The Arctic System Science Data Coordination Center

Playing a Vital Role in ARCSS Knowledge Synthesis

This article was prepared by Nancy Auerbach, ARCSS Data Coordination Center (ADCC), National Snow and Ice Data Center, University of Colorado, Boulder; Yarrow Axford, Institute of Arctic and Alpine Research, University of Colorado, Boulder; Rudy Dichtl, ADCC; Michael Hartman, National Climatic Data Center; Chris McNeave, ADCC; Betsy Sheffield, ADCC; and Keri Webster, ADCC.

The Arctic System Science Data Coordination Center (ADCC) is the central, long-term archive for data collected by the NSF's ARCSS program. The ADCC is located at the National Snow and Ice Data Center (NSIDC) at the University of Colorado, Boulder, a national information and referral center supporting polar and cryospheric research. The primary goals of the ADCC are to collect and preserve ARCSS data and to make those data easily accessible to ARCSS investigators, the scientific community, policy makers, and the general public.

As indicated by the program name, ARCSS promotes the concept of system science, or the integration of knowledge from various Arctic disciplines. This synthesis depends on the accessibility and exchange of data among members of the scientific community. Published research findings focus on final results and target a specific discipline, but researchers can use raw data for reasons that differ from the original reason for collecting the data. Because of this potential, a long-term archive of ARCSS data, with its accompanying metadata, is valuable to the ARCSS program in particular and the scientific research community as a whole.

Having a central point of focus for ARCSS and its data is essential to a program aimed at understanding global change. This is especially true because the ARCSS program focuses on the science of environmental systems and is defined geographically, rather than by discipline. ARCSS supports many projects collecting a diverse set of data.

The ADCC ensures public access to ARCSS research data. It maintains high standards for data management to meet the needs of the current users of the ARCSS data collection, as well as to ensure the long-term viability of the data.

History

NSIDC received its first grant for ARCSS data management in 1990 for support of GISP2. NSIDC also received individual data management grants in 1991 and 1992 for the OAII and LAII programs, respectively. Following the success of these grants, and with the increasing interest in global climate change research, the need for a continuous record of collected environmental data became clear. In 1994, NSIDC received the first grant for creating the ADCC, a distinct entity within NSIDC charged with managing and archiving all ARCSS data. NSF granted a continuation award of the same title in 1997 and a supplemental award in March 2001.

The ADCC remains housed within NSIDC, which has a long and successful history of data management. NSIDC's roles include:

- Serving as one of eight Distributed Active Archive Centers funded by NASA;
- Acting as a national information and referral center supporting NOAA's National Environmental Satellite, Data, and Information Service;
- Providing data and information services to the user community, publishing reports and a quarterly newsletter, and maintaining a large library collection of monographs, technical reports, and journals;
- Contributing to international programs concerning the cryosphere and its role in climate; and
- Supporting the NSF through the ADCC and the Antarctic Glaciological Data Center.

ADCC's relationship with NSIDC is highly beneficial to the ARCSS program. As an umbrella organization, NSIDC has a larger staff that provides operational support, user services support, tape archiving, computer system administration, administrative assistance, writer services, programming support, and other services, on a cost-reimbursable basis. The ADCC also benefits from shared computing and network infrastructure, as well as common data management policies and procedures. Cost sharing and collaboration

with NSIDC extends the ADCC's capabilities far beyond what ARCSS funding could support alone.

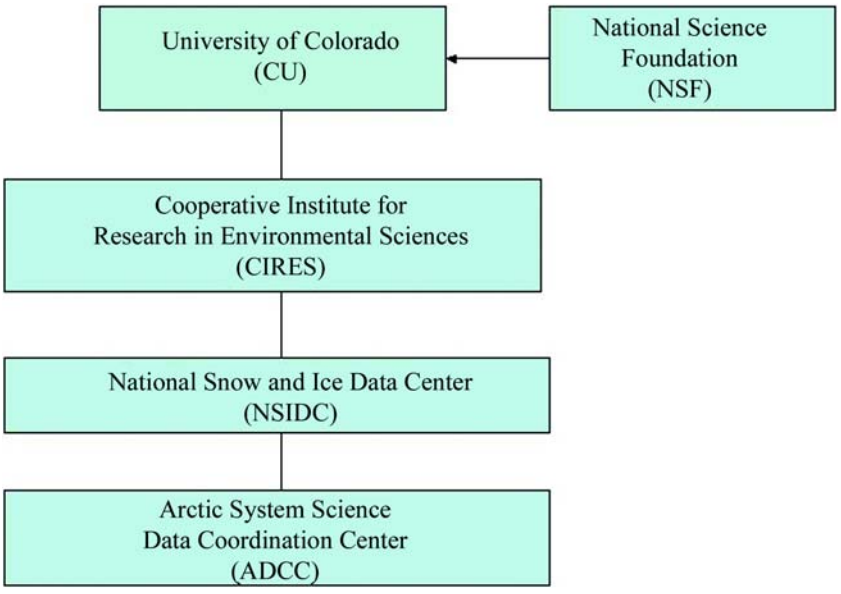
Until the ADCC was founded, data were usually made publicly accessible, at the discretion of the investigator, through structured data centers. Published final results were available, but the raw data were not distributed as frequently. Therefore, the ADCC focused initially on collecting raw data from the then-small community of ARCSS scientists. These data were usually structured or documented in a manner easy for the principal investigator (PI) to understand and use. However, the data were often difficult for other users to understand without the involvement of the collecting investigator to help define variables, column headings, or data location information. As a result the ADCC began to request that ARCSS PIs submit substantive metadata (information about data) to the ADCC along with their data. This shift in focus from collecting only raw data to collecting the supporting metadata as well has helped to ensure long-term viability of ARCSS data.

Data Management and Processing

Researchers must consider data management as early as possible in the development of new research projects. Therefore, the ADCC is involved in all ARCSS science committees. The ADCC participates in planning new ARCSS initiatives to offer data management insights and anticipate future archive needs. When developing data management plans for evolving ARCSS projects, the ADCC considers the broad range of program needs to determine how incoming data will be processed and distributed.

The ARCSS Data Protocol is well established and has been widely adopted by the various ARCSS Science Management Offices (SMOs). The key assertion of the data protocol is that all ARCSS-funded data are considered "community property." ARCSS PIs will retain exclusive use of their data for one year after collection. After the first year, their data are released to other ARCSS investigators, and after two years the data must be released to the public domain via the ADCC. The data protocol is available at http://arcss.colorado.edu/arcss/protocol/protocol.html.

The relationship of the ADCC to the NSIDC and NSF.



Working with PIs

Initially the ADCC conducted project and all-hands meetings as opportunities to solicit data. Although capable of reaching large audiences, this method of contact was inefficient and didn't allow for more personal interactions. As a result, in July 2000 the ADCC began sending "first contact" email messages to PIs of newly awarded grants, offering congratulations and laying the groundwork for later data submission. These messages acquaint PIs with the ADCC and make preliminary recommendations for appropriate data formats to ensure maximum flexibility and data sharing in the future. They also direct PIs' attention to the ADCC web site and the online metadata submission form.

ADCC subsequently sends a second contact letter to all PIs reminding them of the need to submit their data to the ADCC. This occurs about a year after the first contact letter, presumably after the first field season has passed and after the PI has had the opportunity to plan for data submission. In the second letter the ADCC requests information about the expected types of data and their projected date of availability. Once that date arrives, the ADCC contacts the PI to plan for data transfer and to request that the PI submit metadata via the online metadata submission form.

Working with JOSS

The ADCC collaborates closely with the Joint Office of Science Support (JOSS), a University Corporation for Atmospheric Research (UCAR) organization, and the ARCSS SMOs to develop

data management procedures that define the roles of both JOSS and the ADCC and to communicate these procedures to each ARCSS PI. A current example of this collaboration is the coordinated effort to provide the Arctic-CHAMP project SMO with a complete data management plan to meet all of Arctic-CHAMP's requirements. JOSS and the ADCC have unique and complementary services that serve the specific needs of the PIs among the 18 awarded CHAMP projects and the SMO.

PIs who are involved with certain large ARCSS projects, such as SHEBA, ATLAS, and ITEX, initially submit their data to JOSS (http://www.joss.ucar.edu/arcss). JOSS maintains a restricted-access, interim archive for near-real-time data presentation. JOSS's primary role is early data collection to provide field support by redistributing project data back to other project scientists for their immediate review and evaluation. During this period of initial collection, the ADCC and JOSS meet to plan and organize the pending transfer of data and metadata. Often data are in a preliminary state, and the ADCC works closely with JOSS to coordinate the final effort to bring the data up to the standards required for long-term archival. Once a field project has concluded, JOSS transfers its holdings to the ADCC, and the ADCC works with individual PIs to finish bringing data and documentation up to the final standards.

Overall the function of JOSS is to make data physically accessible as quickly as possible and to provide a variety of supporting data from other disciplines (for example, weather forecasts for in-field operational areas). Once data are transferred to the ADCC, the goal is to make data sets intellectually accessible as well as physically accessible. The ADCC ensures that data sets go through a thorough quality assurance process and that the associated metadata and documentation are complete and accurate.

Receiving Data and Metadata

Metadata must accompany data in a long-term archive to ensure the longevity of the data as well as their usefulness to a broad audience. Numerous examples demonstrate that data can become useless if relevant metadata are missing. Metadata may include information such as variables measured, spatial and temporal coverage, data format, and weather conditions during data collection. The ADCC's emphasis on metadata and documentation is one important factor that sets it apart from other archives that lack a long-term perspective.

The ADCC has created an online metadata submission form that streamlines the metadata submission process, making metadata compilation more efficient for PIs and the ADCC. Furthermore, the form *requires* that PIs submit at least a minimum of metadata, ensuring that the ADCC obtains appropriate metadata early in the data submission process. PIs access the metadata submission form through the ADCC web site (http://arcss.colorado.edu/forms/arcss_submit.html). PIs may submit raw data in any form but must use the metadata submission form to send their metadata. The data coordinator reviews both the metadata and the data to ensure that they are complete and that the information provided properly describes the data. The coordinator immediately asks the PI about any incomplete or inconsistent items, so the PI can make corrections.

While the quality control of actual data values is the responsibility of the PI collecting the data, the ADCC performs quality assurance (QA) of the metadata and overall data set. This QA ensures the long-term viability of the data set for users other than PIs who have intimate knowledge of the data or project. The ADCC conducts a thorough check of the data's integrity, accuracy, column headings, and units, comparing them with the documentation that the PI submits with the data.

Creating Documentation

Following the QA process, the ADCC project writer prepares the data set documentation. To assist PIs in meeting metadata requirements, the ADCC produces two types of data set documentation, based on information that PIs submit with their data: a data interchange format (DIF) file and a summary document. Both standardize a data set's metadata, increase ease of use, and allow future access to that data set. The DIF is the metadata file format used by NASA's Global Change Master Directory (GCMD) and is compatible with International Standards Organization (ISO) and Federal Geographic Data Committee (FGDC) metadata standards. The ADCC submits every DIF it writes to the GCMD, which is part of the U.S. Global Change Research Program. Thus, every ARCSS data set appears in the GCMD's searchable database and is visible to the broader global change research community. In writing and submitting DIFs to the GCMD, the ADCC provides

an additional service to PIs, who would otherwise be required by NSF to do so themselves. The DIFs are the internal NSIDC metadata standard and are used to automatically generate the ADCC's online data catalog entries for each project.

The summary documents contain all available metadata for each data set and are often much more comprehensive than the DIFs, providing detailed information about data collection methodology, file structure, tools for accessing data, and other data characteristics. They also describe how to appropriately cite the data set, with the scientists as authors and the ADCC as publisher. Documentation files are packaged with the data and are accessible via the online data catalog entries.

Once the data and documentation are ready for publication, the ADCC team and NSIDC user services personnel review the data set. They identify problems or questions that a user may encounter and take the appropriate corrective steps, in collaboration with the PI, prior to the release of the data on the file transfer protocol (ftp) site or on CD-ROM.

Providing Data Security and Interoperability

The ADCC has established a system for maintaining backups that are safely stored, technologically current, and accessible. This system includes archive files, regularly scheduled backups, off-site backup tape storage, and strategic planning for upgrades. The ADCC creates an archive of each ARCSS data set that includes data and documentation preserved as originally submitted, as well as the quality-assured versions presented on the ADCC web site. A full backup of the ADCC server is performed monthly, and backup tapes are rotated to an off-site location on the University of Colorado's main campus. Additionally, ensuring that a data set remains uncorrupted while it is transferred to ADCC, used, and maintained is important. The ADCC is investigating the use of algorithms to determine the continued integrity of all data sets within its holdings.

Assuring long-term data security is a top priority, so the ADCC participates in NSIDC's long-term archive committee planning. The committee is responsible for defining NSIDC's data management policies, assuring interoperability between systems, and developing and implementing strategic plans for long-term management of NSIDC-hosted archives. Additionally, security plans are in place for all NSIDC systems. System components are reviewed periodically for risk assessment. Systems are monitored daily for intrusion detection, and relevant Computer Emergency Response Team (CERT) bulletins are addressed as soon as possible.

Special Data Handling

In addition to data products archived within the ADCC, various programs and PIs have data that require special handling. Some projects archive their data with alternative permanent archives, such as those maintained for the Long-Term Ecological Research (LTER) program. PIs working on such projects are required to provide the ADCC with enough information so that it can maintain referring links to the alternative archive locations as well as list the data in the ADCC online data catalogs and within the GCMD.

Similarly a PI may have a unique distribution site to which data are frequently added and updated. The ADCC provides a link to this site and lists it in the online data catalogs. However, because of the dynamic nature of such a collection, rather than duplicate the collection and attempt to maintain version control, the ADCC periodically downloads the PI's content for back-up and storage in the event that the originating site ceases operation or experiences unrecoverable loss.

Model output is another special case that the ADCC takes into account. ARCSS research includes the development of environmental models that provide a unique set of challenges for data management. The ARCSS Model Output Protocol addresses these concerns. The policy stipulates that PIs who are developing models should provide the ADCC with descriptions of their models so that other investigators are aware of what work is being done, and PIs should work with the ADCC to determine which model output data are suitable for archiving. For output to be archived at the ADCC, the model must be published in a peer-reviewed publication, fully documented, and in a completed state. If these criteria are not met, or if the output is too voluminous to be feasibly distributed, the ADCC directs users to the PI's web site or an appropriate contact person.

The HARC project is one that has unique data management needs. Handling HARC data involves additional planning by the PIs and the HARC SMO because of the potential sensitivity of social-science-related data and information. PIs,

working with the SMO, must decide which data are to be transferred to the ADCC for open distribution. Additionally, for any HARC data set that is archived at the ADCC, PIs must supply documentation of the informed-consent process and proof that all regulations concerning the protection of human subjects have been met.

The PARCS SMO has its own data management system and archive. That office is responsible for archiving all PARCS data and for interfacing with other ARCSS programs and global change data efforts. The ADCC does not list individual PARCS data sets in its catalog but provides a link directly to the PARCS data archive (http://www.ngdc.noaa.gov/paleo/parcs/data.html).

Non-ARCSS Data

In several cases the ADCC has accepted data that were not collected by ARCSS-funded investigators but have been identified by ARCSS PIs as important adjuncts to their data. For example, the ADCC has accepted some Scientific Ice Expeditions (SCICEX) data for its long-term archive. Even though SCICEX is funded partly by the Arctic Natural Sciences program, there was enough interest in these data among ARCSS scientists to justify collecting them for long-term support. Similarly the SHEBA Reconnaissance Imagery data set (http://arcss.colorado.edu/data/arcss200.html) is a large collection of images that

Sample ADCC data set catalog page.

nsidc.org data projects research the cryosphere news site map

NSIDC catalog search tools help features

Soil Descriptions and Soil Chemistry for LAII/ATLAS Winter Carbon Flux Sites, Alaska, 1992 and 1998-2000

Documentation Access Data

Soil profiles at 27 sites in northern Alaska were described and classified in order to provide baseline soils data for ARCSS/LAII/ATLAS (Arctic Transitions in the Land-Atmosphere System) Winter Carbon Flux studies. The investigators collected samples and field data in the summers of 1992, 1998, 1999, and 2000. Profile descriptions include horizon depth, color, texture, and other morphological properties. Accompanying sample site descriptions provide information about landforms, vegetation cover, and soil parent material. Soils are classified according to USDA soil taxonomy guidelines. Laboratory analyses include measurements of soil water content, bulk density, texture, organic matter content, carbon storage, pH, and chemical components including carbon, nitrogen, calcium, phosphorus, potassium, aluminum, manganese, and iron.

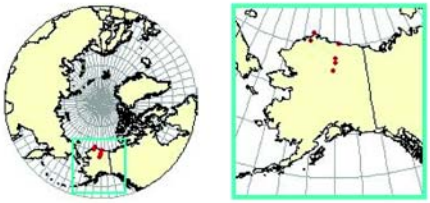
Profile descriptions, laboratory data, and soil pit photographs and profile sketches are available via ftp in MS Excel, MS Word, and .JPG formats.

Data Citation

C.L. Ping, V. Romanovsky, G.J. Michaelson, and W. Lynn. 2002. Soil descriptions and soil chemistry for LAII/ATLAS Winter Carbon Flux Sites, Alaska, 1992 and 1998-2000. Boulder, CO: National Snow and Ice Data Center. Digital media.

Data Location Maps:

Location of sample sites:



See Also:

- User Services

ADCC ARCSS catalog

Data Contributors

- PING, CHIEN-LU
- ROMANOVSKY, VLADIMIR
- MICHAELSON, GARY J.
- LYNN, WARREN

Parameters

- ALPINE/TUNDRA
- CARBON
- NITROGEN
- ORGANIC MATTER
- PHOSPHORUS
- POTASSIUM
- SOIL BULK DENSITY
- SOIL CHEMISTRY
- SOIL CLASSIFICATION
- SOIL COLOR
- SOIL HORIZONS/ PROFILE
- SOIL MOISTURE/ WATER CONTENT
- SOIL PH
- SOIL STRUCTURE
- SOIL TEXTURE

Instruments

- CARBON ANALYZERS
- CHN ANALYZERS > Carbon, Hydrogen, Nitrogen Analyzers
- PH METERS
- SOIL SAMPLER

NSIDC

The National Snow and Ice Data Center (NSIDC)

Supporting cryospheric research since 1976

CIRES, 449 UCB University of Colorado Boulder, CO 80309-0449

were adopted because of their relevance to the SHEBA project. Another example, the Physical and Chemical Properties from Selected Expeditions in the Arctic Ocean data collection (http://arcss.colorado.edu/data/arctic_ocean_expeditions/), is a growing collection of oceanographic data that will eventually include data from more than 20 Arctic Ocean cruises.

An additional non-ARCSS-funded initiative is the recent collaboration between the ADCC and the University of Colorado. The ADCC will acquire, prepare, and release high- and low-resolution imagery and a digital terrain model of Barrow, Alaska. High-resolution imagery will be available to NSF-funded researchers, while low-resolution versions of the data sets will be released to the general public. Data set releases are scheduled for later this year.

Data Presentation and Distribution

The ADCC Web Site

The ADCC web site serves as the primary tool for data archiving and distribution to scientists and the public. The web site provides an online data catalog, information about the ARCSS program and its projects, contact information for all ARCSS PIs, and an invitation to PIs and the public to contact the ADCC.

Entries in the online data catalog uniformly describe each data set, including data contributors, parameters (such as valid variables as defined by the GCMD), instruments, data citations, and related links. Most catalog entries also display sample location maps. Search tools help users locate data sets by project name, NSF grant number, name of PIs and co-PIs, and measured parameters. Additionally an NSIDC-wide data catalog search engine (http://nsidc.org/data/search. html) allows for a free-text search of any words appearing in the DIF metadata files.

Geographic Information Systems

The ADCC uses geographic information system (GIS) software as a tool to spatially represent data location or data visualization displays through maps accessed from the ADCC online data catalog pages. Maps vary, depending on the data set. GIS software is a valuable tool for the ADCC to use for creating data location maps and visualizations of data and for providing improved data access and searching. The ADCC uses the functionality of GIS to enhance the accessibility of data and does not analyze, manipulate, or alter submitted data. The ADCC can customize data sets by producing either static image maps or interactive maps served using an Internet map server (ArcIMS by Environmental Systems Research Institute).

Data Collection Location Maps

Data location maps assist users in quickly determining whether data sets are located in a region of interest. PIs provide spatial coordinates of data collection points or areas when submitting their data to the ADCC. From these coordinates, the ADCC creates standardized location maps and includes them in the online data set catalog entries. More complex data sets require other means of displaying data collection points, for example, in “rollover” maps.

Data set locations may also be displayed via interactive maps served over the Internet. The maps may be queried and subsets of the data accessed interactively through a web browser. At present, two data sets have been enabled on a trial basis using an Internet map server (ArcIMS by Environmental Systems Research Institute): Russian Historical Soil Temperature Data and SCICEX Hydrographic Data.

Data Visualization Displays

Spatial (mapped) data displays assist users in visualizing data sets that are in vector or raster formats. The ADCC produces two types of data visualization displays: thumbnail images of data sets and interactive maps. Both help users assess, before downloading, whether a data set meets their needs.

Thumbnail sketches provide a visual overview of data, whereas interactive maps provide more detailed information. For example, the ADCC has produced various ArcIMS-enabled maps in collaboration with University of Colorado researchers at the Arctic North Slope Climate Impact Assessment (ANSCIA) component of HARC. The interactive maps display spatial data to interested researchers and can inform the public (in this case, the residents of Barrow, Alaska) about research being conducted in their vicinity. The ADCC web site displays data from the ANSCIA program at two scales: the North Slope region and the Barrow vicinity.

To view sample GIS maps from the ADCC, visit the following links:

Rollover maps
http://arcss.colorado.edu/data/docs/arcss/arcss07/iloc079.html

Russian historical soil temperature data
http://adcc.colorado.edu/arcims/website/arcss078/

SCICEX hydrographic data
http://adcc.colorado.edu/arcims/website/arcss064/

Sample thumbnail sketch
http://arcss.colorado.edu/data/docs/arcss/arcss017/access017_tlk500.html

ANSCIA program: North Slope region
http://adcc.colorado.edu/arcims/website/harc_northslope

ANSCIA program: Barrow vicinity
http://adcc.colorado.edu/arcims/website/harc_barrow

CD-ROM Publication

The ADCC uses CD-ROM media to distribute large ARCSS data sets and other information that are too large to transfer online. CD-ROM publication usually coincides with the creation of a special product, acquisition of a large volume of data from one project, or a grouping of data products for ARCSS research.

CD-ROMs have several advantages for some users:

- They enable users to take data with them into the field and to use data in remote locations without Internet connections.
- They can provide educational and outreach materials to classrooms.
- Users with limited computer network capacity for data transfer (in developing countries, for example) can benefit from data provided on CD-ROMs.

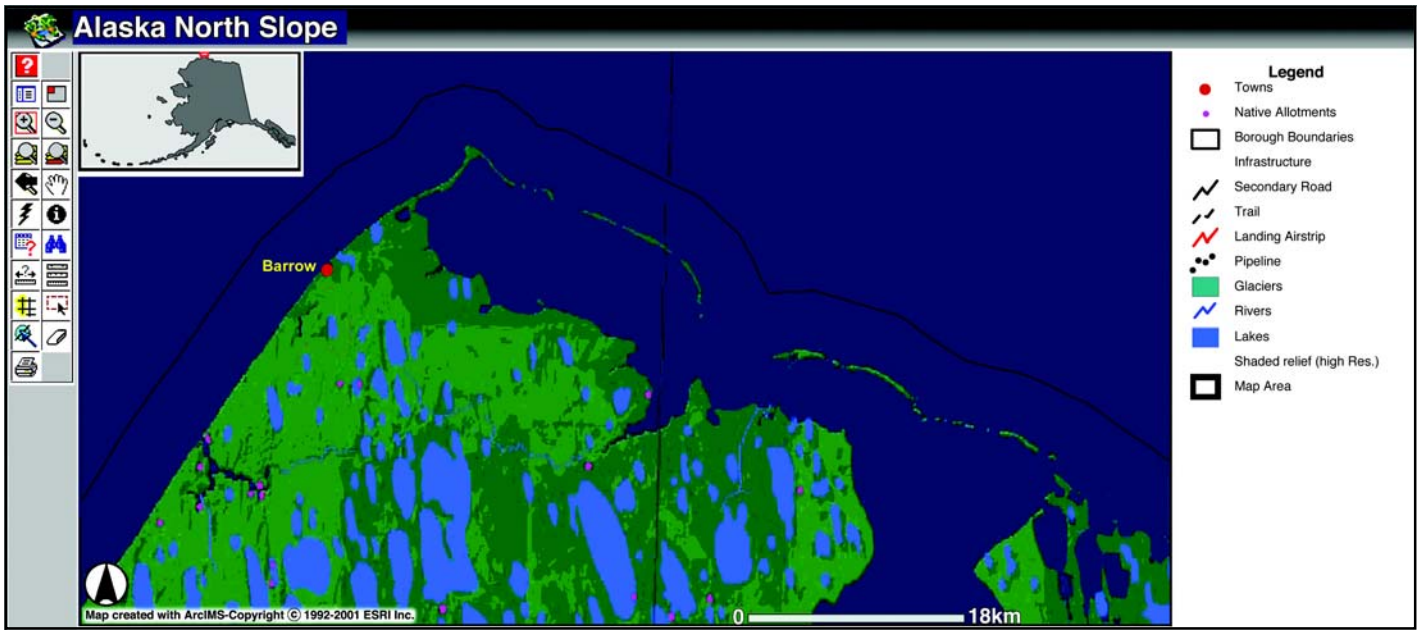
To date, the ADCC has distributed four CD-ROMs:

- *Greenland Summit Ice Cores (GISP/GRIP)*;
- *Into the Arctic*;
- *Circumpolar Active-Layer Permafrost System (CAPS)*; and
- *R-Arctic Net: A Regional Hydrographic Data Network for the Pan-Arctic Region.*

Outreach

In addition to the web site, the ADCC is involved in other efforts to communicate with PIs, the scientific community, and the general public. The ADCC makes frequent presentations at ARCSS-related meetings and general workshops to inform users, potential users, and contributing PIs about the ADCC and its data archiving and presentation goals. These presentations show how ADCC data sets are applied, the importance of complete and thorough metadata submissions, the need for long-term archiving of data sets, and the accessibility and usability of the data sets already on the ADCC web site.

To help promote Arctic education, the ADCC created the *Into the Arctic* CD-ROM. *Into the Arctic* allows teachers and students to access earth science data collected by research scientists studying climate change. This educational product provides data and information acquired from Greenland ice cores and includes lessons and activities appropriate for high school and college earth science, geography, history, social studies, and chemistry courses. This product has been



Sample interactive online map.

accepted by NASA and is listed in its 2002 catalog of approved Earth Science Education Products (http://earth.nasa.gov/education/index.html).

Future Directions

The NSF recently released a 10-year outlook for its Environmental Research and Education portfolio. The report, *Complex Environmental Systems: Synthesis for Earth, Life, and Society in the 21st Century,* was written by the NSF Advisory Committee for Environmental Research and Education. NSF Director Rita Colwell said, "NSF should move in the direction of synthesis of environmental knowledge."

To help promote this synthesis, the ADCC will:

- Continue its support of all newly initiated ARCSS projects. By identifying new initiatives early in the data management process, the ADCC will help new projects and SMOs receive the required data management support as early as possible.
- Increase focus on long-term archiving of data sets, with particular emphasis on data recovery policies and the use of backup media. The ARCSS program management has identified long-term data management as a priority.
- Contribute to the development of a data management plan that will integrate with the ARCSS Program Plan. The ARCSS Committee, made up of active PIs concerned about future research activities, will redefine key components of ARCSS and help promote research among all ARCSS programs. A special subcommittee on data management will create a data management plan, with input from the ADCC, that will integrate with the overall ARCSS Program Plan.

The concept of system science, or the integration of the Arctic sciences, depends on the accessibility and exchange of data among varying scientific disciplines. Long-term preservation of knowledge and the metadata to readily access that knowledge are key to ensuring that data offered have value beyond the goal of the initial collection. The ADCC, through its long-term data archiving and efforts to make ARCSS data easily accessible, will continue to promote the goals of system science.

References

Advisory Committee for Environmental Research and Education (AC-ERE) (2003) *Complex Environmental Systems: Synthesis for Earth, Life, and Society in the 21st Century*. National Science Foundation, Arlington, VA.

Erb, K. (1998) *Guidelines and Award Conditions for Scientific Data*. National Science Foundation, Office of Polar Programs. http://www.nsf.gov/pubs/1999/opp991/opp991.txt. Accessed 25 May 2002.

Michener, W.K., and J.W. Brunt (ed.) (2000) *Ecological Data: Design, Management, and Processing*. Blackwell Science, Oxford, England.

National Research Council (1995) *Finding the Forest in the Trees: The Challenge of Combining Diverse Environmental Data*. National Academy Press, Washington, DC.