Interdisciplinary Research (IDR) Origination Awards

Cover Page

Project Title

Title: An open-access proxy database for enhanced paleoclimate research and climate model evaluation

Principal Investigator(s) (full-time faculty)

Name (PI listed first)	Department	College
Landon Burgener	Geological Sciences	Computational, Mathematical, & Physical Sciences
Albert Tay	Electrical & Computer Engineering	Engineering
Michael Christensen	Statistics	Computational, Mathematical, & Physical Sciences

Track

Track one

Abstract

Quantitative comparisons between paleotemperature and paleoprecipitation reconstructions derived from geological proxies are critical for assessing climate sensitivity and for evaluating the accuracy of climate model simulations of future anthropogenic climate change. Past comparisons have been hindered by a lack of quantitative, gridded, proxy-derived datasets that are directly comparable to climate model output. The immediate aims of this proposed project are to address this lack of standardized paleoclimate datasets by 1) develop a paleoclimate proxy repository that adheres to data formatting and quality-control protocols, and can incorporate data from the entire Phanerozoic (540 Ma to present); 2) create an open-access web portal that facilitates broad use of, and contribution to, the repository; and 3) demonstrate database utility by generating and analyzing new quantitative, global paleoclimate maps for the Paleocene-Eocene Thermal Maximum (PETM) event. Pilot testing of this database has already begun, including publication of an invited peer-reviewed article featuring a small prototype database; however further refinement and development of the database is required, including the creation of a quality-control process for data entry; upgrading the database to a standard format; and creation of a public-facing website for user access. This interdisciplinary project involves faculty with expertise in geology, database development, and Bayesian geostatistics. Award funds will be used to support undergraduate and graduate students in Geological Sciences, Statistics, and Electrical and Computer Engineering, and lead to four external funding applications and an anticipated five peer-reviewed publications.

Summary of Plans for External Funding

Funding Source	Grant Name	Submission Goal
National Science Foundation, EAR	Life and Environments Through Time (LET	Winter, 2026
National Geographic Society	Technology Level II Grant	Summer, 2026
National Science Foundation, EAR	Geoinformatics (GI)	Fall, 2026
National Science Foundation, EAR	Geosciences Open Science Ecosystem (GEO OSE)	Winter, 2027

PROJECT NARRATIVE

1. OVERVIEW AND OBJECTIVES

1.1. An New Paleoclimate Proxy Repository For Improved Paleoclimate Reconstructions

Characterizing the behavior of Earth's climate system during hot greenhouse periods is essential both for understanding how the system has responded to past global climate forcings and for predicting how climate will evolve in the future due to human-induced climate change. Paleoclimate conditions during ancient greenhouse phases can be reconstructed from climate-sensitive minerals or fossils, which are referred to as paleoclimate "proxies", or can be simulated using climate models. Proxy-based paleoclimate reconstructions are particularly important because they provide *direct* evidence of ancient climate, and they can be used to test the accuracy of climate models. Unfortunately, quantitative assessments of proxy data remain constrained because they are often semi-quantitative and spatially and temporally discontinuous. Importantly, the lack of quantitative proxy datasets available within a centralized proxy repository hinders paleoclimate analyses and comparisons between proxy reconstructions and climate model outputs, leading to decreased confidence in model predictions of future climate change and increased uncertainty for policymakers and societies attempting to mitigate human drivers of global warming. *The solution is a standardized repository infrastructure that ensures paleoclimate proxy data are easy to access and analyze using modern geostatistical techniques.* Without such a resource, paleoclimate data will remain isolated and disjointed, hindering future paleoclimate sensitivity analyses, validation of climate models, and refinement of future climate change predictions.

Our *long-term goal* is to use paleoclimate proxy data to answer critical questions regarding Earth's past and future climate conditions. Key field-changing advances that will be enabled by construction of this broad database will include (1) improved estimates of climate sensitivity (Royer, 2016), (2) improved quantification of spatial climate change during major climate transitions (Tierney et al., 2022), (3) refined estimates of the impact of temporal and spatial climate variations on the evolution and extinction of major floral and faunal groups (Burgener et al., 2021), and (4) improved quantitative comparisons of proxy and model paleoclimate estimates (Burgener et al., 2023). To facilitate these future research directions, *our objective here is to develop the digital infrastructure for, and begin populating, an open-access, web-based proxy repository and spatially interpolated paleoclimate datasets that can be quantitatively analyzed and compared with climate model simulations*. To achieve this, we will pursue three specific aims. Aim 1: create the digital infrastructure for a paleoclimate proxy repository that adheres to data formatting and quality-control protocols, and populate the new repository with the preliminary proxy records (n > 50,000) already compiled by PI Burgener and his students. Aim 2: create an open-access web portal that facilitates broad use of the repository and enables future growth of the repository from internal and external data contributions. Aim 3: demonstrate the repository's utility by adding new Paleogene (66 to 35 million years ago) proxy records to the repository and generating/analyzing new quantitative, global paleoclimate maps for the Paleocene-Eocene Thermal Maximum (PETM) event (55 million years ago). Work proposed as part of the IDR will begin population of the database with relevant and needed data, and will continue over many years as we continue to expand and populate the database.

Expected significant scientific impacts resulting from this project include 1) a publicly accessible paleoclimate proxy database and website with a *preliminary* sample size of >50,000 records and a defined process for future growth of the repository; 2) an improved framework for integrating proxy data with climate models; and 3) new quantitative reconstructions and analyses of spatial paleoclimate patterns during the PETM. *Establishing this repository is significant because it will provide the necessary infrastructure to systematically analyze past climate variability at a variety of temporal and spatial scales, and refine future climate change predictions.* The repository will also fulfill a clear need in the paleoclimate community. Previous attempts have been made to create this sort of repository (cf., PALEOMAP Project), but failed for a variety of reasons (e.g., older technology, PI career changes, etc.). Now is the time for BYU faculty to fill this important niche. This project is an ideal candidate for support via the IDR Origination program because its successful completion requires the combination of disparate skills possessed by the three PIs collectively. Additionally, support through the IDR Origination program will lead to applications to four external opportunities, and at least five peer-reviewed publications that will include both graduate and undergraduate co-authors.

1.2. Interdisciplinary Team

The PIs are experts in paleoclimatology, database development and spatial statistics, and well-prepared to accomplish this work. PI Burgener, an expert in proxy-based paleoclimate reconstructions, will lead the project and contribute to Aim 1 by overseeing the transfer of existing records from his lab's databases to the new proxy repository and guiding the overall design of the repository. He will contribute to the conceptual design of the repository website in Aim 2, and identify key functionalities that users will need to select, manipulate and download data. He will contribute to Aim 3 by guiding the development of new geostatistical methods for generating the quantitative paleoclimate maps from the repository data, and will lead the PETM case study. Co-PI Tay, an expert in database construction and management, will contribute to Aim 1 by designing the new repository and ensuring that it meets modern formatting and security standards. Additionally, he will produce a new quality-control workflow to ensure that current and future data uploads are error-free. Co-PI Tay will also contribute to Aim 2 by directing the construction of a website for public access to the repository. Co-PI Christensen is an expert in Bayesian and spatial statistical techniques, and will contribute to Aim 2 via the design of website tools that will allow users to perform simple initial analyses or transformations on subsets of the repository prior to download. Co-PI Christensen will also contribute to Aim 3 by updating and improving the geostatistical techniques pioneered in Burgener et al. (2023) to estimate climate values between known sample locations, and contributing to the creation and analysis of the PETM paleoclimate maps.

2. PRELIMINARY STUDIES

PI Burgener previously developed a preliminary paleoclimate proxy dataset for the Cretaceous Period (145 to 66 million years ago) that incorporated paleotemperature and paleoprecipitation reconstructions from diverse proxy sources. These data included both quantitative (e.g., stable isotope-derived temperature estimates) and semi-quantitative paleoclimate proxies (e.g., climate-sensitive lithologies). The combination of these two proxy types in this small early dataset allowed the former research team to produce the first statistically robust, high-resolution global climate maps for nine key Cretaceous time slices. Using these paleoclimate maps, they conducted quantitative analyses to test long-standing hypotheses regarding Cretaceous climate patterns. They also performed a quantitative comparison between the proxy-based paleoclimate maps and climate model simulations (Valdes et al., 2021). This analysis revealed key areas of agreement between the model and proxy data, and identified regions where model-data discrepancies highlight gaps in our understanding of the climate system.

These previous methods and results were published in an invited, peer-reviewed article (Burgener et al., 2023). Importantly, this publication demonstrate the feasibility and utility of integrating diverse proxy datasets with advanced geostatistical techniques. However, fully realizing the potential of this approach and expanding its temporal application beyond the Cretaceous requires both an interdisciplinary team with expertise in the construction of centralized online-repositories and a website that allows public access to the data. Additionally, the geostatistical techniques that were pioneered in the original study by Burgener et al. (2023) require further development and refinement in order to produce climate maps at higher temporal resolutions. This next phase of research will ensure that global paleoclimate data are standardized, easily accessible, and fully integrated with climate modeling efforts, thus fostering new advances in paleoclimate science.

3. RESEARCH PLAN

3.1. Aim 1: Development and Initial Population of a Standardized Paleoclimate Proxy Database

3.1.1. Introduction. Proxy-based paleoclimate research is hindered by challenges in accessing, standardizing, and integrating existing data. A vast number of valuable records remain trapped in non-digitized sources (Bárdossy, 1982), or are only available in older non-standardized digital formats that impede access (Tabor et al., 2016), or are no longer accessible. Another challenge arises from the diverse nature of paleoclimate proxies. These data fall into two categories: fully quantitative geochemical proxies and semi-quantitative lithologic proxies. Historically, these two classes of data have not been directly comparable, resulting in paleoclimate studies that made use of either one data type or the other (e.g., Boucot et al., 2013). Without a standardized repository that facilitates both the integration of quantitative and semi-quantitative proxy data and methodological consistency, the field remains constrained by fragmented datasets and inconsistent analytical approaches, a situation that limits our ability to conduct robust, high-resolution paleoclimate reconstructions. By building on the Cretaceous paleoclimate proxy database developed by Burgener et al. (2023), Aim 1 directly addresses this issue through the creation of a centralized, open-access paleoclimate proxy repository.

3.1.2. Design. The paleoclimate proxy repository will be built as a *relational database* using *MySQL*, and will be designed to efficiently store, query, and manage diverse paleoclimate proxy records. The repository will utilize a *structured schema* with tables for sample metadata (e.g., location, age, lithology), proxy paleoclimate reconstructions (e.g., mean annual temperature, precipitation, etc.), data sources (e.g., publications, datasets), and additional contextual information.

Database normalization, as proposed for use here, is a systematic approach to structuring a relational database in accordance with a series of normal forms. Each normal form builds upon the previous one, reducing data anomalies such as insertion, update, and deletion inconsistencies. The project can optimize normalized relational data storage, enhance performance, ensure reliable data retrieval and adhere to *FAIR data principles* (Findability, Accessibility, Interoperability, and Reusability) by incorporating standardized vocabularies, controlled terminologies, and persistent identifiers for samples and references. A secure, *RESTful API* will facilitate data retrieval and submission, ensuring compatibility with external databases and analytical tools. Data integrity will be maintained through *automated validation scripts* to flag inconsistencies, along with a *version control system* to track updates.

The repository will support data uploads in standard formats (e.g., csv, xlsx, etc.). To ensure data accuracy and consistency, submitted datasets will undergo a quality control process that includes automated validation checks for formatting errors, missing metadata, and a final review by domain-specific experts who will assess proxy measurement validity, metadata completeness, and adherence to repository standards. Scalability and performance will be prioritized through *optimized indexing and query caching*.

3.1.3. Expected Outcomes and Measures of Success. **Successful completion of Aim 1 will produce a digital repository for storing standardized paleoclimate proxy data, and an initial population of >50,000 records.** This repository will provide a centralized, standardized resource for researchers to access, analyze, and store proxy-based paleoclimate reconstructions spanning the Phanerozoic. This repository will enhance data integration, facilitate large-scale paleoclimate reconstructions, and improve methodological consistency, ultimately advancing our ability to compare proxy data with climate model simulations. The creation of the paleoclimate proxy repository will serve as the main measure of success of Aim 1.

3.2. Aim 2: Develop a Public Portal for Paleoclimate Repository Access and Community Contribution

3.2.1. Introduction. Paleoclimate proxy data remain difficult to access due to outdated storage formats, non-digitized records, and fragmented datasets. Even when available online, many datasets lack standardization or long-term maintenance. Aim 2 directly addresses these challenges by connecting the paleoclimate repository created in Aim 1 to a public-facing website that will be designed to allow users to access, view, and download all—or a specified subset—of the paleoclimate repository, and to submit their own data for addition to the repository. The resulting website and repository will act as a research catalyst, facilitating data discovery, integration, and community-driven contributions and have a fundamental impact on future paleoclimate and climate change discoveries.

3.2.2. Design. The open-access website for the paleoclimate proxy repository will be designed to facilitate data discovery, visualization, and access, similar in structure and functionality to important online geoscience repositories like the Paleobiology Database or EarthChem. The website will be built using a combination of a *frontend framework* (such as React or Vue.js) for user interaction and a *backend system* (such as Django or Node.js) to manage data retrieval and user requests. The repository will be stored in a *relational database* (e.g., MySQL) with an API layer to allow efficient querying and data extraction. The website will include an *interactive map interface* that enables users to explore proxy records by location, age, and proxy type. A *search and filtering system* will allow users to refine datasets based on metadata criteria. The repository will support *downloadable data files* in widely used formats such as CSV or vector shapefiles to accommodate different research needs. Additionally, *integrated analysis tools* will allow users to perform basic statistical analyses and paleo-coordinate calculations. Security and data integrity will be ensured through user authentication for data contributors, version-controlled updates, and adherence to modern *FAIR data principles*. The website's architecture will prioritize scalability, ensuring long-term accessibility and expansion as new data and functionalities are added. The paleoclimate proxy repository and associated website will be hosted on existing hardware in the Geology Department server room. Maintenance and support will be provided by the Geology IT office.

3.2.3. Expected Outcomes and Measures of Success. **Successful completion of Aim 2 will result in a user-friendly web portal providing seamless access to the paleoclimate proxy repository.** This website will enable researchers to search, visualize, and download data. The platform will foster community contributions, enhance interdisciplinary collaboration, and support real-time data integration, ultimately improving the accessibility and usability of paleoclimate data for research and education. The creation of the website and linking it to the paleoclimate proxy repository will serve as the main measure of success for Aim 2.

3.3. Aim 3: Application of the Paleoclimate Proxy Repository - Global Patterns of Climate Change Across the Paleocene-Eocene Thermal Maximum (PETM)

3.3.1. Introduction. Demonstrating the utility of our open-access paleoclimate proxy repository is essential for validating its methodology and encouraging widespread adoption by the scientific community. Aim 3 directly addresses this need by generating and analyzing new quantitative, global paleoclimate maps for the Paleocene-Eocene Thermal Maximum (PETM). These new paleoclimate maps will showcase the repository's ability to integrate diverse proxy records, produce high-resolution climate reconstructions, quantify spatial and temporal paleoclimate patterns and trends, and facilitate direct proxy-model comparisons. These results will establish the repository as a valuable tool for future paleoclimate research, generate a first round of peer-reviewed publications and act as solid preliminary work for future external grants, as described in more detail in Section 4.

3.3.2. Design. Aim 3 will build on the methods of Burgener et al. (2023), with a focus on the periods directly before, during and following the PETM. The paleoclimate proxy data extracted from the new repository will be analyzed using a Bayesian spatio-temporal framework that will allow us to jointly model and interpolate mean annual temperature (MAT), warmest month mean temperature (WMMT) and mean annual precipitation (MAP) fields globally at variable spatial and temporal resolutions, and with accompanying estimates of uncertainty. Climate parameters will be modelled using a multivariate spatio-temporal Gaussian process (or an approximation thereof) with a nonstationary spatial dependence structure to account for interactions between earth's physical geography and the parameters being modeled, as well as nonstationary temporal dependence which will allow the model to produce reliable estimates of model parameters at finer temporal resolutions for the recent past (<100,000 years ago) and at progressively coarser resolutions during earlier periods. The model will be structured to allow for nonlinear and temporally evolving associations between climate parameters and covariates such as latitude and elevation. Semi-quantitative data will be accommodated using a censored likelihood approach within a Markov Chain Monte Carlo framework, and other sources of uncertainty, such as those coming from the variable reliability of different sources of proxy data or dating techniques will be accounted for using methods similar to those found in White et al. (2019) and Zhou and Hanson (2018). Finally, estimated climate fields will be categorized into Paleo-Köppen classifications to infer regional climate patterns across space and time.