Modelling Environmental Data Systems

Lead Supervisor: Julian Kunkel, Department of Computer Science, University of Reading
Email: j.m.kunkel@reading.ac.uk

Co-supervisors: Bryan Lawrence and Annette Osprey (University of Reading, Department of Meteorology)

Large scale environmental data is difficult to handle due to its enormous data volume; modern climate simulations can generate terabytes of simulated data for just one year of one variable and they typically simulate many variables for many years. These data are then exploited by many users who mine the data to understand the physical system and possible impacts on society. The process of data generation and analysis involves user-defined complex workflows that are executed on supercomputers and data centers throughout the world. This combination of volume and large numbers of user workflows means developing environmental data systems is just as difficult as performing the simulations in the first place.

One key part of developing a new system is modelling how a possible system might behave as users interact with it. The modelling covers key behavioral characteristics of the system particularly its performance behavior. That way potential bottlenecks in performance can be identified and alleviated in the system design but also user workflows can be optimized. This requires developing appropriate models for the characteristics which cover both the way users interact with the system, and the way the system responds.

In this project, the student will develop two interlocking performance models, one covering user workflow, and one covering data systems. The goal will be to predict the performance of one or more "possible" data systems under load from a statistical ensemble of users carrying out typical workflows.

The two components will involve:

1. Spending time with meteorologists and climate scientists to understand their workflows, not only what they currently do, but what they might want to do in the future (e.g., exploiting deep learning), and developing
models of those workflows. Besides simpler mathematical models, discrete event simulation will be applied to model the system and user interaction.

2. Measuring how those workflows perform on real computer systems and developing some models of those interactions.

3. Putting the two models together and developing estimates of overall performance on real and future computer architectures -- ranging from cloud computing to future exascale high performance computers.

4. Extending the models to other relevant characteristics like resilience.

This approach will of course find application in more than environmental science. The UK has a grand strategic objective to develop a data driven economy which realises value from connected largescale data. That objective depends on being able to "rapidly analyse" data to generate insights and innovation -- but such rapid analysis depends on suitably configured information systems

**Training opportunities:**

The project will involve survey design, abstract modelling of workflows, modelling of data flows with various techniques, measurement of real workflows on real computer systems, and a significant amount of programming, probably in Python. The skills acquired in this project will find application in any situation which requires modelling of a system and interpreting the results.

**Student profile:**
The project will be suitable for a student with a degree in Engineering, Physics, Mathematics or Computer Science, or related, and the desire and aptitude to work with people and computer systems to understand their interactions.

**Background Information:**

- **Video:** Understanding Discrete-Event Simulation, Part 1: What Is Discrete-Event Simulation?


[http://www.reading.ac.uk/nercdtp](http://www.reading.ac.uk/nercdtp)