Environmental Modelling & Software 96 (2017) 172–180

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

A prototype cloud-based reproducible data analysis and visualization platform for outputs of agent-based models

Xiongbing Jin [a, *], Kirsten Robinson [a], Allen Lee [b], J. Gary Polhill [c], Calvin Pritchard [b], Dawn C. Parker [a]

[a] University of Waterloo, Canada
[b] Arizona State University, United States
[c] The James Hutton Institute, United Kingdom

ARTICLE INFO

Article history:
Received 30 November 2016
Received in revised form
12 April 2017
Accepted 15 June 2017

Keywords:
Reproducibility
Agent-based models
Big data

ABSTRACT

Agent-based models typically have stochastic elements and many potential parameter combinations. This requires that we conduct multiple model runs to sweep the parameter space, creating large quantities of computationally generated, hyper-dimensional, "big data". Understanding the models' implications requires structured exploration of these complex output data. In response to this need, the MIRACLE team has developed a prototype web application that enables researchers who archive their model output data and analysis methods to perform online output data exploration and reproducible, re-parameterizable data analysis. We plan to build on this prototype, integrating with broader reproducibility initiatives in scientific computation and big data, to facilitate improved communication within research groups, and increase access and transparency for external research community and the general public. This paper provides contextual background and a case study of the prototype MIRACLE data storage and analysis web tool.

Software availability

Name: MIRACLE
Availability: MIRACLE is available as an open source project https://github.com/comses/miracle
Year: First available 2015
Minimum hardware requirement: Dual-core CPU, 4GB RAM, 10GB storage
Software requirement: Any operating system that supports Docker (version 0.11.2 or later) and Docker-Compose
Cost: Free

1. Introduction

Pioneering work has demonstrated the scientific utility of agent-based models of coupled human-natural systems (ABM-CHANS) (Gimblett, 2001; Kohler, 2000; Lansing and Kremer, 1993); however, development of ABM-CHANS has posed many challenges (Parker et al., 2003). In the last ten years, model communication and transparency have been dramatically improved in the ABM community, based on improved standards for documenting, validating, verifying, and archiving models (Grimm et al., 2010; Janssen et al., 2008; Müller et al., 2014; Rollins et al., 2014).

Yet, even with these advances, substantial challenges for methods, results and inferential reproducibility (Goodman et al., 2016) of ABM-CHANS remain. Reproducibility requires complete access to model assumptions, code, output data, analysis methods, and specific versions of software dependencies for model and analysis code. Model code archive facilities through online platforms such as the Network for Computational Modeling for SocioEcological Science (CoMSES-Net) Model Library (CoMSES-Net, 2016) are increasingly being used by individuals and academic journals. However, in spite of increasing access to model code, most models are run only by the team that developed them. A preliminary CoMSES-Net analysis of 2367 publications selected with the keyword "agent based model" and published between 1990 and 2014 found that only 10% had made their model source code available online (Janssen, 2016). This means lost opportunities to build on existing models to advance research and decision making. While journal requirements for code archiving can increase the

* Corresponding author.
E-mail address: x37jin@uwaterloo.ca (X. Jin).

http://dx.doi.org/10.1016/j.envsoft.2017.06.010
1364-8152/

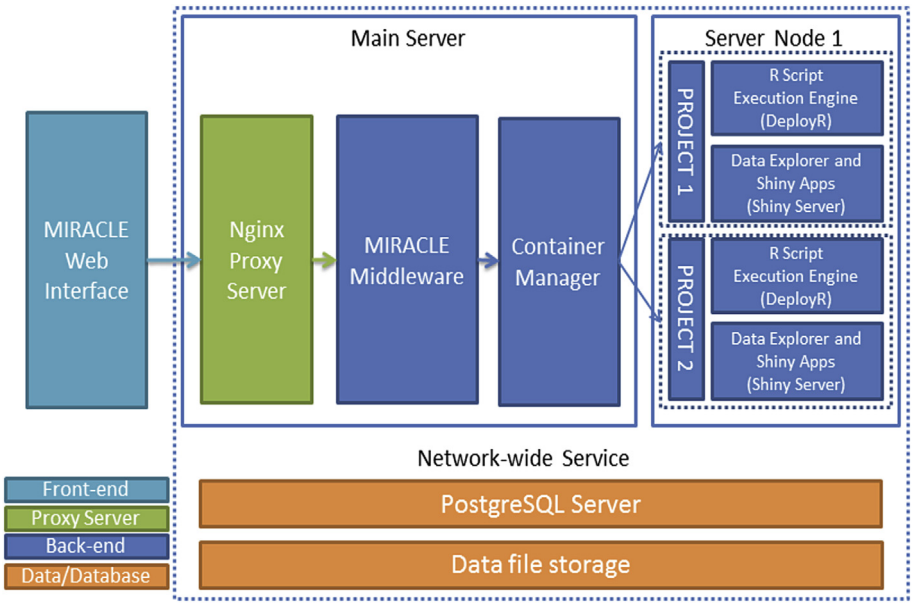

Fig. 1. System architecture plan.

availability of code, its adoption will continue to lag unless we improve accessibility through a combination of manual good practice (Wilson et al., 2016) (e.g., documentation, standard archival formats, improved modularity (Bell et al., 2015)) and supporting cyberinfrastructure for discovering, sharing, and citing code. The complexity and technical challenge of adapting external code is especially relevant for modellers in the social and life sciences who traditionally lack access to computer science training (Allesa et al., 2006).

Of equal importance for ABM-CHANS reproducibility is the availability of model output data and computational workflows that transform those data into statistically significant results. For scientific experimentation, ABM-CHANS are often run using multiple pseudo-random number generator seeds and parameter sweeps to generate large output databases of virtual "big data". Given the complexity of agent decision-making, characteristic heterogeneity and complex spatial and network interactions, ABM-CHANS are capable of producing enormous data outputs. However, as the production of such output data is often time and resource-intensive, without good tools, researchers may be tempted to limit the size of their model runs or to explore only small portions of the output space, reducing the quality of their results. For other research groups, output replication is further challenged by the fact that the model's software environment can be highly specialized, and not necessarily portable (Pignotti et al., 2009). Even if other researchers get the models running, replicating output data is not always feasible without the provision of all setup files that the original researcher used to generate the output on which they based a scientific article. Replicating output data may also be infeasible for pragmatic reasons: lack of access to CPU time and/or disk space. The result is that other researchers cannot test and explore the models' conclusions.

Even with good access to model output data, challenges for analysis and communication remain. Generic data analysis techniques are rarely suitable for agent based model output; rather, specialized analysis methods are needed (Fagiolo et al., 2007). While ABM-CHANS modellers are developing such specialized approaches (Lee et al., 2015), most are not yet part of standard statistical packages. Without appropriate protocols for even within-group sharing and archiving of analysis algorithms, the opportunity for complete reproducibility may disappear with the departure of a research group member or a computer crash.

Increasing the utility of ABM-CHANS models, and ultimately their influence, therefore, depends upon solving the problem of effectively sharing model output data and the data analysis workflows used to transform that data into a research finding of interest. Solutions to store and disseminate model output data are central to improving the communication and transparency of model results and thus the impact of agent-based models.

There is currently no standard for agent-based modellers to store and make available their model workflows, output data, and analysis algorithms. The MIRACLE project, funded through the international Digging into Data initiative (Round 3), has developed a proof of concept prototype web application designed to start a conversation on the research practices needed to archive model output data and analyses that facilitates access, re-use, and further exploration. MIRACLE removes the need to download code, run a model, or replicate data output, and integrates with existing data exploration tools so that other researchers can explore and ask new questions of the data. It thus has the potential to dramatically reduce the barriers to exploring the results from ABM-CHANS research.

2. Context

A number of code and data archival initiatives have sprung up recently alongside calls for greater transparency and reproducibility in the computational sciences (Collberg and Proebsting, 2016; Gil et al., 2016; Vitek and Kalibera, 2012; Stodden et al., 2015). Although these initiatives take a useful first step towards preserving the computational artifacts that a given research finding depends on, they fail to comprehensively address specific challenges outlined above facing researchers who use ABMs or similar complex simulation tools. MIRACLE differentiates itself by addressing these needs while integrating with existing tools. Nothing comparable exists in this space.