Opening the Black Box of a Paleoclimate Reconstruction based on PaleoCAR

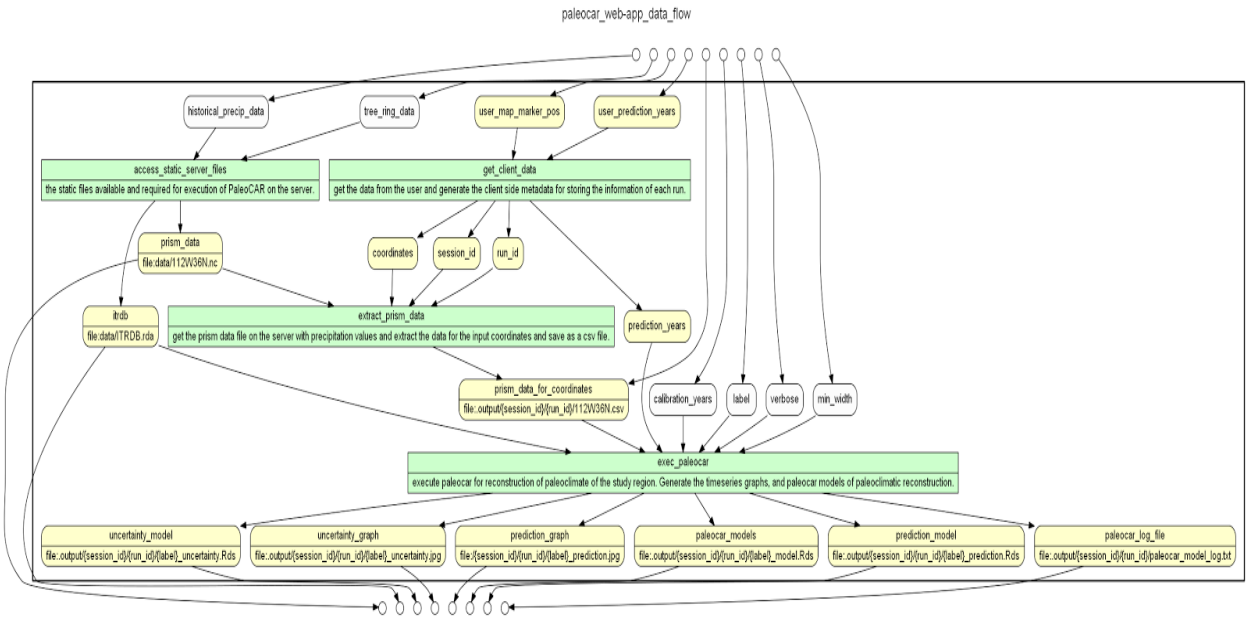
Pratik Shrivastava[1], Timothy McPhillips[1], Kyle Bocinsky[2], Bertram Ludäscher[1]
[1]University of Illinois Urbana-Champaign, [2]Washington State University
pratiks2@illinois.edu, tmcphillips@absoluteflow.org, bocinsky@gmail.com, ludaesch@illinois.edu

Abstract

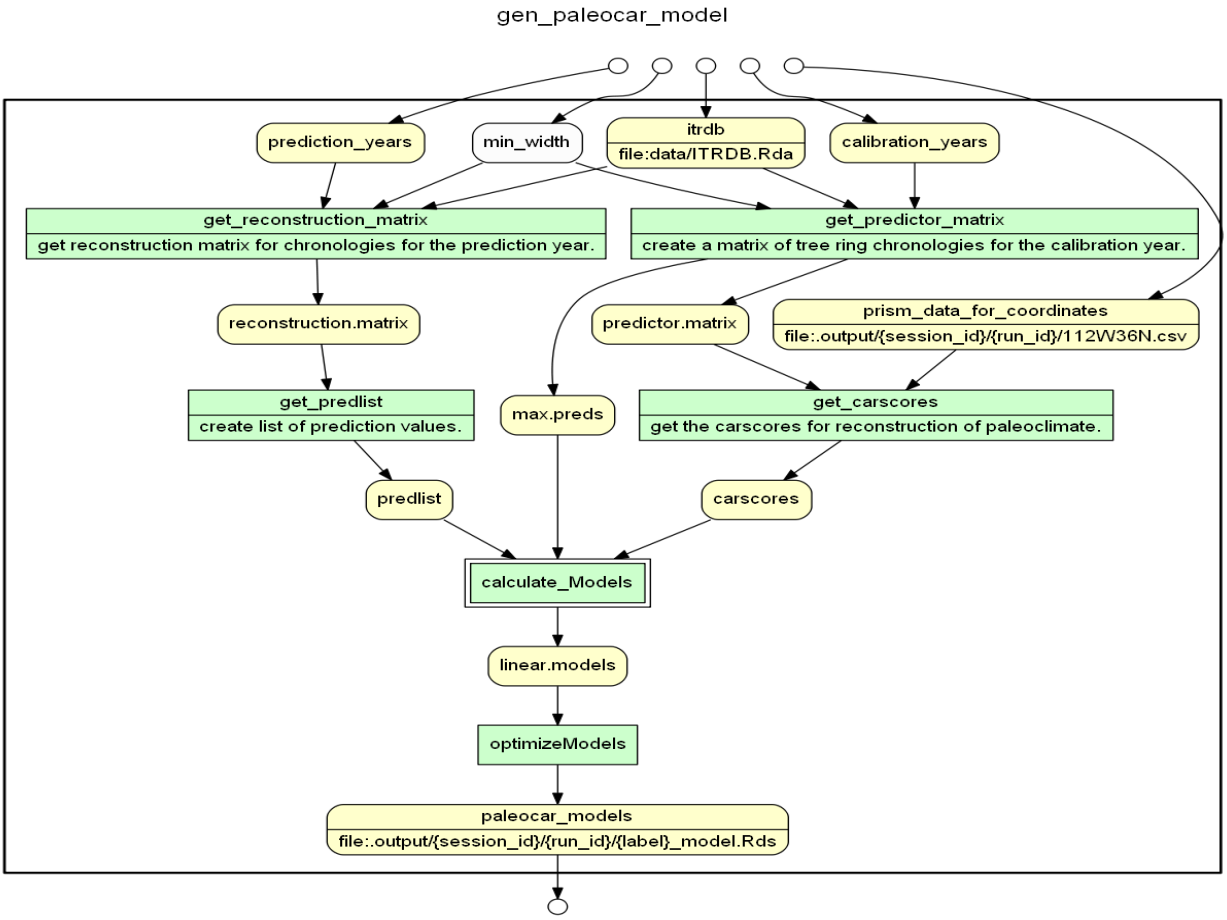
In a scientific study, most of the valuable artifacts and datasets which are used to produce the end results get abstracted in software-based scientific methods. The software comprising a scientific study or method often are black boxes. The web applications can simplify tool usage but may further obfuscate the workings of the underlying software. The data dependencies and the relationship between parameters and the code blocks are not exposed. The lack of provenance information for the data artifacts used and generated in a scientific study or method can raise significant questions about authenticity and reproducibility of the study or method. If adequate provenance information is present it helps in reproducing the research and taking it further forward.

The goal of our project is to represent the provenance information of all the results produced in a scientific study, and we used YesWorkflow to achieve it. Here, we document a paleoclimate reconstruction library for R — called PaleoCAR — using the YesWorkflow system. The YesWorkflow is a software that provides a number of benefits of using a scientific workflow management system without having to rewrite scripts and other scientific software. The users declare scientifically significant steps and reveal data dependencies in the workflow via YesWorkflow annotations, by embedding them directly into scripts or scientific software. The resulting YesWorkflow models (prospective provenance) are then rendered as workflow graph and datalog facts. The prospective provenance then can be linked to the runtime observables, providing the cross-validation and checking opportunities.

The YesWorkflow tool provided the prospective provenance information for the study of PaleoCAR, and it can be linked with runtime observables. Below are the provenance graphs generated for the PaleoCAR. The YesWorkflow graph helped in identification of the pre-requisite datasets and the parameters required for execution of PaleoCAR. The YesWorkflow model helps in answering the interesting questions such as *Which data results that are directly influenced by the input year range? How were the data sets used in every run of the application acquired or (pre)computed with*? The prospective provenance information of the pre-requisite dataset was also generated.



YesWorkflow graph of PaleoCAR web application.



YesWorkflow graph of PaleoCAR models.

References

[1] T. McPhillips, "YesWorkFlow," 30 March 2015. [Online]. Available: https://github.com/yesworkflow-org/yw.

[2] R. K. Bocinsky, "paleocar," February 2016. [Online]. Available: https://github.com/bocinsky/paleocar#paleocar.

[3] Bocinsky R Kyle, Kohler A. Timothy, "A 2,000-year reconstruction of the rain-fed maize agricultural niche in the US Southwest," *Nature Communications,* no. 5618, 21 October 2014.