Data Management Plan

1 Description of expected data products

The research I am proposing here is purely computational, and there will be three types of data produced in these studes: code, raw data, and processed data. All data products will be open-source and all code will be free to use under the GNU General Public License.

1.1 Code

The primary data product that will be produced is the code that is use to run and process simulations. The Dedalus code, on which my simulations will be built, is an open-source Python code licensed under the GNU General Public License (GPL); as a result, the code that I use to run my simulations will be Python files, and will be made open source under the same license. As I have done in the past, each time I publish a new paper, I will create a Zenodo repository of supplemental materials that has a specific doi to preserve the code I used in the production of that work. Day-to-day, the code that I use is developed and stored in version controlled Git repositories, and in addition to the "hard copies" in Zenodo repositories, I will ensure that my cutting-edge Git repositories are open to the public as well. Despite the open-source nature of Dedalus, there can still be a steep learning curve between installation and successfully running simulations, and this hurts reproduceability and transparency in science. These distribution plans should help ensure that even relatively novice users of Dedalus will be able to find our code and recreate our simulations.

1.2 Raw data

Dedalus creates data in an HDF5 output format, which allows for extreme output versatility during simulation runs. As a general rule, storing raw data describing the full flow, temperature, and field data of a simulation with a modest resolution and output cadence easily exceeds terabytes in data. Such large quantities of data are cumbersome to make publically available.

Rather than making the full set of simulation data available for all simulation times, we will instead publish select "checkpoint" data at simulation times which are physically interesting. These checkpoints will not only contain the full state of the simulation at a select time, but can also be easily loaded into the publicly-available code if the curious reader wants to reproduce our results without starting a simulation from scratch. Furthermore, by only publishing select, interesting checkpoint data, we can ensure that the volume of data being published is small enough that these checkpoints can be conveniently found in the same Zenodo repositories as our code, or in the supplemental materials of the published work.

For select state-of-the-art simulations, full raw datasets will be moved into mass storage on NSF XSEDE resources such as DATA OASIS, PYLON, and RANCH. These data will be publically available but will only be given out upon request due to their massive nature.

1.3 Processed data

Often when studying simulations, we focus our efforts on the evolution of scalar quantities or 1D profiles. These data products have a relatively small memory footprint, and so full time traces of these outputs will be made publicly available in Dedalus' native HDF5 format. Time-averaged and statistical scalar quantities used in analysis or figure creation will additionally be made available in a simple .csv format.