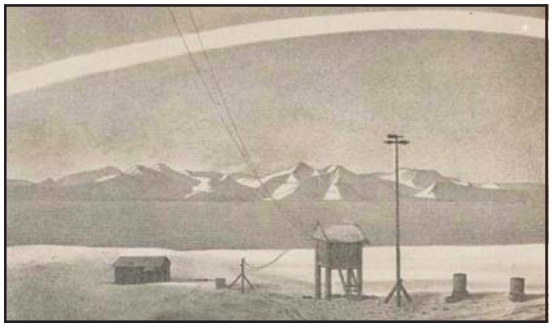
On the Creation of Environmental Data Sets for the Arctic Region

This article was prepared by Florence Fetterer, NOAA Liaison, National Snow and Ice Data Center/World Data Center for Glaciology, Boulder, Colorado, and Igor Smolyar, of NOAA's Ocean Climate Laboratory, National Oceanographic Data Center, Silver Spring, Maryland.

When Karl Weyprecht proposed better coordination of research in 1874, leading to a series of coordinated synoptic observations in the Arctic, little did he think that his ideas would produce scientific data that remains of intense interest to researchers 130 years later. And still less would he have imagined that his proposals, and the resulting International Polar Year of 1882–1883, would inform goals for creating and managing scientific data in the 21st century. Because of advances in observation and data technologies, the questions that Weyprecht addressed have only increased in significance. What constitutes useful environmental data? How are data both a product of research and a catalyst for new research? How should data be managed to ensure continued accessibility and usefulness? The NOAA Arctic research activities described elsewhere in this edition of *Arctic Research of the United States* both use and produce data. This article examines the process of creating environmental Arctic data sets and the symbiosis of research, data, and data management. These data sets may have value beyond that of advancing Arctic research objectives: they may be, for example, monitoring tools for change detection, or they may underlie decision support applications.



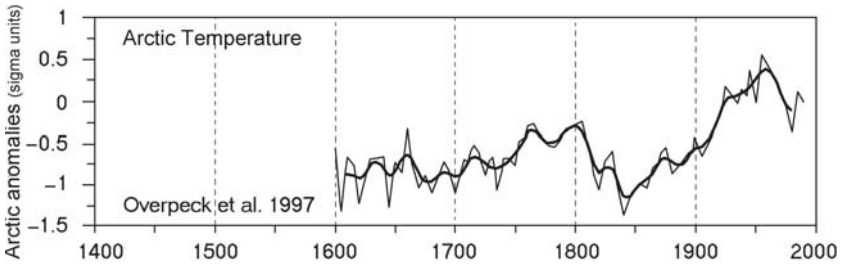
IPY meteorological station, 1882.

This focus on data and data management has become a proper discipline at NOAA in the 130 years since Weyprecht's call for better coordination of research and data resources. Arctic environmental data sets can be found in almost every NOAA line office, but it is the NOAA National Environmental Satellite, Data, and Information Service Data Centers that share a mission of data management. The NOAA National Data Centers' commitment to long-term data management provides institutional support for producing exemplary environmental data sets. Each center has a particular research focus and expertise that adds value to its data management results. After a brief profile of these centers, we will discuss what general characteristics make certain data sets especially valuable and what elements come into play during the production of these data sets, highlighting enough of them here to provide a sense of the breadth of NOAA's Arctic data production activities. An atlas, the *Climatic Atlas of the Arctic Seas 2004*, serves as a case study. We also cite a number of NOAA operational, research, and modeling products as examples of particular aspects of data product creation.

National Data Centers

National Oceanographic Data Center

Located in Maryland, the National Oceanographic Data Center (NODC) is a repository and dissemination facility for global ocean data. Researchers from NODC's Ocean Climate Laboratory (OCL) announced in 2000 that the world ocean has warmed significantly over the past 40 years. Just as the atmosphere has a climate, with variability on different time scales, the ocean's temperature, salinity, and other characteristics change over time. OCL researchers based their conclusions on data laboriously collected, quality controlled, and assembled into a special form of environmental atlas called a climatology. To facilitate comparisons of the past with the present, and to investigate interannual-to-decadal ocean



Arctic temperature anomalies

climate variability, many thousands of raw observations acquired from ships were interpolated to a regular spatial grid and combined over annual, seasonal, and monthly compositing periods. A definitive statement about oceanic warming would not be possible without these climatologies.

In addition to supporting scientific studies, OCL's International Ocean Atlas and Information Series (currently nine in number) exemplify international cooperation. Much of it is taking place through the OCL's World Data Center (WDC) for Oceanography, in Silver Spring, Maryland. International collaboration is an absolute necessity for acquiring a sufficient number of observations for climatologies. The Global Oceanographic Data Archaeology and Rescue (GODAR) Project, for example, has added over six million historical ocean temperature profiles to the archives, as well as a large amount of other data. Initiated by the NODC and WDC, this OCL-directed project was subsequently endorsed by the UNESCO Intergovernmental Oceanographic Commission.

National Climatic Data Center

Among the hundreds of climate data compilations housed at the National Climatic Data Center (NCDC) in Asheville, North Carolina, are Arctic station data from the Global Historical Climate Network, the most comprehensive homogeneous collection of station temperature data available. "Homogeneous" means consistent over the years and from place to place, through changes in instrumentation, acquisition method, and site characteristics, so that scientists may look for trends in the data. Homogeneous data sets require careful quality control. Historical data are made homogeneous with present-day observations by adjusting for non-climatic discontinuities, such as a jump in precipitation that might be caused by a change in instrumentation. An important part of the quality control process is compiling station inventories that detail the history of each station, including changes in instrumentation, changes in location, and changes in surroundings. If a town grows up around a formerly rural station, for example, a heat island effect may be present in the data record.

NCDC also operates the World Data Center for Paleoclimatology (WDC Paleo), located in Boulder, Colorado. Paleoclimatology puts the relatively recent changes in Arctic climate, apparent in the instrumented record, in long-term context. Proxy data from tree rings, ice cores, and lake and marine sediments available from the WDC Paleo were used by an international team of scientists for a circum-Arctic view of surface air temperature changes over the last 400 years.

The WDC Paleo web site provides interpretations of the record: A steep increase in warming between 1850 and 1920 was most likely due to natural processes. Warming since 1920 is more difficult to ascribe to natural forcing alone. For an even longer view, ice core data are valuable. WDC Paleo and the National Snow and Ice Data Center jointly maintain the Ice Core Gateway. Proxy climate indicators from ice cores such as oxygen isotopes, methane concentrations, dust content, and other parameters stretch the record back more than 1000 years.

Ice core samples. Ice cores are taken from ice sheets or ice caps and are used by paleoclimatologists as a record of past climate.

National Geophysical Data Center

The National Geophysical Data Center (NGDC), Boulder, Colorado, contributes significantly to Arctic science through participation in the development of the International Bathymetric Chart of the Arctic Ocean (IBCAO). IBCAO bathymetry provides a detailed and accurate representation of the depth and morphology of the Arctic Ocean seabed. This dynamic database contains all available bathymetric data north of 64°N. It is

maintained as a gridded database, and a version has been published in map form. The IABCO team remapped the Lomonosov Ridge, showing it to be more segmented in structure, wider, and shallower than had previously been mapped. The Lomonosov Ridge is an important topographic barrier that influences deep water exchange between the eastern and western basins of the Arctic Ocean. An accurate seafloor is important for applications including ocean modeling, mapmaking, and other research endeavors. The IABCO effort involves investigators from eleven institutions in eight countries. It has been endorsed and supported by the Intergovernmental Oceanographic Commission, the International Arctic Science Committee, the International Hydrographic Organization, and the U.S. Office of Naval Research.

National Snow and Ice Data Center

Operational products, such as sea ice charts for shipping interests from the NOAA/Navy/Coast Guard National Ice Center, are often laboriously produced by manually interpreting and synthesizing data from many sources, both satellite and in situ. They are generally more accurate than similar products from single sources. The National Snow and Ice Data Center (NSIDC) works with operational groups within NOAA to make these products available to a different user base by archiving operational data, making data available online, providing documentation, and fielding questions from researchers about the data.

Originally founded to manage scientific data from the International Geophysical Year of 1957–1958 (the follow-on inspired by the IPY of 1882–1883), the World Data Center (WDC) for Glaciology is operated by NSIDC in Boulder, Colorado. Today, NSIDC is a NOAA-affiliated data center, designated by NOAA in 1976, affiliated with NGDC, and part of the University of Colorado's Cooperative Institute for Research in Environmental Sciences (CIRES).

The NOAA program at NSIDC supports the WDC and emphasizes data rescue and data from operational sources that can be used for climate research and change detection. NOAA-funded activities complement the activities of NSIDC's Distributed Active Archive Center (DAAC) and its NSF-funded Arctic System Science Data Coordination Center. The latter centers handle large volumes of satellite data (about 80% of NSIDC's funding comes from NASA for operation of the DAAC) and data from individual scientists.

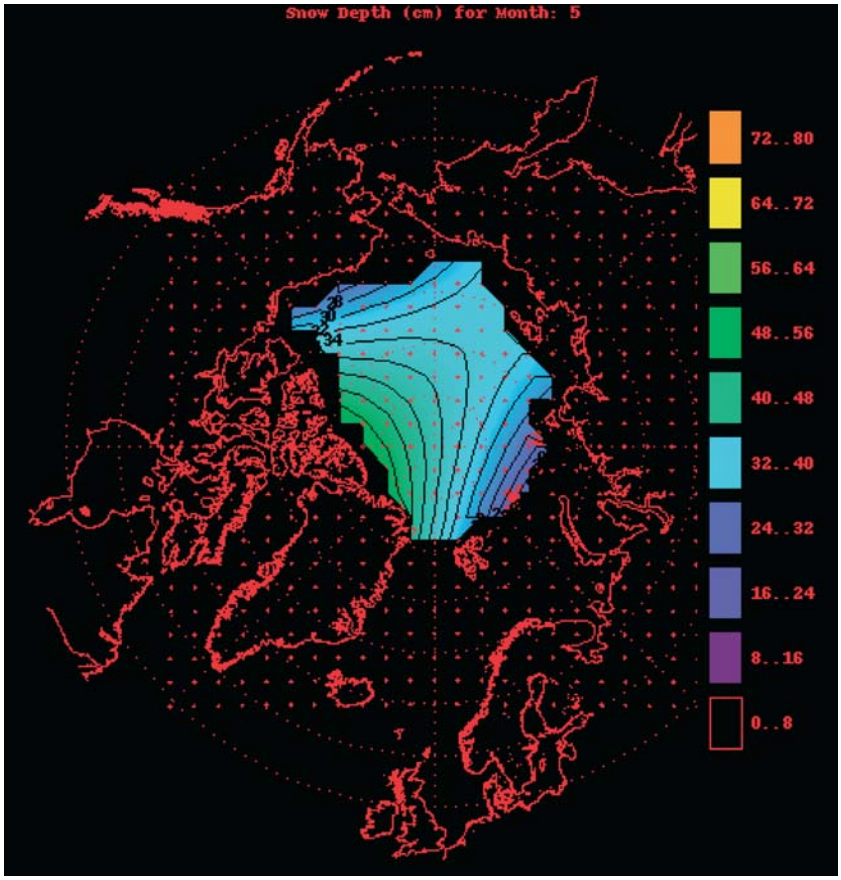
NSIDC has 522 data products in its on-line archive. Excluding satellite data sets, 104 of these are Arctic data sets, and of these about 20% are the more evolved compilations termed Arctic environmental data products.

What Makes a Good Environmental Data Product?

To make a good environmental data product, one starts with raw data and then processes or presents them in such a way that they become information. Data are transformed into a product that can advance a user up a hierarchy from data to information, to knowledge, to wisdom, shortening the user's path from data to knowledge or to the why and how of environmental interactions and change. Following is a summary of some of the data management practices that effect this transformation.

Context

Simply presenting data systematically is sometimes enough to transmit any underlying meaning. That is, even a well-organized collection of raw data can be an environmental data product. Usually, though, data products are more sophisticated. Presenting data in context is important, and the data product creator must decide what "context" means for the particular data under consideration. Temporal context usually means having as long a record as possible, while spatial context may mean covering as much of the Arctic as possible at an appropriate resolution or station density. Context may mean including population data for both predator and prey in an Arctic species survey or including as complete a set of oceanographic hydrochemical parameters as possible. The point is to make significant patterns evident, while committing no "sins of omission" in choosing what to include. Methods are important, as is documenting uncertainty. For example, if the product is a gridded climatology of snow depth on Arctic Ocean sea ice, some analysis should be done to ensure that enough observations are included in each grid cell for an acceptable level of accuracy. Sometimes the way in which data are gridded, interpolated, and presented implies a certain level of precision. For example, a two-dimensional quadratic function fit to snow depth on sea ice tells the user immediately that the snow depth information is not very precise.



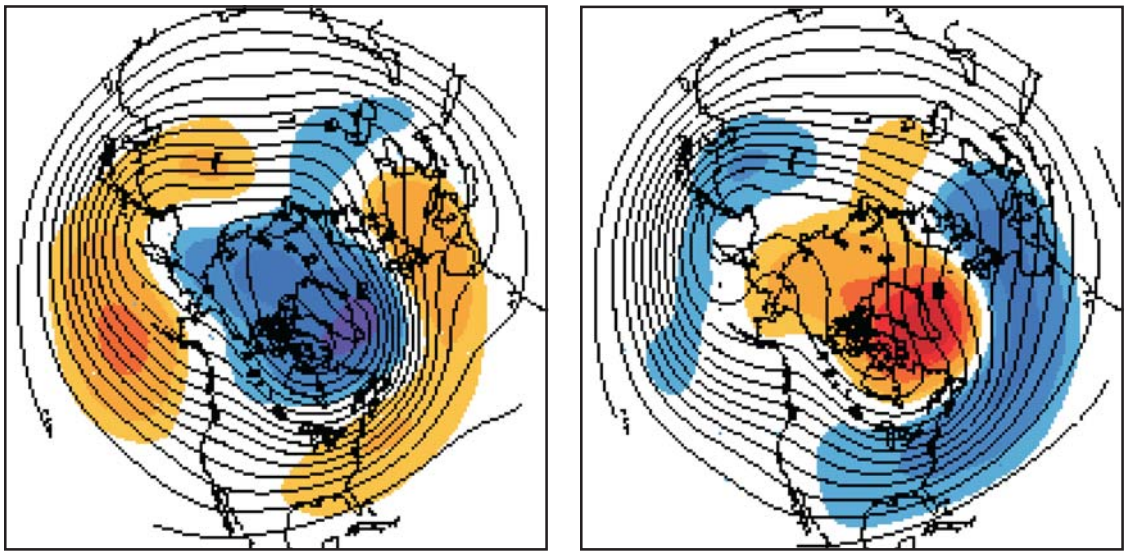
Variation in snow depth on Arctic sea ice, depicted by fitting a two-dimensional quadratic function to available data by month. There are very few measurements of snow characteristics on Arctic sea ice. This representation, from the Environmental Working Group's Meteorology Atlas, is the best possible in the absence of dense station coverage. While it appears unrealistic, product documentation explains why a more sophisticated gridding method for the available data is not appropriate.

Data products may include the results of an analysis, such as an empirical orthogonal function analysis of surface pressure data that shows the Arctic Oscillation pattern (that is, the tendency of pressure near the pole to act counter to pressure at mid-latitudes) or the addition of a trend line or other model fitting to observations. In such cases, the product creators have an extra responsibility to explain the limitations of their method of analysis, since these methods, if appropriately used, draw the pattern in the data rather than leave it to inference.

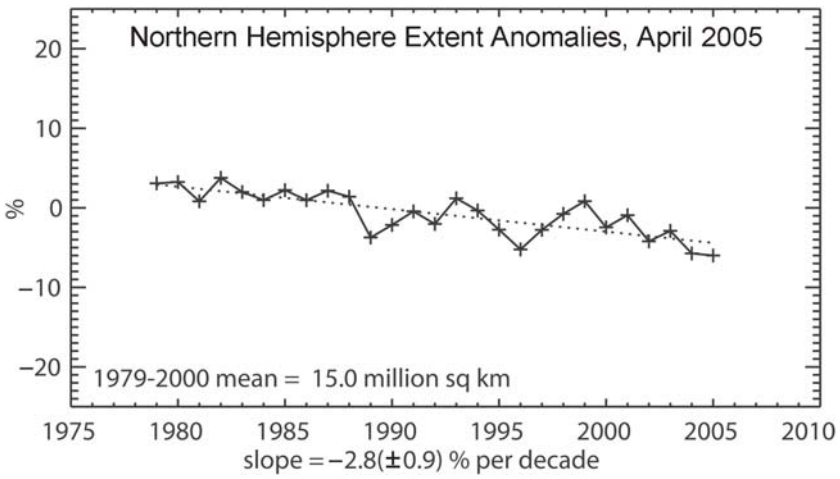
Documentation

Words are often the only way to provide appropriate context. The heat island effect, known only if the weather station history is known, is but one example of the importance of clear and complete documentation. Good documentation is written with the user in mind. For example, if the users are scientists, they will need to know about any known biases in the data record. If the data product is created for the general public, this information is just as important because it influences what a user infers from the data, but the information must be given in non-technical language. Operational users often need today's data irrespective of historical biases and may require little or no documentation.

For example, the NOAA National Weather Service's National Operational Hydrologic Remote Sensing Center (NOHRSC), located in Minneapo-



Examples of the Arctic Oscillation in its positive phase (left) and negative phase (right) from the NOAA National Weather Service's Climate Prediction Center web site. Blue indicates negative pressure anomalies, and orange indicates positive. The Arctic Oscillation is a large-scale atmospheric circulation pattern. Variability in the AO has been implicated in changes such as the recent steep decline in ice extent.



Sea ice extent trends. When data are presented with a trend line, as in this example, data providers should include error bars and document the limitations of the method (linear regression in this case) when it comes to providing information from the raw data.

lis, Minnesota, provides snow information in a variety of products and formats to meet operational forecasting needs. The NOHRSC web site (http://www.nohrsc.noaa.gov/) is designed to serve these users efficiently with interactive products and brief documentation. NSIDC archives assimilation model output (http://nsidc.org/data/g02158.html) from NOHRSC and gives the research community access to this unique data set. Extensive documentation on the NSIDC site, not needed on the NOHRSC site, covers alternative products, data quality and value, and potential research uses of the data. Similarly, the NOAA NESDIS Satellite Services Division Operational Daily Northern Hemisphere Snow Cover Analysis is made available to the operational community at http://www.ssd.noaa.gov/PS/SNOW/ and is archived for the research community at NSIDC (http://nsidc.org/data/g02156.html). Likewise, the NODC web site (http://www.nodc.noaa.gov/) is designed to provide users access to various products with a higher level of documentation than would be needed for operational users.

Graphics and Site Design

Graphical presentation and site design are aspects of information architecture that are especially important for complex environmental data sets that are viewed or used through a web page. Though not an environmental data set per se, the NOAA Arctic Theme Page (http://www.arctic.noaa.gov/) offers an example of good site design.

Careful attention must be paid to the graphical presentation of data. Gridded data to which a color table has been applied, or data smoothed by interpolation, are often subject to misinterpretation. The display resolution of pixels should faithfully represent the underlying resolution of the data, lest a scientist infer regional relevance that is not supported. Color tables should not have sudden jumps in intensity or hue that can draw the eye to a sea surface temperature difference, for example, that is an arbitrary point in a continuum, thus suggesting a pattern that is not there.

Data Integrity

Three main components ensure environmental data product integrity:

- The product must have scientific integrity; peer review of the data and a citation for the data set are needed to accomplish this.
- The data repository must be trustworthy.
- The data must not have been altered since the data were acquired or produced (or any alteration must be well described). The data management concepts of fixity, provenance, and source authentication come into play here.

Often it is the reputation of an individual scientist that imbues his or her data product creation with an aura of integrity. Data centers work with scientists to ensure that the reality measures up. Though it is common in the U.S. for investigators to manage their own data, this is rarely successful over the long term because scientists rarely have the requisite data management background needed to keep their data useful and accessible to the next generation of scientists or the resources to deal with technical issues that keep data secure, such as media migration and off-site backups.

What Makes a Well-Used Environmental Data Set?

Certain attributes will ensure that a data set will have many users. In such cases it is especially important to follow the design precepts above. Data products that include unique data, that are comprehensive collections, that offer continuous coverage over a long time period, that are easy to use, and that provide a synthesis of available information are characteristics of the most popular Arctic data products.

Uniqueness

Upward-looking sonar data from submarines provide the only measurements of ice thickness over a large portion of the Arctic. Ice thickness estimates are critical to estimating the mass bal-

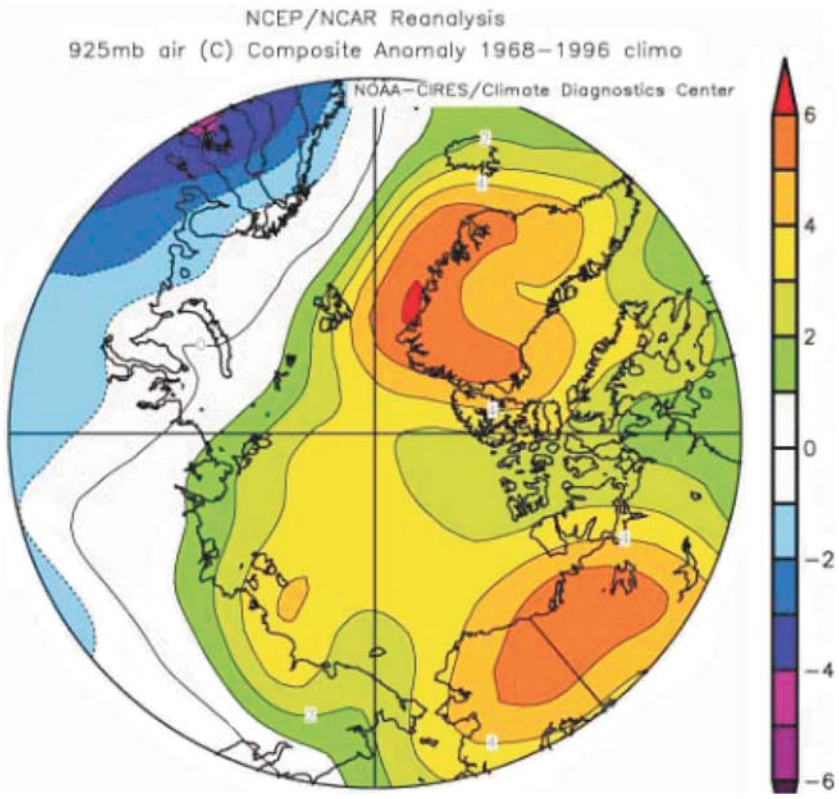
ance of ice in the Arctic Ocean—ice extent and concentration are only two dimensions of a three-dimensional problem. One difficulty in working with original records is that almost all submarine data are classified. Investigators at the U.S. Army's Cold Regions Research and Engineering Laboratory (CRREL) in Hanover, New Hampshire, and the Polar Science Center, Applied Physics Laboratory, University of Washington, worked with the U.S. Navy's Arctic Submarine Laboratory to find a way to "fuzz" the submarine track positions so that the data could be cleared for release by the Chief of Naval Operations. NSIDC distributes the data set, Submarine Upward Looking Sonar Ice Draft Profile Data and Statistics (http://nsidc.org/data/g01360.html). It has been the basis of a number of research articles on the controversial topic, "Is Arctic ice thinning?"

Comprehensiveness

Comprehensive data products offer more value than data sets that must be combined with others in order to have enough data of a single type for a scientific study. It takes a well-funded project, a multi-year commitment, and many individual and institutional partners to assemble, for example, "all" surface marine reports from ships, buoys, and other platform types. As such, the International Comprehensive Ocean Atmosphere Data Set (ICOADS, http://www.cdc.noaa.gov/coads/) of quality-controlled data dating from the late 18th century is a remarkable achievement. An entire body of literature has grown up around topics related to the quality control of historical ship data in ICOADS. For example, sea surface temperatures acquired by throwing a bucket over the side and measuring the temperature of the retrieved water are not the same as temperatures acquired from the engine intake. A "bucket correction" must be applied. This correction is based on modeled heat loss for water in a bucket on deck and should take into account ship speed (and its uncertainty) and the material of the bucket (wood or canvas). Clearly, quality controlling the millions of observations of various types so that they are homogeneous over time is a Herculean task.

ICOADS began as a U.S. project (COADS) in 1981 as a partnership between the NOAA Office of Oceanic and Atmospheric Research's Environmental Research Laboratories and NCDC, CIRES, and NSF's National Center for Atmospheric Research. The NOAA portion of ICOADS is currently supported by the NOAA Climate and Global Change Program. NSIDC makes an Arctic subset available (http://nsidc.org/data/nsidc-0057.html).

A submarine surfacing through sea ice. This photo comes from SCICEX (Scientific Ice Expeditions), a collaboration between the U.S. Navy and civilian scientists for environmental research in the Arctic.



Temperature anomalies for October through December 2002 in the Arctic. This illustration assisted in attributing the causes of the 2002 and 2003 sea ice extent minima to, in part, anomalously warm air temperature. The NOAA-CIRES Climate Diagnostic Center (http://www.cdc.noaa.gov) display tool allows users to choose the month and year to display from the NCEP/NCAR re-analysis products. The color table is intuitive (warm colors are warmer than average temperatures), and the resolution of one degree shown in the color bar is appropriate to the data set.

Continuous Spatial and Temporal Coverage

Products that are as close as possible to continuous in space and time are often desirable because, for example, the danger of aliasing is minimized (that is, there is a smaller chance of

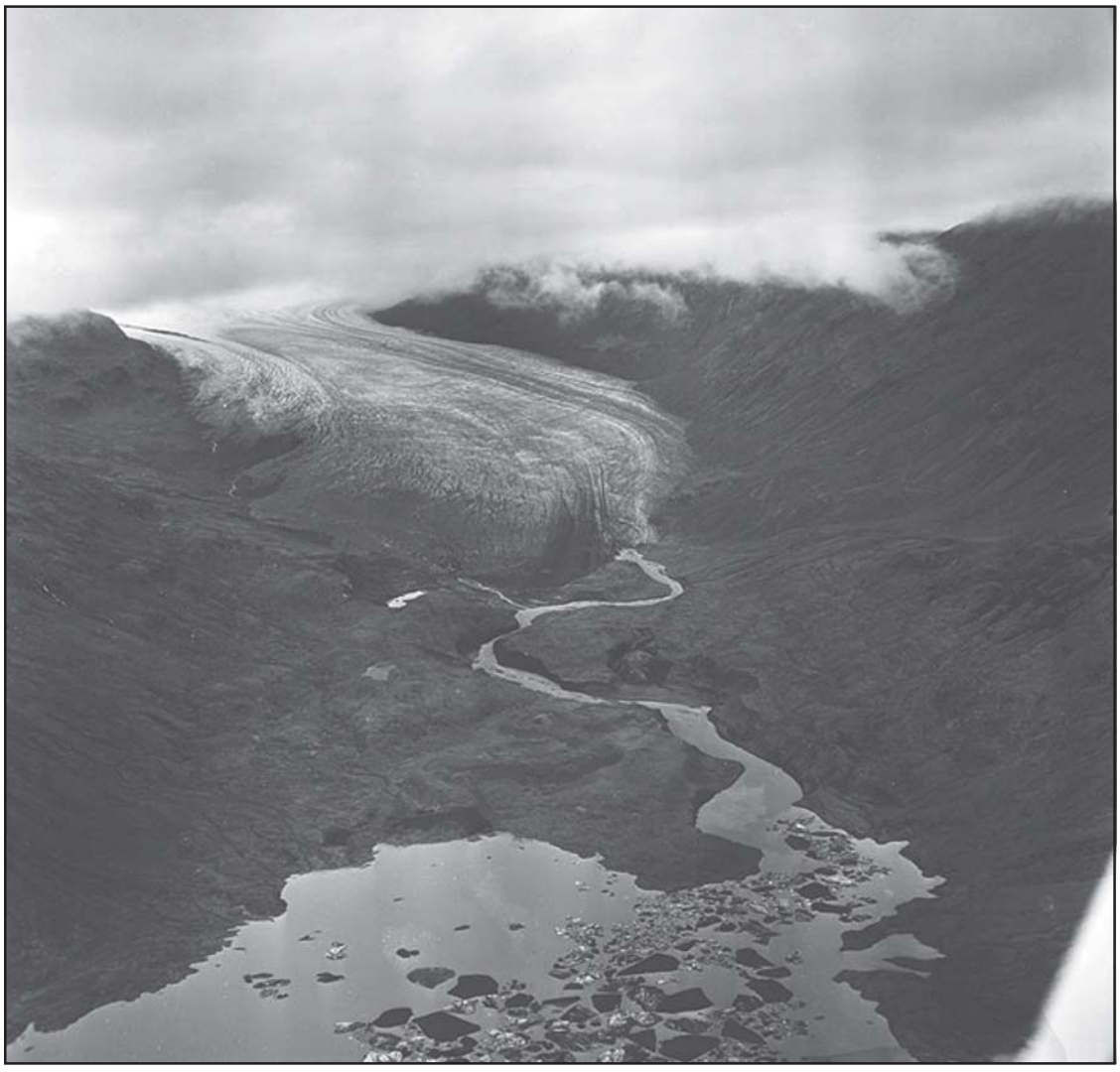
missing a significant event or pattern in the data). The enormously popular "reanalysis" products are an example.

Reanalysis projects take data such as those in ICOADS and assimilate them through a numerical weather prediction model to produce a series of analyses in which parameters such as surface pressure and temperature fields are physically consistent with one another. The fields cover a large area (the Northern Hemisphere, for instance) without gaps and are available at regular time intervals over a long record, making reanalysis output more useful than observations for many applications. One example is studying the spatial and temporal variability of large-scale atmospheric circulation patterns, such as the Arctic Oscillation. NOAA's Arctic Research Office is planning a coupled atmosphere–sea ice–ocean–terrestrial reanalysis optimized for the Arctic region. The description of the climate system it will produce can be used to detect Arctic change and assist in attributing change to specific causes.

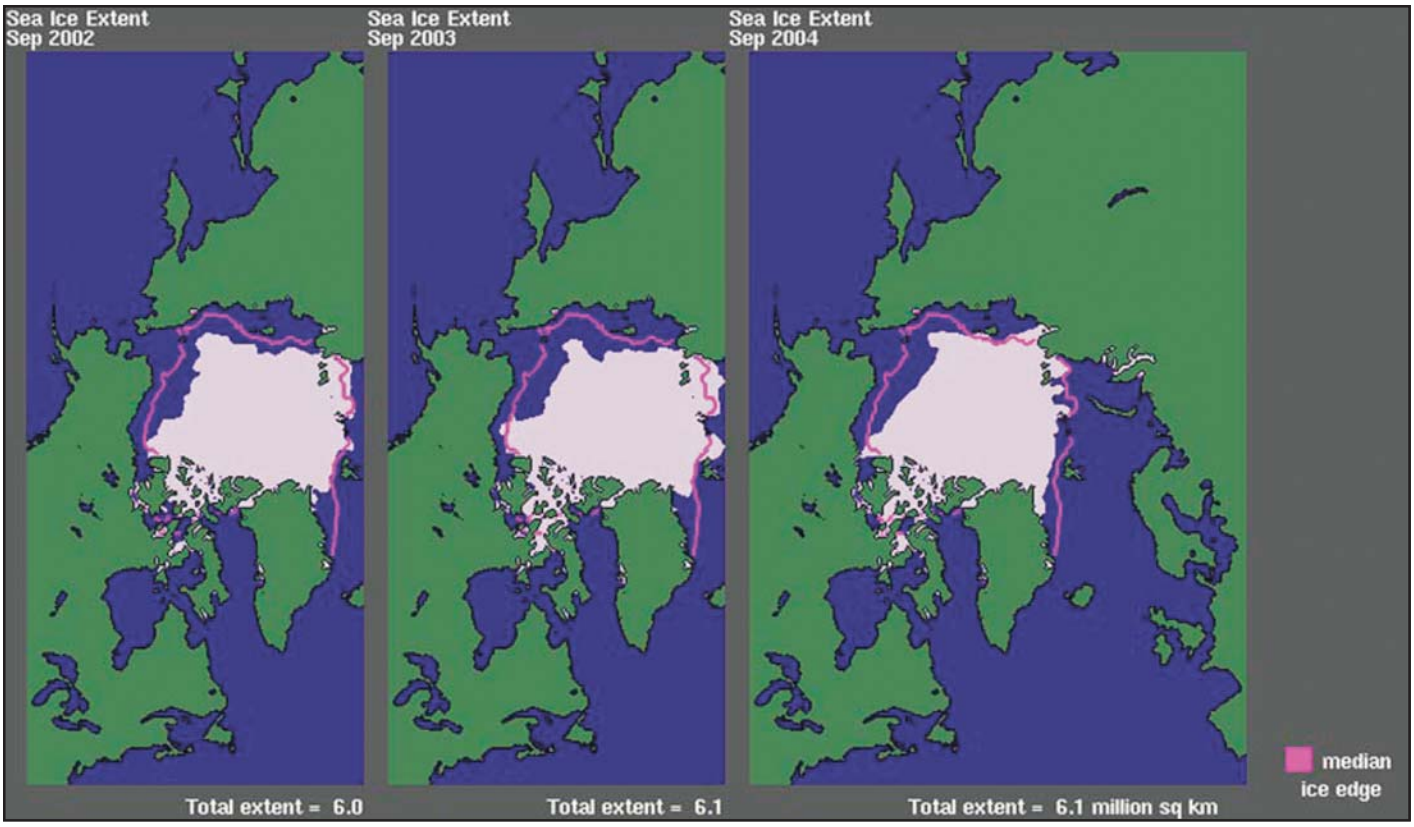
Ease of Access and Use

Many valuable data sets lie unused in archives simply because they are not easily accessible. Formerly, NSIDC's Glacier Photograph Collection of thousands of historical glacier photographs dating from the 1880s saw only a handful of users each year because users had to travel to NSIDC to view the fragile collection of prints. Now, thanks to NOAA's Climate Database Modernization Program funding for scanning the photos, many of the photographs can be viewed on line, and high-quality digital images can be downloaded (http://nsidc.org/data/g00472.html). As a result, the number of users has climbed to about a thousand each month.

Improving access can broaden the user base for a data set. NSIDC's satellite passive microwave



Cushing Glacier, Alaska. This photograph, taken in 1967, is one of thousands that are part of the Glacier Photograph Collection, created by the National Snow and Ice Data Center and available on line at http://nsidc.org/data/g00472.html.



Sample result from the Sea Ice Index, which displays anomalies in ice extent and other ice parameters going back to 1979. Here, recent summer ice extent, which has been the lowest in the data record, is displayed with the median extent (pink line) to give climatological context to the information.

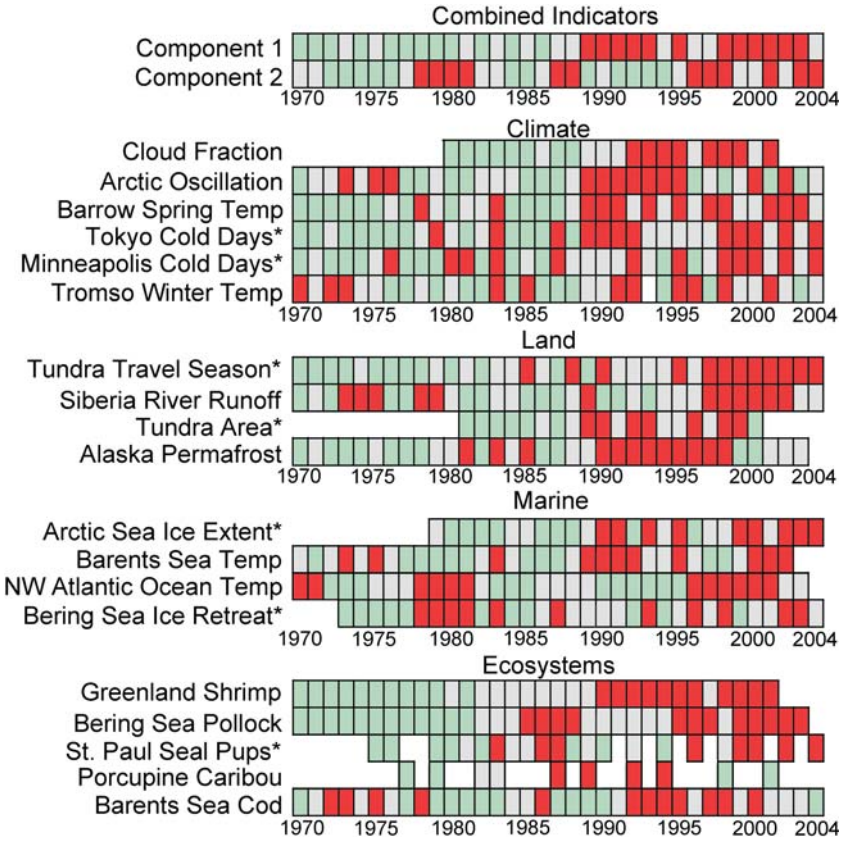
sea ice concentration data are popular among scientists but not geared toward the general public. The data are voluminous, and some technical and scientific sophistication is needed to simply read and interpret the data. To mitigate these issues, NSIDC developed the Sea Ice Index, which provides an easy way to visualize the satellite data. The Sea Ice Index web site lets any user track changes in sea ice extent and compare conditions between years. About 3,000 users visit the Sea Ice Index site every month.

Similarly, the OCL web site (http://www.nodc.noaa.gov/OC5//SELECT/dbsearch/dbsearch.html) allows users to extract data from the World Ocean Database 2001. While contributors to an environmental data product often prefer that the data be compiled on CD-ROM or DVD for reasons of fixity and attribution, data sets are much more likely to be used if they can be easily browsed or manipulated on line with a selection tool to facilitate access.

Synthesis

Data products that offer synthesis—a "big picture" version of the information in the data—are rare because they are difficult to construct. Synthesis products are built by distilling information from multiple sources. An exciting and successful example of this kind of product is NOAA's Near Realtime Arctic Change Indicator web site (http://www.arctic.noaa.gov/detect/), which summarizes "the present state of the Arctic climate and ecosystem in an accessible, understandable, and credible historical context." Designed for decision makers and the general public, it presents a sophisticated 30-year principal components analysis (the synthesis) of 19 climate, land, marine, and ecosystem "indicator" time series, such as the length of the travel season over tundra, the Bering Sea pollock population, the number of extremely cold days each year in cities such as Minneapolis, and the extent of Arctic sea ice.

Taken alone, any one of these time series would not present a compelling account of Arctic change. Taken together, the big picture emerges. The site tracks the rate and extent of changes in the Arctic to facilitate informed decisions concerning the impacts that result. Web pages for each of the indicators give a succinct but complete analysis of the data record in non-technical terms. Changes are given in context, including the context of the human dimension. Links to reports and more detailed data make it a useful resource for scientists as well.



Selection of time series representing Arctic change. The combined indicators are the result of a mathematical analysis (principal component analysis) that resolves the trends in all the time series into two major components. Series noted by an asterisk have been inverted. Red indicates large changes in recent years.

The Arctic Change Indicator web site was developed by NOAA's Arctic Research Office under the stewardship of investigators at the NOAA Pacific Marine Environmental Laboratory. It draws on the work of hundreds of investigators around the globe. A major challenge will be to keep the site updated. As NOAA looks toward building new observing systems, it will be critical to maintain data flow from existing observing stations.

The Climatic Atlas of the Arctic Seas

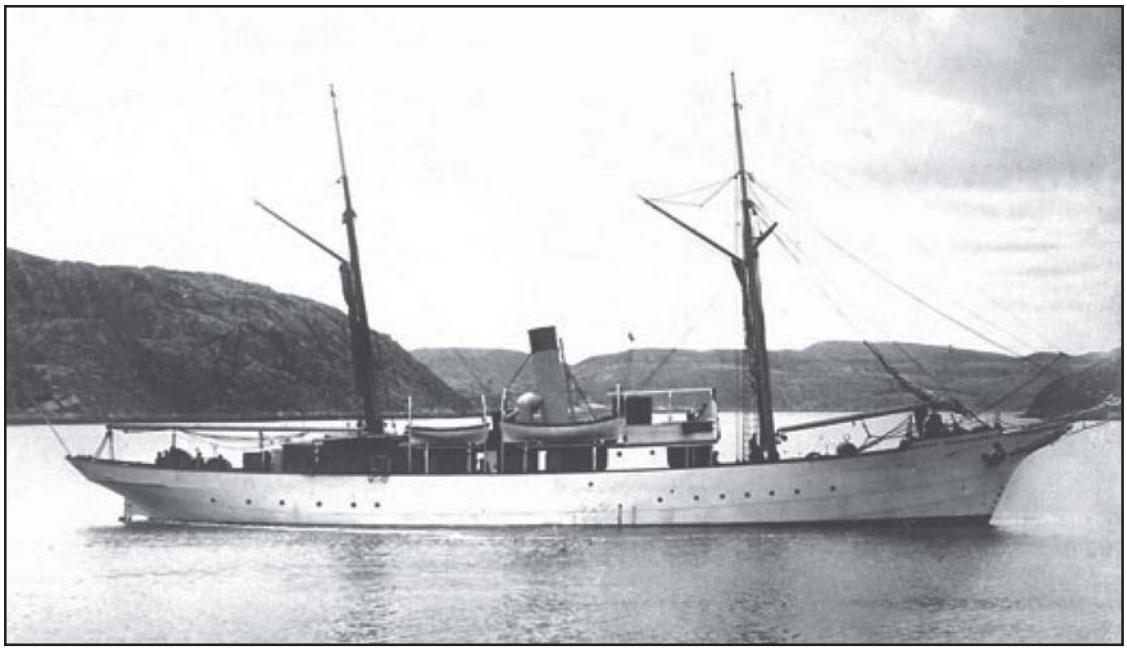
With these principles in mind, we now turn to a case study of active, collaborative data management that produces new knowledge and disseminates this potential across a wide community of researchers.

The WDC for Oceanography in Silver Spring , Maryland, and OCL/NODC have a long history of collaborating with Russian institutions to add more historical data to the OCL World Ocean Database Project. The most recent result is the *Climatic Atlas of the Arctic Seas* on DVD, with meteorology, oceanography, and hydrobiology (plankton, benthos, fish, sea birds, and marine mammals) data from the Barents, Kara, Laptev, and White Seas, collected by scientists from 14 countries during the period 1810–2001. The Murmansk Marine Biological Institute of the Academy of Sciences of the Russian Federation and OCL/NODC prepared the atlas with support from NOAA NESDIS and the Climate and Global Change Program.

The atlas provides historical context for its observations by including a written history of oceanographic observations in the Arctic, as well as scanned copies of selected rare books and articles. A gallery with photos and drawings gives the user some idea of what historical data collection platforms and expeditions looked like.

As is often the case with projects involving data rescue, libraries provided much of the material and documentation; the NOAA Central Library

Marine biologists P. Savitsky and I. Molchanovsky sampling plankton in the Kara Sea as part of an expedition of the Murmansk Marine Biological Institute on the nuclear icebreaker Sovetsky Soyz *in April–May 2002. The* Climatic Atlas of the Arctic Seas *weds early observations with contemporary observations in a seamless package.*



One of the first Russian research vessels, Andrey Pervozvanny*, on an expedition at the beginning of the 20th century, in the Barents Sea.*

(Silver Spring, Maryland), the Slavic and Baltic Branches of the New York Public Library, the New York Museum of Natural History Library, the Dartmouth College Library (Hanover, New Hampshire), the Slavic Library (Helsinki, Finland), and the public libraries of Moscow, Murmansk, and St. Petersburg (Russian Federation) all contributed.

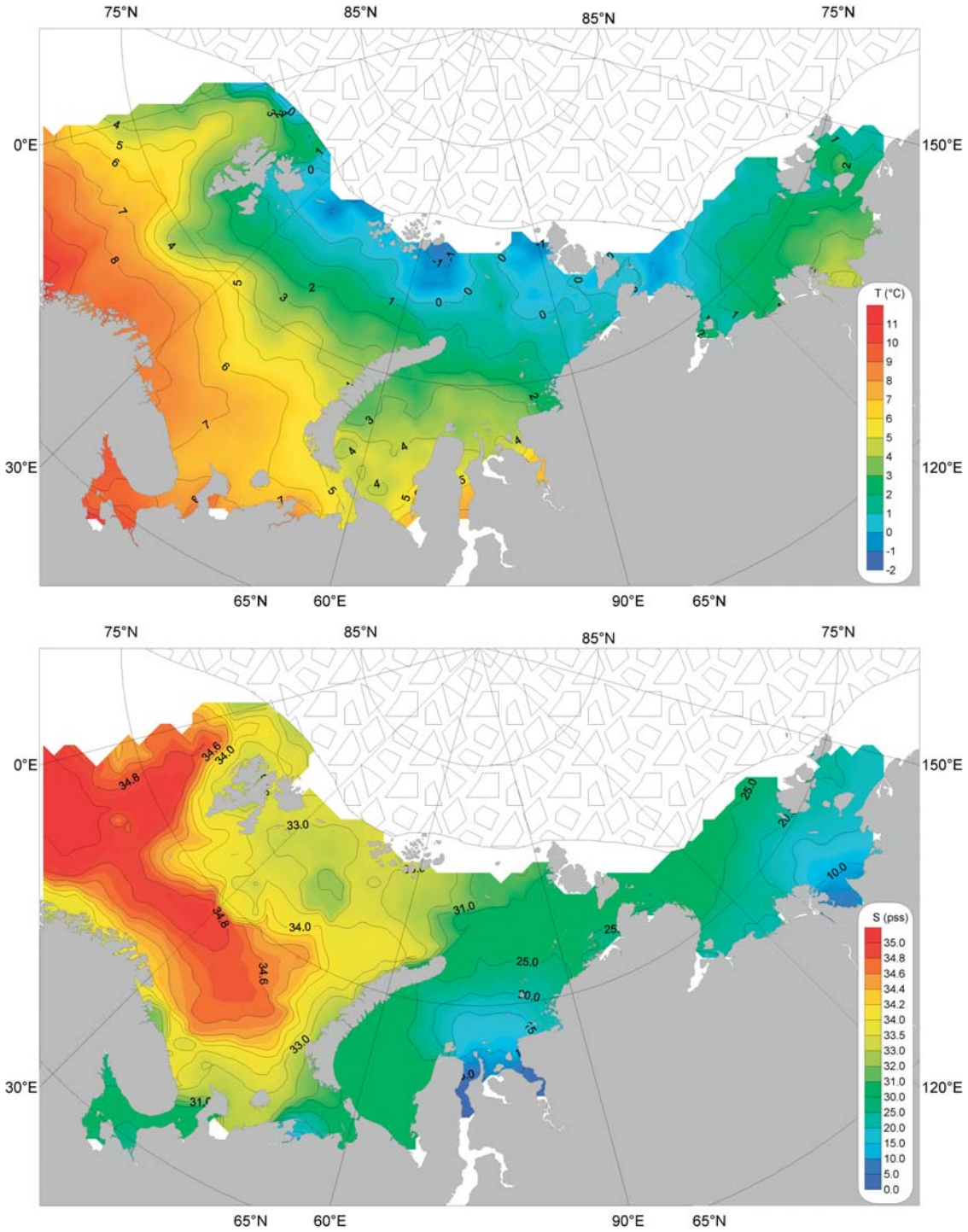
Assembling an atlas on this scale presents a number of challenges. A surprisingly difficult one is the elimination of duplicate stations from different data sources. As databases or parts of databases are shared, metadata are altered. For example, one database may have a station location in degrees, minutes, and seconds, and another may convert to decimal degrees. Rounding errors may give the appearance that these are two stations separated by as much as a few miles. Values of parameters may be presented at observed levels in one data set but interpolated, often by an unknown method, to a standard level in another data set. Another source of uncertainty is converting units of measurement. As a result, the same station data from different sources may differ in coordinates, time of measurement, and values of the parameters themselves. To help choose what station records to include, the atlas authors used a system of priorities: cruise reports, ship logs, and expedition diaries (all original sources) were deemed more reliable that data sources where the data apparently were repeatedly transformed. Elimination of duplicates and "near duplicates" brought the number of stations down from an initial 1,506,481 to a still sizeable 433,179.

Users have two ways of accessing raw observations: by oceanographic cruise or through 1° squares. For every month, a distribution map of stations is generated that allows a user to access data from a chosen square. Data may be easily imported into Excel or other database applications. Access to the actual observations is important for many users. Other users are likely to prefer a climatological presentation, since climatologies provide a convenient representation of average conditions, such as monthly or decadal means. The atlas satisfies both by including mean monthly temperature and salinity distribution fields at five standard depths, using an objective data analysis method.

A Long Journey from the Past: The International Polar Year

The International Polar Year serves as an important milepost for assessing our efforts and establishing stronger standards to carry the value of observations and research far into the future. As we look back over past IPY/IGY efforts and forward to those coming in 2007–2008, those of us who create Arctic environmental data sets have observed some lessons over the years.

Many of us know the tragic story of the Greely Expedition, an American venture sent into the Arctic in 1881 to establish an IPY station that ended in starvation for most of the party, but



Average September surface temperature (top) and salinity (bottom), from the Climatic Atlas of the Arctic Seas*. Note the low salinity at the river mouths. Scientists are working to understand the role of freshwater input from rivers on Arctic Ocean (and global ocean) circulation. Climatologies provide a picture of average conditions against which to evaluate changes.*

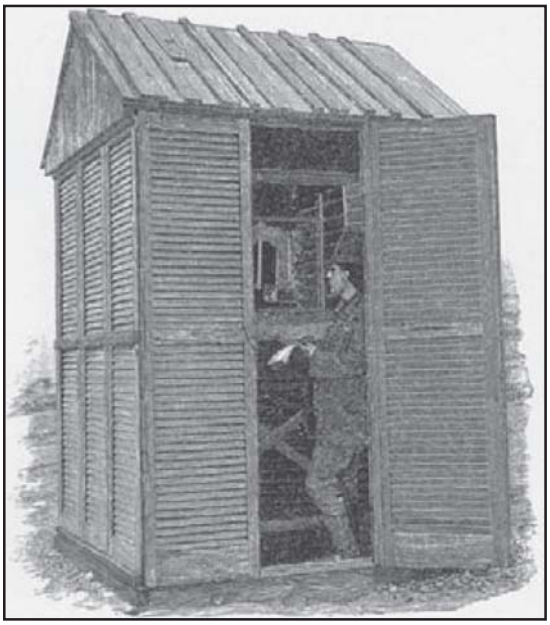
fewer of us are aware of the sad tale of their scientific data. Their observational records should have become a legacy to their efforts and sacrifice, but this is not the case. As Kevin Wood of NOAA writes, narrating the story of their data:

> "Perhaps the most compelling aspect of the Greely tragedy is the utter commitment of these men to preserve their scientific work. Aware that if relief didn't arrive in time they would be left to retreat on their own, Greely began making copies of their scientific work (amounting to some 500 observations per day). When they were forced to abandon Fort Conger in August 1883 they took with them—in lieu of extra rations—these copies sealed in three tin boxes of 50 pounds each, all of the daily journals, 70 pounds of glass photographic plates, and all of the standard thermometers and several other impor-

tant instruments. They also continued a program of scientific observation, in the face of starvation, until just 40 hours before they were rescued."

Surely, Greely and his men hoped that their work would lead to important advances in science. Greely expressed this sentiment when he wrote in his official report, "The conviction that at no distant day the general laws of atmospheric changes will be established, and later, the general character of the seasons be predicted through abnormal departures in remote regions, causes this work to be made public…in the hope that it may contribute somewhat to that great end." The desire to be able to "predict the character of the seasons" still motivates researchers today.

Unfortunately, the research program of the first IPY was never completed as Weyprecht had originally planned. Each nation issued an individual report over the ensuing years, but no systematic study of the simultaneous observations—the heart of the IPY program—was undertaken. The International Polar Commission dissolved, and the data collected at such cost during the first IPY soon fell into obscurity.



Sgt. Jewell recording temperature, Fort Conger, during the Greely expedition, 1881–1883.

Today the original records of the first IPY are widely scattered in various libraries and archives and are often in a perilous state of preservation. Some of the published reports are extremely rare and are very difficult to obtain. The fate of the first IPY records, gained at such high cost, underutilized both then and now, and scattered over the course of time, highlights how important it is to provide for the effective preservation and management of such extremely valuable data.

The scientific legacy of the Greely Expedition and the other expeditions of the first IPY has only with difficulty been preserved. NOAA has recently made meteorological data from the first IPY available in digital format, along with an extensive collection of documentary images (see http://www.arctic.noaa.gov/aro/ipy-1).

There is another kind of legacy that we can create from the experience of the first IPY. As we look forward to a new International Polar Year, we must remain focused on these key lessons about data management.

Lesson 1: Applying the Right Kind and Right Amount of Effort at the Right Time is Imperative

While data rescue is difficult, tedious, and often expensive, it is crucial. The only way to reduce uncertainty in our estimates of past, and predictions of future, Arctic environmental change is to incorporate more, older, and better data into our analyses. NOAA's Climate Database Modernization Program, now in its sixth year, has keyed or scanned and placed on line over 45 million environmental records. More needs to be done, especially in documenting and quality controlling these records, because these last steps require capturing the knowledge of people who know the "rescued" data best, often a cadre that are beyond retirement age.

Lesson 2: Structures that Enable International Collaboration can Dramatically Increase Value

As a result of Arctic geography, the most comprehensive data sets result from international cooperation. GODAR and ICOADS are models for this cooperation. International data-sharing agreements are essential. In contrast to the National Data Centers, the World Data Center system provides a structure within which data sharing can occur with a minimum of diplomatic overhead. WDCs in the U.S. that share Arctic data internationally are the WDC for Glaciology, Boulder (co-located with NSIDC), WDC for Oceanography, Silver Spring (co-located with NODC), the WDC for Marine Geology and Geophysics, Boulder (co-located with NGDC), the WDC for Meteorology, Asheville (co-located with NCDC), and the WDC for Paleoclimatology (affiliated with NCDC).

Lesson 3: Good Data Stewardship is Superior to Untimely Data Rescue

We can avoid expensive and possibly fruitless data rescue efforts in the future by heeding the lessons of the past. The International Polar Year, 2007–2008, will be a catalyst for reinvigorating professional data management. The IPY promises new international collaboration and the potential for synthesis of knowledge under the headings of cross-disciplinary research themes. Good data stewardship will help ensure that this major undertaking will not shortly become a dimly receding spot on the horizon behind us.

Jane Beitler and Ruth Duerr, NSIDC, assisted in editing this article. Kevin Wood, PMEL, provided the material on the first IPY and the Greely expedition.

This effort requires not only attention to the data, but also to capturing the "data about the data" that enables continuing understanding and value. A disciplined effort to define and organize the metadata will enable other researchers to locate, understand, and interpret data for years to come, providing the foundation for long-term coordination and synthesis. In the instance of the first IPY, Weyprecht's vision of coordinated synoptic observations led to the acquisition of a data set that serves as a snapshot of climatic conditions as they existed in that now long-past year. Data collected then can now be compared with conditions as they are today. Making that comparison, Wood and Overland (2005) found that monthly mean air temperatures at IPY-1 stations were generally within recent climatological limits, and spatial patterns in temperature anomalies (departure from the long-term mean) were consistent with Arctic-Oscillation-driven patterns of variability. In a nod to the value of documentation and metadata, Wood and Overland noted that "the qualitative logs are particularly useful in validating climate information."

NOAA will focus its strength in environmental observations and analysis on the polar regions during IPY. NOAA's Arctic Research Office has endorsed a fundamental goal for IPY data management: to securely archive a baseline of data against which to assess future change, and to ensure that IPY data are accessible and preserved for current and future users.

What will this IPY snapshot look like, and how will data be preserved? In contrast to Weyprecht's IPY, most data from the coming IPY will be "born digital." Station logbooks from Weyprecht's day could be preserved in libraries, where they had to be physically protected from destruction by fire, insects, and chemical decomposition of paper and ink. One might think it is easier to preserve digital information, but digital data are not immune from physical destruction, and they require a host of measures to ensure their usability into the future: "digital objects require constant and perpetual maintenance, and they depend on elaborate systems of hardware, software, data and information models, and standards that are upgraded or replaced every few years."*

During the coming IPY, hundreds of investigators and agency programs will produce raw observations, satellite data, and environmental data products in a number and of a complexity that would have been hard to imagine in the late 1880s. To ensure preservation,

- NOAA's Data Centers and the Arctic Research Office will work to advance standards and technologies that support this goal. NOAA advocates the use of the Open Archival Information System (OAIS) Reference Model for metadata. Work on the OAIS model and on technological advances such as GRID computing and interoperable catalogs is happening now at NOAA's National Data Centers.
- Cross-agency support for IPY data management is needed. Because of the international, distributed nature of IPY activities, the data they produce will necessarily be archived and made accessible through distributed data management. This distributes the burden of data management but imposes additional coordination challenges. Within the U.S., the National Academy of Sciences' Polar Research Board has endorsed the concept put forward by the International Council of Scientific Unions' IPY Planning Group of a coordinating IPY Data and Information Service (IPY-DIS). Cross-agency support of the DIS at a national level will ensure that the U.S. leaves a secure IPY data legacy.
- Adequate funding is needed. Funding for the management of data acquired through research programs is often difficult to obtain, either because the importance of data management as a discipline is not recognized or because there are simply not enough dollars to go around. Currently, for every $30 dollars spent nationally on Arctic research, about $1 is spent on Arctic data management.

In the end, it is important to remember that technological advances and digital archives will secure data for future generations of researchers only to the extent that they are successful in capturing what people know about the data. We must also keep today's equivalent of the IPY-I station's "qualitative logs." With them, future researchers will have the appropriate contextual material to turn the coming IPY data into information-filled Arctic environmental data products.

* From *It's About Time: Research Challenges in Digital Archiving and Long-Term Preservation*, final report of a workshop sponsored by NSF and the Library of Congress, August 2003.

References

Arzberger, P., P. Schroeder, A. Beaulieu, G. Bowker, K. Casey, L. Laaksonen, D. Moorman, P. Uhlir, and P. Wouters (2004) An international frame-

work to promote access to data. *Science*, Vol. 303, No. 5665, p. 1777–1778.

Levitus, S., J.I. Antonov, T.P. Boyer, and C. Stephens (2000) Warming of the world ocean. *Science*, Vol. 287, No. 5461, p. 2225–2229.

National Science Foundation and Library of Congress (2003) *It's About Time: Research Challenges in Digital Archiving and Long-term Preservation*, *Final Workshop Report*. Sponsored by the National Science Foundation's Digital Government Program and Digital Libraries Program, Directorate for Computing and Information Sciences and Engineering, and The Library of Congress's National Digital Information Infrastructure and Preservation Program.

Overpeck, J.T., K. Hughen, D. Hardy, R. Bradley, R. Case, M. Douglas, B. Finney, K. Gajewski, G. Jacoby, A. Jennings, S. Lamoureux, A. Lasca, G. MacDonald, J. Moore, M. Retelle, S. Smith, A. Wolfe, and G. Zielinski (1997) Arctic environmental change of the last four centuries. *Science*, Vol. 278, No. 5341, p. 1251–1256.

Peterson, T.C., and R.S. Vose (1997) An overview of the Global Historical Climatology Network temperature database. *Bulletin of the American Meteorological Society*, Vol. 78, No. 12, p. 2837–2849.

Wood, K.R., and J.E. Overland (2005) Climate lessons from the First International Polar Year 1881–1884. Poster presented at the American Meteorological Society 8th Conference on Polar Meteorology and Oceanography, San Diego, California.

Data Set Citations

Berger, V. J., A.D. Naumov, N.V. Usov, M.A. Zubaha, I. Smolyar, R. Tatusko, and S. Levitus (2003) *36-Year Time Series (1963–1998) of Zooplankton, Temperature, and Salinity in the White Sea*. NESDIS Atlas 57, NOAA, Washington, D.C.

Fetterer, F., and K. Knowles (2004) *Sea Ice Index*. Updated from 2002, digital media, National Snow and Ice Data Center, Boulder, Colorado.

Fetterer, F., and V. Radionov (Ed.) (2000) *Environmental Working Group Arctic Meteorology and Climate Atlas*. CD-ROM, Arctic Climatology Project, National Snow and Ice Data Center, Boulder, Colorado.

Lappo, S., Y. Egorov, M. Virsis, Y. Nalbandov, E. Makovetskaya, L. Virsis, I. Smolyar, and S. Levitus (2003) *History of the Arctic Exploration 2003: Cruise Reports, Data*. CD-ROM, International Ocean Atlas and Information Series, Vol. 8, World Data Center for Oceanography, Silver Spring, Maryland.

Markhaseva, E., A. Golikov, T. Agapova, A. Beig, and I. Smolyar (2002) *Zooplankton of the Arctic Seas 2002*. CD-ROM, International Ocean Atlas and Information Series, Vol. 6, World Data Center for Oceanography, Silver Spring, Maryland.

Matishov, G., A. Zuyev, V. Golubev, N. Adrov, V. Slobodin, S. Levitus, and I. Smolyar (1998) *Climatic Atlas of the Barents Sea 1998: Temperature, Salinity, Oxygen.* NESDIS Atlas 26, NOAA, Washington D.C.

Matishov, G. P., Makarevitch, C. Timofeyev, L. Kuznetsov, N. Druzhkov, V. Larionov, V. Golubev, A. Zuyev, V. Denisov, G. Iliyn, A. Kuznetsov, S. Denisenko, V. Savinov, A. Shavykin, I. Smolyar, S. Levitus, T. O'Brien, and O. Baranova (2000) *Biological Atlas of the Arctic Seas 2000: Plankton of the Barents and Kara Seas*. NESDIS Atlas 39, NOAA, Washington D.C.

Matishov, G., A. Zuyev, V. Golubev, N. Adrov, S. Timofeev, O. Karamusko, L. Pavlova, O. Fadyakin, A. Buzan, A. Braunstein, D. Moiseev, I. Smolyar, R. Locarnini, R. Tatusko, T. Boyer, and S. Levitus (2004) *Climatic Atlas of the Arctic Seas 2004: Part I. Database of the Barents, Kara, Laptev, and White Seas - Oceanography and Marine Biology.* NESDIS Atlas 58, NOAA, Washington, D.C.

National Operational Hydrologic Remote Sensing Center (2004) *SNODAS Data Products at NSIDC*. Digital media, National Snow and Ice Data Center, Boulder, Colorado.

National Snow and Ice Data Center (1998) *Submarine Upward Looking Sonar Ice Draft Profile Data and Statistics*. Digital media, National Snow and Ice Data Center/World Data Center for Glaciology, Boulder, Colorado.

National Snow and Ice Data Center/World Data Center for Glaciology, Boulder (Compiler) (2002) *Online Glacier Photograph Database*. Digitized subset of the Glacier Photograph Collection, National Snow and Ice Data Center, Boulder, Colorado.

NOAA/NESDIS/OSDPD/SSD (2004) *IMS Daily Northern Hemisphere Snow and Ice Analysis at 4 km and 24 km Resolution*. Digital media, National Snow and Ice Data Center, Boulder, Colorado.

Serreze, M. (Compiler) (1997) *Comprehensive Ocean–Atmosphere Data Set, LMRF Arctic Subset*. Digital media, National Snow and Ice Data Center, Boulder, Colorado.