Spatial characterization of the meltwater field from icebergs in the Weddell Sea

John J. Helly,a,b,1 Ronald S. Kaufmann, Maria Vernet,c and Gordon R. Stephensonb

aSan Diego Supercomputer Center and bScripps Institution of Oceanography, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093; and cMarine Science and Environmental Studies Department, University of San Diego, 5998 Alcalá Park, San Diego, CA 92110

Edited by Michael Goodchild, University of California, Santa Barbara, CA, and approved November 22, 2010 (received for review August 20, 2009)

We describe the results from a spatial cyberinfrastructure developed to characterize the meltwater field around individual icebergs and integrate the results with regional- and global-scale data. During the course of the cyberinfrastructure development, it became clear that we were also building an integrated sampling planning capability across multidisciplinary teams that provided greater agility in allocating expedition resources resulting in new scientific insights. The cyberinfrastructure-enabled method is a complement to the conventional methods of hydrographic sampling in which the ship provides a static platform on a station-by-station basis. We adapted a sea-floor mapping method to more rapidly characterize the seafloor geophysically and biologically. By jointly analyzing the multisource, continuously sampled biological, chemical, and physical parameters, using Global Positioning System time as the data fusion key, this surface-mapping method enables us to examine the relationship between the meltwater field of the iceberg to the larger-scale marine ecosystem of the Southern Ocean. Through geospatial data fusion, we are able to combine very fine-scale maps of dynamic processes with more synoptic but lower-resolution data from satellite systems. Our results illustrate the importance of spatial cyberinfrastructure in the overall scientific enterprise and identify key interfaces and sources of error that require improved controls for the development of future Earth observing systems as we move into an era of peta- and exascale, data-intensive computing.

Antarctica | remote sensing | hydrography | surface water

This paper describes the results from a geospatial cyberinfrastructure developed to support the characterization of the meltwater field around free-drifting icebergs and to combine it with synoptic, regional-scale data to investigate the role of icebergs in controlling biological productivity in the Weddell Sea. The cyberinfrastructure was formed through the time-based correlation of data sources, computing and communication channels across multiple scales of space and time to produce synthetic results (cf. SI Text). The work was carried out over the course of 2005–2009 including three cruises onboard the National Science Foundation (NSF) ice-capable ships in the vicinity of the Antarctica Peninsula [Fig. 1 (1)]. These results are of interest because the role of icebergs and their meltwater in controlling the freshwater and carbon fluxes in the Antarctic is unclear (2, 3). For example, for a given iceberg, what volume of meltwater, with its terrestrially derived inorganic load from rock scour and atmospheric deposition, is introduced into what volume of seawater and how is it distributed in time and space? An accurate and precise answer to this question requires the spatial estimation of the meltwater field around an iceberg instantaneously as well as integrated over time with local- and regional-scale implications (4). A key goal of this research was to get as close to an instantaneous estimate as possible within the limits of our shipboard capabilities and to do it over large enough spatial extents as to be able to correlate the fine-scale spatial analysis with more synoptic, regional-scale analyses albeit with coarser spatial and temporal resolution.

The characterization of the surface of the ocean with physical, chemical and biological data in the vicinity of a free-drifting iceberg poses a number of challenges. Besides being large, icebergs are affected by geophysical forces that alter their structure and movement: solar radiation, Earth’s rotation, tides and currents, and winds and storms (5). As free-drifting, tabular icebergs proceed through the ocean, they change size, shape, and mass distribution and melt into the surrounding ambient seawater. The relative contributions of geophysical forces on iceberg motion change as ice mass is lost to the surface waters of the ocean, and the losses change the composition of the waters they move through. It is a dynamic and complex system that encompasses a wide range of measurement scales. Our focus has been on small-to-medium-sized icebergs (<250 km²) because they are the most abundant and widely distributed in the Weddell Sea.

Initially, in 2005, we attempted to characterize the meltwater field using hydrocasts distributed around the icebergs and transsects across the near- and far-field normal to the ice face. However, due to the length of time required for these operations, and the dynamic motions of the icebergs surveyed, it was impossible to develop a clear and convincing picture of where the meltwater was; especially whether there was a plume or wake and in which direction it was tending. Hydrocasts typically require 1 h per

SI Text

Fig. 1. Regional map of the study area in the Weddell Sea and Scotia Sea showing the GPS-based ship track for the three research cruises, LMG0514a (red), NBP0806 (yellow), NBP0902 (green), as well two representative RADARSAT-1 image frame boundaries (dashed yellow, wideband mode). Bathymetry is from the Global Topography 11.1 (1).
500 m of depth sampled, and the ship can move at 18.5 km h\(^{-1}\) between stations. A representative near-far, radial transect out to 20 km from an iceberg with a hydrocast every 2-km to 1,500-m depth requires about 31 h. This is a major commitment of ship time, and this is the most scarce resource during a field campaign with many competing demands for it.

An additional limitation on obtaining an estimate of the meltwater field is the difficulty in separating the individual forces of wind, tide, and current and their relative contribution to the motions of a given iceberg and the water it was traveling through because these become integrated as sampling time is extended. We needed a faster sampling method that would enable us to better localize the meltwater field and differentiate the meltwater signal from background ambient conditions as well as to help us choose where to conduct conventional hydrocast transects. We also needed something that would minimize the amount of time and get us as close as possible to an instantaneous estimate. This led to the surface-mapping method (cf. Materials and Methods).

The surface mapping of the meltwater field requires a geospatial cyberinfrastructure with which to organize, combine, and display analytic results coming from a variety of physical and biological measurements. There is no standard definition of spatial cyberinfrastructure so, for the purposes of this paper, we generalize the definition suggested by ref. 6 as follows: Spatial cyberinfrastructure is a complex of information technology and data with four essential elements: (i) computational methods and (ii) communication channels that produce and transport (iii) digital information describing (iv) locations on the Earth. The first three elements are common to most other cyberinfrastructure settings, but it is the fourth, the spatial element, that adds the complexity of an Earth-centric coordinate system.

The additional complexity poses unique challenges in maintaining a reliable quantitative framework for the integration of data while minimizing and managing the propagation of errors in precision and accuracy. This challenge pertains not only to positional information but also to any derived measurements such as area, volume, distance, speed, and orientation. The need to maintain and verify positional accuracy and precision is well-established in geodesy and cartography (7) but not directly addressed by existing geographic information system (GIS) technology (8–10). The dynamic character of Earth systems further complicates spatial measurements as exemplified by the long-standing unsolved problem of defining the coastline due to the variability in topography, tides, and waves (11–13).

Estimation and management of spatial errors are of central concern in conducting scientific research with spatial data, especially as we progress to higher-resolution data and models. Often errors in measurement appear only when data from multiple sources are combined through data fusion. Mismatches in positional information, for example, can arise through limitations in measurement methods and sampling as well as through computational errors. These are frequently revealed when data about a common position on the Earth but from different sources are combined and found to be inconsistent. A rational spatial cyberinfrastructure should account for these major sources of error and facilitate the identification and reconciliation of them.

In spatial analysis, perhaps the most significant purely cyberinfrastructure challenges emerge when large volumes of multisource, multiscale digital data are combined into data fusion products for quantitative analysis. As we continue to expand the use of derivative, data fusion products, it is increasingly important to have a means to establish the provenance and correctness of the information products. The multistep nature of data fusion, along with the difficulty in simply moving and storing large quantities of data of any kind, challenges current methods of verification of data correctness and provenance. It is useful to consider an example of observational data across local and regional scales and use this as a basis to extrapolate to the long-term cyberinfrastructure requirements for future Earth-observing systems. Our results from surface mapping the meltwater field of icebergs provide a good case study.

Results and Discussion

Ship-based sampling of the Weddell and Scotia seas was conducted during 2005–2009 across a range of spatial and temporal scales, from meters to hundreds of kilometers (Fig. 1) and from minutes to weeks using a wide range of physical, chemical, and biological data sources. Through a combination of conventional ship-based sampling methods and a fast, surface-mapping method (Fig. 2, Right), we were able to detect and characterize the meltwater field from free-drifting icebergs on an unprecedented spatial scale (1–10\(^3\) m) that enables the connection of local- and regional-scale, space-based measurements (10\(^2\)–10\(^3\) m). The results of the survey for salinity shown in Fig. 2 (Right), a conservative property of seawater, indicate that the meltwater field is a large fraction (>75%) within the area sampled with salinity values below that of a control site 74 km away (Fig. 3). The surface-mapping grid has 20 × 1 km spacing on the grid pattern. This provides a basis for estimating the structure of the meltwater field geometrically using the percentage area distribution of Fig. 3. In ongoing research, this result will be combined with expendable bathythermograph and Acoustic Doppler Current Profiler (ADCP) data (cf. SI Text) to form an estimate of the volume of ambient seawater affected by the meltwater. These results are beyond the scope of this report and will be presented elsewhere.

These results provide direct evidence of the spatial characteristics of the meltwater field from free-drifting icebergs and its effects on the ambient seawater. Changes in measures of biomass as well as physical and chemical properties of the seawater are clearly and consistently revealed by surface mapping across the biogeochemical domains (Fig. 2, salinity, and chlorophyll-a) providing confidence that the surface maps captured real features and not sampling artifacts. Consequently, we now have a sound basis for more effectively sampling the influence of the meltwater on carbon flux from the surface waters to the sea-floor sediments using more time-consuming but versatile sampling methods such as hydrocasts. Current measurements made during the surface-mapping survey (Fig. 2, Bottom Right) were combined at a series of discrete depths resulting in a set of maps around the iceberg to a depth of approximately 250 m. These were used to produce animations showing the vertical structure of the circulation and wake turbulence in the near and far field relative to the location of an iceberg. Repeated sampling of the same geographic region revealed that the meltwater field persisted for at least 10 d after the iceberg moved through the area (4).

Because icebergs and their environs are so dynamic and the areas involved are so large (e.g., 10\(^{10}\)–10\(^{13}\) km\(^2\) in area, 1–20 km in longest dimension) compared to the scale of ship-based sampling techniques, a static sampling strategy, such as a Cartesian grid of sample locations using hydrocasts, is impractical. Using the fast surface-mapping method, we were able to focus our sampling by mapping the areas influenced by the iceberg meltwater first, followed by targeted sampling of the water column and biota, stopping or operating the ship at slow speeds (Fig. 2, Bottom Right) as required. Surface mapping enables us to define the important length scales and regimes of interest around an iceberg and helps to direct the location and spacing of water column sampling for physical and biological variables using the surface expressions of water mass properties detected by surface mapping.

The rapidity of surface mapping also enables extensive surveys that provide overlap with regional-scale space-based measurements across multiple pixels. This provides a sound basis for data fusion across local and regional scales of measurement (cf. SI Text). The overlap provides an essential means of connect-
ing iceberg-scale processes to synoptic, regional-scale oceanic processes that cannot be observed effectively from a ship. The local-scale maps provide subpixel information within the more coarse, regional-scale maps and densify the measurements within the extent of the local-scale maps. This cross-mapping spans 2–3 orders of magnitude resulting in both a locally high-resolution and regionally synoptic dataset.

The fused spatial products also span interdisciplinary domains, chemical, biological, and physical, each with unique sampling and scale characteristics that must be considered in the context of a cyberinfrastructure for spatial data exploitation. A key focus of this Antarctic research is to connect the fine-grained local measurements in the vicinity of icebergs (e.g., plant productivity, animal abundance and distribution, and seawater chemistry and physics) to the coarser-grained, regional measurements of ocean surface features (e.g., sea-surface temperature, chlorophyll, iceberg positions, and track history). This connection is made using time to index the position of the ship using global positioning system (GPS) time to combine continuously recorded physical oceanographic data and discretely sampled chemical and biological measurements (Fig. S1). By fusing these data, we are able to develop a spatially detailed picture of the interactions of oceanic water masses, iceberg meltwater, iceberg motion and relationship to local currents, tides, and wind. The label C18A identifies the location of the large tabular iceberg discussed in this paper. A partial western perimeter of the iceberg can be seen in the surface-mapping images in the right column.

Fig. 2. Correspondence of regional and local data products across temporal scales and spatial extent and resolution. The local-scale measurements on the right reveal subpixel processes within the regional-scale raster data on the left. (A) Phytoplankton biomass is estimated from Moderate Resolution Imaging Spectroradiometer (MODIS) chlorophyll-a products compared to fluorescence measurements from surface mapping of shipboard measurements of fluorescence. (B) MODIS sea-surface temperature (SST) is used to investigate thermal fronts versus near-field iceberg-melt affects on SST from shipboard thermosalinograph. (C) JASON-2 sea-surface elevation data is correlated with shipboard ADCP current data to investigate the geostrophic forces on iceberg motion and relationship to local currents, tides, and wind. The label C18A identifies the location of the large tabular iceberg discussed in this paper. A partial western perimeter of the iceberg can be seen in the surface-mapping images in the right column.
mapping also supported research on the state of biological communities (e.g., plankton biomass and community composition) with a sampling effort of one to several hours per sample. This approach saves time and provides more meaningful data and demonstrates the importance of spatial information in the joint interpretation of physical, chemical, and biological data as well as focusing the sampling strategy itself.

In the results presented here, positional accuracy and precision are controlled by the underlying accuracy and precision of the clocks used within the computers of the onboard sensor systems, the temporal and positional errors in GPS measurements at these high latitudes (also clock-based), and the temporal precision of the computational method of “binning” the data from multiple sensor systems with the navigational data (cf. SI Text). Time is the controlling factor because it is what is used as the selection and merge-key for associating sensor data with positional data (Fig. S1B). In principle, these sources of error are knowable, but in practice they are generally unknown and not explicitly accounted for. In the data fusion products, combining local-scale maps with regional-scale maps, the positional uncertainties in the field data are confounded with those inherent in the data products from the satellite sensor systems (Fig. S1A) for which the spatial errors are also poorly represented and accounted at the pixel level. This is the level that data fusion is done in gridded maps of the meltwater field on an areal and percentage basis. The control value is also plotted (black dashed line).

The open-source software development model exposes the programming language source code to continuous challenges by domain-expert user communities and allows careful analysis of each data transformation and fusion step. Very importantly, these codes typically run on a wide range of computational platforms and can be embedded in scalable computations and data analyses. This provides the opportunity for ongoing development of increasingly robust, automated error management schemes with explicit representations of sensor system measurement time and global system time, at the pixel level. Many instrument manufacturers already provide external time source interfaces by supporting the National Marine Electronics Association standard message from external GPS receivers to enable GPS time to be embedded in their sensor outputs. Ideally, all time references would be derived from a common global clock, but in lieu of that there is a need for capturing the local measurement time and the global system time simultaneously in a geospatial cyberinfrastructure. This will make it possible to postprocess the data for spatial errors using whatever is chosen as the global system clock and navigational reference system. This is GPS for many current research applications but could be other systems such as the Russian GLONASS or the future European Galileo system. Such pathways to improvement are not available in proprietary GIS software but fortunately alternatives to proprietary GIS exist and can be found in systems such as GRASS (http://grass.itc.it/) and QGIS (http://www.qgis.org/).

We also need improvements in data formats to represent positional uncertainty as part of the data and metadata structures in such a way that the error ellipse for any given georeferenced pixel can be easily accessed and correctly propagated in computations. This requires the representation of both time and position and the errors in each to be expressed explicitly in a data structure. An analogy that suggests itself is the method of computation with complex numbers where the real and the imaginary parts of the numbers are explicitly carried through the computation. This requires both specialized data types and operators to carry out this class of computations. Another interesting example is the use of orbital elements to precisely specify the position of an orbiting spacecraft. Although these “two-line” element sets may seem burdensome with respect to earth science data, they do provide a clear example of how space and time information are encoded to enable highly precise and accurate position determinations to be made.

There are related problems having to do with the georeferencing of place-names as well, for example, the Weddell Sea. When we talk about the Weddell Sea, it is generally understood to be defined by both physical oceanographic properties of the seawater as well as by the bathymetry and boundaries of the region but not by a well-defined polygon described in geographic coordinates. One approach to addressing this would be to carry along with the place-name a polygon with explicit geospatial error terms, as measures of uncertainty due to ambiguity in definition rather than measurement, or physical properties of the water mass along a given polygon edge. This way researchers would be better equipped to know if they are talking about the same place.

In our ongoing research, we are examining methods of properly representing the error budget in our surface-mapping calculations and generalizing their representation but have no general solution to these problems at present. Our current practice is to constrain the mapping using ship tracklines and designing sampling plans with this explicit purpose in mind. However, more work needs to be done in methods for mapping and statistically summarizing both navigational and observational errors.

The interdisciplinary nature of this research and the associated wide variety of sensors and sampling methods results in a large
number of data files: about 20,000 files for one research cruise with a moderate volume of data. Using the data from these three expeditions as a basis, we find that the minimum data volume (i.e., onboard data acquisition system, satellite data, ADCP, and derived analytic data) for each cruise is ~200 gigabytes (10^9 bytes) (Table S1). A typical field research program could have four of these cruises over a 3-y program resulting in a data load of ~3 terabytes (TB, 10^12 bytes) of data per research program, including both field and laboratory data. If this were extended to a 50-y monitoring project, a single copy of the complete data record is ~40 TB. This is not an extraordinary amount of data in contrast to the amounts generated by climate models, but it is enough to pose a significant challenge to the individual researcher and is well beyond the capability of many GISs to handle for reasons such as limited memory, sequential processing, and 32-bit addressing. Some of these limitations will be gradually removed through technology evolution and commercial development, but that is driven by business needs, not scientific needs. Focused efforts are needed within the scientific community to ensure we are prepared to handle these types and quantities of data.

Manipulating this data record using data fusion methods requires that multiple sources of data come together within a computer at some point regardless of whether the data are stored in a centralized or decentralized manner for archival purposes. This is because computation requires that data operands be held within primary computer memory (random access memory) at the same time. The implication is that the data must be contained in file systems accessible by a given processing unit for the duration of the computation however those file systems are deployed and connected to the processors. As our ability to produce high volumes of georeferenced data have grown, the demands on communication and data transfer capacities have become a rate-limiting step in scientific progress. On land, in the laboratory, this is a problem, and at sea it is a bigger problem due to much more limited shipboard resources and options.

Actual computation with data typically requires at least twice the disk space required to store it and more typically 3× including copying from one system to another and work space required for the processing of intermediate computational results. At present, and for the short-term future such as the next 5 to 10 y, there is no reasonable expectation of sufficient network communication infrastructure to enable the virtualization of this processing except in special cases. We are only just beginning to provision this type of computing within the research community through programs such as the National Science Foundation (NSF) TeraGrid XD (for extreme data). Despite these efforts, tera-scale data analysis resources will be scarce for the next 10 y. This scarcity will be exacerbated by the need within the geospatial cyberinfrastructure for improvements in spatially enabled data representations. A rational, coherent, spatial cyberinfrastructure for multidisciplinary research should be planned into Earth-observing systems from the outset, and these should be consciously organized around community standards and best practices. This will facilitate the development of better methods for error representation and propagation management as well as a broader understanding of them. Until we can achieve these goals, the capability for routine spatial and temporal fusion and visualization of data, in the range of terabytes, is needed to provide observing systems with the analytic capability to make scientific progress using the current ad hoc methods in the face of poorly characterized systematic and nonsystematic errors in observational and model data.

Other key features that should be factored into future observing systems include network interfaces that provide on-demand data access from within scientific intranets to host-platform data acquisition systems. For example, some existing policies of network management on the ships make it difficult to obtain access to real-time and near-real-time data. This can be overcome by designing the data acquisition systems with “demilitarized zones” in the network architectures where open access is encouraged rather than grudgingly permitted. Increasingly, security requirements in multiuser computing systems limit the ability of scientists to access and control critical parts of an operating system to verify the integrity of measurements or to rapidly access the most recent measurements as part of decision-support analyses. Networks should be designed to service science requirements without sacrificing security. The present models suffer from the retrofitting of post hoc approaches that often succeed only in keeping out the real users. We need new approaches to provide reliable and convenient access to data.

System services such as timekeeping and file-system interoperability must be reliable and explicit in order to enable internal and external consistency-checking especially when dealing with a multiplicity of data types, large numbers of files, and large volumes of data. Connecting natural phenomena across spatial scales requires careful attention to time and position and demands an understanding of the limitations of timekeeping and data interoperability in computing systems. For example, most common computing systems do not provide easy access to time at a precision of less than 1 s and have limited means for preserving accuracy by correcting time bias across clocks. Loss of accuracy and precision in time or position (e.g., temporal binning, truncating decimal degrees of longitude or latitude in data conversions) throw away important location information and introduce hard-to-discover ambiguities in position that are often revealed during data fusion in a trial-and-error fashion.

Finally, sharing data within a large, multidisciplinary science team requires a systematic method of storing, cataloguing, archiving, and distributing data to ensure that each member can unambiguously identify the data they are working with. The international community has recognized this problem and reached a multinational consensus on the need for change in the management of long-term scientific data (16). We have the same need within observational systems as the data are being produced. To address these issues we have developed the Digital Library Framework (DLF) and have been using it to support a number of scientific research projects through the California Coastal Atlas (http://CaliforniaCoastalAtlas.net) as well as other NSF-supported projects. The DLF is a data publication system that is designed to support the scientific enterprise as illustrated in Fig. S1 and employs the use of digital object identifiers to integrate the data holdings with the publication industry’s standard cross-referencing systems. The system is portable and is used onboard the ships to manage the onboard system data for analysis and interoperates by design with the shore-based system to provide the basis for an editorially controlled publication process for all types of scientific data but especially geospatial data. Spatial cyberinfrastructure efforts such as these are at the frontier in the discovery, development, and implementation of effective solutions to these long-standing and ongoing scientific needs.

Materials and Methods
In November–December 2005, during a research expedition (designated LMG0514a) on the RV (Research Vessel) Laurence M. Gould, we made ship-based observations that demonstrated the existence of a near-field to far-field gradient in several biological and geochemical parameters around a free-drifting iceberg (3). In an effort to better understand the distribution of meltwater around icebergs, a second expedition (NBp0806), on the RVIB (Research Vessel/ice Breaker) Nathaniel B. Palmer, in June 2008, was devoted to the development of improved sampling methods and exploring approaches to characterize the environment around icebergs. This focus led to a method of sampling the surface waters that we dubbed surface mapping because of its similarity to seafloor swath mapping and the commonly used “mowing-the-lawn” sampling pattern used with onboard multibeam sonar (Fig. 2, Right). The third expedition (NPb0902), in March–April 2009, also on the RVIB Nathaniel B. Palmer, was the first operational use of the surface mapping as a routine part of our iceberg characterization. Using the surface mapping sampling method in combination with surface water
measurements from a flow-through thermosalinograph instrument, we were able to characterize large areas \((10^2–10^3 \text{ km})^2\) near and around a large tabular iceberg, designated C18A, at full speed \((18.5 \text{ km h}^{-1} \text{ knots})\) without stopping.

Conclusions

The ability to correlate multisource measurements accurately and precisely across spatial and temporal scales requires careful management of the georeferencing and georegistration of the data and a clear understanding and control of the sources of error inherent in data processing. Because timekeeping is fundamental to geolocation systems, the relationships between measurement instrumentation and sources of time used in geolocation must be accounted for more effectively. This can be achieved by integrating geolocation–service system time into computing and instrumentation systems and by adding interfaces between them for synchronization. This level of sophistication in timekeeping is found in the geodetic community but not broadly in other scientific disciplines that need it. Improvements to the production of metadata, now widely recognized as important in scientific data management, need to move forward more rapidly so that they are automatically produced according to well-defined standards.

Metadata are not only important in documentation and archival but they are increasingly important in machine-to-machine interoperability and in our ability to automate the highly labor-intensive data processing necessary for high-resolution, high-volume data from future observing systems. The specification of spatial reference system metadata has enabled the geospatial community to achieve a high level of data interoperability. It needs to be expanded to explicitly provide for error management. In addition, temporal reference system metadata are needed to provide the equivalent for the time domain.

The determination and control of the propagation of temporal and geospatial errors must be improved. We need new models of these errors, and we can learn from our colleagues in the astronomical community. Models of errors in telescope performance have been used to make progress in observing system design and performance for decades. End-to-end modeling of measurements not only provides insight into the parameterization and analysis of error propagation, but it can also be used as an aide in planning data collection (17). Because the resultant models simulate subgrid (subpixel) processes, they can be integrated with geophysical models to make important connections between the geophysical, chemical, and biological domains at local scales of measurement.

As global and regional geophysical models approach the ability to resolve processes at local spatial and temporal scales, new observing systems should plan for the data interoperability needed to correlate observations with model predictions. This is essential to the effective assessment of the predictive skill of models and to advance the use of data assimilation with geophysical models (18). There is a need for this integration to be accounted for as part of the future modeling and instrumentation programs and this will require new interactions and roles between and within the modeling and observing communities.

Acknowledgments. We thank all the shipboard scientific personnel and crew on the R/V Laurence M. Gould (LMG05-14A) and the R/VIB Nathaniel B. Palmer (NBPO082, NBPO092) for excellent support. We especially thank Chief Scientist Kenneth L. Smith who ensured that we had sufficient ship time to develop and employ these methods during each of the cruises. We also thank two anonymous reviewers whose comments significantly contributed to improvements to the manuscript. This research was supported by NSF Grants ANT-0636723, ANT-0636809, ANT-0636440, ANT-0636543, ANT-0636719, ANT-0636813, ANT-0636730, ANT-0529815, ANT-0650034, and OCE-0327294, and by the David and Lucile Packard Foundation. RADARSAT-1 and PALSAR images were provided by National Aeronautics and Space Administration (NASA) through the Alaska Satellite Facility and a NASA investigator grant (to J.J.H.). NASA MODIS data were provided by the Goddard Space Flight Center Ocean Color system. Support for the Digital Library Framework was provided by NSF Grant OCE 0607372.

Supporting Information

Helly et al. 10.1073/pnas.0909306108

SI Text

Geospatial Cyberinfrastructure. The geospatial cyberinfrastructure used to support the surface mapping effort is shown in Fig. S1A and B. Fig. S1A depicts the cyberinfrastructure as part of an end-to-end research workflow and Fig. S1B details the data sources, data interfaces, and software used in the shipboard laboratory to produce the surface maps. Fig. S1A also depicts the research phases (precruise, in-cruise, and postcruise) to provide a context for the discussion of the data transportation requirements and the diversity of data sources needed to provide the regional and global-scale context for the results measured in the field.

The lines with arrows in Fig. S1A represent data transportation pathways and, as indicated in the legend, include both Internet connections and physical transportation of disk drives. The only Internet connection to the ships is via the INMARSAT satellite, which has limited bandwidth and is costly to use. Consequently, Internet connections are usually established only twice daily for the purposes of obtaining e-mail for cruise participants and to support administrative tasks. Exceptions are made for the reception of larger data transfers, up to about 1 Mbyte, but this is done only on an occasional basis by prior arrangement and approval by National Science Foundation due to cost. Consequently, large data transfers must be done by shipping external disk drives or computers between the ship and university laboratories.

Fig. S1B shows the detailed data flow necessary to georeference the time-series data collected by the shipboard sensor systems on the left (thermosalinograph, transmissometer, pCO₂, fluorometer) with the Global Positioning System (GPS) data received by the ship. The georeferencing is accomplished by algorithms that search the navigation data for the time interval that matches the sample time for a given measurement and extracts the longitude and latitude values for that time. Each sensor system has its own computer with its own internal clock. The clocks are synchronized using the network time protocol (NTP) method within the shipboard local area network using the time from one of the onboard GPS receivers. Each sensor system has its own sample rate, which is less than the GPS sample rate that is sampled at a variable rate ranging typically between 0.02–0.86 s.

Each sensor system is georeferenced by a special-purpose MATLAB™ code that is written to accommodate its particular data format, units, and sample rate, which is typically slower than once per second. Consequently, there may be multiple GPS time points for a given sensor time point. When this occurs, the average of the GPS locations is computed and assigned to the sensor time point observation. This is referred to as “binning” the time and represents a loss of positional precision of the observation. If you are traveling at one knot, in 1 s you move 0.514 m. So, at 10 knots, a 1-s sample rate represents a loss in positional accuracy of more than 5 m. This is in addition to the expected 10- to 15-m GPS uncertainty in positional accuracy that cannot be corrected due to lack of GPS base stations in this part of Antarctica. We have not been able to determine how much this expected midlatitude positional inaccuracy is affected by the high latitude location of our field sites.

Once the data have been georeferenced, they are visualized as surface maps of the parameter field and can be fused with similarly projected data from satellite sensors to resolve regional-scale features. The data fusion provides a quality control function as well as an analysis function because disparities and inconsistencies at the different scales of measurement tend to quickly show errors in the data or in the processing.

Table S1 presents an estimate of the data capacity requirement for a 3-yr research program utilizing the surface mapping approach combined with the laboratory research program to integrate the field data into the regional and global contexts. This table excludes other data sources such as video from a remotely operated vehicle or other continuously operating systems such as multibeam sonar and towed vertical profiling systems that would likely be employed in a future research program.

The physical measurement capability of the ship was central to the provision of the geophysical background to biological and chemical sampling. It was of particular benefit that the surface mapping method could be executed rapidly to provide hourly to daily information on the local to regional extent of surface water features in at least five important measurements relating directly to biological and chemical processes: sea-surface temperature, salinity, fluorescence, transmissivity, and partial pressure of carbon dioxide (pCO₂).

In addition to the surface parameters, we sampled the water column using hydrocasts [i.e., conductivity, transmissivity, and depth (CTD), water bottles] and expendable bathythermographs (XBTs). Sixty-nine (69) CTD casts were taken at various distances from C18A. XBTs were used during surface mapping surveys as well as during a circumnavigation of the iceberg to develop a map of the temperature “curtain” around the iceberg during a circumnavigation within 0.9 km of the iceberg. Another approach used XBTs deployed at regularly spaced times during a surface mapping survey near the iceberg; this survey was repeated 10 d later, after the iceberg had passed to examine the longevity of the temperature signal from the iceberg.

A 150-kHz Acoustic Doppler Current Profiler (ADCP) was used to measure currents in 8-m depth bins from 50 m below the surface to a maximum depth of 300 m. These measurements provide data on temperature, stability, and density. We also employed XBTs to obtain vertical temperature profiles while under way during surface mapping. This combination of sampling methods enabled us to combine the 2D surface mapping results with information about the physical properties of the water column to construct 3D images of the water mass near and around an iceberg.

Also during NPB0902, a 10-m² MOCNESS (Multiple Opening/Closing Net and Environmental Sensing System) trawl, comprised of six separately deployable nets for sampling particular depths, was towed parallel to the C18A at distance from 0.4 to 18.5 km. Although primarily intended for biological sampling, the MOCNESS was equipped with a CTD producing sawtooth profile in the top 300 m of the water column as the trawl was raised and lowered during the course of each tow. The CTD-based measurements and GPS ship position and time provide a geophysical and spatial connection between the biological and geophysical domains.

Regional synoptic imagery was obtained from a variety of sources. Moderate Resolution Imaging Spectroradiometer (MODIS) chlorophyll-a and sea-surface temperature (SST) were obtained from the National Aeronautics and Space Administration (NASA) Ocean Color team during the time we were in the field via e-mail attachments but images in Fig. 2, Left) are from data obtained after the expedition due to communication limitations while at sea. Because of bandwidth limitations onboard the Antarctic research vessels, only lossy, compressed images can be received during the expedition time. These are not suitable for quantitative analysis because of the lossy compression and are not georeferenced. Uncompressed data are obtained from the

Helly et al. www.pnas.org/cgi/doi/10.1073/pnas.0909306108 1 of 4
NASA archives for postcruise analysis. The GPS and DMSP (Defense Meteorological Satellite Program) are both received directly by onboard receivers.

RADARSAT-1 and PALSAR data were provided pre- and postcruise by the Alaska Satellite Facility at the University of Alaska, Fairbanks, with support from NASA and the US National Science Foundation. These files are also too large to transmit to the ships, but are used to examine the distribution of icebergs by size, mass, and abundance in the laboratory. Advanced Very high Resolution Radiometer (AVHRR) and National Oceanic and Atmospheric Administration (NOAA) satellite data are transmitted to the ships via direct readout but, like the MODIS data, are obscured by clouds and are useful only during rare periods of visibility from space. QuikSCAT scatterometer data was used by Dr. David Long at Brigham Young University during all research phases and is critical to our ability to navigate directly to icebergs of interest. Tracking reports and small images were transmitted on a daily basis to the ships during the in-cruise phase of the research.

Algorithms for data fusion between the local-scale and regional-scale data proceed by selecting a pixel off the regional-scale data (Fig. 2, Left) and mapping the neighborhood of the local-scale data to it. This is a spatial analog to the temporal binning performed in georeferencing the time-series data from the onboard sensors. A general expression for this algorithm is

$$\{ x(l, m), y(l, m) \mid p(i, j), q(i, j) \mid \forall l, m \subseteq i, j : R(l, m, i, j) \leq r \},$$

where $x(l, m), y(l, m)$ are points within the local-scale data lattice and the $p(i, j), q(i, j)$ are indices into the regional-scale data lattice. $R$ is a distance metric from the $(lm)$ indexed lattice points to the $(ij)$ indexed lattice points such that values at the $(lm)$ points within distance $r$ of $(ij)$ points are combined, using statistics such as mean or median, to correlate and compare values across spatial scales. This algorithm permits flexibility in defining the distance measure $R$ as well as the method of combining subpixel data.
**Fig. S1.** Research workflow (A) and shipboard laboratory data flow (B). The solid lines in A indicate relatively high-speed Internet connections between national and international satellite data resources and university research laboratories. The thin dotted line represents the INMARSAT Internet connection to the National Science Foundation's Antarctic research ships, the R/V Laurence M. Gould and the RVIB Nathaniel B. Palmer. The thick dotted line represents the physical transportation of disk drives or computers between the ships and the university laboratories. The research phases (precruise, in-cruise, postcruise) are labeled with representative durations. The collection of satellite data occurs during all research phases [indicated by the left- and right-pointing arrows from the non-Department of Defense (DoD) sources] but only a small component of this is provided to the ship during field campaigns to aid in the search for suitable icebergs to study (QuickSCAT) and to provide information regarding the SST and chlorophyll-a background when cloud cover permits (MODIS). ASF, Alaska Satellite Facility; GSFC, Goddard Space Flight Center; BYU, Brigham Young University; AVHRR, Advanced Very High Resolution Radiometer.
Table S1. Summary of the minimum data loading required to support a 3-y research program employing surface mapping based on results from 2009 cruise (NBP0902)

<table>
<thead>
<tr>
<th>Research component</th>
<th>Scale</th>
<th>Source</th>
<th>Data volume per sample, bytes</th>
<th>Samples per day</th>
<th>Number of days</th>
<th>Total data volume, Gbytes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field</td>
<td>Regional</td>
<td>PALSAR</td>
<td>2.3E-08</td>
<td>10</td>
<td>40</td>
<td>92.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MODIS</td>
<td>6.0E-08</td>
<td>4</td>
<td>40</td>
<td>96.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>QuikSCAT</td>
<td>5.0E-05</td>
<td>1</td>
<td>40</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DMSP</td>
<td>5.0E-06</td>
<td>1</td>
<td>40</td>
<td>0.20</td>
</tr>
<tr>
<td>Local</td>
<td>Shipboard</td>
<td>5.0E-08</td>
<td>1</td>
<td>40</td>
<td>20.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ADCP</td>
<td>2.0E-07</td>
<td>1</td>
<td>40</td>
<td>0.80</td>
<td></td>
</tr>
</tbody>
</table>

Total per cruise: 209

Total Field Program (4 cruises per year for 3 y): 2,508

Laboratory Monthly Time-series Sampling Twice per Month for 3 y

<table>
<thead>
<tr>
<th>Research component</th>
<th>Scale</th>
<th>Source</th>
<th>Data volume per sample, bytes</th>
<th>Samples per day</th>
<th>Number of days</th>
<th>Total data volume, Gbytes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regional</td>
<td>PALSAR</td>
<td>2.3E-08</td>
<td>10</td>
<td>72</td>
<td>165.60</td>
<td></td>
</tr>
<tr>
<td>MODIS</td>
<td>6.0E-08</td>
<td>4</td>
<td>72</td>
<td>172.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QuikSCAT</td>
<td>5.0E-05</td>
<td>1</td>
<td>72</td>
<td>0.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DMSP</td>
<td>5.0E-06</td>
<td>1</td>
<td>72</td>
<td>0.36</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total: 339

Combined 3-y total: 2,847

*Only the data sources used in that cruise are listed, so this table represents a lower bound on the data capacity for the surface mapping cyberinfrastructure. The table is organized into field and laboratory categories as well as regional- and local-scale data sources within each of those categories. The QuikSCAT mission ended on November 23, 2009 and a comparable alternative source will be used in future research.*