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

School of Information Sciences
The iSchool at Illinois

Challenges

- ❑ Software comprising a **scientific** study or method often are **black boxes**.
- ❑ Web applications can simplify tool **usage** but may further **obfuscate** the workings of the underlying software.
- ❑ Information about prerequisite, intermediate, and the result dataset remains screened.
- ❑ The information about overall dataflow between code blocks also remains hidden.
- ❑ The relationship between parameters and the code block is not exposed.

Inputs for Web Application

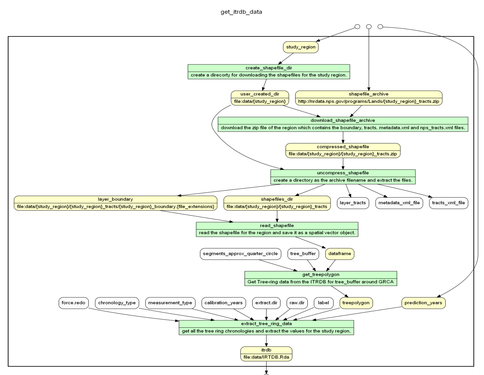
Location coordinates:

Year Range

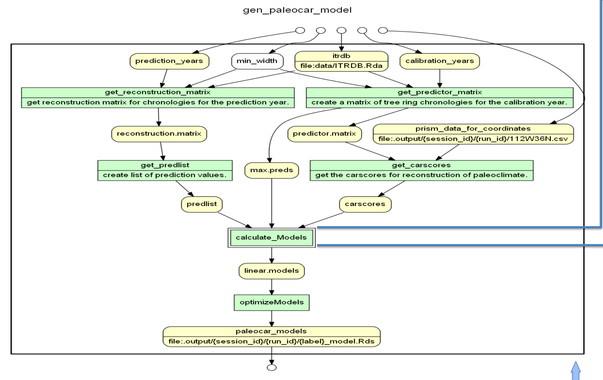
Prediction Years:

1800:2100

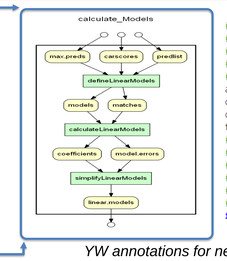
Results of PaleoCAR



YW graph of tree ring data



YW graph of PaeloCAR Models




```
# @BEGIN calculate_Models
# @IN predlist
# @IN carscores
# @in max.preds
# @OUT linear.models
allModels <- data.table::data.table(cell
coefs=numeric())
complete.cell.years <- data.table::data.
times <- vector('numeric',max(prednums))
# @BEGIN defineLinearModels
# @IN predlist
# @IN carscores
# @in max.preds
# @OUT models
# @OUT matches
for(i in 1:maxPreds){
```

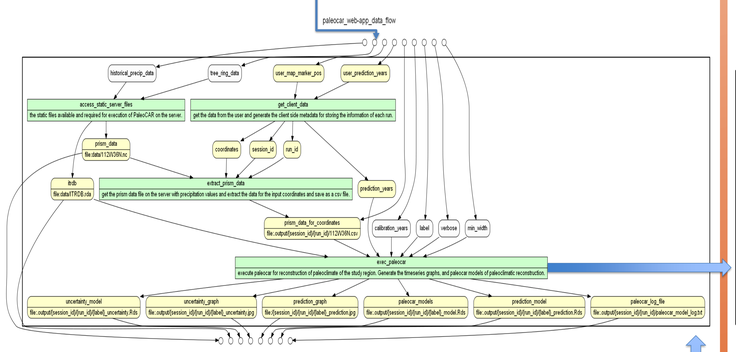
YW annotations for nested Sub-workflow

YesWorkflow (YW)

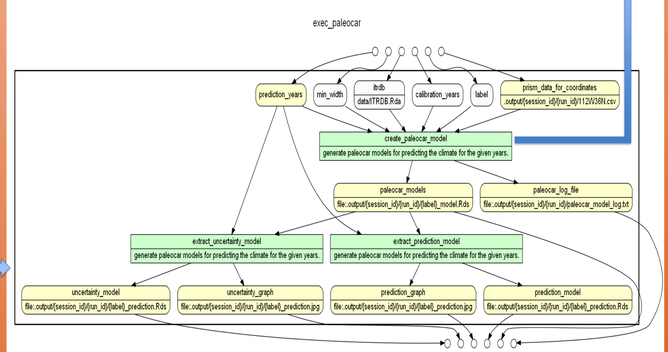
- ❑ **YesWorkflow** helps in uncovering shrouded information from the software-based scientific methods.
- ❑ Users **declare scientifically significant steps** and **reveal data dependencies and dataflow** via YW annotations, typically embedded in script comments.
- ❑ The resulting **YW models** (a.k.a. prospective provenance) can be rendered as a workflow graph, showing **what kinds of provenance graphs can be expected** after execution.
- ❑ The prospective provenance graph can be linked with **retrospective** (runtime) **observables,** providing additional **cross-validation** and checking opportunities: the observed provenance then either corroborates the declared YW model or indicates possible modeling errors.

What is PaleoCAR?

- ❑ **PaleoCAR** implements a correlation-adjusted regression of tree-ring series with 100+ years of contemporary data modeled by PRISM at an 800-m scale to retrodict climatic variables, notably precipitation and temperature over the last 2000 years.
- ❑ PaleoCAR is an **R package**, which consists of the functions that helps users to recreate the spatiotemporal paleoclimate reconstructions.
- ❑ The information generated by PaleoCAR is stored in **R object** (*.rds)



YW graph of PaleoCAR Web Application



YW Graph for exec PaloeCAR block

```
// @begin exec_paleocar @desc execute paleocar for reconstruction of paleoclimate of the study region. Generate the timeseries
graphs, and  paleocar models of paleoclimatic reconstruction.
// @in prediction_years @desc period for reconstruction of the paleoclimate using paleocar.
// @in data @as prism_data_for_coordinates
// @param tree_ring @as itrdb
// @param calib_year @as calibration_years  @desc period for calibrating the information for predicting the climate.
// @param label @desc user entered label for the study region.
// @param min_width
// @param verbose


// @out pred_model @as prediction_models @uri file:.output/{session_id}/{run_id}/{label}_prediction.Rds @desc  R model of the
paleocar reconstruction of prediction.
// @out pred_plot @as prediction_graph  @uri file:/{session_id}/{run_id}/{label}_prediction.jpg  @desc timeseries plot of
prediction model of the paleocar reconstruction.
// @out uncertain_model @as uncertainty_model @uri file:.output/{session_id}/{run_id}/{label}_uncertainty.Rds  @desc R model of
the paleocar reconstruction of uncertainties.
// @out pal_model @as paleocar_models   @uri file:.output/{session_id}/{run_id}/{label}_model.Rds  @desc R model generated for
the paleoclimatic reconstruction.
// @out uncertain_plot @as uncertainty_graph  @uri file:.output/{session_id}/{run_id}/{label}_uncertainty.jpg  @desc timeseries
plot of uncertainty model of the paleocar reconstruction.
// @out log_file @as paleocar_log_file @uri file:.output/{session_id}/{run_id}/paleocar_model_log.txt  @desc timeseries plot of
uncertainty model of the paleocar reconstruction.
// @end exec_paleocar

// Execute  PaleoCAr for the Vector region for now.
var cmd_exe_paleocar = 'Rscript  '+ curr_dir + 'Rscript\\exec_paleocar.R ' + curr_dir + ' ' +
                       test_dir + ' ' +  out_file_prism_data + ' ' + user_label + ' ' +  calibration_yearss + ' ' +
                       prediction_years.get() + ' ' + '1' + ' ' + "n" ;

Meteor.call('exec_Rscript',cmd_exe_paleocar,function(error, result)
  {
    if (error)
    {
```

Scripts with YW annotations

Approach

- ❑ Built a new **web application** for running PaleoCAR.
- ❑ Users can execute PaleoCAR for a **single location of GRCA region** and reconstruct the paleoclimate for the user entered **year range**.
- ❑ **YW annotations** are embedded in the web application file and in the PaleoCAR to expose the information of the data used and produced while reconstruction of the paleoclimate.
- ❑ The **YW graphs** are integrated with the web application.
- ❑ The data artifacts generated during the run are exposed to the user which can be **compared** with the YW graphs for better assessment and understanding.
- ❑ Creation of **datalog facts** from the YW model, for querying prospective and retrospective provenance information.
- ❑ Creation of the retrospective provenance information such as the tree-ring chronologies or species of trees used for reconstruction of the paleoclimate using PaleoCAR.

Interesting Questions that YW graphs helps to answer.

- ❑ The data results that are directly influenced by the input year range.
- ❑ The data used by application for every run.
- ❑ Which parameters were required for each and every run.
- ❑ How were the data sets used in every run of the application acquired or (pre)computed?

Provenance Queries.

EQ3 : What out ports are qualified with URIs?

eq3(uncertainty_model).
eq3(paleocar_models).
eq3(uncertainty_graph).
eq3(prediction_model).
eq3(paleocar_log_file).
eq3(prism_data).
eq3(prism_data_for_coordinates).
eq3(itrdb).
eq3(prediction_graph).

Findings & Future Work:

- ❑ The web application YesWorkflow graph tallies with working of the web application which integrates PaleoCAR.
- ❑ YesWorkflow graph helped in identification of the pre-requisite dataset and the parameters required for execution of PaleoCAR.
- ❑ The parts which are executed once or multiple times by changing the user input can be easily distinguished.
- ❑ The data dependencies are tracked using graph and provenance queries.
- ❑ The prospective provenance information of the pre-requisite dataset is also generated.
- ❑ YesWorkflow can facilitate querying of the prospective provenance.
- ❑ YesWorkflow can be used to reconstruct retrospective provenance information.
- ❑ Enable YW to extract retrospective provenance from R data files (analogous to log file extraction in YW now).
- ❑ Ability to view the actual code corresponding to a particular script or code block via the web app.

References

- ❑ Bocinsky R Kyle, Kohler A. Timothy. (2014, October 21). A 2,000-year reconstruction of the rain-fed maize agricultural niche in the US Southwest. *Nature Communications*(5618). doi:10.1038/ncomms6618
- ❑ Bocinsky, R. K. (2016, February). *paleocar*. Retrieved from github: https://github.com/bocinsky/paleocar#paleocar
- ❑ McPhillips, T. (2015, March 30). *YesWorkFlow*. Retrieved from GitHub: https://github.com/yesworkflow-org/yw-prototypes/wiki

Github Repository

- ❑ WholeTale Internship 2017 GitHub Repo : https://github.com/idaks/wt-prov-summer-2017

WHOLE TALE

Acknowledgments

Supported by NSF OAC-1541450, SMA-1637155

NSF

ILLINOIS