**eplusr: A framework for integrating building energy simulation and data-driven analytics**

*Hongyuan Jia, Adrian Chong*

Dear Prof. Jian-lei Niu,

We would like to thank the reviewers and you for the positive and constructive comments and suggestions. The manuscript has been revised accordingly. Our responses to reviewers are given in a point-by-point manner below. Changes are shown in yellow in the revised manuscript.

We hope the revised version is not suitable for publication and look forward to hearing from you.

Sincerely,

Adrian Chong

National University of Singapore

**Responses to reviewers' comments:**

***Reviewer #1:***

1. This research develops a useful package to extend the application of EnergyPlus into the R environment. This package would be helpful for parametric energy analysis in the field of building energy analysis since the R has a lot of packages on advanced statistical functions. The examples added in the appendix are also a good start point to be familiar with this new package. This is a well-structured paper to clearly present useful information on this new R package.  
   **Response**: Thanks for the kind comments. To the authors' knowledge, this is the first framework focusing on bridging the gap of integration between the building energy simulation and data science domains. With seamless integration, we hope the framework can help ease the efforts to conduct productive and credible BES studies and advocate the BES domain to embrace a reproducible workflow.

***Reviewer #2:***

1. The motivation for integrating data-driven analytics with BES is not well described in section 1. For instance, what are data-driven approaches and why/how they can be integrated with BES. A very concise summary will be necessary for the readers as general background.  
   **Response**: Thanks for the comments and suggestions. We have merged the State of the art section into the Introduction section. A summary of current research gaps has been added, including why data-driven analytics should be used when large parametric BES is involved. Please see page 1 and 2 in the revised manuscript with and without changes.
2. An overall summary of state-of-art will be necessary to form up the research gaps.  
   **Response**: Thanks for the suggestion. In the merged section, we have emphasized the motivations of this work and current research gaps have been added. Please see page 1 and 2 in the revised manuscript with and without changes.
3. Why choose R package particularly? What is the situation of other language. Does Python has similar package?  
   **Response**: Thanks for the questions.

Over the last decade, the R programming language has become a vital tool for implementing data analysis algorithms in many fields. It has extensive libraries of statistical modeling, data and database manipulation, wrangling, and visualization. New advanced statistical methods are usually first enabled through R libraries. This makes it a perfect choice for data-driven analytics. Another reason for choosing R is because of its good support for reproducible research, including literate programming, pipeline toolkits, project workflows, code/data formatting tools etc. Those facilities can be easily adapted to BES domain with the proposed R-based framework.

In the Introduction section, Table 1 gives a summary of the situation of other programming languages, including Ladybug & Honeybee and eppy packages (Python based), Modelkit (Ruby based), EpXL (Excel and VBA based), MLE+ (Matlab based). However, to the authors' knowledge, the proposed framework is the first one that focuses on seamless integration between BES and data-driven analytics, and BES reproducibility.

1. Can the developed framework be transferred into other language for larger application?  
   **Response**: Thanks for the question. Yes, it can. The developed framework is based on R which has a rich ecosystem of cutting-edge interface packages to communicate between open-source languages. Also, the proposed computational environment for BES is based on Docker, which is easy to be deployed for larger applications, including cloud computing.
2. Will the tool be possible for optimisation of building and energy systems? If so, how does it work?  
   **Response**: Thanks for the question. The proposed framework is leaned on the R language and its ecosystem. R offers rich facilities for solving various kinds of optimization problems, including optimization infrastructure packages, general-purpose solvers, mathematical programming solvers, and multi-objective optimization. A good example is the epluspar package mentioned in the Applications section, which integrates the proposed framework with multi-objective optimization solvers existing in R. With the flexible and extensible API, it is also possible to optimize building and energy systems if the problem can be abstracted and coded into R.

***Reviewer #3:***

1. 1. Abstract: the functionality of the R package should be provided.  
   **Response**: Thanks for the suggestion. We restructured the abstract to describe the functionalities of the eplusr package more clearly. Please see page 1 in the revised manuscript with and without changes.
2. State of the art: it is better to merge the State-of-the-art section into the Introduction section to point out the knowledge gaps of the proposed study.  
   **Response**: Thanks for the suggestion. We have merged the original State-of-the-art section into the Introduction section. We also added a dedicated subsection describing the aim and objectives. Please see page 5 in the revised manuscript with and without changes.
3. Methodology: it is better to show the functionality of the R package rather than the detailed software structure and coding. The authors may also need to point out the novelty or academic contribution of each feature.  
   **Response**: Thanks for the suggestion. The main functionality of the proposed framework has been described through elaborating 3 main components together with their features. In the revised version, we added more description pointing out the novelty of each component feature. Please see page 5 in the revised manuscript with and without changes.
4. Figures 8, 11, 16 are not easy to understand.  
   **Response**: Thanks for the comment. Figure 8, 11, and 16 are mainly used to demonstrate the workflow of each application using the proposed framework. We added a short description below each method/function to make clear the context and the goals. Please see Figure 8 on page 14, Figure 11 on page 15, Figure 13 on page 17, and Figure 16 on page 19 in the revised manuscript without changes. And Figure 8 on page 14, Figure 11 on page 16, Figure 13 on page 18, and Figure 16 on page 20 in the revised manuscript with changes.
5. Figure 9: this figure can be easily drawn based on the data in the eplus.html result file. The authors need to change the example to demonstrate the necessity of using the R package. For example, the users can use this R package to easily create some result figures, while it will require far longer without using the R package. Same comments for sections 4.2, 4.3 and 4.4.   
   **Response**: Thanks for the comments. We agree that Figure 9 may not be a good example for showcasing the capabilities of integration. However, even for such a simple pie chart, current existing tools still lack an easy way to extract data from the eplusout.html. The main focus of Section 3.1 is to showcase how the proposed framework streamlines and boosts the initial data exploratory analysis using structured output that meets the Tidy data principle.  
     
   In the revised manuscript, we (1) replaced the original pie chart in Figure 9 with a stacked area plot of monthly breakdown of electricity consumption and (2) replaced the original bar chart in Figure 10 with an energy signature diagram of outdoor temperature against electricity computation. Descriptions have been added to showcase the advantages of structured tidy time-series output when doing data aggregation and visualization. Please see Figure 9 on page 14 and Figure 10 on page 15 in the revised manuscript without changes, and Figure 9 and Figure 10 on page 15 in the revised manuscript with changes.  
     
   For Figure 12 in Section 3.2, Figure 14 in Section 3.3, and Figure 18 in Section 3.4, they are generally not extremely complex visualizations. However, without the streamlined workflow combing the proposed parametric simulation manager and tidy data interface, the data extraction and preprocessing of those large parametric simulations will require much more time and effort to achieve. For instance, there are in total 2000 annual energy simulations in the multi-objective optimization application in Section 3.3. Based on the proposed framework, parameter inputs and fitness values for all population and Pareto sets are stored in a tidy data frame. Thus, it becomes intuitive and effortless to perform further data-driven analytics.
6. "Lines 59 - 124 in Listing 1", "line 42 in Listing 1", "only 15 lines of codes (line 60 - 82 in Listing 1)": It may not be necessary to show such detailed coding information. Fewer lines of code do not mean that they are easier to write.  
   **Response**: Thanks for the comments. We agree that fewer lines of code do not mean that they are easier to write. Nevertheless, one core contribution of the proposed framework is that it brings reproducible research to the building energy simulation world. We believe that it is necessary to include the corresponding code of the 4 example applications since the package itself highlights the importance of reproducibility. We put all the code in the Appendix to avoid distractions to the audiences.
7. The authors may need to show more about the "data-driven analytics" as indicated in the title.  
   **Response**: Thanks for the suggestion. In Section 3.1, we included an energy signature analysis based on the time-series output of outdoor temperature against electricity computation. It further demonstrates how the structured output extraction using the proposed framework can ease the effort to perform further data-driven analytics. Please see page 14 in the revised manuscript without changes and page 14-15 in the revised manuscript with changes.

***Reviewer #4:***

1. This paper presents a framework for integrating building energy simulation and data-driven analytics (eplusr). In summary there is some good underlying work here, but the paper does not follow a typical peer reviewed paper format. It is presented as a final report. For example, there is no stand alone section to talk about the exact research methodology, results, and discussions; specially about figures/tables.  
   **Response**: Thanks for your comments and suggestion.

In the revised manuscript, we updated the paper structure to make it follow a typical peer-reviewed paper, including:

* 1. The State-of-the-art section has been merged into the Introduction section. The research gaps have been move clearly discussed there.
  2. The novelties of the proposed framework have been added into the Methodology section.
  3. A Discussion section has been added to compare the proposed framework with existing solutions, including the advantages and current limitations.

Please see page 1-5 and 21 in the revised manuscript without changes, and also page 1-5 and 22 in the revised manuscript with changes.

In terms of Figure 9, 10,12, 14 and 18 in the Applications section, we agree there are definitely rooms for the artwork, including colors, legends, fonts etc. However, they are all generated and fully reproducible using the code in the Appendix. The main focus is to demonstrate how the structured output using the Tidy data interface can ease the efforts for further data-driven analytics of BES results. Aesthetic tweaking will need more codes and is usually done in the final analysis, but is not necessary for exploratory data analysis.

1. In my opinion, the article does not present any innovative contribution as it is mainly based on already existing research, and the significance of the results is quite limited. The results are not critically discussed in the context of the state of the art and uncertainties, strengths and weaknesses of the work are not discussed.  
   **Response**: Thanks for your comments.

The main contribution of this research is the software development of a framework for seamless integration between BES and data-driven analytics and reproducible BES research. To the authors’ knowledge, this is the first research trying to bridge this gap.

In the revised manuscript, we added a discussion section to discuss the strengths and weaknesses of the proposed framework, compared to existing solutions. Please see page 21 in the revised manuscript without changes, and page 22 in the revised manuscript with changes.

1. Quality assurance on EnergyPlus simulations needs to be addressed carefully in this paper. Working with simulations facilitates mistakes, which are seldom found without equally thorough quality assurance procedures. Please describe which sort of tests and measures were adopted to assure the quality of the presented results.  
   **Response**: Thanks for the comments.

The focus of the proposed framework is to provide solutions for seamless integration between the EnergyPlus simulation engine and the R-based data-driven analytics environment. We agree that quality assurance on EnergyPlus simulations is quite important but is a separate topic from this research.

Despite that, the framework indeed contains features for EnergyPlus input validation and simulation error extraction. These include:

1. a rule-based model data validator to avoid any possible input errors before modifying EnergyPlus models. It includes 13 different checks and is triggered whenever any methods are called to modify an IDF. This is also true for the parametric model generation. Issues with detailed reasons and possible solutions are directly reported to the users.
2. an EnergyPlus simulation error (ERR) file parser to show simulation debug information and summarize error messages. The simulation manager has an *errors* method to collect error messages from single simulation and parametric simulations. Also, the Tidy data interface will check if the simulation completes successfully and will issue a warning message otherwise to remind users that the collected data is not reliable.

Besides, we pay serious attention to the quality assurance of the proposed framework. The R package follows the Test-Driven Development (TDD) process. Currently, there are more than 3600 unit tests which covers around 90% of the total codebase. The released version of eplusr is distributed via CRAN which runs all the tests automatically on Windows, macOS, Linux and Solaris at each eplusr submission and also each new release of the R language itself. The development version is held in a GitHub public repository with Continuous Integration (CI) using GitHub Actions which runs all the tests whenever any code changes occur. We added a subsection in the Methodology section to describe the code quality assurance in more detail.

A Quality assurance and quality control subsection has been added in the Methodology section to further discuss those efforts. Please see page 12 in the revised manuscript without changes and page 12-13 in the revised manuscript with changes.

1. The topic is interesting and worthy of investigation, but it is not clear how this study contributes to advance the state of art about methodology, data processing, and reliability of results.  
   **Response**: Thanks for the comments. As mentioned above, to the authors’ knowledge, this is the first research focusing on providing seamless integration between BES and data-driven analytics, and solutions for reproducible BES research. Further discussion on the novelty and advantages has been added in the new Discussion section. Please see page 21 in the revised manuscript without changes and page 21-22 in the revised manuscript with changes. The Applications section, together with the code in the Appendix section, demonstrates the benefits of the tidy-formatted BES results which can be easily fed to various data-centric analytics using existing tools in R. The efforts of quality assurance and result reliability of the proposed framework has been added in the Methodology section. Please see page 12 in the revised manuscript without changes and page 12-13 in the revised manuscript with changes.