community members using the Gateway.

Cyberinfrastructure Leadership and Community Service

Unidata community members look to the UPC not only for technological solutions, but for guidance on emerging trends in cyberinfrastructure and to represent their interests in collaborations with standards bodies and organizations that work across scientific disciplines. As standards-based solutions have become increasingly important to the conduct of international science and interoperability, Unidata has assumed a central role in identifying and articulating standards, conventions, and data formats. Unidata's standards efforts have enabled ongoing collaboration with dozens of international organizations – especially those represented in the OGC MetOceans, Earth System Science, and Hydrology Domain Working Groups. Similarly, Unidata's work in defining and explaining scientific workflows that take advantage of cloud-based computing resources is designed to help community members make the most of remote-computing opportunities to lower their costs and IT infrastructure needs.

In addition, Unidata undertakes a wide variety of activities aimed at fostering a shared vision for and community ownership of the program and encouraging community input into its operation and direction. Bringing the community and stakeholders together to share knowledge and address problems that are important to them through meetings, workshops, conferences, and other venues is a key aspect of Unidata's community service mission, as are efforts to disseminate information of interest to community members.

Results of Prior Support Under NSF 1344155

A great majority of activities at the UPC are sponsored under the five-year core-funding award "Unidata 2018: Transforming Geoscience through Innovative Data Services" (NSF 1344155). The period of performance for that award is 1 April 2014 to 31 March 2019.

***Note**: A delay in funding after NSF 1344155 was approved shifted the period of performance to begin in April 2014, but Unidata began the activities outlined in the plan in the fall of 2013. Dates in this retrospective section reflect the choice to begin work before the official period of performance began.*

The core-funding proposal identified four strategic goal areas as the focus of Unidata's efforts:

1. Enabling widespread, efficient access to geoscience data
2. Developing and providing open-source tools for effective use of geoscience data
3. Providing cyberinfrastructure leadership in data discovery, access, and use
4. Building, supporting, and advocating for the geoscience community

During the period of performance for the current award, Unidata met most of the goals described in that proposal, accomplishing many of the stated objectives. Unidata has increased the volume and variety of data provided to the community, expanded and enhanced its software offerings, developed and deployed new tools and services, and actively engaged its growing user community. The program's contributions and accomplishments have had a demonstrable and sustained impact on the geosciences community, empowering faculty, students, and researchers to be more productive and enhancing their ability to advance science and learning.

Key Achievements 2013-2018

The following list highlights, in no particular order, Unidata's most significant accomplishments during the past five years.

- **Data Delivery.** The volume of observational data and model output delivered to Unidata community members and institutions in near real-time continues to grow. As of August 2018, Unidata's Internet Data Distribution (IDD) clusters deliver roughly 35 Terabytes per day to downstream systems, up from roughly 32 Terabytes per day in 2017. As a result of the GOES-16 spacecraft coming on-line and the data becoming operational in late 2017, the volume of data served via remote access methods (most notably ADDE) has increased from roughly 659 GB/day to roughly 1.58 TB/day.
- **Python Activities.** Unidata has increased its support for the use of the Python programming language in the atmospheric sciences tremendously in the past five years. Program Center staff lead development of the open source MetPy project, which has been downloaded over 46000 times from the Python Package Index and Anaconda Cloud, and which has an active and growing community of users and developers. Similarly, the Siphon project, a Python library to facilitate access to remote data sets, has been downloaded over 39000 times. Staff have led 14 Python-focused training workshops during the period, reaching more than 380 learners at universities, scientific conferences, and at the UPC. Unidata has also contributed numerous enhancements and fixes to upstream Python packages, benefitting the entire scientific Python ecosystem. The MetPy and Siphon projects have also greatly benefited from community participation; individuals outside the UPC have submitted more than 130 bug fixes and feature contributions.
- **AWIPS.** Unidata has refactored and repackaged the National Weather Service's Advanced Weather Interactive Processing System (AWIPS) into a configuration suitable for use at

universities and other organizations with limited computing resources. Since 2013, more than 80 universities have installed the software — as well as numerous research institutions, commercial sites, and individuals. In addition, dozens of universities take advantage of the cloud-based EDEX server operated by the UPC in the NSF Jetstream cloud. This allows them to provide courses using the AWIPS client software (CAVE) without the need to expend resources to build and configure a local EDEX data server. Unidata's creation of a MacOS version of CAVE allows the package to reach an even wider audience.

- **Containerization of Unidata Technologies.** In recent years, the cloud computing industry has settled on Docker containerization technology as a standard for packaging and deploying software in cloud environments. Users of containerized packages need only deploy the container to reap the benefits of "professional" installation. Docker technology also helps create reproducible computing environments through the creation of "Dockerfiles" that capture all of the technical knowledge required to install and run a package correctly. Containerization of the Weather Research and Forecasting (WRF) model provides a working example of the value of this technique (Hacker 2017). Unidata has devoted significant resources to containerizing Unidata technology offerings; the UPC now maintains containers for the LDM, the TDS, ADDE, and RAMADDA.
- **Unidata Science Gateway.** The Unidata Science Gateway collects Unidata-related technologies hosted on the NSF's Jetstream Cloud. The gateway demonstrates a workflow involving combining cloud-based resources — perhaps coupling them with XSEDE HPC resources — to create end-to-end scientific computing workflows. One of the most exciting tools in the Unidata Science Gateway is the JupyterHub server. Targeted at students and educators, this server provides Jupyter notebooks from three Unidata projects: the Unidata Python Workshop, the Unidata Notebook Gallery, and Unidata Online Python Training. These notebooks include pre-built Python environments needed to run them, and can be used for instruction and as templates for scientific experimentation.
- **Cloud Computing.** Unidata's activities to support the use of cloud-based computing resources in the geoscience community have taken many forms beyond the containerization and Science Gateway projects. Unidata's CloudStream technology provides a way to move legacy desktop software into the cloud without requiring significant engineering resources, enabling data-proximate visualization and analysis. UPC staff have provided technical support to numerous projects undertaken by the university community to evaluate the use of cloud-based computing resources in education.
- **NetCDF.** Unidata released multiple updated versions of the netCDF library and data model over the period of performance, most notably adding support for remote access to netCDF-4 model data via the DAP4 protocol. The NOAA GOES-R series satellite program makes all L2 and L2+ products available in the netCDF-4 format, which is also used in disciplines including chromatography, neuro-imaging, molecular dynamics, and fusion research. All in all, netCDF is by far the most widely used Unidata technology, with more than 140,000 downloads in the most recent year. NetCDF development has benefited greatly from its embrace of the GitHub open source software development platform, with more than 300 bug fixes and new features submitted by community contributors.
- **GOES Data.** In 2016, NOAA/NESDIS funded installation of equipment to ingest imagery and products from the GOES-16 spacecraft. In 2017 the UPC continued its collaboration with NOAA to propose a similar installation for the next satellite in the series, GOES-17. While GOES-17 is not certified operational at this writing, data from GOES-16 and GOES-17 are flowing to the community through Unidata ingest equipment and data servers.

- **IDV Advances.** Unidata's IDV has received numerous updates over the past five years. New features include a Time Matching feature to enable seamless combination of datasets with different time-series values, new Grid 3D and Volume trajectory visualizations, an Adaptive Resolution display feature to support efficient visualization of very large datasets at different scales, a new Match Display Region feature to streamline inspection of subsets of large areas, a new feature for visualizing Isentropic surfaces in three dimensions, a new Scatterplot feature, and facilities for visualizing GOES-16/17 Global Lightning Mapper (GLM) data.
- **TDS Advances.** The THREDDS project has seen many advances during the current period of performance. The efficiency in which GRIB 1 and GRIB 2 data are handled by netCDF-Java and served by the TDS has significantly increased, and collections of these data are now highly scalable and available as GRIB FeatureCollections. In a beta-test phase as of this writing, TDS version 5 reduces the memory footprint of TDS catalogs, simplifies the configuration process, and allows adding/modifying TDS catalogs on-the-fly.
- **Equipment Awards.** During 2014-2018, the UPC provided 27 equipment grants to 26 universities, encouraging new members from diverse disciplinary backgrounds in the geosciences to join the Unidata community and to allow existing members to continue and enhance their active participation. In keeping with Unidata's intention to encourage universities to experiment with modern workflows, during this period the UPC has entertained proposals for grants to fund purchase of cloud-computing resources in addition to traditional computing hardware.
- **EarthCube Activities.** In addition to Unidata director Mohan Ramamurthy's direct involvement with the EarthCube initiative in his role as director of the EarthCube Science Support Office (also located among the UCAR Community Programs), numerous UPC staff have been involved in EarthCube-related projects from software development to work on community data standards.
- **Community Workshops.** In addition to software training workshops held at the UPC and at host universities, Unidata organized and hosted triennial Users Workshops in 2015 and 2018, each of which brought approximately 75 community members together to share tools, techniques, and educational strategies. The UPC also organized community workshops to discuss modeling research in the cloud and to further work towards Climate and Forecast (CF) metadata standards for netCDF.
- **Regional Workshops.** In response to community concerns about the cost of traveling to the UPC for software training workshops, Unidata has significantly expanded its offerings of workshops held on university campuses, serving nearly 400 attendees at 14 different workshops in the past five years
- **Internships and Mentoring.** In the years 2014-2018, the UPC has hosted a total of eleven summer student interns who built skills in software development and science communication. Unidata summer interns have presented their work at AMS annual meetings and have served as student ambassadors for the Unidata program. Most recently, one of our 2016 student interns returned as a presenter at the 2018 Unidata Users Workshop.
- **Building Open Source Community.** Unidata has wholeheartedly embraced modern open source development strategies. Source code for most Unidata software packages is now accessible via GitHub; this reinforces our commitment to making Unidata software freely available and provides a robust and streamlined platform for integrating contributions from community members.
- **The NOAA Big Data project (BDP).** The BDP is an initiative to increase public availability of large volumes of environmental data collected and generated by NOAA. As part of the BDP,

Unidata works in collaboration with Amazon Web Services (AWS) to provide access to a more than twenty years of archived NEXRAD Level II radar data — augmented continuously with new, real-time data via the LDM — stored in Amazon's Simple Storage Service (S3) environment. NOAA now estimates that 80% of the requests it receives for NEXRAD data are handled by cloud servers (Kearns, 2017).

Additional Highlights 2013-2018

In addition to the above-listed achievements, the following sections provide brief summary of some of the other accomplishments in the four strategic goal areas outlined in NSF 1344155.

Enabling widespread, efficient access to geoscience data

Objective	Accomplishments
Distribute atmospheric and other geoscience data in real time	The volume of geoscience data distributed via the IDD has expanded from roughly 13TB/day in 2013 to over 35TB/day in 2018. Improvements to the THREDDS Data Server, and specifically the addition of GRIB FeatureCollections, have made remote access to data more efficient. The Unidata Science Gateway has TDS and ADDE servers providing a good portion of the IDD with a five day archive.
Develop innovative cyber-infrastructure solutions to facilitate dissemination of scientific data	Unidata has enabled domain-specific applications running within Jupyter notebook environments hosted in the cloud, facilitating data proximate workflows that reduce the volume of data distributed over the network. The Rosetta data translation tool makes it easy to create standards-compliant datasets.
Work with data providers to make geoscience data freely available	NCEI, HYCOM, IOOS, several other government agencies, as well as numerous universities use the TDS to make data publicly available. The TDS is also in use by the private sector; some organizations make their data freely available (e.g., PlanetOS).
Develop and maintain Unidata's computing and networking infrastructure	Unidata's IDD data servers have an uptime of 99.96%, and a network topology that helps ensure the flow of data to universities even in the event of an outage. Additional satellite hardware for reception of GOES satellite data has been installed, and work is underway to install a Unidata-controlled IDD cluster at the NCAR Wyoming Supercomputing Center.

Providing open-source tools for effective use of geoscience data

Objective	Accomplishments
Analyze, integrate, and visualize geoscience data in two, three, and four dimensions	MetPy provides a robust set of tools in Python for reading data, performing geosciences calculations, and plotting the results. The Integrated Data Viewer (IDV) has added a number of new display and analysis tools including Grid 3D and volume trajectories and isentropic and scatter analysis.
Enable visualization and effective use of very large data sets	The IDV Adaptive Resolution feature allows users to visualize very large data sets through on-the-fly adjustments to data sampling parameters, resulting in the highest quality image at any resolution but making the most efficient use of computing and network resources.
Access, manage, and share collections of data from diverse sources	Both Siphon and netCDF-Java facilitate access to data sets using a variety of access methods that allow downloading of subsets of data, helping to make accessing large datasets manageable.

Providing CI leadership in data discovery, access, and use

Objective	Accomplishments
Develop useful data models, frameworks, and protocols for geoscience data	The Python-AWIPS library allows access to AWIPS EDEX data stores from Python programs, and Jupyter notebook scripting examples show users how to query and render a variety of AWIPS data types.
Advance geoscience data and metadata standards and conventions	Led efforts to advance the Climate and Forecast (CF) metadata conventions for netCDF (netCDF-CF) via two community workshops.
Evaluate emerging cyberinfrastructure trends and technologies, providing information and guidance to community members	Unidata's direct engagement of the scientific Python ecosystem helps ensure its continued relevance and utility to the geosciences. In some cases, this involves the contribution of fixes and changes to other projects to address our community's needs. Instances of the TDS running in cloud-computing environments (Jetstream, AWS) have provided valuable insights into how Unidata technologies can take advantage of cloud resources.
Facilitate discovery mechanisms for quickly finding and accessing geoscience data	Worked with ESSI-Lab's GI-CAT and pyCWT to make sure the ncIso service in the TDS provides sufficient information to enable data discovery.

Building, supporting, and advocating for the geoscience community

Objective	Accomplishments
Represent the academic community in partnerships with agencies and stakeholders	The Unidata AWIPS distribution is the result of Unidata's advocacy for the university community with the National Weather Service and the lead AWIPS developers (Raytheon). Unidata has worked extensively with NOAA to ensure university access to GOES-16/17 data streams.
Foster interactions between community members	Unidata hosted the OGC Technical Committee meeting in Boulder, CO in June, 2015, and has organized several NSF-funded workshops exploring cloud-computing and community data standards.
Present Unidata community perspectives	We use conferences like AMS, AGU, and EGU to disseminate developments at Unidata, and in the geoscience ecosystem as a whole, to our community.

Intellectual Merit

In this proposal, we describe three strategic focus areas for Unidata's continuing efforts:

- managing geoscience data by making it accessible, discoverable, and usable by our community;
- providing software tools that promote efficient modern scientific computing workflows, reproducibility, and open science; and
- supporting the researchers, educators, students, and others who make up our community by providing technical support, building capacity through training and open source development processes, leading advanced cyberinfrastructure initiatives, and advocating for community data needs.

Our vision of "Geoscience at the speed of thought through accelerated data discovery, access, analysis, and visualization" can only be realized if we are successful in working with our community and collaborators to foster the use of modern, scalable, and data-proximate tools and workflows. We

propose to build on our recent successes in adding cloud-based technologies to the existing geoscience software toolset to make it possible for a much broader range of community members to access and easily use the very large and multidisciplinary datasets that have until recently been available only to those with significant dedicated computing resources. Our overarching goal is to help our community begin transitioning to a *Science as a Service* model, as described more fully in "Vision for the Future" (page 17). Initially, this will mean establishing a Unidata Science Gateway on the NSF-funded Jetstream Cloud facility (an endeavor that is already underway), and augmenting that gateway to provide a host of cloud-native data and software services, advanced workbench capabilities to access and execute data-proximate analysis tools and workflows, and supporting instructional training materials. We will enlarge and adapt Unidata's toolset to provide open, free, and easily shareable modern computational methods in support of Open and Reproducible Science. In addition, we will continue to work with our partners to guide the community as it adapts to these socio-technical changes through training, demonstration, and ongoing support interactions. As a result of these efforts, we hope to empower our community to tackle multidisciplinary problems, develop the profession's human capacity, and transform the conduct of science.

Broader Impacts

Service to the geoscience community is at the core of all of Unidata's activities. As shown in Table 4 (page 3), Unidata's products and services are in use by over 2200 organizations including over 1600 academic institutions in 204 countries. Unidata technologies benefit many dozens of research labs, weather services around the world, national and international agencies, and scientific projects. Across a wide range of geoscience disciplines, scientists and service providers routinely use Unidata's scientific software libraries to conduct their research and share data with others — Unidata technologies are built into more than 20 commercial and 100 open source software packages.

This section touches on a few of the ways Unidata's activities extend beyond our service to our core atmospheric science community to benefit the broader geoscience community and society at large.

Impact on Research

The impact of Unidata's data systems, software, and services on scientific research extends far beyond our core atmospheric science community. The number and diversity of publications in peer-reviewed journals that cite Unidata or its software are compelling measures of Unidata's impact on research.

For example, a search of American Meteorological Society journals over the five-year period 2013-2017 retrieved 196 papers referring to Unidata software or data services; 267 papers referring to Unidata software or data services were published in American Geophysical Union journals during the same period. Many of these papers refer to Unidata software or data services without mentioning the Unidata Program itself; this correlates with anecdotal evidence of widespread use of Unidata products (especially netCDF) beyond the communities traditionally served by Unidata.

A review of citations reported by the Google Scholar search engine revealed that between 2013 and 2017, Unidata software and data services were cited 2970 times in the full range of scholarly literature encompassed by the search engine. Interestingly, Google Scholar returned an additional 8250 articles mentioning netCDF, but which did not include mention of its connection to Unidata.

Impact on Education

Unidata software and data services are in use at over 500 U.S. colleges and universities and more than three times that many in other countries; the list includes universities in all 31 EPSCoR jurisdictions, including many institutions that have a large number of students from underrepresented communities. We continue to reach out to these institutions by providing equipment grants, local software training workshops, and data services.

The UPC's most recent survey of its core academic community members, in 2017, provided insights into the current use of Unidata software and data services in education. Among the approximately thirty institutions who responded to the survey, responses indicate that more than 4400 undergraduates and 2200 graduate students used Unidata provided data, services, or software in the past five years. Of these, more than 1600 students used these data, services, or software in research in undergraduate honors theses, graduate theses, or Ph.D. dissertations and 438 received MS or PhD degrees.

At the University of Missouri, near-real-time data fed via the LDM and analysis software from GEMPAK to AWIPS have been instrumental to educational outcomes. From the 2017 survey:

> *Many graduate research projects have been sparked by the availability of Unidata data and software... Daily forecasting activities for the public and field research projects have been facilitated by GARP, NMAP, and now a transition to AWIPS.*

Educators at Valparaiso University noted:

> *The success of our meteorology program has been built on the backbone of data provided through Unidata software, specifically the LDM used in conjunction with GEMPAK and Python technologies. The ease with which faculty and students are able to access relevant, current, worldwide data is a service unequaled by any other entity.*

Community Equipment Awards

Each year, the UPC sets aside $100,000 to fund the Unidata Community Equipment Awards program. The program provides funds to encourage new geoscience departments to join the Unidata community and to allow existing members to continue and enhance their participation. During the past five years, Unidata made 27 awards to institutions ranging from major U.S. research universities to community colleges. Projects funded in the past five years include:

- Cloud-based projects supporting student use of AWIPS at Embry-Riddle Aeronautical University received support. The Emergency Response Meteorology (ERM) certificate program lets meteorology students develop skill and experience analyzing gridded forecast model data and weather observations in real time from any location where mobile Internet service is available.
- The University of North Dakota used award funding to build out data infrastructure (a THREDDS server) to integrate with research and education initiatives already being conducted as part of the NSF Big Weather Web project. As a part of their grant project, they also investigated the use of containerization (Docker) and cloud-based computing resources for ensemble modeling.
- Oregon State University used award funding to implement a Jupyter Notebook server to enable meaningful climate/geospatial data experiences for undergraduate students.

A complete list of projects funded under the Community Equipment Awards program is available online on Unidata's web site.

Impact on Other Organizations and Projects

Unidata-developed cyberinfrastructure, in addition to being used widely in universities to advance education and research, also provides a substrate for other stakeholders in federal agencies, the private sector, and many non-governmental and international organizations. For instance, many data services in NOAA, NWS, NASA, USGS, DOE, DOD, NCAR, ECMWF, EUMETSAT, CMA, and CPTEC are built on the formats, software, and data systems that Unidata has developed. Within NOAA alone, at least 26 distinct divisions or groups are known to be using Unidata technologies to access, distribute, or archive data. Unidata systems and technologies are integral parts of numerous large-scale projects, including the WMO's THORPEX, the NSF-sponsored EarthScope project, the IPCC, and others.

The letters of commitment to collaborate from some of Unidata's partners, provided in Appendix A, provide a view of how Unidata's work is having a positive impact on those organizations. Some additional examples not previously mentioned:

- The **Oceanographic In-situ data Interoperability Project** (OIIP) is a collaboration between NASA's Jet Propulsion Laboratory (JPL), Unidata, and the Large Pelagics Research Center (LPRC) at the University of Massachusetts, Boston. The project uses the Rosetta data translation tool to help address key interoperability and data challenges associated with oceanographic in-situ datasets, focusing on marine animal electronic tagging data as a representative use case.
- **Meteoclim**, based in Palma, Spain, is a spinoff of the University of the Balearic Islands. They use the TDS in their work of developing atmospheric, climate, and hydrological predictions for public and private enterprises.
- Closer to home, Unidata actively supports the COSMIC program by distributing data from the university-based **SuomiNet** real-time Global Positioning System (GPS) network for atmospheric research and education via the IDD.
- Unidata's **network common data form (netCDF)** is used by scientific organizations around the globe. A sampling includes: Asia-Pacific Data Research Center, CSIRO Marine and Atmospheric Researcher Remote Sensing (Australia), AVISO (France), Federal Waterways Engineering and Research Institute (BAW) (Germany), International Centre for Theoretical Physics (Italy), and European Climate Assessment (ECA) (Netherlands).
- The **National Water Model** (NWM) simulates observed and forecast streamflow over the entire continental United States. Several Unidata technologies are used in connection with the NWM and at the National Water Center (NWC) in Tuscaloosa, Alabama, including netCDF, used for output from the NWM, and LDM, used to acquire data for NWM initialization and to transfer model output to NOAA's National Operational Model Archive and Distribution System (NOMADS).
- The Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) **HydroShare project** uses netCDF to represent multidimensional content and a THREDDS data server to add value to multidimensional data netCDF format in HydroShare.
- **Use of TDS in the ocean community.** Because the TDS easily supports a variety of data types (gridded, point, trajectory, etc.) that are widely used when studying the ocean environment, it has been enthusiastically embraced by the ocean community for delivering data both internally (for local researchers) and to the public. Some examples of active TDS maintained by ocean researchers are: USGS Woods Hole Science Center, Integrated Ocean Observing System (IOOS), Ocean Observatories Initiative (OOI), The HYbrid Coordinate Ocean Model (HYCOM), and the international Argo program.

Building Community

Unidata undertakes a variety of activities with the goal of building a vibrant community in the geosciences and beyond. The following are a sampling of activities not previously mentioned:

- **Scientific Society Meetings.** Unidata staff convene sessions and making presentations at AGU, AMS, and EGU meetings as well as at other national and international conferences and workshops. UPC staff members helped create AGU's Earth and Space Science Informatics session in 2004; participation in the AGU ESSI program has grown to over 700 abstracts for the 2017 Fall Meeting. The EGU ESSI Division was formally launched in 2008 with the active involvement of UPC staff; it has grown significantly, receiving nearly 400 papers at the 2018 European Geosciences Meeting.
- **Scientific Python.** As the scientific Python ecosystem grows and matures, Unidata continues to contribute in many ways beyond development work on the MetPy and Siphon packages. Staff members have worked to evangelize the use of Python toolchains in forums like the Symposia on Advances in Modeling and Analysis Using Python at AMS Annual meetings, and have served on the organizing committee of the SciPy scientific computing in Python conference.
- **Aiding with Community Outreach.** Unidata technologies continue to help members reach out to their own communities, facilitating the provision of meteorological data through dozens of popular web sites. In the above-mentioned 2017 survey of Unidata's academic community, respondents reported that in excess of 100,000 people outside of their colleges or universities had viewed public outreach products that relied on Unidata provided data, services, or software in the past five years.

Synergistic Activities

For the benefit of the geoscience community, the UPC participates in projects and undertakes activities that are funded through non-core awards. These projects are always consistent with Unidata's mission, and synergistic with ongoing core-funded activities. We believe that these projects and the resulting collaborations are essential to maintaining a healthy program, and that many benefits accrue to our community. In addition to the intrinsic merit of the projects, non-core funding brings modest additional resources to the UPC (contributing about 10% of overall funding), relieving some of the pressure on Unidata's budget. Both the Unidata Strategic Advisory Committee and NSF have encouraged Unidata to participate in synergistic opportunities as appropriate. The UPC undertakes non-core projects only after careful analysis of their merit and benefits to our core community, and endorsement by the Strategic Advisory Committee.

Currently, the UPC is participating in the following projects that are not funded by NSF 1344155:

- The NASA ROSES program is funding a project to enhance Rosetta services and capabilities to work with science domain outside Rosetta and Unidata's traditional communities.
- NOAA has funded two projects to build capacity for data reception related to the GOES-R (GOES-16) and GOES-S (GOES-17) satellites.
- The NSF EarthCube program has funded a project to enhance the netCDF Climate and Forecast (CF) metadata standard and promote its use in the geoscience community.
- EarthCube has also funded a project to develop tools that allow the Python ecosystem to use the advanced visualization technologies of the IDV.
- The NOAA Integrated Ocean Observing System has provided funding to develop TDS capabilities for unstructured grids and for more general aggregations.

Vision for the Future: Enabling "Science as a Service"

In the more than three decades since Unidata's inception, advances in information technology have had a profound effect on the scientific and educational landscape in which Unidata's community members operate. From the early days of computer-assisted analysis to today's milieu of up-to-the-minute satellite observations and cloud computing, Unidata has worked to provide its community with data, tools, and services that both meet their current needs and prepare them for the changes that are already visible on the horizon. In this spirit, Unidata has focused resources on making cloud-computing technologies useful and relevant to its community and on helping members navigate the changing data management requirements of funding agencies and meet Office of Science and Technology Policy mandates for public access to research results (Holdren 2013; 2014). Geoscientists are increasingly pursuing collaborative, transdisciplinary research agendas, as recognized in NSF's *Dynamic Earth: GEO Imperatives & Frontiers 2015-2020*, which prioritizes developing "community-driven cyberinfrastructure to advance data-enabled science and education" (NSF 2014). This, too, meshes well with the feedback Unidata receives from its community members about their technology needs and the challenges they face.

Transdisciplinary data synthesis, facilitated through increased interoperability of heterogeneous, multidisciplinary, and multiscale datasets, is essential to solving grand challenge environmental science problems like climate change and prediction of natural hazards and extreme events. Unidata's vision is to serve as a catalyst to accelerate geoscientists' ability to characterize, understand, and predict the complex Earth System by enabling data-centric as well as data-proximate scientific workflows.

Recognizing widespread public interest in the openness and accessibility of the scientific process, the National Science Foundation's Strategic Plan for 2018 – 2022, *Building the future: Investing in Discovery and Innovation*, emphasizes the use of best practices to ensure that research is reproducible, including emphasizing the open availability of results and the data that support them (NSF 2018). The urgent need to promote open and reproducible science is echoed in the recently released National Academy of Sciences report *Open Science by Design: Realizing a Vision for 21st Century Research* (NAS 2018). Unidata has long been a champion of open, community-focused techniques, and our plans for the coming years revolve around a concept of reproducibility that naturally integrates data, code, workflows, and other investigative artifacts into the scholarly record. Emerging ideas like the FAIR Data Principles (Findability, Accessibility, Interoperability, and Reusability) (Wilkinson 2016) serve as guideposts as Unidata works to help make the scientific process more open and available to all geoscientists. Broadening participation in Science, Technology, Engineering, and Math (STEM) fields is another key driver of our planning, and Unidata is committed to helping universities prepare a diverse and globally engaged STEM workforce by enhancing access to geoscience data, tools, and training, and lowering the barriers to the use of cyberinfrastructure.

Given all of these motivating ideas, combined with the challenges inherent in working with very large datasets, distributed research teams, and educational environments with constrained computational resources, we have gravitated to the concept of *Science as a Service* as a promising approach to making data- and user-centered end-to-end workflows more easily accessible, customizable, and reusable. The vision for *Science as a Service* is a future in which research is more efficient, creative, open, and collaborative,

Our *Science as a Service* vision draws together Unidata's ongoing work to provide geoscience data and software for analysis and visualization with access to workflows designed to take advantage of cloud computing resources. The approach we envision encompasses a wide range of scenarios — from essentially turn-key online resources for classroom use, relying on data and software provided directly by the Unidata Science Gateway in NSF's Jetstream cloud — to highly customizable, scalable, portable, and reproducible workflows for researchers and teams, drawing together data and software from disparate sources and melding them into scientific artifacts that can be efficiently shared with others and archived for the benefit of future investigators. We see this vision as a natural evolution of our work in the previous five-year period of performance, which aimed at making cloud-based tools and resources more available to the Unidata community.

Two ideas that go into the *Science as a Service* concept are particularly compelling. One is the opportunity this paradigm affords to bring together disparate existing and emerging data sources so that they can be studied in conjunction with one another easily and without the need to move data around the network. We see data-proximate analysis as a key concept for the usability of these (potentially very large) datasets. Second, we consider the idea of a digital "science workbench" or "makerspace" to be key to democratizing data access, science education, and research. The online environment we envision will provide users with access to data and common tools (computing resources and software) as a starting point, but also allow them to bring their own data and tools as well. In an environment where use of common tools and data is the norm, participants will be encouraged to share techniques, ideas, and results with each other directly.

To be clear: while the *Science as a Service* idea is well suited to use of cloud-based resources, nothing in Unidata's plans require that community members work in any specific computing environment, cloud-based or otherwise. We will continue to make data and software available for download for community members who prefer to work in a local environment, and we will continue to strive to support workflows that can combine use of local and remote resources.

While we are working towards these solutions with today's technologies — the Unidata Science Gateway is already making tools like Jupyter notebooks and Docker containers useful to Unidata community via Jetstream resources — we recognize that the technological environment is in constant flux, and that the solutions we end up with may be very different from what appears possible today. The core concept, however, remains unchanged: that Unidata can provide tools and workflows that make scientific research and educational activities easier, less expensive, and more robust for our large community, letting them get down to the business of doing and teaching science rather than spending time trying to manage data and software.

Unidata envisions a future in which researchers spend less time finding, accessing, and processing data, allowing them to focus on analysis and interpretation — and publication of their results. We believe that our vision of "Geoscience at the speed of thought through accelerated data discovery, access, analysis, and visualization" dovetails with the NSF's *Harnessing the Data Revolution* initiative. If Unidata is successful in working with its community and collaborators to foster the use of modern, scalable, cloud-optimized, data-proximate tools and workflows, the tools and techniques developed will fit seamlessly into NSF's vision of a national-scale research environment.

Plan of Action

For Unidata, balancing our community's requirements for data, tools, and services in the present with the activities needed to ensure they will be able to meet future challenges is an ongoing, everyday process. We strive to allocate resources to maintain and enhance programs our community finds useful, while embarking on forward-looking initiatives that capitalize on advances in technology and

respond to the evolving scientific landscape. When resources are limited, there is often a temptation to prioritize today's needs over plans for tomorrow. But without constant innovation and experiment, the tools and services in use today can become obsolete or irrelevant.

Unidata's strategic plan defines the program's mission, vision, and goals, and informs our decisions about how to allocate resources between these two imperatives: to continue doing what is useful now and to figure out what will be important tomorrow. This section describes Unidata's strategic focus areas, along with actions planned for the near term and activities envisioned to take place over a longer time span. (You can read the strategic plan online on Unidata's web site if you wish.) The descriptions below indicate whether the activity supports an ongoing service (sustaining), adds new features to an existing package or service (enhancing), or adds a new package or service to complement Unidata's existing portfolio (innovating).

Focus Area: Managing Geoscience Data

Data volumes have expanded tremendously in the past decade. For example, the volume of data emanating from each GOES-R class satellite is at least 20 times larger than that the older GOES-IP satellites. Numerical Weather Prediction models have increased their resolution in both spatial and temporal domains, as well as generating more realizations of those predictions, with resulting increases in output volume. Unidata has worked to help its university community members, who have been integral participants in the push-based IDD network, adapt to rising data volumes by fine-tuning the data sent to them and by taking advantage of pull-based remote access workflows based on the OPeNDAP and other remote access protocols.

Making Geoscience Data Accessible

Distributing real-time meteorological data is a foundational Unidata activity and remains one of its highest priorities. Unidata will work to make geoscience data more accessible by:

- Continuing to operate a robust IDD network, distributing NWS data, high resolution model output, operational satellite data, volume scan radar data, and other relevant data sources. (sustaining)
- Promoting remote access methods that make it possible to use the relevant portions of a dataset without the need to store the entire dataset locally. To this end, we will continue to improve and promote technologies including AWIPS/OGC Data Delivery, the TDS, and ADDE. (sustaining, enhancing)
- Improving the reliability and scalability of remote access servers like the TDS to better handle terabyte- and petabyte-scale datasets. Enhancements such as GRIB FeatureCollections, parallel processing workflows, and data streaming protocols promise to allow access by a wider range of users. (enhancing, innovating)
- Incorporating additional storage technologies like cloud-based object store, zarr, or tiledb into the netCDF data model, enabling native access to cloud storage systems and providing alternative file-based access methods. (enhancing, innovating)
- Working to implement recent NSF-supported Unidata research on multicast technologies, potentially allowing any source on the IDD network to reliably supply real-time geoscience data to an arbitrarily large number of recipients. (innovating)

Making Geoscience Data Discoverable

Unidata will work to make existing geoscience datasets easier to find by:

- Fostering community adoption of standard data discovery services to help users locate and acquire data appropriate for their projects. (enhancing)
- Working to ensure that data available from Unidata-managed services are discoverable through standard data discovery mechanisms. (enhancing)
- Investigating incorporating techniques like those in Google's Dataset Search and EarthCube's Project 418 that help expose geoscience data to common search engines into Unidata services, using Schema.org markup of datasets. (enhancing)

Making Geoscience Data Usable

Unidata will work to make geoscience data more usable by:

- Promoting the use of data standards like netCDF with the Climate and Forecast (CF) conventions that allow scientists to quickly understand the shape and provenance of datasets. (enhancing)
- Using technologies like Jupyter notebooks and features of the DAP4 protocol to improve data-proximate workflows, in which processing and analysis are done close to the data rather than moving data close to the processing tools. (enhancing, innovating)
- Enhancing analysis and visualization tools to make interactive exploration of results fast and easy, with outputs like IDV bundles, AWIPS procedures, and browser-viewable 3D models. (innovating)
- Enhancing the IDV's ability to visualize and analyze complex dynamic and physical processes in three dimensions. (enhancing)
- Investigating ways that Unidata tools can take advantage of "Linked Data" best practices for exposing, sharing, and connecting datasets across networks. Collaboration with EarthCube's Project 418 offers an avenue for progress. (innovating)

Enhancing Community Access to Data

Unidata's ongoing efforts to ensure access to important geoscience datasets will continue, with special emphasis on:

- Collaborations with government science agencies to gain community access to datasets like those generated by the GOES-R series satellites. (sustaining, enhancing)
- Collaborations with Unidata university researchers to secure access to value-added datasets like those being developed for GOES-R GLM lightning data at Texas Tech University. (enhancing)
- Creating a testbed for algorithms designed by community researchers to create Level 2 products for GOES-R class satellites. Such a testbed could lower the barriers to testing the viability of new algorithms and products. (enhancing, innovating)
- Continued operation of downlinks for operational GOES-East and -West platforms, along with dissemination of all products to the community via existing push (IDD) and pull (ADDE, TDS, and AWIPS EDEX) services. (sustaining)
- Collaboration with NCAR Computational & Information Systems Laboratory (CISL) to enhance community access to NCAR-hosted datasets. (sustaining, enhancing)

- Collaborations with commercial-sector data providers to secure access to datasets available through public-private partnerships like NOAA's Big Data Project. (sustaining)
- Continuing Unidata's strong collaboration with the NSF Jetstream project to make ready-to-use cloud computing resources available to the university community. (enhancing, innovating)
- Advocating with stakeholders such as scientific societies, universities, and standards organizations like OGC and ISO for clear community standards around data sharing and access. (sustaining)
- Extending key existing tools such as Python libraries to facilitate ease-of-access to publicly accessible data stores that may or may not be backed by Unidata technology. (enhancing)
- Enhancing the IDV to consume GOES-R GLM data available through remote access servers such as the TDS and ADDE. (enhancing)

Focus Area: Providing Useful Tools

To pursue meaningful scientific inquiry using the abundance of data available, researchers and educators need usable and freely available software to access, analyze, manage, and visualize data. Evidence indicates that the most robust solutions come about through "agile" software development processes — that is, incremental and iterative evolution involving feedback from an engaged user community. Unidata has long served as such a center of collaboration, building open source and freely available software tools in a process guided by input from researchers and educators who use Unidata tools daily.

One refrain we hear from our community is make the process of accessing data simpler. Unidata has focused on building tools that let practitioners spend less time on tasks like searching for data or managing software and more time studying data and investigating the information it contains. Our efforts to bring robust and standardized tools for working with geoscience data to the Python ecosystem, to expand the capabilities of the National Weather Service's AWIPS package for use in research and education, and to help researchers design smooth and useful data management workflows that promote collaboration and reproducibility are all examples of Unidata's efforts aimed at reducing time to science.

Promoting 21st Century Scientific Workflows

Unidata will work to broaden community adoption of modern analysis and visualization tools by:

- Enhancing offerings of web-based tools that minimize the need for local hardware and software. (enhancing)
- Expanding offerings of software that promotes the use of cloud-computing workflows. (enhancing, innovating)
- Improving the ability of Unidata technologies to take advantage of high-performance computing environments. (enhancing)
- Providing professional-grade and cloud-hosted software tools like AWIPS, IDV, and geoscience-focused Python libraries that leverage modern cyberinfrastructure such as Docker containerization technologies and Jupyter notebooks. (enhancing, innovating)
- Investigating the creation of fully web-based data analysis tools, potentially leveraging open-source Python and JavaScript libraries and the Jupyter notebook. (innovating)
- Enhancing our existing suite of geoscience-focused Python libraries to ensure that advancements in the broader ecosystem of scientific Python and Jupyter notebooks are available as simply as possible to educators, researchers, etc. (enhancing)

- Continuing to enhance the Unidata Science Gateway on Jetstream with features such as a JupyterHub server with proximate access to IDD data holdings. (enhancing, innovating)
- Designing “installation ready” Cloud-based solutions where complex software environments can be configured by Unidata experts for easy use by community members. (enhancing)

Creating Modern Data Management Workflows

Unidata can lead the community by developing tools and workflows geared toward promoting open and reproducible science through data-proximate analysis and visualization and robust management of project data. Unidata will work to foster adoption of modern data workflows by:

- Continuing to advance data server technologies such as OPeNDAP and the THREDDS Data Server that facilitate remote data access. (enhancing)
- Collaborating with NCAR CISL to explore new services such as a “Geoscience Digital Data Resource and Repository Service” (GeoDaRRS) to support NSF-funded researchers in meeting data archiving requirements. (innovating)
- Building, demonstrating, and promoting workflows that facilitate long-term reproducibility of scientific analysis, by making it easier for others to replicate the analysis. (enhancing)
- Extending existing tools, such as Rosetta, to enable long-tail data providers to produce files that conform to widely adopted, community driven data and metadata standards. (enhancing)
- Exploring the creation of cloud-native data access services, based around common cloud technologies like object store, hosted databases, and “serverless” technologies to that simplify access to collections of large data in the cloud. (innovating)
- Continuing to create and maintain containerized versions of Unidata tools such as the TDS, LDM, and IDV. (sustaining, enhancing)

Supporting Legacy Workflows

Unidata will continue to support existing scientific workflows by:

- Providing ongoing technical support for existing packages and processes that are currently in active use in the geoscience community. (sustaining)
- Exploring the transition of legacy workflows to use cloud-computing resources and browser-based solutions. (innovating)
- Demonstrating, and advocating for, techniques for archiving software to enhance long-term reproducibility of scientific research. (enhancing)

Focus Area: Supporting People

Since its inception, Unidata has focused on supporting community members in their scientific endeavors. Our training and technical support activities have helped researchers and educators build the technical skills needed to advance their data-intensive scientific inquiries. By working to expand the Unidata community, we hope to increase access to the tools and techniques enabling scientific discovery. Experience shows that community-based standards — processes that arise from practicing scientists doing real work — are most likely to benefit a broad spectrum of participants.

Direct participation in Unidata’s governance by members of the academic geoscience community grounds the program in the realities of teaching and doing science. In turn, this high level of involvement by researchers and educators allows Unidata to advocate effectively with science agencies, standards organizations, and other entities to benefit the broad community.

Unidata also serves in the role of convener, bringing stakeholders at many levels together to discuss data-related issues, community needs, and large-scale trends in the cyberinfrastructure landscape. Using venues like scientific meetings and data-focused conferences, we seek to bring people together to address issues faced by all.

Providing Technical Support

Direct interaction between users and software developers at the UPC benefits all phases of the development process. Unidata is committed to maintaining the high level of technical support on which community members have come to rely (sustaining, enhancing). We will do this by:

- Providing authoritative replies to support inquiries
- Establishing and managing community e-mail lists
- Participating in open source development forums like GitHub and StackExchange
- Interacting with community members one-on-one and during visits to community sites
- Providing individualized technical support for core Unidata sites on a case by case basis
- Providing Unidata Science Gateway tools and technical support for workshops and meetings using Unidata Jetstream resources.
- Enhancing software documentation and training materials

Building Capacity through Open Source Software Development

Software developed by the UPC has always been freely available. Unidata aims to increase community participation in this development by:

- Encouraging community participation in software development, testing, and documentation through mechanisms like GitHub. (sustaining, enhancing)
- Participating in Free and Open Source Software (FOSS) initiatives and advocacy efforts. (sustaining)
- Ensuring robustness of the scientific computing ecosystem by participating in development of key upstream open source projects, both as core developers (e.g. Matplotlib, CartoPy, and VisAD) as well as contributors. (sustaining)

Building Community Cyber-Literacy

Unidata's long-standing software training and cyber-literacy programs continue to evolve. For the benefit of our user community, and with an eye toward the development of the next generation of geoscientists, we will:

- Provide software training workshops at the UPC and at member institutions. (sustaining)
- Find new ways to support community members in attending Unidata training. (enhancing)
- Enhance our distance learning and online training materials, with an emphasis on interactive, programming-based materials such as Jupyter Notebooks. (enhancing)
- Continue to convene the community through meetings and workshops to discuss and study cyberinfrastructure topics. (sustaining)
- Continue our Summer internship program to bring students to the UPC for immersive software development training. (sustaining)
- Establish a Visitor program to bring students and faculty to the UPC to collaborate with staff for shorter time periods. (enhancing)

Setting Priorities

Unidata is an established program with a large and diverse member community. Members are accustomed to receiving a variety of data streams, software tools, and support services, and each student, professor, department, or institution can have different ideas about what is most important for the program to provide. Accordingly, our Plan of Action touches on numerous projects that can be justified by one of the following statements:

- Our community depends on Unidata to provide reliable and robust data services, tools for analysis and visualization, and user support services. Meeting these needs is our highest priority.
- In order to continue meeting these needs with limited resources, we must develop and introduce our community to new ways of accessing data and using software tools.
- Collaborating with community members to develop software can both bring additional development resources to bear and focus efforts on projects that are most important to users.

While we would like to be able to accomplish everything we set out to do — AND all of the additional things that will present themselves as the technological landscape continues to evolve — we recognize that resources must be assigned to the projects that are most important to our community first. The following table attempts to delineate activities described in our plan that we deem to be critical to our community, along with those we feel could be delayed if resources are constrained.

Focus Area	Activities Critical to the Core Effort	Activities to Delay if Necessary
Managing Geoscience Data	• Continue management of the IDD, adding new data streams as needed. • Incorporate cloud-based technologies such as object storage into Unidata tools. • Improve community access to data-proximate workflows.	• Development and deployment of multicast LDM to IDD network • Work with collaborators to advance data discovery capabilities. • Creating an algorithm design testbed
Providing Useful Tools	• Support community use of AWIPS. • Enhance offerings of web-based data analysis and visualization tools. • Enhance existing suite of geoscience-focused software libraries written in C, Python, and Java.	• Improving Unidata tools for HPC environments • Promote workflows that facilitate long-term reproducibility of scientific analysis
Supporting People	• Provide expert assistance on software, data, and technical issues. • Encourage community participation in software development, testing, and documentation. • Convene the community through meetings and workshops.	• Provide Unidata Jetstream resources for classroom and workshop use • Extensive participation in “upstream” open source projects.

When making key decisions on setting priorities, allocation of resources, and evolving or phasing out products and services, Unidata will continue to seek input from its users as well as guidance from its governing committees in order to achieve the best outcome, maximizing benefits to our core community while ensuring that progress is made toward the vision articulated in this proposal.

Closing Remarks

Midway through its fourth decade of providing geoscience data, tools, and services to the university community, Unidata continues to evaluate and respond to changes in the scientific, educational, and technological environment in which our community members operate.

Modern cyberinfrastructure, which we envision as seamlessly integrating data, software, and services, has become a critical element in today's data-intensive world of geoscience research. Just as crucially, cyberinfrastructure-enhanced learning, discovery, and collaboration methods hold promise for broadening participation and increasing diversity along individual, geographical, and institutional dimensions.

The plan articulated in this proposal positions Unidata to serve as a catalyst, accelerating geoscientists' ability to characterize, understand, and predict the complex Earth System by enabling data-centric as well as data-proximate scientific workflows. Through the proposed activities, Unidata will work toward a future in which researchers spend less time finding, accessing, and processing data — allowing them to focus on analysis, interpretation, and publication of their results. Our vision for *Science as a Service* is a future in which research is more efficient, creative, open, and collaborative.

As we work to lower the barriers to effective use of contemporary cyberinfrastructure and facilitate scientific progress, Unidata will:

- Enable discovery of and provide access to rich collections of geoscience data
- Promote modern scientific workflows, including data-proximate analysis and visualization
- Provide an ecosystem of cloud-optimized tools and a framework for data access, use, and sharing
- Contribute to scholarship in data science
- Facilitate open and reproducible science
- Foster community-based open source software development
- Broaden participation through concerted efforts to introduce Unidata resources and capabilities to universities with underrepresented student populations
- Advance NSF and community goals to increase data, cyber-, and geospatial literacy, contributing to the development of the next-generation workforce

As evidenced by the letters of collaboration included in Appendix A, Unidata maintains its commitment to deepening collaboration with its broad array of partners. As we work toward the bold and ambitious goals outlined in the proposal, we will leverage those collaborations — and continue to draw upon our highly effective governance system — to advance our goals, our profession, and our community.

We are confident that the result of the goals Unidata is setting for itself will be a geoscience community that is better able to use and learn from the rapidly expanding volume of Earth System data. Wider adoption of integrated, data-centric workflows will promote transparency, reusability, and long-term stability of scientific knowledge. We believe that Unidata's plans are in close alignment with the goals of NSF's *Harnessing the Data Revolution* Big Idea, and through our activities to expand our profession's human capacity as a "21st century data-capable workforce" we will come closer to the vision elucidated in the National Science Foundation's Strategic Plan of

"A Nation that is the global leader in research and innovation."