Developing and Deploying SKOPE: Synthesizing Knowledge of Past Environments

Project Summary

Overview:

Achieving systematic understandings of the long-term interactions of human and natural systems is a major focus of research in the social and natural sciences. Research on fundamental changes in historical social systems must always either implicate--or argue against the relevance of--environmental variability. Research on the long-term sustainability of human systems must account for both the effects of climatic variation on human societies and the substantial impacts of humans on ancient and modern environments. Contemporary research must accommodate the increasingly obvious fact that environments are not stable and that today's environments were not replicated in the past; scholars need environmental knowledge specific to their spatial and temporal research contexts. They are likely to find, though, that current data on past environments are difficult or impossible to discover and even harder to integrate and interpret. SKOPE (Synthesizing Knowledge of Past Environments) will be an online resource for paleoenvironmental data and models. SKOPE builds on an 18-month effort by this same team--including needs assessment, design, and prototyping--to make such data accessible.

Intellectual Merit:

By enabling scholars to easily discover, explore, visualize, and synthesize knowledge of environments in the recent or remote past, SKOPE will enhance research in such diverse disciplines as anthropology, archaeology, ecology, economics, geography, political science, sociology, and sustainability. Given a location and temporal interval, SKOPE will offer access to diverse sources of long-term, high-resolution environmental data. SKOPE will be a dynamic resource; it will allow users to rerun models with different inputs and it will seamlessly accommodate new datasets and models. SKOPE addresses two critical challenges to contemporary science: increasing access to publicly funded research; and ensuring that scientific results are transparent and reproducible. SKOPE will provide robust support for reproducible scientific research requiring paleoenvironmental data. It will not just enable discovery and access to paleoenvironmental data; it will provide researchers with an unprecedented ability to explore the data's provenance--a detailed, comprehensible record of the origin and computational derivation of the supplied data. Central to SKOPE's support for transparency and reproducibility will be further development of YesWorkflow, a system for revealing the fine-grained provenance of data produced by scripts, programs, and computational pipelines without adapting software to run within a scientific workflow management system and without the overhead of a runtime provenance recorder. This work also addresses the problem of integrating multiple sources of provenance information, indicating when provenance query results are ambiguous due to incomplete or conflicting information.

Broader Impacts:

Infrastructure: SKOPE will enhance the infrastructure for research and education. It will transform vast amounts of prior data collection and research into readily usable environmental knowledge. SKOPE will substantially enhance scholars' ability to execute reproducible research on a broad range of social and natural science topics involving long-term interactions of humans with their environments and will facilitate ongoing improvement of paleoenvironmental reconstructions. Enhancements to YesWorkflow will make new provenance inference capabilities also available outside SKOPE to the many researchers who use conventional scripting and logging approaches. *Education*: Students will have free access to high-quality environmental scenarios in which to situate their studies. Members of the general public will be able to discover how ancient environments differed from those of today. *Public Policy*: SKOPE will make available publicly funded paleoenvironmental modeling and data to users in academia, industry, and government. It will provide greatly superior information and eliminate the need for countless heritage management and environmental assessment projects to do their own reconstructions. Planners will be able to use SKOPE's easily accessible long-term environmental reconstructions to investigate vulnerabilities in infrastructure not revealed by recent history. Expanding the community able to use paleoenvironmental information adeptly and wisely increases public scientific literacy and public engagement with science and technology, ultimately contributing to the well-being of individuals in our society.

Developing and Deploying SKOPE: Synthesizing Knowledge of Past Environments

1. Vision and Rationale

Achieving systematic understandings of the long-term interactions of human and natural systems is a major focus of research in the social and natural sciences [KAB+14,RCG+10]. Research on the long-term sustainability of human systems *must* account for both the effects of environmental variation on human societies and the substantial impacts of humans on ancient and modern environments. Almost without exception, historical social science research on such fundamental topics as the development of sociopolitical complexity, migration, exchange, and societal collapse implicates (or must demonstrate the irrelevance of) environmental variability affecting the ability to satisfy subsistence needs, produce a surplus or extract taxes. Similarly, our understanding of current landscape structure and species distributions—and our projection of their future states—must account for legacies resulting from past environments and past human-environment interactions.

Scholars examining anything other than short very recent intervals can neither assume that the environment is stable nor that today's environment was replicated in the past; they need environmental knowledge specific to their spatial and temporal problem contexts. However, they are likely to find that relevant, state-of-the-art data on past environments are difficult or impossible to discover and even more difficult to integrate, process, and interpret.

In response, we propose to develop and deploy SKOPE (**S**ynthesizing **K**nowledge **o**f **P**ast **E**nvironments), a dynamic online resource for paleoenvironmental data and models that will enable scholars in diverse research communities (e.g., anthropology, archaeology, ecology, economics, geography, political science, sociology, and sustainability) to easily discover, explore, visualize, and synthesize knowledge of environmental factors most relevant to humans in the recent or remote past. SKOPE will deliver data in both raw form or preprocessed according to established protocols.

SKOPE will provide robust support for *reproducible scientific research* that requires paleoenvironmental data. To this end, SKOPE will not just offer tools that enable discovery, access, visualization, and analysis of paleoenvironmental data; it will provide unprecedented ability to explore the data's *provenance*—a detailed, comprehensible record of the computational derivation and origins of the supplied data. Moreover, SKOPE will allow users to rerun models using different parameter settings, and to contribute new and revised models as they become available.

SKOPE will facilitate answering a wide range of fundamental questions arising in the historical social sciences. Generic versions of many of these have recently been identified as prominent Grand Challenges facing archaeology today [KAB+14a,KAB+14b]. Examples of questions that could be addressed include:

- Why did famous archaeological phenomena such as Chaco and Cahokia rise and collapse? To what extent were climatically driven environmental changes causal?
- How tightly coupled are variability in surplus production and increases/decreases in sociopolitical scale, and how can we explain regional or cultural differences in the strength of coupling?
- How does incidence of violence relate to variability in production?
- To what degree did climatic teleconnections entrain relatively simultaneous sociocultural changes in widely separated regions? At what distances do correlations (mutual information) among cultural/demographic sequences in different regions decay, and how does that change through time?
- What are the temporal, spatial, and cultural barriers to correspondence among different regional sequences?

Answering questions like these requires confronting paleoenvironmental and demographic/cultural data within or across regions. Historical social scientists have rarely operated at such scales—or have done so in a non-rigorous, anecdotal fashion—largely because of the difficulty of developing the necessary datasets. SKOPE will substantially improve researchers' ability to marshall the environmental and social data required to answer such questions.

This project is an important step towards a rigorous continent-wide understanding of culture history, process, and demography that honors local detail and climate history. It will therefore contribute to other critical research such as understanding the impact of human-induced land-use/vegetation change on regional and continental climates.

2. User Interactions with SKOPE

Given a location and a temporal interval, SKOPE will offer easy access via a standard web browser to diverse sources of long-term, high-resolution environmental data (primary, and derived/reconstructed) relevant to humans. The proffered environmental information will be supplied with provenance and assessments of its resolution and accuracy. SKOPE will be a dynamic resource designed to seamlessly accommodate new datasets as they become available, incorporate classes of environmental data not initially included, and continuously expand the analytical, modeling, and inferential operations employed.

This project does not include any new collection of raw paleoenvironmental data. Instead, it invests in expanding and making more accessible existing reconstructions of environmental variables important to human societies. *Thus, we build on vast amounts of prior data collection and previous research, transforming those data into readily usable environmental knowledge.*

While SKOPE will be readily extensible and has no inherent spatial or temporal limitations, this grant focuses on paleoenvironmental reconstructions and models developed for the contiguous 48 US states (CONUS) for the last 2000 years. The US Southwest will be particularly well represented due to its high-precision tree-ring record and we will extend some established reconstructions for the Southwest to other states insofar as available proxies allow us to remain within reasonable limits of accuracy and precision.

Personas. SKOPE is designed to serve three (overlapping) types of users that we term "researchers" "tinkerers," and "modelers." *Researchers* are scholars in the social and natural sciences who want browser-based access to the best-available reconstructions of key environmental variables for a given location and temporal interval. *Tinkerers* are researchers with a more focused interest in a specific reconstruction model, who wish to adjust parameters and rerun the models. Finally, paleoenvironmental *modelers* are researchers who seek to build and offer broad access to new or modified models and/or retrodicted environmental data through SKOPE. Although *researchers* and *tinkerers* constitute our primary audience and main focus, *modelers* will enhance SKOPE and enable it to remain current.

Example User Stories. A ***researcher*** is investigating the comparative resilience of societies that practice household *vs*. communal (village-level) storage of surplus food. This archaeologist has identified several areas in which each storage strategy was employed and that differed in their long-term settlement persistence. While today these areas are all environmentally similar, she wants to further control for climatic variability that would have influenced agricultural success over the time periods for which she has settlement data. She uses a browser to navigate to the SKOPE application, which she heard about through a webinar hosted by the Society of American Archaeology (SAA), and is greeted with a familiar, map-based search tool. She zooms in on each study area and sets the beginning and ending dates for the area's occupation. A window pops up showing a graph of tree-ring-reconstructed precipitation and growing season growing degree days for each year in the interval. Through SKOPE, she downloads graphs and the reconstructed values for each area, so she can do some additional quantitative analysis. That analysis demonstrates that observed differences in resilience associated with the two storage strategies are *not* accounted for by differences in climatic variability during the occupation periods. The researcher is further pleased to find that each downloaded archive was automatically associated with citations of the source data and model that she can include, with selected graphs, in her publication.

A historical geographer, having recently read about a new method of reconstructing temperature and precipitation from tree-rings, realizes that he could use it to extend by several hundred years his historic-era research relating demographic scale to agricultural productivity. To further study this technique, he teams up with a tree-ring climatologist, who finds that this new reconstruction method is a core component of SKOPE. This ***tinkerer*** goes to the SKOPE web site, and registers for a free account and

associates it with one of his existing user IDs (e.g., his ORCID or Google ID). He is then presented with a choice of tools for exploring the details of the model. He views a visualization of the model's workflow (the model's data inputs, transformations, and outputs), explores the model run's intermediate data products, and inspects the model code. During this review he notes that the input included one tree-ring chronology he considers unreliable. After removing the suspect tree-ring chronology from the input dataset, he submits the model to SKOPE to be recomputed. A brief time later an email informs him that the reconstruction is complete, and he returns to SKOPE and reviews the results from his user account. Finding that excluding the questionable tree-ring chronology had little impact on the results, he and his geographer colleague gain confidence in the validity of the published reconstruction.

Finally, a paleoenvironmental ***modeler*** has developed a new method that successfully retrodicts agricultural potential (using aboriginal technologies) for the US Southeast. The modeler wants to make both her high-resolution retrodiction dataset and the model code itself available to other scholars through SKOPE. She realizes that providing access to users in the scientific community will substantially increase the scholarly impact of her work, as evidenced by citations in articles using the data or model. The modeler signs up for a free user account with SKOPE, and registers her model by entering her source code's GitHub location. She also registers the dataset with the retrodicted agricultural potential estimates by providing a link to its location in a trusted database (such as the NOAA Paleoclimatology database) along with metadata sufficient to allow SKOPE to serve the data to its users. At that point, all users can discover and access the dataset and model. Encouraged by wide usage of her dataset (as tracked by SKOPE) she undertakes the additional step (guided by SKOPE documentation) of annotating her code so that SKOPE can actually execute the model. With those changes, all users of the data or the model are able to obtain detailed provenance information, and *tinkerers* using SKOPE can run it with different parameters or extend it to different areas. Some time later, the modeler is notified that another SKOPE user has explored her model and has posted a comment suggesting a parameterization that reduces uncertainty, allowing her to refine the model and improve the dataset. In these ways SKOPE simultaneously enhances transparency, reproducibility, and meaningful cumulation of her research.

3. Overview of SKOPE Structure and Function

Researchers: Discovery, Exploration, Visualization, and Download. SKOPE will provide a *researcher* interface designed for users seeking simplicity or quick access. We expect that a large fraction of user interactions will be satisfied in this mode. Using a *web-mapping interface*, users can pan and zoom in on, and select a point and/or area of interest. The user will specify a time interval and select from a list of available reconstructions that temporally and spatially overlap their specifications. The selected paleoenvironmental reconstruction will be provided online through visualization and as downloadable tabular or geospatial datasets. Where available, SKOPE will provide access to the model source code.

On-screen and downloadable visualizations will include *time series graphs* of data values for a given location and *animations* that display mapped data through time. SKOPE will also produce *difference plots* that facilitate comparison of reconstructed values at different locations or values produced by different reconstruction models. SKOPE will apply alternative temporal and spatial *smoothing algorithms* to map and chart visualizations, enable selection of reference layers (e.g., soils, elevation) for plotted maps, and adjust basic display parameters (such as transparency) for overlaid map layers. Each map view will generate a unique URL that can be shared with other investigators and collaborators. Our SKOPE prototype application (Figure 1, [SKOPE16b]) illustrates some of this functionality.

For transparency and reproducibility, SKOPE places particular emphasis on comprehensive metadata, including detailed *data provenance* [BKW01,BMC+06,DBE+07,LP15]. SKOPE's results will hyperlink the model's name to a model metadata page including author, title, description, date, version, citation, links to the source code, and additional documentation. This metadata page will also provide links to the input data files and their associated metadata: title, description, date, version, temporal scale, regional scale, and citation. If the input file is derived, metadata will include links to its source datasets.

Whenever supported by the model, "deep provenance" of the reconstructed paleoenvironmental data will be exposed. Provenance visualizations will display model parameter values, identify the run's input, intermediate and output datasets and will permit detailed exploration of its computational procedures. Model results will include machine-readable provenance records that can address complex queries (e.g., to recursively "chase" data lineage and derivation history [ABL09,MBB+15], or to automatically create data citations in an "attribution report").

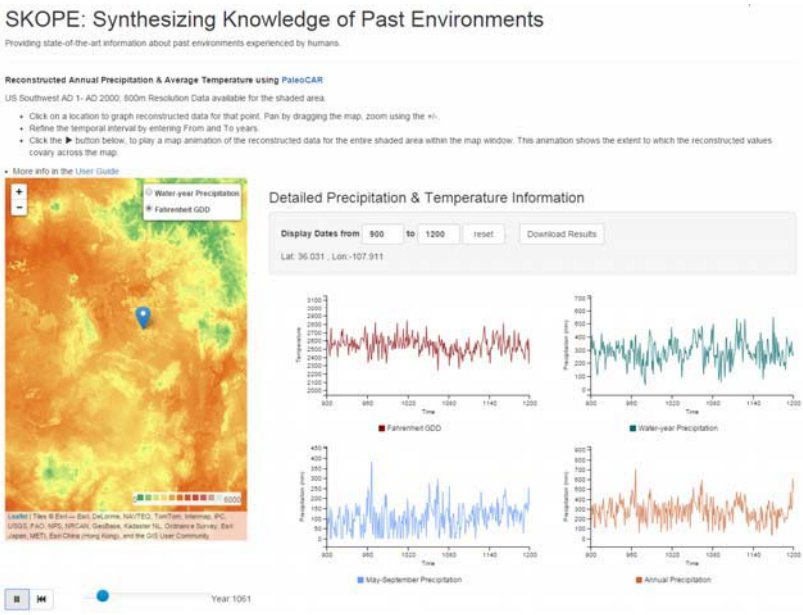

Figure 1. Screenshot of SKOPE prototype showing graphs of reconstructed annual, water year, and growing season precipitation and Fahrenheit growing degrees days for AD 900-1200 in Chaco Canyon, New Mexico.

In addition to the reconstructions and models that can be explored and visualized directly, SKOPE will facilitate discovery of and access to key sources of paleoenvironmental sample data (e.g., pollen). Users will be able to search such data by sample type, geographical area, time period, and researcher.

Tinkerers: Execute Tuneable Models. For key retrodicted or interpolated datasets (e.g., retrodicted temperature and precipitation), a *tinkerer* can rerun the model, having adjusted parameter values or substituted compatible input datasets. For example, a maize productivity model developed for one location could be adjusted to suit other crops for application in another area. User execution of tuneable models will be possible only in cases where SKOPE has access to the source data and model code, and is able to execute the entire computational workflow. Users will be reminded that any new outputs they generate by running customized models have not been reviewed.

How quickly the run is executed will depend on the runtime required. An *on-the-fly* run (<10 sec.) will be executed immediately. More intensive *small runs* (< 20 min.) will be queued to run on an HPC (high performance computing) cluster with the user notified when the results become available. *Large runs* with greater runtime or storage requirements (e.g., high-resolution or large spatio-temporal extent) will run like small runs, but may require a request for an explicit allocation of HPC resources that we expect would be granted for substantial research projects with publication potential. Finally, SKOPE will allow advanced users to download an executable environment (using *container* technologies described below) and run a model on other resources such as desktop computers, local clusters, or cloud computing systems.

Modeler Interface. In addition to the implementation of paleoenvironmental reconstruction models and the inclusion of core datasets, SKOPE will provide an interface that will allow a technically sophisticated *modeler* to register datasets and computational procedures in SKOPE so that they can be accessed by a broad user community. Modelers will also be able to supply user-oriented documentation and provenance sufficient to enable the responsible scientific use of uploaded datasets. SKOPE will build on ideas developed by DataONE [DONE16], which already allows data submissions with embedded provenance information.

Data Sources. SKOPE will provide an online and programmatic interface to a variety of paleoenvironmental datasets. Table 1 lists sources to which we expect to provide access of some sort. SKOPE's key datasets will be the spatial retrodictions of environmental variables based, e.g., on tree-ring

chronologies. We expect SKOPE also to provide easy access to atemporal spatial datasets (e.g., soils, surface hydrology, and elevation), drawn from recognized sources and to *measured* time-series data that are spatially interpolated (e.g., PRISM, [DHS+08]). SKOPE will provide access to paleoenvironmental *point reconstructions* (e.g., streamflow) and *sample data* that apply to single locations. Sample data include fauna and pollen from dated archaeological or environmental contexts.

Table 1. Paleoenvironmental, Environmental and Demographic Data SKOPE Expects to Offer. (Modelers can add more.) CONUS: 48 contiguous US states; SW: AZ,CO,NM,UT; 30 arc-sec: ~800m. **Currently operational models in bold.**

Information	Spatial Scope	Spatial Resolution	Temporal Scope	Temporal Resolution	Model/Source
SKOPE-generated Retrodictions with Tuneable Models					
Precipitation	**CONUS**	**30 arc-sec**	**AD 1–2000**	**Annual**	**PaleoCAR/tree-ring chronologies [BK14]**
High-frequency Temperature	**CONUS**	**30 arc-sec**	**AD 1–2000**	**Annual**	**PaleoCAR/tree-ring chronologies [BK14]**
Low-frequency mean temperature/warmest month	CONUS	30 arc-sec	Holocene	~10^3 years (~10^2 years last 2K years)	MAT/pollen (stretch goal) [OWP85,VLG11]
Crop Niche Temp/Precip	CONUS	30 arc-sec	AD 1-2000	Annual	PaleoCAR (extension)
Maize Productivity	SW	30 arc-sec	AD 1-2000	Annual	PaleoCAR/DSSAT [JHP+03]
Staple Wild Plant Habitat	SW	30 arc-sec	AD 1-2000	50 years	LTVTP
Biome Reconstruction	CONUS	30 arc-sec	Holocene	~10^3 years (~10^2 years last 2K years)	Biomisation [PGH+96] or pseudobiomisation [FRW10]
SKOPE-facilitated Access					
Precipitation	CONUS	30 arc-sec	1895-	Annual	PRISM [DHS+08]
High-frequency temperature	CONUS	30 arc-sec	1895-	Annual	PRISM [DHS+08]
PDSI	CONUS	0.5 degree	AD 1-2000	Annual	NADA [CK04]
Elevation	CONUS	⅓ arc-sec	Modern	–	NED [USGS16a]
Surface Hydrology	CONUS	Vector	Modern	–	NHD [USGS16b]
Soils	CONUS	Vector	Modern	–	SSURGO [NRCS16a]
Soil characteristics	CONUS	⅓ arc-sec	Modern	–	gSSURGO [NRCS16b]
Human Settlement Distribution	Selected E-US states	Usually, by county	Varies	Varies	DINAA [WKK+14]
Human Population	SW	TBD	AD 1250-1500	50 years	Coalescent Communities Database [HCD+10]
Human Population	~680km^2 strata in CO & NM	TBD	AD 600-1500	~40 years	VEP [Ort16, SBO+16]
Streamflow Reconstructions	CONUS	Point	Varies	Annual	TreeFlow [TREE16]
Tree-ring Chronologies	Global	Point Samples		Annual	ITRDB [GF97]
Tree-ring Dating Samples	SW			Annual	LTRR [KB15]
Pollen Samples	Global			Varies	Neotoma [Neo16a], tDAR [TDAR16], DataONE [DONE16]
Macrobotanical Samples	Global				
Faunal Samples	Global				
Radiocarbon samples	W hemisphere				CARD [GMP+11]
Dated Fire Events	Global				PaleoFire & Fire History Analysis and Exploration System [BVS+16]

4. Project Foundations: Results of Prior NSF Support

This proposal relies heavily on our NSF-funded planning and design grant ("BCC-SKOPE") and leverages NSF's previous investments in the National Center for Supercomputer Applications (NCSA) at one of our collaborating institutions. It further builds on other recent projects led by project PIs.

BCC: Collaborative Research: Designing SKOPE: Synthesized Knowledge of Past Environments (PIs: Kintigh, Kohler, Ludäscher; Co-PI: Kinzig; SMA 1439591 Arizona State University (ASU); SMA 1439603 University of Illinois Urbana-Champaign (UIUC); SMA 1439516 Washington State University; (WSU) $393,089 total; 9/1/14 – 2/29/16). The signature outcome of the grant, as anticipated by the NSF ***B**uilding **C**ommunity and **C**apacity* program announcement, was development of the present proposal and the BCC-SKOPE prototype on which it is based. Over the 18-month grant period, team interactions included 24 videoconference meetings, four, two-day face-to-face workshops (one hosted by SFI, the Santa Fe Institute), and extensive independent work at the collaborating institutions. Through these efforts our team was able to (1) determine the needs of diverse academic and professional users, (2) define a vision and a realistic scope for the project in terms of time, space, and content, (3) experiment with and solve key technical challenges through the prototype, (3) develop a design through which the proposed implementation can expeditiously proceed, and (4) construct the present proposal.

As an initial step, the project sponsored two needs-assessment workshops for a broad range of professionals. Participants enthusiastically proposed a great diversity of potential uses for a platform providing well-documented, high-resolution paleoenvironmental data. While the expressed interests included an enormous range of environmental variables and sample types, strong consensus pointed to reconstructed precipitation and temperature as the highest priorities. To better address key development issues that SKOPE will encounter, we developed a prototype (Figure 1, [SKOPE16b]) that delivers paleoenvironmental reconstructions of those two most-requested variables for the last 2000 years for the four Southwest US states [BKB+16]. Deploying the prototype required us to solve the challenge of delivering responsive animations of reconstructed values through 2000 time steps over an arbitrarily-zoomed portion of a map consisting of more than 2 million 30 arc-second cells. We used NCSA's HPC resources to convert the model's output into pre-rendered tiled maps quickly loadable by the client.

Intellectual Merit. High-resolution, regional-scale paleoenvironmental datasets enable archaeological syntheses of subsistence practices, migration [BK14], and evolution of social and economic systems [BRK+16,SBO+16]. Refinement of methods in [BK14] enabled the creation of paleoenvironmental precipitation and temperature reconstructions covering Arizona, Utah, Colorado, and New Mexico at sub-km resolution. These reconstructions are readily available through the SKOPE prototype, and source software resulting from this effort has been deposited in open repositories [Boc15,Boc16], including an R package, *FedData* [Boc16], that already has >4000 downloads. The project web site [SKOPE16a] describes the project components, provides a useful list of web-accessible environmental data resources, and provides access to the SKOPE prototype.

The BCC grant also funded development of YesWorkflow [MBB+15,MSK+15,YW16], a software system that brings the advantages of scientific workflow automation to researchers using scripting languages such as R, Python, and Bash. YesWorkflow enables script writers to reveal the computational steps and flow of data within their scripts, i.e., *prospective* provenance, by annotating their code with special comments. YesWorkflow extracts and analyzes these comments, represents the scripts in terms of entities based on the typical scientific workflow model, and provides graphical renderings of the scripts. YesWorkflow additionally enables researchers to reconstruct *retrospective* provenance of data products used by scripts, and to query prospective and retrospective provenance jointly. Together these capabilities allow users clearly to see the actual computational steps that occurred in runs of models and other scripts, and the data that passed between those steps to yield the final outputs. YesWorkflow was used to document paleoenvironmental reconstruction scripts employed in the SKOPE prototype and will be used extensively in SKOPE to collect, report, and enable users to explore and query the provenance of data provided by SKOPE tools, thus strongly contributing to scientific transparency and reproducibility.

Broader Impacts. The BCC grant focused on *Enhancing Infrastructure for Research and Education*. The prototype developed and the SKOPE system proposed here serve unmet needs for easy access to high-quality paleoenvironmental information by a diversity of researchers in academia and industry, while at the same time greatly enhancing the reproducibility of their results. The easy-to-use system will *increase public engagement in science* by allowing members of the public to explore the environments of places and periods as they have existed over the long term.

SI2-SSI: CyberGIS Software Integration for Sustained Innovation (PI: Wang; Co-PIs: Nyerges, Wilkins-Diehr, Anselin, Bhaduri; ACI-1047916, $4,804,821, 10/1/10-9/30/16.) CyberGIS is Geographic Information Science and Systems (GIS) based on advanced cyberinfrastructure (CI; [WAB+13]).

Intellectual Merit: The grant has developed multiple leading-edge cyberGIS software tools and published 75+ peer-reviewed papers in advanced CI, GIS, geography and social sciences, and geosciences [Cyb16]. One of the software tools, CyberGIS Gateway, simplifies access to advanced CI [LPW15], serving over 1,300 registered users and enabling collaborative geospatial problem solving based on Gateway applications across numerous subjects (e.g., hydrology [FYW+14], bioenergy [HLL+15], and emergency management [JWS16]). The Gateway integrates the Structured Participation Toolkit to scalably support asynchronous participation, feedback, and decision making [RT13].

Broader Impacts: The innovative cyberGIS software tools have provided straightforward access to advanced CI, including XSEDE, ROGER, the Open Science Grid, and cloud computing systems. Training activities have engaged over 200 participants. The CyberGIS Fellows program's open-access educational materials address the gap in access to the rapidly advancing state of the art in cyberGIS.

Other Strongly-Related NSF Funding. We briefly summarize four other NSF-funded efforts involving project PIs that are integral to SKOPE. Two archaeology-focused NSF Coupled Natural and Human Systems (CNH) projects in the US Southwest developed and applied powerful methods of environmental analysis for investigating long-term human ecodynamics. By incorporating refined and generalized versions of these methods in SKOPE, we profitably exploit these investments. SKOPE itself is a direct offshoot of a third grant that developed recommendations for NSF cyberinfrastructure investments. A fourth NSF grant (from the NSF-ABI program) funded research on provenance-enabled workflow automation for data curation. These technologies will be central to the power and success of SKOPE.

CNH: Coupled Natural and Human Ecosystems over Long Periods: Pueblo Ecodynamics (PI: Kohler; Co-PIs: Allen, Kobti, and Varien; DEB-0816400, $1,506,988; 1/1/2009-12/31/2015.) The Village Ecodynamics Project (VEP II) synthesized data on known archaeological sites in two large areas of the Pueblo Southwest between AD 600 and 1760 and developed empirically based population profiles through time [Ort16,SBO+16]. It also developed agent-based models to predict optimal household locations through time, assuming that households minimize energy devoted to agriculture, hunting, and acquiring water and fuelwood. Comparison of model outputs with known site distributions yielded inferences about key social and natural processes in the study areas [KV12,KBC+12]. By reconstructing population size and production through time, the VEP contributed fundamental new findings on processes provoking violence in the ancient Southwest [KOG+14], causes and consequences of population growth in Neolithic societies [KR14], drivers of sociopolitical evolution [KCB+15], and causes of the famous depopulation of the Four Corners in the late AD 1200s [SBO+16]. The VEP developed R scripts to reconstruct, through time, the extent of the agricultural niche for maize [BK14]. That code was incorporated in the SKOPE prototype and will play a key role in the proposed SKOPE system.

CNH: The Complexities of Ecological and Social Diversity: A Long-Term Perspective. (PI: Nelson; Co-PIs: Kinzig, Anderies, Hegmon, Norberg; BCS 1113991; $1,425,000; 9/1/2011-2/28/15.) Through intensive comparisons of four Southwest US archaeological study areas, the project expanded knowledge of the conditions under which ecological and social diversity are advantageous or disadvantageous for long-term societal resilience, attending especially to collapse and major social transformations [HPK+08, Nel11,NKA+10,NHK+11,NHK+12,NID+16]. In much of the semi-arid Southwest, social groups

mitigated the risk of frequent crop failure in part by storage and in part through the establishment of exchange relationships. The project developed and implemented the concept of "risk landscapes"—a map showing the degree to which, in each location, an exchange relationship could be beneficial to a target location (i.e., their key climate parameters are anti-correlated based on the reconstructions provided by the SKOPE prototype [BRK+16]). A related effort reconstructed the spatio-temporal distributions of habitat suitability of wild plant species that ethnographically served as dietary staples. Both the risk landscape and wild plant species habitat suitability retrodictions will be included in SKOPE.

Planning Archaeological Infrastructure for Integrative Science (PI: Kintigh; BCS 1202413 $49,999, 1/15/12-6/30/13), developed recommendations for investments in computational infrastructure that would enable archaeology to better serve the needs of the scientific community and contemporary society [KAK+15]. In order to prioritize these investments, the project developed 25 grand challenges for archaeology, using both crowdsourcing and a workshop of distinguished scholars. The challenges that emerged [KAB+14a,KAB+14b] are not exclusively archaeological; they are social science questions whose answers demand knowledge on temporal and spatial scales that only archaeology can provide. The SKOPE proposal is a direct response to a recommendation concerning how emerging computer science research can empower the synthetic research demanded by the grand challenges.

Collaborative Research: ABI Development: Kurator: A Provenance-enabled Workflow Platform and Toolkit to Curate Biodiversity Data (DBI-1356751, PI Ludäscher, UIUC, $748,931.00; DBI-1356438, PI Hanken, Harvard, $896,967; 9/1/14-8/31/17). This project [Kur16] is developing workflow automation and provenance management technologies for data curation and cleaning for the biodiversity community [LMS+15,MLK+15]. SKOPE will build on our experience in achieving Kurator objectives, including: automated workflows for accessing and cleaning biodiversity data; an Akka-based workflow engine; a web application for composing and running workflows; Docker containers for running the Kurator web application on dedicated servers, cloud platforms, and personal computers; containers for isolating runs of workflows with conflicting software dependencies; support for using Python scripts as workflow components; YesWorkflow feature development (collaboratively with BCC-SKOPE via McPhillips and Ludäscher); and an agile development process suitable for a distributed engineering team.

5. Paleoenvironment Model and Dataset Development

SKOPE will be a dynamic, online community resource that delivers paleoenvironmental data and access to the underlying reconstruction models. Initially, SKOPE will be populated with numerous publicly available (but often not easily accessible) sources of environmental information (Table 1). It will also include refined, extended, and generalized methods of paleoenvironmental reconstruction developed by the archaeology-focused CNH projects (including the Southwest US precipitation and temperature data provided by the prototype). A model for the reconstruction of past biomes from pollen cores will fill a major gap. YesWorkflow will enable provision of detailed provenance information for all project models.

Extension of Temperature, Precipitation and Crop-Niche Retrodiction to the US. We will extend the prototype's methods of reconstructing precipitation, temperature, and temperature/precipitation-bounded crop niches for the Southwest to as much of CONUS as possible while maintaining high standards for accuracy and precision as judged by model performance on an annually resolved cell-wise basis. SKOPE will enable users to compare results with other available reconstructions of precipitation/drought such as the North American Drought Atlas [CK04] and northern hemispheric temperature anomaly curves based on tree-ring and multiproxy reconstructions for the last two millennia (e.g., [LKB+12,MJ03,MSH+05]).

Our default methods have passed rigorous peer review in *Nature Communications* [BK14] and *Science Advances* [BRK+16], and through SKOPE will be open to easy modification by others. As new tree-ring sequences are added to the ITRDB, temperature and precipitation reconstructions will automatically improve and the models can be run on larger or differently screened datasets. A modeler proposing an alternative model (e.g., incorporating PDSI in the determination of the maize dry-farming niche) would find that SKOPE will help them develop the model, visualize its outputs, and share it easily with others.

Paleoproductivity Models. Using existing precipitation, temperature, maize-niche models, and other models generated by the VEP, SKOPE will present agricultural paleoproductivity estimates for the Southwest and will attempt to extend these reconstructions to a larger area. These estimates are critically important to begin to understand the direct impacts of past environmental change on agrarian societies; some societies in the US Southwest, for example, relied on maize for ~90% of their calories [Mat15]. Our approach will integrate the Decision Support System for Agrotechnology Transfer (DSSAT) suite of models into SKOPE, and develop systems for integrating soils, precipitation, and temperature data into DSSAT. We anticipate that this may inspire and guide others in contributing similar models for other regions, including the gridded DSSAT forecasts currently being produced by researchers in the Agricultural Model Intercomparison and Improvement Project (AgMIP, [EMD+15]).

Staple Wild Plant Habitat Suitability. SKOPE will incorporate a model developed by the Complexities of Ecological and Social Diversity CNH project that retrodicts distributions of habitat suitability for wild plant species that ethnographically served as dietary staples in the Southwest ([But15] presents related work). In addition to retrodicting the habitat suitability at 800-m spatial and 50-year temporal resolution, it produces related species-diversity estimates. This model is calibrated with modern PRISM data and uses using SKOPE's existing precipitation and temperature reconstructions. Incorporation of this model will entail adaptation to the platform and incorporating YesWorkflow annotations.

Biome Reconstruction. The North American Pollen Database NAPD [Neo16b] made available by Neotoma [Gri08,Neo16a] contains 619 dated pollen cores in CONUS. These extremely valuable data can be used for vegetation (biome) reconstruction using the biomisation model, implemented in an R function developed by Prentice's climate group at Imperial College London (originally described in [PGH+96]). Modification of the model will be necessary to accommodate the range of biomes in CONUS (see [MCH+09] for a Latin American example). We will also consider implementing an alternative procedure making fewer assumptions, called pseudobiomisation [FRW10]. In either case, a user will be able to select an area and date and generate expected biomes at locations with pollen datasets meeting the requirements of an acceptable age model that overlaps the date specified. We will also provide access to the other resources of Neotoma, working closely with consultants Grimm and Williams to strategically expand the pollen datasets in the NAPD to fill spatial/temporal gaps, enabling better biomisation.

We will strive to achieve two additional goals with the pollen database. First, we will explore the possibility of spatial interpolation across biomes at reconstruction points, leveraging correlations between the PRISM dataset and contemporary biomes. Second, we will experiment with and assess the plausibility of implementing one of the several approaches to paleoclimate field reconstruction based on pollen, especially the Modern Analog Technique and its many recent variants [OW12]. A low-frequency temperature record derived in this way could potentially be used to modulate the high-frequency temperature reconstruction derived from tree-rings via a regionalized application of wavelet modulation [MSH+05,VLG11], with possible application in the maize niche and productivity models.

Archaeological Tree-Ring Dating, Pollen, and Demography-Relevant Datasets. Working closely with Peter Brewer of the University of Arizona Laboratory of Tree-ring Research (LTRR; also Chair of the International Tree-Ring Databank [ITRDB] steering committee) SKOPE will expand the available database of archaeological tree-ring dates, which currently contains 29,311 dates from AD 500–1400 [BRK+16,KB15]. Brewer will also advise on our usage of ITRDB tree-ring chronologies. We will make available the settlement location/date data in the Digital Index of North American Archaeology (DINAA) for 11 southeastern states (the addition of several additional states is imminent), and two Southwest US demography-relevant datasets listed in Table 1.

6. SKOPE Technology and Software Development

6.1 System Overview

Web Application and Researcher Interface. Users will interact with SKOPE (Figure 2) via supported web browsers. The interface will provide functionality for data discovery, filtering, and visualization. We will employ widely used web application and web server technologies.

Map Tile and Video Streaming Services. Users will be able to rapidly pan and zoom 2-D views of reconstructed paleoenvironmental conditions overlaid on maps and visualize changes through time with animated maps. The computational load and data bandwidth requirements for user systems and web browsers will be minimized by rendering 2-D image tiles and compressed video streams on SKOPE system resources and streaming them to the users' web browsers.

Data Catalog. All datasets accessible through SKOPE will be registered in the data catalog, along with rich metadata including spatial and temporal coverage, provider, unique identifier (e.g., URI or DOI), and the URL from which the data can be obtained. SKOPE will adopt or adapt metadata representation approaches developed by DataONE [MAB+12,DONE16].

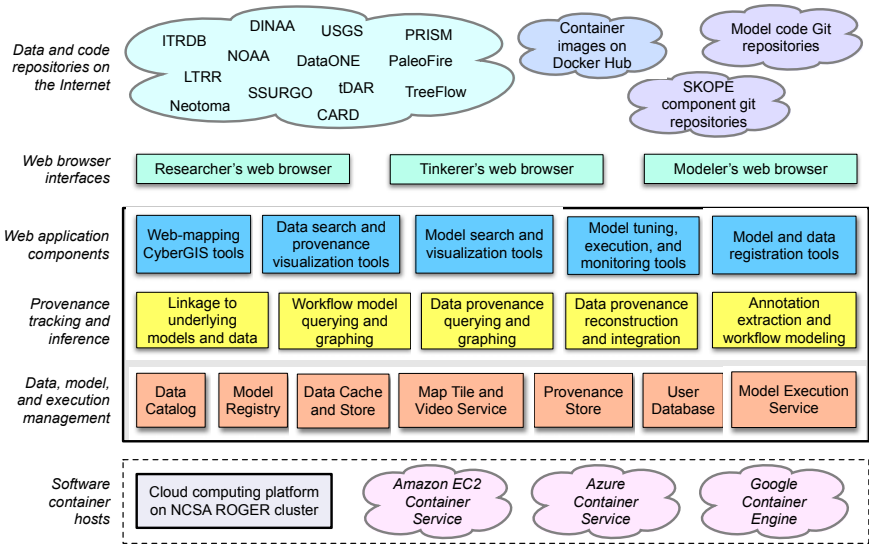

Figure 2. Envisioned full deployment of SKOPE. The core system (center) comprises modules and services for web interaction; provenance tracking and inference; and data, model and execution management. Early SKOPE deployments will comprise functional subsets of these loosely-coupled components. We will develop and deploy the production system on the NCSA ROGER cluster; ultimately it will be possible to deploy part or all of SKOPE on cloud-based computing resources. SKOPE will use data sources directly accessible to users, and store models and application code in public repositories.

Model Registry. Scripts and programs implementing retrodiction models will be registered in the SKOPE Model Registry and stored in a directly accessible public Git repository. The registry will also store the (YesWorkflow) workflow model of the code to facilitate queries and visualizations of models.

Provenance Store and Query Service. Data lineage information will be maintained in the provenance store to provide full transparency and to facilitate reproducibility. Each provenance relationship asserts that a dataset was derived from one or more other datasets via a run of a specific version of a particular model or other program with given parameter values. Dataset references will be to entries in the SKOPE Data Catalog, and model references will be to the Model Registry. The provenance store will be wrapped in a provenance query service to allow the full lineage of any dataset to be returned from a single query.

Data Cache and Data Store. Datasets frequently accessed by users, datasets needed to run available reconstruction models, and datasets whose archived formats require substantial preprocessing for effective use will be stored in a local data cache (and rebuilt as needed) to increase the speed of model execution and user interaction in the web application, and to improve system reliability by minimizing real-time accesses of remote data sources. The SKOPE data store will hold retrodicted datasets produced by SKOPE model runs that are not yet published to a stable repository.

User Account and Activity Database. Most SKOPE functionality will be available to unauthenticated users, including interactive exploration and visualization of datasets and their provenance. Users who wish to maintain records of what they have done in the SKOPE environment and to have access to retrodicted datasets produced on their behalf will log in by providing their credentials through a federated identity service (CILogon [BFG14,CIL16]). For registered users, the system will maintain an activity history and links to their saved reconstructions and visualizations. Authenticated users are also allowed to run more computationally intensive scenarios than unauthenticated users.

Model Execution Service. *Tinkerers* seeking new paleoenvironmental reconstruction model output can request that models be executed on their behalf. This scenario includes applying a model to a new region or time period, or using different model parameters than the published model's default. The service will take as input references to input datasets, model parameter values, and a reference to a software container image that provides the computing environment for running the model. The service will run the container, which in turn acquires needed datasets, runs the model on the input parameters, and saves outputs to the SKOPE Data Store from which they can be visualized or downloaded using the web application. The container will run on the ROGER CyberGIS cluster; advanced users can download a container image that includes the needed computing environment and run it on their own (e.g., on their desktops or on cloud computing resources). In Year 2 we will provide and test cloud deployments for key reconstruction code.

ROGER - CyberGIS Computing Cluster. The SKOPE web application, databases, registries, model execution, and other services will run on the ROGER system managed by NCSA. This NSF-supported system [NCSA16] is designed for projects such as SKOPE by integrating HPC (high-performance computing), cloud computing, and big data storage. As a part of our sustainability strategy, the SKOPE system will run its services in software containers, which will provide reproducible and portable environments that will run on ROGER, other academic or commercial clusters, or desktop computers.

6.2 Comprehensive Provenance Management with YesWorkflow

To support transparency and reproducibility of research, SKOPE will: (1) automatically capture and store the provenance of all reconstructed datasets that SKOPE has computed; (2) enable researchers easily to explore, visualize, and query the provenance of the data, models, or other programs that produced them; and (3) facilitate the discovery, exploration, and analysis of data in terms of its provenance. YesWorkflow is the key technology we will employ and further extend to achieve these goals [MSK+15,YW16].

Revealing Prospective Provenance. Provenance information can be prospective or retrospective. *Prospective* provenance reveals in advance how—during the execution of a workflow (comprising a defined sequence of one or more scripts or programs)—the output data *will* be derived from the input.

We will employ YesWorkflow to declare the dataflow structure of the retrodiction models and other scripts available through SKOPE. YesWorkflow annotations, when applied to a script or program written in any text-based programming language, declare the computational steps and dataflow links between them. These annotations enable YesWorkflow to infer, in detail, the prospective provenance of the program's expected outputs [MBB+15]. For example, the prospective provenance can be queried to determine which input parameters will affect which computational steps in a script, or what input files will be used in computing a particular output dataset. YesWorkflow's graphical views of prospective provenance (cf. Figure 3) enable insights into the model's internal processes and the data it employs and produces. The annotations further serve to declare the input data types and parameter values for a model, and the types of data it outputs. These declarations will enable the SKOPE model execution service to correctly bind input datasets to model runs and to properly integrate output files with other SKOPE application components, even when contributed models are written in diverse programming languages and vary in their input and output data requirements. The declarations also will enable SKOPE to expose model-specific parameters to tinkerers requesting a customized model run. Since prospective provenance reveals the high-level, conceptual workflow underlying a set of executable code (e.g., a model), it also serves as documentation for its author and anyone wishing to use it.

Reconstructing Retrospective Provenance. *Retrospective* provenance comprises the records of computational steps and intermediate data production events *that actually occurred* during a run of a workflow and that led to one or more data products. Although YesWorkflow (YW) does not directly record these events while a script is executing, one development goal for YW is for it to fully *reconstruct* these events using a combination of: (1) prospective provenance, declared via YW annotations in the code; (2) actual data products left behind by a run of the workflow; and (3) other information recorded or produced during the run (e.g., log files, metadata stored in data file headers, or program reports). For example, by comparing the names and locations of files produced during a model run with variable-containing path templates declared in script comments, YW can infer the values of path-differentiating variables at specific points in the script and so answer queries about the computational lineage of data products [MBB+15]. We will enhance YesWorkflow's ability to infer retrospective provenance by similarly correlating information in script log files and data file headers with YW annotations that declare what information is stored in these files. For scripts written in R, MATLAB, or Python we will enable YesWorkflow to import additional provenance information from the recordr R package [SJJ16], the DataONE MATLAB Toolbox [Mat16], and noWorkflow [MBC+14], respectively. Integrating multiple sources of provenance information reliably requires detecting when provenance query results are ambiguous, e.g., due to incomplete (under-constrained) or conflicting (over-constrained) information. Addressing this problem likely will require techniques from *Answer Set Programming* [DRL13] and related logic programming approaches explored in the NSF-funded *Euler Project* led by Ludäscher [FCY+16].

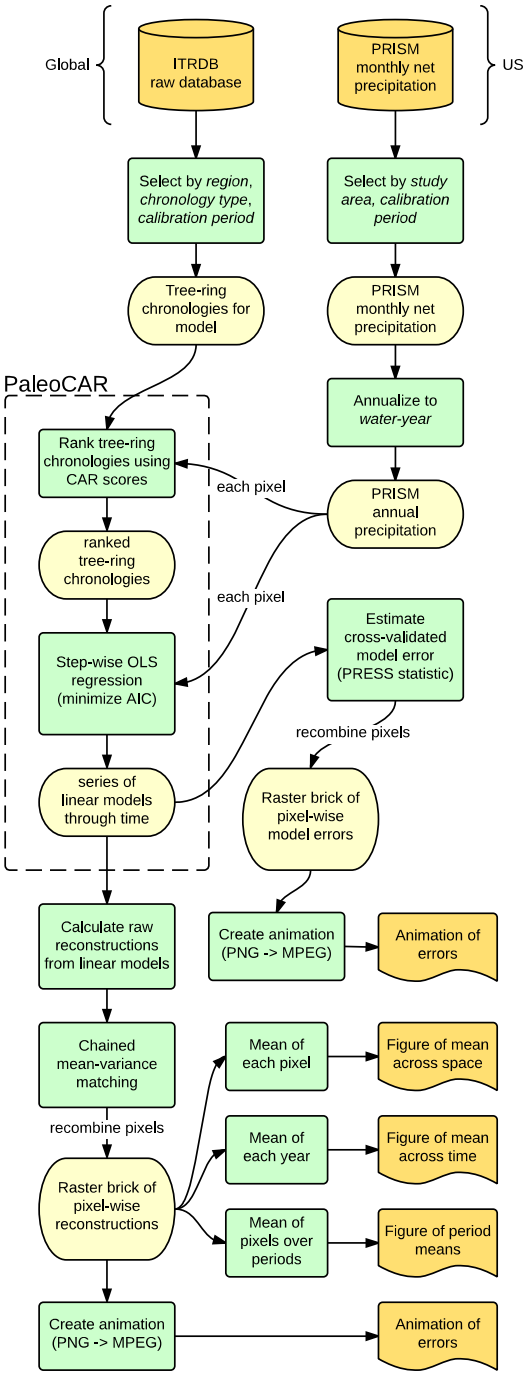

Figure 3. Provenance of a PaleoCAR reconstruction and visualizations. YesWorkflow will generate similar diagrams illustrating *prospectively* how SKOPE derives new data products, and *retrospectively* how existing data products were derived. These diagrams include both scripted operations that occurred during model runs, as well as manual operations carried out interactively through the SKOPE web application.

Complete Retrospective Provenance for All Data. The SKOPE application will record the sequences of model runs and data management operations used to compute all derived and retrodicted datasets. By combining this script-run level provenance with the fine-grained retrospective provenance reconstructed by YesWorkflow, users will have access, via visualizations and queries, to complete and highly detailed lineages for every data product computed by SKOPE.

Exporting Provenance for Further Research. SKOPE users will be able to export YesWorkflow data that capture the full history of selected data products (as a self-contained SQLite database, or a file to be imported into another relational DBMS). Users who adopt YesWorkflow in their day-to-day research on their own computers will be able to query the combined provenance of data products obtained through SKOPE and other data products, including those subsequently derived from SKOPE-generated datasets. This enables continuous, queryable, and visualizable provenance chains spanning SKOPE-provided *and* researcher-owned computer resources.

6.3 SKOPE Support for Reproducible Research

Reproducible Computing Environments. A key objective of SKOPE is full transparency and reproducibility [NAB+15] of operations, computations, and data products. Paleoclimate retrodiction models will be executed within Docker containers to ensure that the computing environments and dependencies are fully documented and reproducible. Researchers will be able to retrieve Docker images from Docker Hub and independently run the models on their own computer systems or public cloud computing resources. Because the full source code for models will be hosted in public Git repositories, model runs will be identified with specific versions of the models, changes between model versions will be evident, and users will be able to acquire, run, and modify the models independently. The source code for SKOPE application and service components will also be hosted in public Git repositories, and Docker images to run the complete SKOPE web application and computational back end will be shared publicly on Docker Hub [DOCK16]. This will allow any organization to reproduce the complete computing environment required to run SKOPE on their own computer systems or on public cloud resources.

Reproducible Interactive Workflows. Finally, each scientifically meaningful operation in SKOPE will have an associated YesWorkflow annotation. SKOPE will record the sequence of operations leading to a particular visualization, dataset, or other data product and enable this record later to be viewed and queried using YesWorkflow. For example, a researcher might decide to perform a reconstruction of precipitation from tree-rings using a different set of chronologies than used by the default model—she selects a new set of tree-rings (perhaps using a geographic bounding box) then submits the model with revised inputs to the SKOPE system. Once the latter run completes, the researcher will be able to visualize and query how the final dataset was produced, and the resulting query results will include the interactive operations (i.e., defining the bounding box to select the tree-ring chronologies) as well as the computations that occurred during the model runs that preceded and followed the interactive session. For users' convenience we also expect to provide capabilities for editing and rerunning recorded user interaction sequences as automated workflows within the SKOPE web application; such workflows, however, are likely to be fragile in the long run and thus will contribute less to reproducibility than will visualizations and queries of records of the original interactions and model runs.

7. Dissemination Plan

Publication/Tool Announcement. In early 2018, we will publicize SKOPE's release and illustrate use-cases with short technical articles in open-access periodicals in (for example) archaeology, historical geography, sustainability, paleoclimatology, and paleoecology. We will advertise the availability of SKOPE through online and newsletter outlets associated with relevant social and natural science professional organizations and we will attempt to place notices in blogs and newsfeeds of DataONE, Neotoma, tDAR, and SKOPE's participating institutions.

Online Seminars (Webinars). In spring 2018, Bocinsky, Brin, Rush, and McPhillips will present Webinars—webcast presentations and interactive sessions where participants will learn about the functions available in SKOPE and will use the tool while SKOPE personnel are available (digitally) to answer questions. We expect that it will be possible to do this, at a minimum, in SAA and DataONE webinar series. The webinars will be recorded and freely available on the SKOPE website.

The first webinar will focus on *researchers* who explore and download paleoenvironmental data and *tinkerers* who will also adjust and rerun the models. We will introduce SKOPE and guide participants through the processes of selecting a point or area of interest, viewing graphs, time-lapse videos and summary statistics, downloading the data for further statistical or GIS processing, and adjusting model parameters for alternative reconstructions. The workshop will highlight use cases that demonstrate how SKOPE can be used to answer actual research and policy questions. We will also discuss the methods that underlie key reconstructions and how users can delve into the provenance of the data they want to use.

The second webinar will focus on *modelers* who may edit model source code, download a model to run it locally, or contribute a model or dataset to SKOPE. In this workshop, Bocinsky, Rush, and McPhillips

will review SKOPE's basic functionality, and delve deeper into the provenance-recording and model-download functions. They will also cover registering new models and data on the SKOPE platform.

Conference Presentations and Exhibition Booth. In addition to the webinars, we will co-host a booth in the exhibition hall at the 2018 SAA meetings in Washington, D.C. SKOPE personnel will assist potential users in exploring SKOPE using laptops at the booth. Throughout the project, key project personnel will provide presentations on different aspects of the project at relevant professional meetings.

Web Metrics. We will track web metrics on page-views, downloads, and model runs using Google Analytics. Source code downloads and forks will be tracked using GitHub's repository tracking systems.

8. Project Personnel and Responsibilities

Project personnel have international reputations in archaeology, computer science, paleoenvironmental modeling, and ecology. Having worked effectively together for 2+ years on the BCC-SKOPE pilot, we are prepared to immediately initiate the project. Consultants are identified in the Technical Plan.

Keith Kintigh (PI) is Associate Director of the School of Human Evolution and Social Change and Professor of Archaeology at ASU. He serves on the Board of Directors of Digital Antiquity and is a former president of the SAA. Kintigh led the initial development of tDAR (the Digital Archaeological Record; [Kin06,MK10]), an international digital repository for archaeological data. He has extensive experience leading and participating in transdisciplinary research and will be responsible for overall project leadership and coordination.

Timothy Kohler (PI) is Regents Professor of Anthropology at WSU, external professor at SFI, and Research Associate at Crow Canyon Archaeological Center (CCAC). He directed a joint WSU/University of Washington IGERT training grant on evolutionary methods and theory in anthropology and biology. For 12 years he coordinated the Village Ecodynamics Project, funded by NSF's Biocomplexity and CNH competitions (see Section 4). Kohler will coordinate the project's paleoclimate, paleovegetation, and paleodemographic efforts, working with Bocinsky, the consultants, and the WSU RA.

Bertram Ludäscher (PI) is Professor and Director of the Center for Informatics Research in Science and Scholarship in the Graduate School of Library and Information Science at the University of Illinois, Urbana-Champaign (UIUC). He is a faculty affiliate at NCSA and the Department of Computer Science. Ludäscher is a leader in scientific data and knowledge management research, focusing on workflows, provenance, data integration, and knowledge representation. He co-founded the open-source Kepler project, and serves on the DataONE leadership team. Ludäscher will oversee the overall technical direction, focusing on models and techniques for provenance analytics.

R. Kyle Bocinsky (Co-PI) is the Director of Sponsored Projects at CCAC and Adjunct Research Faculty at WSU. He has extensive experience developing paleoenvironmental reconstruction tools. He led development of *PaleoCAR*, the SKOPE prototype's paleoclimate model, and *FedData,* which downloads data from federated repositories and is essential to SKOPE workflows. With WSU researchers and our external collaborators Bocinsky will extend *PaleoCAR* to the CONUS, develop a SW maize paleoproductivity reconstruction, and implement the other reconstruction methods listed in Table 1.

Ann Kinzig (Co-PI) is Professor in the ASU School of Life Sciences, Chief Research Strategist in ASU's Global Institute of Sustainability, and was co-PI of one of the CNH projects SKOPE is leveraging. As an ecologist, Kinzig will provide key oversight of project modeling efforts and will assist in adapting the risk landscape and ethnobotanical species presence models to SKOPE.

Timothy McPhillips (Co-PI) is a scientific software developer at UIUC. He designed the *Blu-Ice/DCS* software used at synchrotron light sources around the world; co-developed the *RestFlow* workflow system used by the *AutoDrug* fragment-based drug design platform; and is the primary developer of *YesWorkflow*. McPhillips will lead the SKOPE engineering team, continue developing YesWorkflow, and pursue opportunities for accelerating *PaleoCAR* and other retrodiction models using GPUs.

Shaowen Wang (Co-PI) is Professor of Geography & Geographic Information Science at UIUC, Associate Director of NCSA, and Founding Director of CyberGIS Center for Advanced Digital and Spatial Studies. He was a Councilor of the Open Science Grid. He is President-Elect of the University Consortium for Geographic Information Science, and a member of the National Academies' Board on Earth Sciences and Resources. He has led several multi-institutional NSF grants concerned with cyberGIS and scientific problem solving in numerous disciplines. Wang will supervise cyberGIS project staff, work closely with McPhillips, and oversee NCSA development with a particular emphasis on sustainability.

Adam Brin is Director of Technology for ASU's Center for Digital Antiquity. He will ensure that tDAR repository content is effectively integrated in SKOPE, provide expertise in archaeological informatics and assist in ensuring that SKOPE metadata and data meet technical challenges.

Johnathan Rush is Education, Outreach, and Training coordinator for the UIUC CyberGIS Center for Advanced Digital and Spatial Studies. During the design grant he assisted in the prototype's implementation on advanced cyberinfrastructure [BRK+16]. He will contribute experience in geovisualization and conduct outreach to the geography and cyberinfrastructure communities.

8. Broader Impacts

Enhanced Infrastructure for Research and Education. This project is directly focused on enhancing infrastructure and expanding its use by researchers in academia and industry. SKOPE builds on vast amounts of prior data collection and previous research, transforming them into readily usable environmental knowledge. SKOPE will substantially enhance scholars' ability to execute reproducible research on a broad range of social and natural science research topics, especially those involving long-term interactions of humans with their environments. SKOPE will also facilitate ongoing monitoring and improvement of included models for compatibility, efficiency, and accuracy. Inside and outside SKOPE, researchers will be able to use YesWorkflow's provenance-tracking capabilities.

Education and the General Public. All users will have free access to high-quality environmental scenarios in which to situate studies in diverse subjects. As members of the general public see how ancient environments differ from modern ones in places of interest to them, SKOPE may serve to highlight the magnitude of both climate change and human impacts.

Increased Partnerships between Academia and Industry. SKOPE will make available results of NSF- and other publicly-funded paleoenvironmental data both to the academy and to industry. To take a single example, every year, small-scale paleoenvironmental reconstructions are created by thousands of heritage management and environmental assessment projects conducted in response to US law. SKOPE would provide vastly superior information at no cost, freeing funds for other uses.

Public Policy. Finding appropriate societal responses to long-term climate change and accelerating anthropogenic impact on landscapes is a critical public policy priority [EKF+13]. Weather and streamflow records from the historic calibrated period are far too short in duration to use in planning [COC+12]. Preparing for impending climate change is increasingly seen as critical to the welfare of the country. [WWL+10] have shown how model-based decision support systems can incorporate long-term environmental records of the sort provided by SKOPE. With SKOPE, planners could much more easily use long-term environmental data to investigate vulnerabilities of existing infrastructure to drought or temperature conditions not experienced in recent history. For example, had long-term environmental data been available in 1922 when the Colorado River Compact was negotiated, their use could have avoided a public policy disaster. Reconstructed streamflow data have shown that the early 20th century—on which water allocation for seven western states is now based—had "the highest sustained flows in the entire record, 1520 to 1961…. In effect, water that was not likely to be in the river on a consistent basis was divided among the basin states" [WGM06].

References Cited

[ABL09] Anand, M. K., S. Bowers, and B. Ludäscher. (2009) A Navigation Model for Exploring Scientific Workflow Provenance Graphs. In *Proceedings of the 4th Workshop on Workflows in Support of Large-Scale Science*, p. 2. ACM, New York.

[BFG14] Jim Basney, Terry Fleury, and Jeff Gaynor. (2014) CILogon: A Federated X.509 Certification Authority for CyberInfrastructure Logon. In *Concurrency and Computation: Practice and Experience*, 26(13):2225-2239. http://dx.doi.org/10.1002/cpe.3265.

[BK14] Bocinsky, R. Kyle, and Timothy A. Kohler. (2014) A 2,000-year reconstruction of the rain-fed maize agricultural niche in the US Southwest. *Nature Communications* 5:5618. DOI: 10.1038/ncomms6618.

[BKB+16] Bocinsky, R. Kyle, Timothy A. Kohler, Adam Brin, Keith W. Kintigh, Bertram Ludäscher, Timothy McPhillips, Johnathan Rush. (2016) *SKOPE User's Guide*. http://www.envirecon.org/skope-prototype-users-guide (accessed 5 February 2015)

[BKW01] Buneman, Peter, Sanjeev Khanna, and Tan Wang-Chiew. (2001) "Why and where: A characterization of data provenance." In *Database Theory—ICDT 2001*, pp. 316-330. Springer Berlin Heidelberg.

[BMC+06] Bowers, S., T. McPhillips, B. Ludäscher, S. Cohen, and S. B. Davidson. (2006) A Model for User-oriented Data Provenance in Pipelined Scientific Workflows. In *Provenance and Annotation of Data*, pp. 133–147. Springer Berlin Heidelberg.

[Boc15] Bocinsky, R. Kyle. (2015) *PaleoCAR: Paleoclimate Reconstruction from Tree Rings using Correlation Adjusted correlation*. R package version 2.1. https://github.com/bocinsky/PaleoCAR/archive/2.1.tar.gz.

[Boc16] Bocinsky, R. Kyle. (2016) *FedData: Functions to Automate Downloading Geospatial Data Available from Several Federated Data Sources*. R package version 2.0.4. https://github.com/bocinsky/FedData.

[BRK+16] Bocinsky, R. Kyle, Johnathan Rush, Keith W. Kintigh, and Timothy A. Kohler. In Press. Exploration and exploitation in the macrohistory of the prehispanic Pueblo Southwest. *Science Advances* (accepted 11 February 2016)

[But15] Butterfield, Bradley J. (2015) Environmental filtering increases in intensity at both ends of climatic gradients, though driven by different factors, across woody vegetation types of the southwest USA. *Oikos* 124:1374–1382. doi: 10.1111/oik.02311.

[BVS+16] Brewer, P.W., Velásquez, M.E., Sutherland, E.K. and Falk, D.A. (2016) *Fire History Analysis and Exploration System (FHAES)*. http://www.fhaes.org (accessed 21 February 2016)

[CIL16] CILogon 2.0 - An Integrated Identity and Access Management Platform for Science. http://www.cilogon.org/news/cilogon20 (accessed 21 February 2016)

[CK04] Cook, E. R. and P. J. Krusic. (2004) The North American Drought Atlas. Lamont-Doherty Earth Observatory and the National Science Foundation. http://iridl.ldeo.columbia.edu/SOURCES/.LDEO/.TRL/.NADA2004/.pdsi-atlas.html.

[COC+12] Cutter, S., B. Osman-Elasha, J. Campbell, S.M. Cheong, S. McCormick, R. Pulwarty, S. Supratid, and G. Ziervogel (2012) Managing the risks from climate extremes at the local level. In C. B.

Field, V. Barros, T. F. Stocker, D. Qin, D. J. Dokken, K. L. Ebi, M. D. Mastrandrea, K. J. Mach, G.-K. Plattner, S. K. Allen, M. Tignor & P. M. Midgley (Eds.), *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change (IPCC)*,pp. 291-338. Cambridge University Press, Cambridge, UK, and New York, NY, USA.

[Cyb16] CyberGIS. (2016) *Project Publications*. http://cybergis.cigi.uiuc.edu/cyberGISwiki/doku.php/papers (accessed 24 February 2016).

[DBE+07] Davidson, S. B., S. C. Boulakia, A. Eyal, B. Ludäscher, T. M. McPhillips, S. Bowers, M. K. Anand, and J. Freire. (2007) Provenance in Scientific Workflow Systems. *IEEE Data Eng. Bull.* 30(4):44–50.

[DHS+08] Daly, Christopher, Michael Halbleib, Joseph I. Smith, Wayne P. Gibson, Matthew K. Doggett, George H. Taylor, Jan Curtis and Phillip P. Pasteris. (2008) Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology*. DOI: 10.1002/joc.1688.

[DRL13] Dey, S., Riddle, S. and Ludäscher, B., (2013). Provenance analyzer: Exploring provenance semantics with logic rules. *5th Workshop on the Theory and Practice of Provenance* (TaPP), Lombard, IL.

[DOCK16] Docker (2016) Build, Ship, & Run Any App, Anywhere. https://hub.docker.com/ (accessed 21 February 2016).

[DONE16] DataONE. (2016) *DataONE: Data Observation Network for Earth*. https://www.dataone.org (accessed 21 February 2016).

[EMD+15] Elliott, J., C. Müller, D. Deryng, J. Chryssanthacopoulos, K. J. Boote, M. Büchner, I. Foster, M. Glotter, J. Heinke, T. Iizumi, and R.C. Izaurralde. (2015) The global gridded crop model intercomparison: Data and modeling protocols for phase 1 (v1.0). *Geoscientific Model Development* 8(2):261–277.

[EKF+13] Ellis, Erle C., Jed O. Kaplan, Dorian Q. Fuller, Steve Vavrus, Kees Klein Goldewijk, and Peter H. Verburg. (2013) Used planet: A global history. *PNAS* 110(20):7978–7985. DOI: 10.1073/pnas.1217241110.

[FCY+16] Franz, N.M., Chen, M., Yu, S., Bowers, S. and Ludäscher, B., (2016) Names are not good enough: reasoning over taxonomic change in the Andropogon complex. *Semantic Web Journal–Interoperability, Usability, Applicability-Special Issue on Semantics for Biodiversity*, 7(6), to appear.

[FRW10] Fyfe, R. M., N. Roberts, and J. Woodbridge (2010) A pollen-based pseudobiomisation approach to anthropogenic land-cover change. *The Holocene* 20:1165–1171.

[FYW+14] Fan, Y., Y. Liu, S. Wang, D. Tarboton, A. Yildirim, and N. Wilkins-Diehr. (2014) Accelerating TauDEM as a Scalable Hydrological Terrain Analysis Service on XSEDE. In *Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment,* p. 5. ACM.

[GF97] Grissino-Mayer, Henri D. and Harold C. Fritts. (1997) The International Tree-Ring Data Bank: An enhanced global database serving the global scientific community. *The Holocene* 7:235–238.

[GMP+11] Gajewski, K., S. Munoz, M. Peros, A. Viau, R. Morlan, and M. Betts. (2011) The Canadian Archaeological Radiocarbon Database (CARD): Archaeological ^{14}C Dates in North America and their Paleoenvironmental Context. *Radiocarbon* 53(2):371–394.

[Gri08] Grimm, Eric C. (2008) Neotoma: An Ecosystem Database for the Pliocene, Pleistocene, and Holocene (Draft). http://www.neotomadb.org/uploads/NeotomaManual.pdf.

[HCD+10] Hill, J. Brett, Jeffrey J. Clark, William H. Doelle, and Patrick D. Lyons (2010) Depopulation of the Northern Southwest: A Macroregional Perspective. In *Leaving Mesa Verde: Peril and Change in the Thirteenth-century Southwest*, edited by T. A. Kohler, M. D. Varien, and A. M. Wright, pp. 34-52. University of Arizona Press, Tucson.

[HLL+15] Hu, H., T. Lin, Y. Y. Liu, S. Wang, and L. F. Rodríguez. (2015) CyberGIS-BioScope: a cyberinfrastructure-based spatial decision-making environment for biomass-to-biofuel supply chain optimization. *Concurrency and Computation: Practice and Experience*, 27(16): 4437-4450.

[HPK+08] Hegmon, M., M. Peeples, A. Kinzig, S. Kulow, C.M. Meegan, and M.C. Nelson. (2008) Social Transformation and Its Costs in the Prehistoric US Southwest. *American Anthropologist* 110(3):313–324.

[JHP+03] Jones, James W., Gerrit Hoogenboom, Cheryl H. Porter, Ken J. Boote, William D. Batchelor, L. Anthony Hunt, Paul W. Wilkens, U. Singh, Arjan J. Gijsman, and Joe T. Ritchie. (2003) The DSSAT cropping system model. *European Journal of Agronomy* 18:235–265.

[JWS16] Jeong, M.-H., S. Wang, and C.J. Sullivan. (2016) Analysis of dynamic radiation level changes using surface networks. In *Advancing Geographic Information Science: The past and Next Twenty Years*, Needham, edited by H.J. Onsrud and W. Kuhn, pp. 199-212. MA: GSDI Association Press.

[KAB+14a] Kintigh, Keith W., Jeffrey H. Altschul, Mary C. Beaudry, Robert D. Drennan, Ann P. Kinzig, Timothy A. Kohler, W. Fredrick Limp, Herbert D.G. Maschner, William K. Michener, Timothy R. Pauketat, Peter Peregrine, Jeremy A. Sabloff, Tony J. Wilkinson, Henry T. Wright, and Melinda A. Zeder. (2014) Grand Challenges for Archaeology. *American Antiquity* 79(1):5–24.

[KAB+14b] Kintigh, Keith W., Jeffrey H. Altschul, Mary C. Beaudry, Robert D. Drennan, Ann P. Kinzig, Timothy A. Kohler, W. Fredrick Limp, Herbert D.G. Maschner, William K. Michener, Timothy R. Pauketat, Peter Peregrine, Jeremy A. Sabloff, Tony J. Wilkinson, Henry T. Wright, and Melinda A. Zeder. (2014) Grand Challenges for Archaeology. *Proceedings of the National Academy of Sciences* 111(3):879-880.

[KAK+15] Kintigh, Keith W., Jeffrey H. Altschul, Ann P. Kinzig, W. Fredrick Limp, William K. Michener, Jeremy A. Sabloff, Edward J. Hackett, Timothy A. Kohler, Bertram Ludäscher, and Clifford A. Lynch. (2015) Cultural Dynamics, Deep Time, and Data: Planning Cyberinfrastructure Investments for Archaeology. *Advances in Archaeological Practice* 3(1):1-15. DOI: 10.7183/2326-3768.3.1.1.

[KB15] Kohler, Timothy A. and R. Kyle Bocinsky. (2015) Compiled Tree-ring Dates from the Southwestern United States. tDAR id: 399315; doi:10.6067/XCV86974XW.

[KBC+12] Kohler, Timothy A., R. Kyle Bocinsky, Denton Cockburn, Stefani A. Crabtree, Mark D. Varien, Kenneth E. Kolm, Schaun Smith, Scott G. Ortman, and Ziad Kobti. (2012) Modelling Prehispanic Pueblo Societies in their Ecosystems. *Ecological Modelling* 241:30–41. DOI: 10.1016/j.ecolmodel.2012.01.002.

[KCB+15] Kohler, Timothy A., Stefani A. Crabtree, R. Kyle Bocinsky, and Paul L. Hooper. (2015) Sociopolitical Evolution in Midrange Societies: The Prehispanic Pueblo Case. SFI Working Paper 2015-04-011, DOI: 10.13140/RG.2.1.1737.3204. Chapter in volume tentatively titled *Complexity and Society: An Introduction to Complex Adaptive Systems and Human Society*, edited by Jeremy Sabloff et al., submitted to Princeton University Press, Princeton, NJ.

[Kin06] Kintigh, Keith W. (2006) The Promise and Challenge of Archaeological Data Integration. *American Antiquity* 71(3):567–578.

[KOG+14] Kohler, Timothy A., Scott G. Ortman, Katie E. Grundtisch, Carly M. Fitzpatrick, and Sarah M. Cole. (2014) The Better Angels of Their Nature: Declining Violence Through Time among Prehispanic Farmers of the Pueblo Southwest. *American Antiquity* 79(3):444–464.

[KR14] Kohler, Timothy A. and Kelsey M. Reese. (2014) Long and Spatially Variable Neolithic Demographic Transition in the North American Southwest. *PNAS* 111(28):10101-10106.

[Kur16] Kurator. (2016) *Kurator.* Project web page, accessed 02/24/2016 at https://opensource.ncsa.illinois.edu/projects/KURATOR.

[KV12] Kohler, Timothy A., and Mark D. Varien (editors). (2012) *Emergence and Collapse of Early Villages: Models of Central Mesa Verde Archaeology*. University of California Press, Berkeley.

[LKB+12] Ljungqvist, F. C., P. J. Krusic, G. Brattström, and H. S. Sundqvist. (2012) Northern Hemisphere temperature patterns in the last 12 centuries. *Climate of the Past* 8: 227. doi:10.5194/cp-8-227-2012.

[LMS+15] Ludäscher, B., T.M. McPhillips, T. Song, J. Hanken, D. Lowery, J.A. Macklin, P.J. Morris, and R.A. Morris. (2015) Kurator: An Extensible, open-source workflow platform for users and makers of data curation tools. *Society for the Preservation of Natural History Collections, 30th Annual Meeting.*

[LP15] Ludäscher, Bertram, and Beth Plale(editors). (2015) *Provenance and Annotation of Data and Processes: 5th International Provenance and Annotation Workshop, IPAW 2014, Cologne, Germany, June 9-13, 2014. Revised Selected Papers*. Vol. 8628. Springer.

[LPW15] Liu, Y., A. Padmanabhan, and S. Wang. (2015) CyberGIS Gateway for enabling data-rich geospatial research and education: CYBERGIS GATEWAY. Concurrency and Computation: Practice and Experience,27(2): 395–407. http://doi.org/10.1002/cpe.3256.

[MAB+12] Michener, W.K., S. Allard, A. Budden, R.B. Cook, K. Douglass, M. Frame, S. Kelling, R. Koskela, C. Tenopir, and D. Vieglais. (2012) Participatory Design of DataONE—Enabling Cyberinfrastructure for the Biological and Environmental Sciences. *Ecological Informatics* 11:5–15.

[Mat15] Matson, R.G. (in press) The nutritional context of the Pueblo III depopulation of the northern San Juan: Too much maize? *Journal of Archaeological Science: Reports*. http://dx.doi.org/10.1016/j.jasrep.2015.08.032.

[Mat16] Matab DataONE Toolbox. https://github.com/DataONEorg/matlab-dataone.

[MBB+15] McPhillips, T., S. Bowers, K. Belhajjame, and B. Ludäscher. (2015) Retrospective Provenance Without a Runtime Provenance Recorder. *7th USENIX Workshop on the Theory and Practice of Provenance* (TaPP'15), July 8-9, Edinburgh, Scotland.

[MBC+14] Murta, L., V. Braganholo,, F. Chirigati,, D. Koop, and J. Freire. (2014) noWorkflow: Capturing and analyzing provenance of scripts. In *Provenance and Annotation of Data and Processes,* pp.71-83. Springer International Publishing.

[MCH+09] Marchant, R., A. Cleef, S. P. Harrison, H. Hooghiemstra, V. Markgraf, J. van Boxel, T. Ager, L. Almeida, R. Anderson, C. Baied, H. Behling, J. C. Berrio, R. Burbridge, S. Bjorck, R. Byrne, M. Bush, J. Duivenvoorden, J. Flenley, P. De Oliveira, B. van Geel, K. Graf, W. D. Gosling, S. Harbele, T. van der Hammen, B. Hansen, S. Horn, P. Kuhry, M.-P. Ledru, F. Mayle, B. Leyden, S. Lozano-Garcıa, M.

Melief, P. Moreno, N. T. Moar, A. Prieto, G. van Reenen, M. Salgado-Labouriau, F. Schäbitz, E. J. Schreve-Brinkman, and M. Wille. (2009) Pollen-based biome reconstructions for Latin America at 0, 6000 and 18 000 radiocarbon years ago. *Climate of the Past* 5:725-767. doi:10.5194/cp-5-725-2009.

[MJ03] Mann, Michael E. and Phil D. Jones. (2003) Global surface temperatures over the past two millennia. *Geophysical Research Letters* 30(15). doi:10.1029/2003GL017814.
[MK10] McManamon, Francis P., and Keith W. Kintigh. 2010. Digital Antiquity: Transforming Archaeological Data into Knowledge. *SAA Archaeological Record* 10(2): 37-40.

[MLK+15] P.J. Morris, B. Ludäscher, S. Köhler, J. Hanken, D. Lowery, J.A. Macklin, T.M. McPhillips, P.J. Morris, R.A. Morris, and T. Song. (2015) A scientific workflow tool for targeted data quality improvement of natural science collections data. *DemoCamp: Society for the Preservation of Natural History Collections, 30th Annual Meeting.*

[MSH+05] Moberg, A., D. M. Sonechkin, K. Holmgren, N. M. Datsenko, and W. Karlén. (2005) Highly Variable Northern Hemisphere Temperatures Reconstructed from Low- and High-Resolution Data. *Nature* 433:613–617.

[MSK+15] McPhillips,Timothy, Tianhong Song,Tyler Kolisnik, Steve Aulenbach, Khalid Belhajjame, Kyle Bocinsky, Yang Cao, James Cheney, Fernando Chirigati, Saumen Dey, Juliana Freire, Christopher Jones, James Hanken, Keith W. Kintigh, Timothy A. Kohler, David Koop, James A. Macklin, Paolo Missier, Mark Schildhauer, Christopher Schwalm, Yaxing Wei, Mark Bieda, and Bertram Ludäscher. (2015) YesWorkflow: A User-Oriented, Language-Independent Tool for Recovering Workflow Information from Scripts. *International Journal of Digital Curation* 10(1):298–313. DOI: 10.2218/ijdc.v10i1.370.

[NAB+15] Nosek, B.A., G. Alter, G.C. Banks, D. Borsboom, S.D. Bowman, S.J. Breckler, S. Buck, C.D. Chambers, G. Chin, G. Christensen, and M. Contestabile. (2015) Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility. *Science (New York, NY)*, *348*(6242):1422. (Available from https://cos.io/top/).

[NCSA16] NCSA. (2016) *ROGER System Information.* https://wiki.ncsa.illinois.edu/display/ROGER/ (accessed 24 February 2016).

[Nel11] Nelson, Margaret C. (2011) Synthesis: Vulnerability, Traps, and Transformations—Long Term Perspectives from Archaeology. *Ecology and Society* 16 (2): 24. (online) URL: http://www.ecologyandsociety.org/vol16/iss2/art24/.

[Neo16a] Neotoma. (2016) *Neotoma Paleoecology Database*. http://www.neotomadb.org (accessed 21 February 2016).

[Neo16b] North American Pollen Database. http://www.neotomadb.org/groups/category/napd (accessed 21 February 2016).

[NHK+11] Nelson, M. C., M. Hegmon, S. R. Kulow, M. A. Peeples, K. W. Kintigh, and A. P. Kinzig. (2011) Resisting Diversity: A Long-term Archaeological Study. *Ecology and Society* 16(1):online 25. URL: http://www.ecologyandsociety.org/vol16/iss1/art25/.

[NHK+12] Nelson, M.C., M. Hegmon, K.W. Kintigh, A.P. Kinzig, B.A. Nelson, J.M. Anderies, D.A. Abbott, K.A. Spielmann, S.E. Ingram, M.A. Peeples, S. Kulow, C.A. Strawhacker, and C. Meegan. (2012) Long-term Vulnerability and Resilience: Three Examples from Archaeological Study in the Southwestern US and Northern Mexico. In *Surviving Sudden Environmental Change,* edited by J. Cooper and P. Sheets, pp. 193–217. University Press of Colorado, Boulder.

[NID+16] Nelson, Margaret C., Scott E. Ingram, Andrew J. Dugmore, Richard Streeter, Matthew A. Peeples, Thomas H. McGovern, Michelle Hegmon, Jette Arneborg, Keith W. Kintigh, Seth Brewington, Katherine A. Spielmann, Ian A. Simpson, Colleen Strawhacker, Laura E.L. Comeau, Andrea Torvinen, Christian K. Madsen, George Hambrecht, and Konrad Smiarowski. (2016) Climate Challenges, Vulnerabilities, and Food Security. *PNAS* published online before print December 28, 2015, 113(2):298-303. DOI:10.1073/pnas.1506494113; http://www.pnas.org/content/113/2/298.

[NKA+10] Nelson, M. C., K. Kintigh, D. R. Abbott, and J. M. Anderies. (2010) The Cross-scale Interplay between Social and Biophysical Context and the Vulnerability of Irrigation-dependent societies: Archaeology's Long Term Perspective. *Ecology and Society* 15(3): 31. [online] URL: http://www.ecologyandsociety.org/vol15/iss3/art31/.

[NRCS16a] NRCS: Natural Resources Conservation Service, United States Department of Agriculture. (2016a) *Soil Survey Geographic (SSURGO) Database*. http://sdmdataaccess.nrcs.usda.gov/ (accessed 21 February 2016).

[NRCS16b] NRCS: Natural Resources Conservation Service, United States Department of Agriculture. (2016b) *Gridded Soil Survey Geographic (SSURGO) Database*. http://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/geo/?cid=nrcs142p2_053628 (accessed 21 February 2016).

[Ort16] Ortman, Scott G. (2016) Uniform Probability Density Analysis and Population History in the Northern Rio Grande. *Journal of Archaeological Method and Theory* 23:95-106 DOI 10.1007/s10816-014-9227-6.

[OW12] Ohlwein, Christian, and Eugene R. Wahl. (2012) Review of probabilistic pollen-climate transfer methods. *Quaternary Science Reviews* 31:17-29.

[OWP85] Overpeck, J. T., T. Webb III, and I. C. Prentice. (1985) Quantitative interpretation of fossil pollen spectra: dissimilarity coefficients and the method of modern analogs. *Quaternary Research* 23:87–108.

[PGH+96] Prentice, I. C., J, Guiot, B. Huntley, D. Jolly, and R. Cheddadi. (1996) Reconstructing biomes from palaeoecological data: a general method and its application to European pollen data at 0 and 6 ka. *Climate Dynamics* 12:185–194.

[RCG+10] Reid, Walter V., D. Chen, L. Goldfarb, Heide Hackmann, Y.T. Lee, K. Mokhele, Elinor Ostrom, Kari Raivio, Johan Rockström, Hans Joachim Schellnhuber, and A. White. (2010) Earth system science for global sustainability: grand challenges. *Science* 330:916–917.

[RT13] Roderick, M. J. and T. L. Nyerges. (2013) Structured Participation Toolkit: An enabler for knowledge production in Science Gateway, Cluster Computing (CLUSTER), *2013 IEEE International Conference*, Indianapolis, IN, 2013, p. 1-3. doi: 10.1109/CLUSTER.2013.6702699.

[SBO+16] Schwindt, D. M., R. K. Bocinsky, S. G. Ortman, D. M. Glowacki, M. D. Varien, and T. A. Kohler. (2016) The Social Consequences of Climate Change in the Central Mesa Verde Region. *American Antiquity* 81(1):74–96.

[SJJ16] Slaughter, P., M. B. Jones, and C. Jones. (2016) *recordr: Provenance tracking for R*. R package. https://github.com/NCEAS/recordr.

[SKOPE16a] SKOPE. (2016) SKOPE: Synthesizing Knowledge of Past Environments. http://envirecon.org/ (accessed 9 February 2016).

[SKOPE16b] SKOPE. (2016b) SKOPE: Synthesizing Knowledge of Past Environments Prototype. http://demo.envirecon.org/browse/ (accessed 9 February 2016).

[TDAR16] TDAR: The Digital Archaeological Record. (2016) *tDAR: The Digital Archaeological Record, A Service of Digital Antiquity*. http://www.tdar.org(accessed 21 February 2016).

[TREE16] TreeFlow. (2016) *TreeFlow: Streamflow reconstructions from tree rings*. http://treeflow.info(accessed 21 February 2016).

[USGS16a] USGS: U.S. Geological Survey. (2016) *The National Elevation Dataset*. http://ned.usgs.gov(accessed 21 February 2016).

[USGS16b] USGS: U.S. Geological Survey. (2016) *The National Hydrography Dataset*. http://nhd.usgs.gov(accessed 21 February 2016).

[VLG11] Viau, A.E., M. Ladd, and K. Gajewski. (2011) The climate of North America during the past 2000 years reconstructed from pollen data. *Global and Planetary Change* 84-85:75–83. doi:10.1016/j.gloplacha.2011.09.010.

[WAB+13] Wang, S., L. Anselin, B. Bhaduri, C. Crosby, M.F. Goodchild, Y. Liu, and T. L. Nyerges. (2013) CyberGIS software: a synthetic review and integration roadmap. International Journal of Geographical Information Science, 27(11): 2122–2145. http://doi.org/10.1080/13658816.2013.776049

[WGM06] Woodhouse, C. A., S. T. Gray, and D. M. Meko. (2006) Updated Streamflow Reconstructions for the Upper Colorado River Basin. *Water Resources Research 42*(5).

[WKK+14] Wells, J. J., E.C. Kansa, S. W. Kansa, S. J. Yerka, D. G. Anderson, T. G. Bissett, K. N. Myers, and R. C. DeMuth. (2014) Web-based discovery and integration of archaeological historic properties inventory data: The Digital Index of North American Archaeology (DINAA). *Literary and Linguistic Computing* 29(3):349-360. DOI: 10.1093/llc/fqu028.

[WWL+10] White, D. D., A. Wutich, K. L. Larson, P. Gober, T. Lant, and C. Senneville. (2010) Credibility, salience, and legitimacy of boundary objects: water managers' assessment of a simulation model in an immersive decision theater. *Science and Public Policy* 37(3):219-232.

[YW16] YesWorkflow Source Code Repository https://github.com/yesworkflow-org/

Data Management Plan

Overview and Principles. SKOPE builds on vast amounts of prior data collection and modeling, with the goal of making paleoenvironmental knowledge readily available and highly useable. SKOPE will *not* collect new observational data, but instead make available paleoenvironmental reconstruction datasets that result from running (typically, but not necessarily, published) models. These models are part of scientific workflows (implemented in R, MATLAB, or Python, for example) that access, transform, and analyze data from established, primary sources by means of well-documented computational models such as PaleoCAR [BK14,Boc15]. SKOPE will also modestly incentivize entry of existing data derived from pollen cores and archaeological tree-ring-dated samples into repositories accessible through SKOPE.

Principal aims in developing and deploying SKOPE (prototyped in the BCC-SKOPE pilot) are to provide *transparency* and *reproducibility* of the underlying computations, workflows, and data products. We have built into the design of SKOPE mechanisms to capture, manage, and employ provenance information to facilitate transparency and reproducibility. New computational data products and models created through SKOPE are not only shared freely through the web site, but are also thoroughly documented by provenance information that identifies the underlying source data and the computational methods and software employed.

This rich provenance information will be easily accessible through the web site ("provenance for others"). We can publish and share this provenance due to SKOPE's use of tools in support of "provenance for self." The YesWorkflow tools, employed and extended by SKOPE, allow authors of computational models to capture and query prospective and retrospective provenance during model development and tuning. In this way, by the time models and data products are ready to be shared through SKOPE, the model documentation (prospective provenance) and the data-processing history (retrospective provenance) are readily available as well.

Documentation, Tutorials, and Metadata in the form of README files, build documentation, Dockerfiles, and build files will describe the production of executables and container images from source code. Software engineering artifacts may include standard documentation formats (e.g., UML, JavaDoc) but also custom formats and tools, in particular YesWorkflow to support advanced modeling, querying and visualization of provenance information. For data exchange we employ well-known standards, e.g., from FGCS[1], and follow best practices[2]. For exchanging provenance information, we will employ W3C PROV; and its DataONE extension ProvONE when combining prospective and retrospective provenance.

Plans for Archiving and Preservation. In support of an ongoing software development community, source code will be maintained in a public open-source repository such as GitHub or GitLab. Docker containers will be shared via DockerHub. In addition, snapshots of major releases of the source code and published model-run configurations will be archived at an institutional repository. All digital objects will be stored in preservation-friendly formats. Special attention will be paid to keep software executable despite changes to the code and the underlying platforms, as described in the Technical and Sustainability Plans. SKOPE will be deployed initially on ROGER which uses resilient data storage (RAID) technology. For performance reasons, a SKOPE data cache is used, but content can always be recreated via the primary resources and published model-run configurations. The latter are inherently small in size (few MBs) and will be backed up regularly to institutional file shares, to a Box.net account provided under contract from Box to UIUC (250GB free to research groups), and to additional locations if needed.

Policies for Data Sharing, Public Access, and Confidentiality. All data and software products will be shared following established best practices for data sharing and software engineering (see also the Technical Plan and Sustainability Plan supplemental documents). Other products such as research reports and scientific publications will be made accessible via institutional repositories and as peer-reviewed

[1] http://www.ngs.noaa.gov/FGCS/tech_pub/1984-stds-specs-geodetic-control-networks.htm

[2] https://www.dataone.org/best-practices/metadata

publications. Confidentiality of precise archaeological site locations currently enforced by DINAA will be preserved (for the most part DINAA aggregates their locations by county), and we will work with personnel from CARD, the LTRR, the VEP, and the Coalescent Communities Database to obscure precise site locations to their specifications.

All SKOPE software development will be open source under a permissive license such as the Apache Software License Version 2.0, or a similar BSD- or MIT-style license. The choice of such a license will enable SKOPE to leverage other open-source components with compatible licenses. This will allow our development to focus on SKOPE-specific features, while leveraging functionality provided by widely-used libraries and web apps, or by other open source efforts providing specialized but relevant features for SKOPE (e.g., DataONE tools for provenance capture, data sharing, and indexing).

Roles and Responsibilities. The build manager (one of the two developers at Illinois) will be responsible for technical aspects of release of the code and metadata documenting the build system. The project management team will be responsible for oversight of the release and licensing of the codebase.

Data Access to repository content is provided through a web interface with basic and advanced (spatiotemporal) search capabilities. Data published through SKOPE will be open access and accompanied with DOIs or other persistent identifiers whenever available and practicable. All data downloads will include appropriate citation information. Transitive credit and attribution will be supported through reporting tools based on data provenance queries.

Technical Challenges and Solutions. The nature of SKOPE includes unique technical challenges to data management, which will be solved as part of the main development effort (see Project Description and Technical Plan). For example, custom model runs produce results that are not immediately in an official data repository. These data should be available through the SKOPE web application for visualization and as input to additional model runs. If models are run outside of the initial application context (ROGER), how can models access the data cache? We will develop the necessary solutions as part of the SKOPE core development, and will leverage existing technologies wherever possible and practical. For example, tDAR and NCEI/NOAA already are or will shortly become DataONE member nodes. We plan to leverage DataONE approaches (e.g., for data access, sharing, provenance recording) whenever practical, and set up a SKOPE member node at NCSA to simplify data sharing across the DataONE federation. Data replication services (for redundancy or efficiency) are also available through DataONE and through iRODS, as used by other projects at NCSA.

Technical Plan

Synergy of Engineering and Science Efforts. We envision the project comprising two parallel, strongly unified sets of activities: (1) the development and deployment of the SKOPE web application, model execution services, and data and provenance management facilities; and (2) the incorporation of paleoclimate data sets, building and adaptation of paleoclimate reconstruction models, and the application of these data and models to scientific problems via the SKOPE application. These engineering and scientific efforts will proceed concurrently, with the science requirements and users' feedback informing and driving the engineering schedule. The aim will be for each SKOPE system feature to serve a real and immediate need at the time it is developed, so that it is evaluated and thoroughly exercised by researchers investigating actual research questions.

Project Planning and Communication. Because project activities will occur at a rapid pace in multiple physical locations, frequent effective communication will be critical. Working together over the last two years, our experience demonstrates both that biweekly project steering teleconferences (using WebEx) are effective in maintaining project momentum and that periodic face-to-face meetings are essential for strategic brainstorming and making key decisions. In addition to the biweekly teleconferences (monthly in Year 3), five two-day face-to-face meetings will be spread over the three years of the project. The engineering team will hold at least one additional meeting each week to plan work and solve technical problems. During each project steering teleconference the engineering team will demonstrate the latest SKOPE application features and gather feedback on the features just implemented as well as on those planned for development over the next two weeks. This approach will enable the entire project team to steer the engineering and to maximize the effectiveness of the system as it is developed. As with the BCC grant, Kintigh at ASU will provide overall leadership and coordination of the project efforts.

Engineering Process. The SKOPE engineering effort will be centered at UIUC, led overall by Ludäscher and managed on a day-to-day basis by McPhillips. The engineering team will include McPhillips, Wang, Rush, and an additional full-time NCSA software developer, with assistance from Brin. The team will employ an engineering process, developed as part of the Kurator project, that accommodates geographically distributed team members while maintaining focus on shared development goals. In support of this agile process, the SKOPE team will use the Jira[1] issue tracking and project management system running on the NCSA Open Source servers[2] to maintain the backlog of user stories, features, bugs, and tasks. The team will then address the most pressing subset of this backlog during successive two-week development sprints. A single, integrated backlog of development, deployment, and user support issues will facilitate a DevOps approach to integrating all aspects of planning, implementing, running, and troubleshooting the SKOPE application and the underlying infrastructure. The team will automate building, testing, and packaging of all software components using the Bamboo[3] system running on NCSA servers and using free, hosted continuous integration systems such as Travis CI[4]. We will provide public access to the backlog and to all build and test results. All source code and other artifacts developed in the project will be hosted publicly in git repositories and will be made freely available for unrestricted use by others via an Apache 2.0 or similar open source license.

Science Process. In parallel with the engineering effort, Kohler will lead the paleoenvironmental reconstruction model development at WSU, working with Bocinsky at the CCAC. Building the new models and migrating the existing models and datasets (see Section 5, Project Description) will involve several domain-specialist consultants: ***Peter Brewer*** (Ph.D. Reading 2003; Laboratory of Tree-Ring Research (LTRR), Arizona: Tree-Ring Data (archaeological and climatic); ***Bradley Butterfield*** (Ph.D.

[1] https://www.atlassian.com/software/jira

[2] https://opensource.ncsa.illinois.edu

[3] https://opensource.ncsa.illinois.edu/bamboo/

[4] https://travis-ci.org/

ASU 2009 Plant Biology; Northern Arizona University: Staple Wild Plant Habitat Suitability); ***Eric C. Grimm*** (Ph.D. Minnesota 1981; Earth Sciences: Neotoma and Pollen Data Usage and Expansion); ***John W. (Jack) Williams*** (Ph.D. Brown 1999; Geography; Neotoma and Pollen Data Usage. Letters of Collaboration provided in Supplemental Materials). This project will also support a modest data entry effort at the LTRR to increase the number of archaeological tree-ring dates available in the database (coordinated by Brewer) and a parallel effort, coordinated by Grimm, to increase the number of dated pollen cores, focusing on regions with poor representation in the NAPD [Neo16b].

Student Training: The project will support one WSU graduate student. The student will participate in all project meetings as part of their duties and training. Under Kohler's mentorship, the student will also work closely with Bocinsky, Grimm, and Williams on developing and applying the biomisation model. The WSU RA will also closely collaborate with Brewer on inputting archaeological tree-ring dates without compromising exact site locations, and on illustrating productive uses for the corpus.

Development and Release Schedule. The parallel engineering and scientific development activities will be concentrated in the first two years of the three-year funding period, culminating in the release of a feature-complete SKOPE deployment at the end of Year 2. We will deploy working features of SKOPE incrementally throughout the first two years, enabling researchers to make productive use of each iteration of SKOPE's application, data, and models. Feedback from the researcher community—including feature requests, bug reports, technical support rendered, and our own observations of how users actually interact with the system—will be gathered by both the engineering and the science teams and used to direct the ongoing development effort. During Year 3, team efforts will be devoted to aggressively ramping up use of the system, maintaining the tool in production, and enhancing the system to support long-term sustainability. A schedule of key project milestones is presented in Table 2.

Table 2. SKOPE Milestones by Quarter. Engineering milestones: plain text. *Science milestones: italics.*

Y1 Q1	Continuous delivery infrastructure deployed. Prototype ported to publically accessible production web environment. PaleoCAR runnable from web environment for project members and collaborators.
Y1 Q2	Data registry and cache operational. Model execution service binds model parameters, inputs, and outputs named and typed using YesWorkflow annotations. Desktop release of YesWorkflow for researchers.
Y1 Q3	Web GIS tools implemented. YesWorkflow visualizations of models and prospective provenance queries through web application. *Precip/temp/crop-niche models extended.*
Y1 Q4	Datasets for SKOPE-facilitated access (see Table 1) registered. SKOPE site open for user registrations. YesWorkflow imports provenance from recordr, MATLAB DataONE Toolbox, and noWorkflow.
Y2 Q1	Video rendering service supporting animated maps. Reconstructions, visualizations, and queries of retrospective provenance. *Staple Wild Plant Habitat Suitability, Risk Landscapes developed/implemented.*
Y2 Q2	Dataset provenance can be downloaded and imported into YesWorkflow. Maize *paleoproductivity model developed/implemented*.
Y2 Q3	Registration of SKOPE-generated Retrodictions with Tuneable Models (see Table 1). Tinkerers can perform customized model runs. *Biome reconstructions developed/implemented.*
Y2 Q4	Downloadable containers to run models on external resources. Users may register new models.
Y3	Outreach, training, sustainability/hardening; limited new features in response to user feedback.

Dissemination. Ongoing SKOPE feature availability will be communicated to our multidisciplinary user community through our current website [SKOPE16a], to be continuously enhanced beginning in Q2 of Y1. In early 2018, we will publicize SKOPE's release through notices, blogs, and short technical articles in newsletters and periodicals, related projects, and participating organizations. We will offer online training materials and webinars that we expect to be hosted by DataONE and SAA. In addition, project participants will offer professional conference presentations on the project. We will track web metrics on page-views, downloads, and model runs, e.g., using Google Analytics. Source code downloads and forks will be tracked using GitHub's repository tracking systems.

Sustainability Plan

SKOPE is being designed to have a light sustainability footprint—a core strategy is to minimize what it is that we have to sustain. SKOPE need *not* serve as a long-term data repository as source datasets will be obtained only from federal agencies or other stable data repositories such as NCEI/NOAA, tDAR. directly, or via a data federation such as DataONE. For software, SKOPE will use an externally-hosted distributed version control system shared via a public repository. SKOPE's design strives to address three key dimensions of sustainability: social (building a strong user base), technological (computational resources, software maintenance, and adaptability to evolving computational environments), and economic (revenue sources).

Social Sustainability. SKOPE, by way of its prototype, begins with an engaged community of scholars that it plans to grow through the course of the project. During our BCC design grant, we undertook two needs-assessment workshops to explore and prioritize potential capabilities. We also involved key data providers in team planning meetings at ASU and SFI. A first SKOPE prototype is already available; the first release of SKOPE will be halfway through grant Year 1; and a feature-complete version will be released by the end of Year 2. Year 3 is devoted to outreach, training, user-driven improvements and sustainability. Throughout, we will draw on an engaged community of users for feedback to improve the system. SKOPE should not require user support beyond the project-produced training and documentation.

Technological Sustainability. All software produced by the project will be made available under open source licenses, allowing it to be freely used, modified, and shared by the user community. The data and model registries, software container definitions, and the rest of the SKOPE system's source code will be versioned and made publicly-accessible through Git repositories. Maintenance time of SKOPE is minimized by employing a sustainability-friendly design from Day 1, which is a direct result of our design for reproducible research (Section 6.3). By deploying the SKOPE functionalities in software containers, the system as a whole may be migrated to other infrastructures if needed, and users will be able to download SKOPE containers and run them on their desktops or in a cloud. SKOPE will initially be deployed on NCSA's NSF-supported ROGER HPC cluster. Resources have been budgeted in Year 3 for our team to work with NCSA's Innovative Technology Services (ITS) group to ensure that SKOPE software is secure and ready to migrate to continuing infrastructure. NCSA commits to run SKOPE for at least 2 years beyond the end of the project as part of this infrastructure.

Economic Sustainability. SKOPE will provide substantial research value to diverse communities of users and contributors concerned with paleoenvironments and human-environment interactions. Demonstrated user demand and research impacts should enhance our ability in Year 3 to obtain long-term support for sustainability through federal agencies or private foundations. Ideally, SKOPE use will remain free. Absent securing continued grant funding, we will consider alternative strategies while we are still funded under this grant. Alternative long-term options may include a model where data searching, visualizing, and downloading is free, but users pay for (e.g., cloud-based) computing resources required to run large reconstructions. SKOPE might be adopted by an agency or long-term project that recognizes its value. Or, we could adopt the business model of some disciplinary repositories, in which fees would be charged for registering new models and datasets with the tool, but user access is free.

Operational Cost Estimate. The ITS group at NCSA runs numerous services as continuing infrastructure, e.g., web servers, databases, version control, Kerberos, etc. ITS can run appropriately designed and hardened production software for at least two years at negligible cost. By investing in security hardening in the third year of funding, SKOPE can then be expected to run for two more years without additional funding. To continue the same system *beyond* that time, we expect another investment of a month of security review from ITS and up to a month of time from a member of the project team would prepare SKOPE to run for two more years. Based on current salary and benefit rates, this brings the average total annual cost after Year 5 to $7,500 (2 person-weeks/year for ITS alone) or to $15,000 (for 4 person-weeks/year for ITS & project updates, if the project-developed software keeps changing). There are no additional costs for running SKOPE using ROGER or continuing infrastructure at NCSA.

TPI 7670128