COLLABORATIVE RESEARCH: A TRANSPARENT-MIDDLE-LAYER COMPUTATIONAL AND DATA MANAGEMENT INFRASTRUCTURE FOR SYNOPTIC APPLICATIONS OF COSMOGENIC-NUCLIDE GEOCHEMISTRY

This is the 'Project Description' section of a proposal written by Greg Balco and Ben Laabs and submitted to the "Geoinformatics" program of the US National Science Foundation in August, 2019. Some parts of the project description required by NSF proposal rules but not particularly relevant to the main point (e.g., the 'results of prior work' sections) have been removed.

The goal of this proposal is to build the computational and data management infrastructure necessary for applications of cosmogenic-nuclide geochemistry to synoptic geochronology, paleoclimate, and surface processes research.

Building this infrastructure is important because cosmogenic-nuclide geochemistry underpins a broad range of research into both Earth history and Earth processes, including, for example: reconstructing polar ice sheet contributions to past and present sea-level change; diagnosis of climate dynamics from paleoclimate records; evaluation and validation of paleoclimate and ice sheet models; seismic hazard analysis for major fault systems; and multifactor analysis of tectonic and environmental forcing of erosion and sediment transport rates. The proposed computational infrastructure is critical for this field as it would be for any field of Earth science: it enables the visualization, synthesis, and analysis of large data sets needed to advance global-scale applications.

This proposal is also potentially more broadly valuable in Earth science because cosmogenic-nuclide geochemistry is exemplary of many areas of geochemistry and geochronology in that interpreting a geochemical measurement as some geologically useful parameter -- an age or a process rate -- requires an interpretive middle layer of calculations and ancillary data, both of which are the subject of active research and evolve rapidly. Any synthesis of data from more than one study requires continual recalculation of ages or rates from a constantly growing data set of raw observations, using a constantly improving calculation method. Although this challenge is not unique, especially in geochronology, cosmogenic-nuclide geochemistry stands out because of the complexity of the middle-layer calculations and the rapid development of both the calculation methods and the data set of calibrated or measured parameters needed to apply them. Thus, managing this problem in cosmogenic-nuclide geochemistry may provide knowledge, experience, and implementation models applicable throughout Earth science.

In the text below, we provide: (i) *an introduction to cosmogenic-nuclide geochemistry* and its important applications in Earth science, and how computational infrastructure development can transform and enable these applications; (ii) *background on some prototype computational infrastructur*e, with evidence that it is already valuable to users and facilitating new applications; and (iii) *a plan to develop an effective, professional, inclusive, and enabling computational infrastructure* that will support innovation in research and education, both in technical aspects of the field and for broader Earth science applications. Finally, throughout the proposal we highlight some ways in which this project would differ from existing cyberinfrastructure projects for geochemistry and geochronology. We relate these differences to the motivating ideas of the project, and argue, based on evidence from the prototype implementations, that they may lead to innovations that have not so far been possible.

This is a "Catalytic Track" proposal.

I. Science motivation: Introduction to cosmogenic-nuclide geochemistry and its applications.

I.1. Cosmogenic-nuclide geochemistry is a broad term for geochemical methods applied in Earth science research that rely on the production of rare nuclides by cosmic-ray interactions with rocks and minerals at

Earth's surface. Cosmic-ray-produced nuclides are geologically useful because the cosmic-ray flux is nearly entirely stopped in the first few meters below the surface, so the nuclide concentration in a rock sample is related to whether that sample has ever been exposed at the surface, how long it was exposed, and, if the sample is not now at the surface, when the exposure happened. This enables many applications in dating geologic events and measuring rates of geologic processes that move rock from the subsurface to the surface, or from the surface into the subsurface (see review in Dunai, 2009). Geochronological applications include "exposure dating" of landforms and surficial deposits to determine, for example, the timing of glacier and ice sheet advances and retreats (e.g., Balco, 2011; Jomelli et al., 2014; Johnson et al., 2014; Schaefer et al., 2016) or fault slip rates and earthquake recurrence intervals (e.g., Mohadjer et al., 2017; Cowie et al., 2017; Blisniuk et al., 2010); as well as "burial dating" of clastic sediments for studies in tectonics, paleoclimate, and human evolution (e.g., Shen et al., 2009; Granger et al., 2015). Surface process applications include quantification of weathering, erosion, and sedimentation rates and understanding landscape evolution processes generally (e.g., Portenga et al., 2011; DiBiase et al., 2011; Larsen et al., 2014). Many of these applications not only rely heavily on cosmogenic-nuclide data, but would be nearly impossible without them. Overall, the ability to make cosmogenic-nuclide measurements in surface rocks and sediments has played an important role in the ongoing transformation of geomorphology and surface process studies from qualitative to quantitative fields, and has become a central element of nearly all areas of Earth science that focus on surface processes.

The raw observations for all these applications are measurements of the concentrations in certain minerals of trace nuclides that are diagnostic of cosmic-ray exposure. Mostly these are rare isotopes of common elements produced in common minerals: widely-used nuclide-mineral pairs are ^{10}Be in quartz, ^{26}Al in quartz, and ^{3}He in pyroxene and olivine; less common pairs include ^{21}Ne in quartz, ^{3}He and ^{10}Be in other targets, and ^{36}Cl in feldspars and other minerals. Measurements of most of these nuclides require substantial physical research infrastructure, in particular accelerator mass spectrometers, but large US and international investments in this infrastructure over several decades have established cosmogenic-nuclide analysis as a fairly routine and widely available capacity.

The basic approach to, for example, measuring an exposure age, is fairly simple. One measures the concentration of one of these nuclides, estimates the rate at which it is produced by cosmic-ray interactions, and divides the concentration (e.g., atoms/g) by the production rate (atoms/g/yr) to obtain the exposure age (yr). For erosion rate measurement, a similar concept applies: the surface erosion rate is inversely proportional to the surface nuclide concentration, and the constant of proportionality is defined by the production rate. However, significant complications to these simple relationships arise because the cosmic-ray flux, and therefore the production rate, varies with position in the atmosphere and the Earth's magnetic field, and the production rate also depends on the chemistry and physical properties of the mineral and rock matrix. Production rate calculations are geographically specific, dynamic (because the Earth's magnetic field changes over time) and require not only a model of the cosmic-ray flux throughout the Earth's atmosphere, but an array of ancillary data including atmospheric density models, paleomagnetic field reconstructions, nuclear interaction cross-sections, and others. In addition, production rate models are empirically calibrated using large sets of "calibration data," which are nuclide concentration measurements from sites whose true exposure age is independently known.

I.2. The interpretive middle layer for even the simplest cosmogenic-nuclide applications, therefore, includes all these ingredients: physical models for geographic and temporal variation in the production rate; geophysical and climatological data sets; physical constants measured in laboratory experiments, and calibration data used to tune the production rate model. All these elements have been, and continue to be, the subject of active research. Historically, new production rate scaling models and magnetic field reconstructions have been developed every few years. New calibration data are continually generated by research groups worldwide and published with a frequency of 1-2 papers annually. The result of this continuous development is that nearly all cosmogenic-nuclide exposure ages, burial ages, or process rate estimates in published literature have been calculated with obsolete production rate models, physical parameters, or calibration data. It is rare (although possible) for improvements in middle-layer calculations to completely nullify or supersede the conclusions of previous research, but, regardless, any use of

published data that are more than 1 or 2 years old, or any comparison at all of data generated at different times or by different research groups, requires complete recalculation of any derived quantities from the raw data.

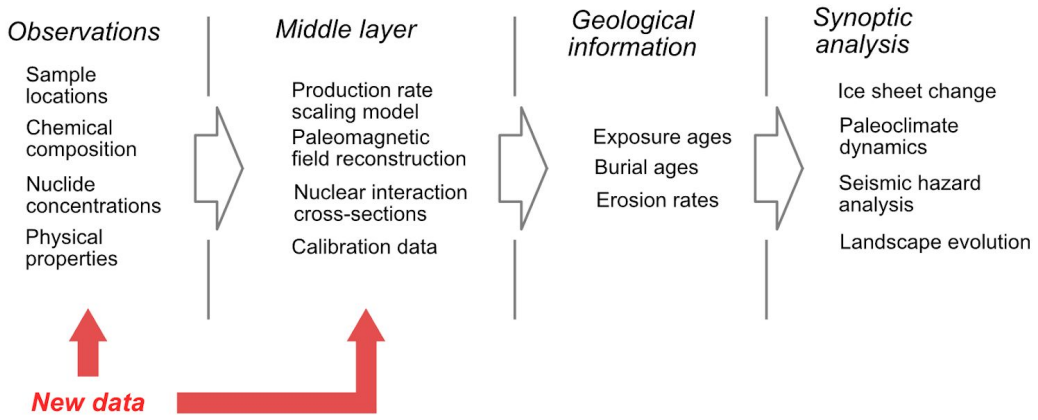

Figure 1. Conceptual workflow for applications of cosmogenic-nuclide geochemistry (or, in principle, nearly any other field of geochronology). Any large-scale analysis of ages or process rates needs to continually assimilate a growing observational data set and improving middle-layer calculations...or else it will be immediately obsolete.

I.3. Synoptic applications. As there are tens of thousands of cosmogenic-nuclide measurements in the published literature, the need for continuous recalculation of ages and process rates from raw data as the observational data set and middle layer evolve is a major challenge to the use of these data for any sort of synoptic research. The key motivation for this proposal, and also the motivation for developing the initial steps toward a prototype computational infrastructure for the field that we describe below, is that many of the most important applications of cosmogenic-nuclide geochemistry involve regional or global analysis of large data sets. These include, among others, analysis of regional and global glacier change (e.g., Young et al., 2011; Jomelli et al., 2011, 2014; Shakun et al., 2015; Heyman et al., 2016), global compilations of erosion rates (e.g., Portenga et al., 2011; Codilean et al., 2018), and analysis of ice sheet change and sea level impacts (e.g., Clark et al., 2009; Whitehouse et al., 2012), and they highlight the necessity for an effective computational and data management infrastructure. These studies were based on painstakingly compiled spreadsheets of hundreds to thousands of exposure ages or erosion rates that were calculated using methods available at the time. Because calculation methods and supporting data have improved, and the size of relevant data sets is increasing rapidly (for example, the amount of published exposure-age data relating to alpine glacier change worldwide has increased by at least 25% since the 2015 Shakun study), these compilations -- and possibly the overall conclusions of some of the studies -- are now obsolete. It is not possible to assimilate new data without repeating the entire manual compilation exercise, so no one knows whether the conclusions of the studies are still valid. Also, these data sets are redundant -- for the most part, each research group developed separate, variably formatted and organized, and likely mutually inconsistent compilations of raw data -- and, although “snapshots” of some of these compilations are archived with publications, they are, in general, not easily accessible for public viewing and data checking, or easily reusable for updated or new analysis. These are all serious obstacles not only to reproducing, updating, and validating these past studies, but to carrying out new research or exploratory data analysis that would take advantage of rapidly growing data sets.

I.4. A computational infrastructure to enable synoptic analysis. The motivation for developing a modern data management and computational infrastructure for this field is that we *can* do better than this, and the amount of cosmogenic-nuclide data now available for studies like these has become large enough that we *must* do better than this. Synoptic research in these areas and all other Earth science applications supported by cosmogenic-nuclide geochemistry could be enabled and, likely, transformed, by:

1. *A single source of raw observational data* that can be publicly viewed and evaluated, is up to date, and is generally agreed upon to be a fairly complete and accurate record of past studies and publications, beneath:
2. *A transparent middle layer* that serves derived geologic information such as exposure ages or erosion rates, calculated with a consistent and up-to-date method, to:
3. *Any Earth science application* that needs the complete data set for analysis or interpretation.

This basic structure allows continual assimilation of new observational data into part (1), and assimilation of improved calculation methods and new calibration data into (2), such that any analysis in (3) is always acting on a data set of exposure ages or erosion rates that is up-to-date, internally consistent, and uses the best available calculation method. This overall vision has inspired the prototype computational infrastructure described in the next section, and, we propose, would enable and transform the use of large data sets based on cosmogenic-nuclide measurements as paleoclimate proxy records, climate or ice sheet model validation targets, or portrayals of global landscape evolution processes. To summarize, cosmogenic-nuclide geochemistry is a critical part of nearly all Earth science research that focuses on processes active at the Earth's surface. It is supported by an extensive international physical research infrastructure that enables the generation of large data sets. The aim of this proposal is to catalyze a comparable computational and data management infrastructure that can move the field from site-specific to global-scale research.

II. Initial steps in prototype computational infrastructure.

II.1. Online exposure age calculators. Production rate calculations during the initial development of cosmogenic-nuclide geochemistry relied on simple computer programs (Clapp and Bierman, 1996) or spreadsheet templates (Phillips, 1996; Vermeesch, 2007) that were distributed among researchers. These became less viable ca. 2004-2006, after the development of several time-dependent production rate scaling models (Dunai, 2001; Desilets et al., 2006; Lifton et al., 2005) that required iterative exposure-age calculations not easily performed in spreadsheet templates. This, in turn, led to the first serious attempt to deal with middle-layer calculations using cloud computing, an "online exposure age calculator" developed by Greg Balco (co-PI of this proposal) in 2006, and described in detail in a later publication (Balco et al., 2008; http://hess.ess.washington.edu). A similar system was developed in 2007 and described by Ma et al. (2007), but was not widely used and to our knowledge is no longer functional. Essentially, the online exposure age calculator provided an input form, viewable in a web browser, into which one could cut-and-paste from a spreadsheet the raw data needed to compute an exposure age. A web server passed the input data to a script written in MATLAB code that computed exposure ages using a variety of scaling methods, and returned a web page with results formatted for easy pasting into the user's spreadsheet. Later iterations added an erosion rate capability, and, importantly, the ability to enter an arbitrary set of production rate calibration data, fit production rate scaling models to those data, and then subsequently calculate exposure ages using the customized production rate calibration. In 2014, it was migrated from MATLAB to Octave (the corresponding open-source environment) and since that time has been periodically updated with new production rate scaling models and background data.

Most evidence suggests that the online exposure age calculator was transformative for the field, and played an important role in developing exposure dating and erosion rate measurements from an obscure area of geochemistry to a routine capability underpinning global surface process research. It has been used in hundreds of studies and performs more than 100,000 exposure age calculations annually. It made the results of exposure-dating studies more credible, reproducible, and accessible to non-specialists by allowing readers to verify the accuracy of published results simply by pasting the data into the online calculator themselves. It substantially improved the quality of data reporting in publications by creating a *de facto* standard that any paper must include all data needed to recalculate ages or erosion rates using the online calculator. The examples of large-scale, synoptic Earth science research cited above used the online exposure age calculator to generate internally consistent data sets, and might not have been possible without it.

Recently, two similar online exposure age (but not erosion rate or production rate calibration) calculators have been developed: the “CRONUSCalc” and “CREp” systems (Marrero et al., 2015; Martin et al., 2017), and the typical best-practice workflow for exposure-age research at present relies on manual, asynchronous use of one or more of these services. Most researchers interested in comparison or analysis of exposure-age data (i) maintain a spreadsheet of their own and previously published observational/analytical data; (ii) cut-and-paste from this spreadsheet into an online calculator; (iii) cut-and-paste calculator results into the spreadsheet; and (iv) proceed with analysis of the results. Although the ability to use the online calculators in this way to produce an internally consistent set of results has made analysis of large data sets drawn from multiple sources possible, it creates redundancy and inconsistency among separate compilations by many researchers, relies on proprietary data compilations that are, in general, not available for public access and validation, interposes many unnecessary steps between data acquisition and downstream analysis, and creates many obstacles to dynamic data assimilation into either the source data or the middle layer. The workflow for erosion rate applications is not as well standardized: although use of the online calculator for erosion rate compilations is common (e.g., Portenga, 2011), software that embeds erosion rate calculations into GIS packages is also available (Mudd et al., 2016).

II.2. Laboratory information management system (‘LIMS’). This is not a primary focus of this proposal, but we discuss it here for completeness in reviewing prototype computational infrastructure. LIMS are common commercial/industry software systems that capture laboratory data -- sample weights, processing actions, reagent or spike additions, etc.. -- and store them in a database. The advantages of a LIMS over the typical alternative of recording process data in a notebook or spreadsheet are that (i) all information is recorded in a single location, so error correction easily and accurately propagates into downstream results, (ii) calculations rely on a single code, which makes errors less likely relative to distributed spreadsheet calculations, and (iii) observational data and results can be seamlessly transferred to downstream applications. Although most cosmogenic-nuclide laboratories have not adopted LIMS and continue to use spreadsheets for this purpose, a LIMS (‘ChemDB’) for the most commonly measured nuclides (^{10}Be and ^{26}Al) was developed in 2005 at the University of Washington by Greg Balco and Zach Ploskey, is open source and publicly available (https://github.com/zploskey/chemdb), and in use at several laboratories. Balco has also developed a similar system for ^{3}He and ^{21}Ne analysis at BGC (although this code has not been distributed), and a recently funded NSF-Earthcube project (EAR-1740694) aimed broadly at integration of LIMS and database systems for many areas of geochronology may also generate new LIMS capabilities applicable to cosmogenic-nuclide geochemistry.

II.3. A prototype transparent-middle-layer database. The next step towards modern data management infrastructure was the ICE-D (“Informal Cosmogenic-nuclide Exposure-age Database”) project developed by Balco in 2014 (http://www.ice-d.org). Its original purpose was simply to manage exposure-age data collected in Antarctica by Balco and colleagues, but project development coincided with the widespread availability of free or low-cost cloud computing services from providers such as Google Cloud and Amazon Web Services, and it grew into a test of several ideas that are important for this proposal. Specifically, it is the first (to our knowledge) attempt in geochronology to implement the idea of a transparent, dynamic middle layer between a database of raw observations and a resulting feed of derived ages. It includes four elements that run on inexpensive or free cloud computing services: (i) a relational database (MySQL on Google Cloud SQL) that stores raw observations like nuclide concentrations and sample information; (ii) an API for the online exposure age calculator (running on a Linux virtual machine on Google Compute Engine) allowing programmatic access to the Octave code that computes exposure ages; (iii) improved code for the exposure age calculator that is fast enough to support dynamic serving of exposure ages; and (iv) a web server (running on Google App Engine). The web server responds to browser requests by obtaining raw data from the database, dynamically acquiring exposure ages from the online calculator, and returning web pages displaying exposure ages with site and sample information, geographic context, and some downstream analysis of the ages (Figure 2). Thus, a user can explore and evaluate large data sets of exposure-age data, originally collected at many times and by many people, in an internally consistent form with all ages calculated using the same up-to-date method. It replaces the conventional asynchronous workflow, in which each user must

maintain their own data set and manually update it when data or calculation methods change, with a transparent system that manages data storage and all middle-layer production rate and age calculations, allowing users to focus entirely on analysis of the exposure ages. Because only raw observations are stored and all age calculations are dynamic, changes to the database are immediately propagated into exposure-age results. The ICE-D prototype, therefore, provides a proof of concept for our overall vision of a system that can maintain a single, publicly viewable database of observational data and serve an internally consistent set of derived ages through a transparent middle layer.

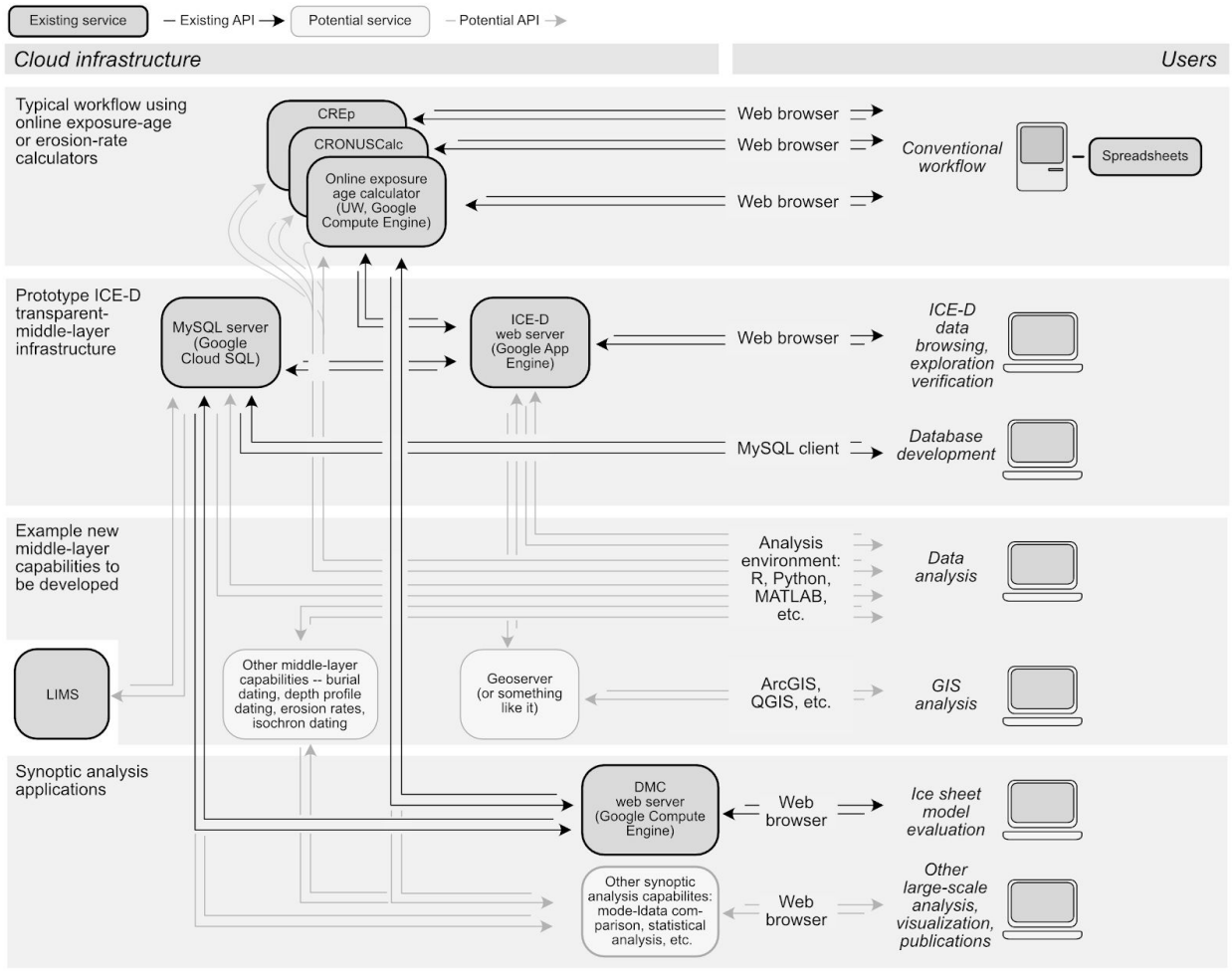

Figure 2. Generalized topology of existing prototype cloud infrastructure and some examples of proposed new developments. Cloud computing services interact with each other to supply raw data, calculated exposure ages, and other derived products to users at various stages of analysis. Server applications such as the DMC:ANTARCTICA prototype can serve live, dynamically updated results of analysis operations to websites intended for data exploration or visualization, or potentially to online publications.

At present, the ICE-D project includes several focus areas, each of which includes a small fraction of the total inventory of published cosmogenic-nuclide data, but represents the complete, or almost complete, data set needed for a particular synoptic application. The first focus area is ICE-D:ANTARCTICA (http://antarctica.ice-d.org), which contains nearly all known cosmogenic-nuclide data, both published and unpublished, from the Antarctic continent. Although this is a relatively small amount of data (~4600 measurements on ~3400 samples), it is the complete data set needed for continent-scale applications such as reconstructing past changes in Antarctic ice sheet volume and consequent sea-level impacts, or validation of numerical models for Antarctic ice sheet change. Subsequent ICE-D projects include

ICE-D:CALIBRATION (calibration.ice-d.org), which consolidates data from sites with independently known exposure ages that are necessary for calibration of production rate scaling models, and ICE-D:ALPINE (alpine.ice-d.org), which is the most ambitious part of the project so far (more than twice the size of the Antarctic data set with ~10,000 measurements) and is intended to include all exposure-age data relevant to alpine glacier change worldwide and facilitate use of these data as a global climate proxy record.

The ICE-D project is distinct in several ways from other data management or cyberinfrastructure projects in geochronology or geochemistry. It is an entirely "grassroots" project initiated mainly by one individual with no significant training or work background in software development, and further developed by researchers and students with expertise in exposure dating, but limited or zero software development background. It has no institutional sponsor, no dedicated funding source (see discussion below), and no technical support resources. Instead of starting from a formal planning or design process, or specification of user requirements, it developed organically in response to research needs. Its focus on fully populating data collections that are relatively small but nevertheless sufficient for certain applications enables more rapid realization of research value than projects that aim to archive a certain type of analytical data regardless of area or application. Unlike many geoscience data management projects, it is strictly forward-looking rather than backward-looking in two ways. First, it aims to be a dynamic research tool rather than an archive, and replaces elements such as formal peer review of included data, versioning, and strict metadata requirements with continuous, unmanaged addition and update of data by multiple users, coupled with a "trust-but-verify" approach that requires some level of trust in data quality on the part of users, but allows them to justify (or not) their trust with complete, granular, and public data viewability. It also looks forward rather than backward in that it strictly separates observational data (which are not made obsolete by improvements in middle-layer calculations, and are therefore important to preserve), from derived exposure ages (which are assumed to be rapidly obsolete, and are therefore not important to preserve), and makes no effort to reproduce or archive past published age calculations.

Either in spite of or because of these aspects of the project (we argue "because"), available evidence indicates that the ICE-D project has been surprisingly successful, has become a critical resource for several user groups, and is potentially transformative for the field of exposure-dating. No attempt has been made to gather conventional usage metrics (e.g., hit rates, user tracking, etc.) for the ICE-D websites and APIs, so to support this assertion we rely on several lines of evidence:

Adoption as reference database for Antarctic ice sheet change researchers worldwide. Personal interactions, discussion, and contacts with researchers involved in cosmogenic-nuclide research in Antarctica indicate that ICE-D:ANTARCTICA has been rapidly adopted by the international research community in this field as a common resource. Balco receives multiple requests each year to add newly generated data. Published papers (e.g., Johnson et al., 2019; Small et al., 2019; Nichols et al., 2019) and conference presentations in the field now commonly reference it as a data source.

Inclusion in new proposals. ICE-D:ANTARCTICA was a focus of a recent successful NSF proposal by BGC postdoctoral researcher Perry Spector (OPP-1744771; Balco was PI of this proposal for administrative reasons) to evaluate ice sheet models using the continent-wide exposure-age data set. We know of at least four recently submitted NSF proposals that would either use ICE-D databases as central project components or contribute data to them, and at least three proposals to funding agencies in other countries (France, New Zealand, Australia) that highlight integration with ICE-D databases as positive elements of the proposal.

Supportive reviews of related proposals. Although there has been no effort to formally gather user input on the ICE-D project, peer review of the Antarctic data-model comparison proposal most likely sampled the ICE-D:ANTARCTICA user pool. Three out of six reviews specifically highlighted the ICE-D project as a valuable resource, including very supportive language such as:

> *"...the database ICE-D:ANTARCTICA...is widely admired for its consistency and quality of the data."*
>
> *"...the ICE-D:ANTARCTICA database of cosmogenic nuclide observations [is] very positive both within academic circles and in communication with the public..."*

"Broader impacts of the proposal include...further development of the ICE-D database for the scientific community...very impressed with this superb resource and hopeful that this project may spur development of a global TCN database."

User engagement. Researchers and students worldwide who are interested in analysis of exposure-age data have expressed strong interest in participating in the ICE-D project. We have facilitated this participation through the idea that the basic, commonly-used, industry standard software tools used for the ICE-D infrastructure allow significant progress simply through giving researchers and students, even those without any software background, enough basic knowledge about the computational infrastructure to contribute to project development. In effect, because the project does not have resources for dedicated developers, we have adopted a model where the users are co-developers. In three small workshops during the past two years, PI Balco trained a total of 6 US and international researchers, mostly early-career, and 4 graduate students in aspects of the software infrastructure, including working with the MySQL databases and ingesting data into downstream analysis frameworks such as MATLAB and ArcGIS.

Monetary contributions. An extremely unusual aspect of the ICE-D project is that, because the project has no dedicated funding, hosting costs on Google cloud services have been paid by Balco personally. Although costs are very low ($15-$30/mo), they total $1168.08 since the project start in 2014. To defray these out-of-pocket costs, Balco solicited contributions via the ICE-D website, which generated $785 in offsetting funds. Although some of this is from institutional funding, several researchers and students made personal contributions of $10-$100 to support the project. Frankly, we can propose no stronger evidence that this project has value to the research community than the observation that its members are willing to make personal financial contributions to sustain it.

II.4. Other databases. The "OCTOPUS" database is a 2018 compilation of catchment-scale erosion rates calculated from cosmogenic-nuclide measurements on fluvial sediments worldwide (Codilean et al., 2018). It has no connection to any of the other projects described in this section and was developed with Australian government funding. Although similar compilations were distributed in spreadsheet form in the past (Portenga et al., 2011; Willenbring et al., 2013), this one is accessible through a web interface and as an online GIS coverage. It is similar to the ICE-D project in that it aims to collate a relatively small collection of cosmogenic-nuclide data for a specific synoptic analysis application and deliver it online, but it differs in that (i) so far, it is static, versioned, and has not been regularly updated, and (ii) it archives both observational data and derived erosion rates calculated with one production rate model, so does not have the transparent-middle-layer property of the ICE-D project.

II.5. Funding history and current support for computational infrastructure. Initial development of the first online exposure age calculator in 2005-06 was supported by several months of postdoctoral salary for Balco provided by "CRONUS-Earth," a five-year, multi-PI NSF project aimed at improving capabilities in cosmogenic-nuclide geochemistry. Balco was not a PI of that project, and there has been no further support for the initial online calculator, except that the U. of Washington has continued to maintain the Linux server that hosts it. The later "CRONUSCalc" online calculator (Marrero et al., 2016) was developed by Ph.D. student Shasta Marrero and colleagues with CRONUS support. The CRONUS project ended in 2011, and we are not aware of any continuing support for CRONUSCalc. Development of CREp was supported by French government funding and, as far as we know, it has continuing support. OCTOPUS was supported by an initial grant from the Australian government, and we believe that continued support has been applied for. The ICE-D project has no dedicated funding, although a map interface for ICE-D:ANTARCTICA was contributed by the NSF-funded Polar Geospatial Center. As noted, continuing monetary costs are supported by Balco personally, and development activities by BGC through Balco's salary. ICE-D workshops were partly supported by small contributions from Lawrence Livermore National Laboratory and the 'Earthchem' program, but most participants attended at their own expense.

II.6. Overall situation. From the perspective of this proposal, the critical lessons from this section are:

First, the complexity of the middle-layer calculations, and the need for continuous recalculation, in exposure-age and erosion-rate applications led the cosmogenic-nuclide community to become early adopters of rudimentary cloud computing services such as the online exposure age calculators. These capabilities, although quite simple, made possible the few large-scale analyses of cosmogenic-nuclide data that highlight the potential value of these data for important Earth science applications in surface processes, paleoclimate, and tectonics. However, these capabilities are inadequate to support similar analyses on rapidly growing data sets that now include tens of thousands of measurements.

Second, the ICE-D project prototypes a transparent-middle-layer infrastructure for solving this problem and enabling efficient, dynamic, and scalable analysis of these data sets. This could be a significant step forward, not only for the specific field of cosmogenic-nuclide geochemistry, but also for other fields of geochemistry and geochronology that utilize equivalent middle-layer calculations.

Third, the rapid expansion of the ICE-D prototype has created a situation where the project is too large for one person, who is not a trained software developer, to manage effectively in its current form, fully develop existing capabilities, or create new capabilities. The rapid adoption of the ICE-D:ANTARCTICA database as an important resource by Antarctic researchers worldwide has also created the precarious, and, frankly, rather bizarre, situation in which an important community resource -- that is an element of significant funded and proposed research projects in several countries -- is at risk of disappearing completely if Balco's personal credit card is not accepted for payment by Google.

Therefore, the premise of this proposal is that the prototype transparent-middle-layer computational infrastructure represented by the ICE-D project has been extraordinarily successful as a grassroots effort with very little institutional or funding support. It has been enthusiastically adopted by users and shows potential for major progress in synoptic analysis of cosmogenic-nuclide data. The purpose of this catalytic-track proposal is to move the project forward from a prototype with great potential to a fully realized infrastructure that can support this potential.

III. Research plan.

This is a 'catalytic' proposal that aims to begin with the prototype transparent-middle-layer infrastructure represented by the ICE-D project, professionalize and fully develop it using industry-standard, open-source software, and expand it into a set of organized, documented, and well maintained research databases, APIs, cloud services, analytical tools, and software development projects that can scale as data sets become larger and calculation methods become more complex. This will then form the basis for realizing the potential of the transparent-middle-layer concept for dynamic and scalable visualization, exploration, and analysis of geochronological data.

The research plan has three elements. One, middle-layer software and database development based on the ICE-D prototype. Two, development of example research applications. Three, engagement and training workshops, modeled on the existing ICE-D workshop program and aimed at embedding transparent-middle-level concepts and infrastructure into research workflows, for researchers and graduate students. Here we outline the specific tasks and resources associated with these objectives.

III.1. Project team. The expertise needed to execute this project includes (i) full-stack web and database development using industry-standard software (for this project, the Linux-Apache-MySQL-Python or "LAMP" stack); (ii) management of open-source software development projects; (iii) scientific programming, mainly as applied to production rate calculations, data analysis, and model fitting; (iv) Earth science applications of cosmogenic-nuclide geochemistry; (v) Earth science and geospatial technology education, and (vi) overall scientific guidance. The main challenge in assembling this expertise is the need for an experienced software developer with experience in industry-standard full-stack development, and we considered several models for this. Neither PI has access to in-house programming staff who could be tasked to this project. Some programming tasks could potentially be outsourced via freelance/gig marketplaces, without the need to hire dedicated staff. However, this approach requires expertise in

specification, bidding, testing, and acceptance of contract software, and we do not have this expertise or access to it. Hiring a person with such expertise would be costly. Thus, we propose a structure in which two dedicated project staff will be located at BGC and supervised by Balco: a software developer to be responsible for design, construction, and management of the overall software infrastructure, and a postdoctoral researcher, with expertise in scientific programming and cosmogenic-nuclide applications, to focus on middle-layer calculations and downstream applications. At NDSU, Laabs will supervise graduate and undergraduate students who will build databases relevant to our initial application focus in paleoclimate and glacier change, link them to geospatial data and analyses, and integrate them into research and education. Thus, the project team will consist of the following:

- Greg Balco (co-PI) is a research scientist specializing in cosmogenic-nuclide geochemistry and its applications, currently focusing on Antarctic ice sheet change research. As noted above, Balco has been responsible for much of the existing computational infrastructure that this proposal builds on.
- Ben Laabs (co-PI) is a geoscience faculty member supervising graduate and undergraduate research and education, with expertise in geospatial technology and glacier change applications of cosmogenic exposure dating. Much of his research involves applying geospatial and geochronological data pertaining to mountain glaciers to regional analysis of glacier change and paleoclimate.
- A professional software developer with experience in full-stack web development, cloud computing infrastructure, and open-source project management.
- A postdoctoral researcher with expertise in computational aspects of cosmogenic-nuclide geochemistry.
- Graduate and undergraduate students at NDSU. Students supervised by Laabs typically combine research and coursework in geosciences and geospatial technology, and students in this project will be further trained in database systems and management through computer science courses. The aim is to build a cohort of students with both Earth science research experience and a background in computational technology and data analysis who can contribute to all aspects of this project.

One significant potential risk to this plan is that the software developer is critical to many of the project goals. This person will not only be expected to perform specific programming tasks, but also to engage with the scientific aspects of the project, participate in overall project design, and train and interact with students and workshop participants who may have minimal programming experience. This expertise has high market value outside academia, and hiring this person represents a significant part of our budget. Also, market demand for these skills creates a risk that it might not be possible to find a highly qualified person for this position at all. However, we see no reasonable alternative to this approach: we cannot build a modern, effective, and scalable software infrastructure without professional-level skills.

III.2. Research plan part 1: Software development. The overall goals of this element of the project are to (i) professionalize and make scalable the basic elements of the prototype software, (ii) build out APIs to enable connectivity between databases, middle layer, and analysis or data visualization software used for applications, and (iii) develop the capabilities of the middle layer for applications beyond simple exposure-dating. This section breaks down these overall goals into more specific tasks.

- *Robust and scalable web server* (developer). The current ICE-D web server software is functional, but does not meet basic standards for software design and documentation, and is not open source. We will replace it with professionally designed and built, scalable, and well-documented software, with development managed as an open-source project. At present, we expect this will continue to use the LAMP stack (which is entirely open-source) running on Google cloud services.
- *Robust, scalable, standardized, and well documented APIs* (developer). This focuses on connectivity between database server, web server, middle-layer calculations, and user applications software (e.g., Fig. 2). We will produce well-documented, industry standard APIs that not only allow all aspects of the server infrastructure (exposure age calculator, web server, etc.) to interact programmatically, but also allow user access via analytical (e.g., MATLAB, R, etc.), visualization, or GIS software to all levels of raw data and calculated results.

- *Scalable middle layer* (Balco, postdoc, developer). This aspect will focus on developing a scalable capability to incorporate a range of middle-layer calculations into a single architecture. The prototype ICE-D middle layer includes only exposure age and erosion rate calculations based on the MATLAB/Octave code underlying one of the online exposure age calculators. We will build a more generalized infrastructure to allow existing and to-be-developed code for other applications of cosmogenic-nuclide geochemistry, for example burial dating, depth profile dating (e.g., Hidy et al., 2010), cosmogenic-nuclide paleothermometry (Tremblay et al., 2014, 2019), or inversion of exposure-age data sets to constrain landform evolution models (Applegate et al., 2010) to plug into the middle layer. The aim is that researchers who develop such code can make it available through a reasonably standardized and secure scheme that will support a diverse set of programming environments (e.g., R, Python, Julia, etc. in addition to MATLAB/Octave).
- *Production rate calibration* (Balco, postdoc, developer). We will link the ICE-D:CALIBRATION database of production rate calibration data to the middle layer to enable validation and exploratory analysis of scaling methods as well as assimilation of new calibration data into downstream analyses.
- *Open-source project management* (developer). We will manage all aspects of software development as public open-source projects using standard, open-source, distributed version control systems.
- *Database management infrastructure* (developer). At present, the ICE-D prototype manages database editing by multiple contributors using only basic tools available in MySQL itself, which is inconvenient and risky. As we expect that continuing the workshop program will result in steady growth in the number of contributors and the number of focus area databases, we will develop an improved user-contribution management system and programmatic data checking scheme that will make it easier to enter and validate data, as well as recover from errors. In addition, we will develop a structure for storing results of some calculations in the database to facilitate search on calculated parameters such as exposure ages.
- *Documentation* (Balco, Laabs, postdoc, developer). We will set up a documentation scheme, most likely based on open-source wiki packages, to house documentation for database structure, APIs, etc., and attempt to generate documentation that is as complete as possible.
- *Links to LIMS, other databases* (Balco, developer). We will explore/implement links between the ICE-D databases and external LIMS (ChemDB, others) as well as other geoscience archives or databases (e.g., NEOTOMA, USPRR, OCTOPUS, etc.). This is not a core element of initial project development, but we anticipate that it will be important in improving measurement traceability (LIMS) and in enabling downstream analysis applications (other databases).

III.3. Research plan part 2: database development. This aspect of the project will continue the prototype strategy of developing collections of cosmogenic-nuclide data that are relevant to specific larger-scale applications.

- *Continued buildout of ICE-D data collections* (Balco, Laabs, postdoc, students, workshop participants). We will continue and expand the existing program of improving, maintaining, and developing focus area data collections with groups of researchers in each focus area who can take responsibility for data compilation and maintenance. A 2019 workshop initiated a collection of exposure-age data relevant to Greenland Ice Sheet change (http://greenland.ice-d.org). We have had discussions with other researchers about developing comparable data collections for (i) exposure-age data related to the Laurentide and Scandinavian Ice Sheets and (ii) diverse cosmogenic-nuclide data used to constrain earthquake frequency and slip rates on major fault systems worldwide. We expect that initial future developments will be in these directions.
- *Links to ancillary/geospatial data* (Laabs, Balco, postdoc, students, workshop participants). We will link existing and future ICE-D data collections to ancillary data needed for downstream applications. Laabs and students will focus on linking a geospatial database of Pleistocene glacial features with exposure-age data, including compiling new and existing mapping of moraines or other landforms, glacier area reconstructions, and linking these geospatial data to the ICE-D:ALPINE database, with the aim of enabling paleoclimate applications that require both geochronological and geospatial data. Linking marine radiocarbon data that constrain the extent of the Antarctic Ice Sheet (e.g., Bentley et

al., 2014 and references therein) to ICE-D:ANTARCTICA is another example, and we anticipate that additional similar needs will arise from workshops.

III.4. Research plan part 3: Applications development for research and education. This element focuses on the end uses of cosmogenic-nuclide data sets coupled to middle-layer calculations. A prototype of what is possible here is a project being developed by BGC postdoc Perry Spector which allows comparison of observed cosmogenic-nuclide data at sites in Antarctica with predictions from ice sheet models (http://dmc.ice-d.org). Here we identify similar concepts that will be the main focus of graduate and undergraduate research supervised by Laabs and will guide initial work in this area. In addition, an important element of our proposed workshop program is that participants will bring new ideas and needs for synoptic applications that we have not already planned, and we expect that work on this part of the project will evolve as these needs emerge.

- *Glacier change applications* (Laabs, Balco, postdoc, developer, students, workshop participants) Laabs and students will focus on applications involving (i) reconstructing paleoclimate information from exposure-age chronologies for past glacier change, and (ii) comparison of paleoclimate model predictions to exposure-age chronologies. This will involve the work described above in linking exposure-age data to geospatial data needed for glacier reconstruction, as well as the additional step of using these data to infer equilibrium line altitudes and other paleoclimate parameters from glacier reconstruction. In effect, this represents an extension of the middle-layer calculations to serve not only exposure ages, but also climate proxy data associated with the exposure ages. The aim of this work will not only be to better enable use of glacier chronologies as paleoclimate proxy data, but also to (i) integrate cyberinfrastructure development into undergraduate and graduate Earth science education to make students better prepared to succeed in both science and other fields, and (ii) build the capability to link small-scale undergraduate and graduate research (e.g., mapping or geochronology of a single glacier system) to the wider-scale importance of this work (regional- or global-scale paleoclimate reconstructions).
- *Educational resources* (Laabs, developer, students) The present ICE-D database infrastructure supports the association of exposure-age data with photos, videos, data sets, or anything with a URL. Given proper interface design, this association could potentially support rich classroom exercises aimed at relating geological information to chronological data; for example, exercises could explore the relation between landform age and geological observations inferred from photos, video, or digital elevation data, or include 'virtual field trips' aimed at providing students with a sense of the overall age structure of mountain landforms. Laabs, with NDSU students and other project staff, will develop scalable educational resources that integrate exposure-age data with imagery, geospatial data, and geologic context, and test them in geology classes at NDSU (Climate Change, Glacial Geology).

III.5. Research plan part 4: engagement strategy. Our engagement strategy supports our goals of (i) enabling and transforming synoptic applications of cosmogenic-nuclide geochemistry, and (ii) building a community of researchers who can embed transparent-middle-layer infrastructure in their research workflow, and who can participate in building and maintaining the databases and the middle-layer calculations in the future. It builds upon the past ICE-D training workshops and focuses on creating positive incentives for researchers and students to work with and take ownership of the infrastructure we are developing. In this section we highlight three guiding principles for our engagement strategy and list the steps we will take to implement it.

Principle 1: build positive incentives for users. An often noted obstacle to participation in community data management infrastructure (e.g., Fleischer et al., 2011; van Noorden, 2013; Fowler, 2016) is the conflict between the broad, generalized incentive for the overall community to develop centralized infrastructure, and the immediate incentives of researchers who may view individually authored publications as more critical to career-development objectives. Although resolving this conflict is well beyond the scope of the present proposal, our approach is not to rely on a diffuse community incentive to drive engagement, but instead to provide positive, immediate incentives for individuals. To do this, we will identify what potential users view as their most important goals (e.g., publishing high-impact papers), identify the tasks needed to accomplish those goals that are most time-consuming or difficult

(e.g., statistical analysis, generating statistical or graphical comparisons of large quantities of new or published data, comparing data with complex model predictions), and make it so that engaging with our proposed data management and computational infrastructure makes it easier to carry out those tasks. User engagement should represent a trade: a user provides a service to the community by making data or tools available, and in exchange is provided with services that help to fulfill their own individual goals faster, better, and more easily.

Principle 2: engage influencers. We will reach out to specific individual scientists, identified by their publications, meeting presentations, or other research activity, who are best positioned to benefit from improved computational infrastructure, and whose work will highlight how improved infrastructure can facilitate new discoveries. Although we do not know exactly who these "influencers" will be, analogy with those already engaged with the ICE-D project suggests that they will be early- or mid-career researchers who generate and/or use large data sets of cosmogenic-nuclide data in paleoclimate or ice sheet change studies or quantitative geomorphology. We will invite these researchers to participate in our workshop program and guide our efforts in developing new data collections, new middle-layer capabilities, or new links to downstream analytical software that will help them to accomplish their research goals. We think that successful examples of using the proposed computational infrastructure for high-impact research that arise from this interaction will be the best advertisements for engagement.

Principle 3: Users are co-developers. The existing approach of ICE-D workshops, which have focused on giving Earth science researchers enough basic technical information about the ICE-D infrastructure to contribute to it and make use of it, evolved by necessity due to the lack of dedicated software development resources. However, this 'users-are-co-developers' model has important advantages in relation to a conventional 'users-are-clients' model that interposes a layer of IT professionals between researchers and data. Specifically, creating a large group of researchers who participate in development reduces sustainability risks, and, more importantly, developing the skills and mindset to integrate small-scale studies into synoptic analyses using modern cyberinfrastructure provides critical elements of a modern Earth science education.

Workshop program. Our primary means of implementing this engagement strategy will be to expand the ICE-D workshop program. We will support regular (budgeted at 2/yr) small-group workshops at BGC and NDSU, that will include both invited "influencers" (see above) and an open application process advertised via geoscience meetings and other channels. To facilitate participation by students and early-career researchers, travel support for workshop participants is included in this proposal. As in the 2018-19 workshops, the training element will focus on providing students and researchers interested in regional- or global-scale analysis of cosmogenic-nuclide data an understanding of the elements of the computational infrastructure and the skills needed to work with them at multiple levels. More broadly, from the perspective of developing a community of researchers working together with a modern computational infrastructure, the workshops will have the following goals:

- *Establish groups of researchers who collaborate on database development and maintenance*. At the most basic level, this goal can be expressed as moving from a situation where all researchers maintain separate, inconsistent, and redundant spreadsheets to one where everyone is at least working with the same visible, verifiable, and generally accepted data set. So far, we have used this model successfully for both the ICE-D:ALPINE and ICE-D:GREENLAND data collections. We aim to continue this approach in developing new data collections.
- *Establish a group of researchers who can build out the middle layer.* This goal will involve people who have developed or are developing code for specific applications of cosmogenic-nuclide geochemistry, e.g., depth profile analysis, erosion rate analysis, or burial dating. The aim is to build a group of researchers who can work together to expand the transparent-middle-layer concept beyond the present focus on exposure dating.
- *Expand the group of researchers who are using downstream applications.* This aspect will target (i) researchers who are perhaps not specialists in cosmogenic-nuclide geochemistry, but who would use data sets derived from cosmogenic-nuclide data for broader applications in, for example, paleoclimate, ice sheet change, or tectonics; and (ii) researchers who would be likely to train their own students or colleagues in using the proposed infrastructure, thus expanding the user group beyond workshop participants alone.

IV. Sustainability plan.

As this is a 'catalytic track' proposal focused on developing research capabilities that do not yet exist, it does not include a detailed financial plan for long-term maintenance of specific capabilities or resources. Regardless, in this section we highlight design features intended to ensure resource sustainability. First, the proposed infrastructure, like the existing ICE-D prototype, will utilize inexpensive, industry-standard cloud computing services rather than physical servers. As cloud computing costs are expected to decrease over time, this minimizes sustainability costs. The use of simple, standardized, and modular software tools as in the ICE-D prototype also reduces "lock-in" risk associated with development of custom or proprietary code bases; development using these tools is rapid, easily reconfigurable, and can take advantage of numerous widely used frameworks and development environments. Second, our users-as-co-developers strategy of building a community of researchers with the knowledge needed to improve database and middle-layer elements of the infrastructure minimizes "key-man" risk that would be incurred by a users-are-clients model reliant on one or more software developers. Once the overall infrastructure exists, users will be able to create and maintain databases and middle-layer capabilities with minimal or zero assistance. Third, our plan to manage all software development tasks as public open-source projects has a similar function of enabling community participation in infrastructure development and therefore minimizing sustainability risk. Basically, our goal is not to function as a service provider with members of the cosmogenic-nuclide research community as clients, but instead to develop a community of researchers who are capable of contributing to infrastructure development and maintenance and also motivated to do so into the future.

V. Management plan

V.1. Overall management. This is a small project involving only two PIs and two dedicated staff. Our management plan is correspondingly simple and flexible, and relies mainly on close collaboration among a small number of personnel. The main management challenge for this project is the need to integrate a wide variety of relatively small tasks -- professionalization of existing software, API development, database development and management -- into a coherent overall structure that is scalable and extensible in support of science goals that we know about now and that will emerge in the future, and addressing this challenge will require close collaboration between science staff of the project (Balco, Laabs, postdoc) and the software developer. Management of the initial set of development tasks needed to professionalize the existing prototypes will involve mainly Balco (who wrote most of the existing code that needs to be improved), the developer, and the postdoc. Developing connectivity to downstream analysis applications will expand this to include Laabs and his students.

V.2. Roles of the workshop program. The proposed workshop program serves two important roles in project management. First, it provides a "build-measure-learn" feedback pathway to incorporate the experiences of end users into project management. Second, it fills a community oversight role by guiding and prioritizing development tasks, so that ideas and application needs from the workshops are incorporated as the project evolves. This model does not include a formal external oversight board empowered to review and direct the project, but in our view this would not be appropriate for a 'catalytic track' proposal focused on initial development of a capability that, as shown by evidence for engagement with the ICE-D prototypes, is already known to have significant potential value to the community.

VI. Intellectual merit.

The most important contributions of cosmogenic-nuclide data to large-scale Earth science questions so far, and likely in the future, have involved synoptic analysis of large data sets for applications in ice sheet change and sea-level impacts, paleoclimate, tectonics, and surface processes. Three decades of investment in physical infrastructure for cosmogenic-nuclide geochemistry has produced a research tool capable of generating the global-scale data sets that support these applications and, more broadly, nearly all Earth science research focused on processes active at the Earth's surface. The transparent-middle-

layer infrastructure we propose would provide the corresponding computational and data management infrastructure needed to make existing and future global-scale applications of cosmogenic-nuclide data sets extensible, reproducible, scalable, and accessible to broader areas of Earth science investigation and modeling, and in addition potentially provide a proof of concept for computational infrastructure development in other areas of geochemistry and geochronology.

VII. Broader impacts.

Elements of this project focus on enabling Earth science research and education, including (i) our users-as-co-developers strategy for engagement and project guidance, (ii) our strategy of incorporating synoptic analysis and visualization into teaching resources, and (iii) our proposed program of training Earth science researchers and students in computational and geospatial skills. Integrating cyberinfrastructure development into undergraduate and graduate Earth science education and research will build a community of students and researchers who are better prepared to succeed in both Earth science and other fields, as well as building the capability to link small-scale field research to the broader regional- or global-scale data set of similar observations. Finally, innovative aspects of this project, including our emphasis on building a forward-looking research infrastructure in contrast to a backward-looking archive, our incentive-based engagement structure, our users-as-co-developers strategy, and our focus on rapid population of data collections geared toward specific applications, may provide models for how to enable scientific investigation through cyberinfrastructure development in other areas of Earth science.