Hi Quantum Black partners,

Thank you for collaborating with us on our drought prediction project. I am sending this email on behalf of our team to provide a summary report on our current progress and our roadmap for the remainder of the project. Additionally, we have attached our research paper for your reference. We welcome your feedback and any additional ideas you may have.

In the past several weeks, we have implemented different types of models and experimented with different combinations of features, training and prediction horizons. After switching to predicting categorical variables (drought versus no drought, six categories of drought levels) other than numerical variables, we achieved better prediction accuracy. We find, in general, the best results from neural networks models, such as MLP, LSTM, CNN, suggesting that machine learning is useful in the drought prediction case. Our next step is to fine-tune our better models and implement RNN as our last exploration. Meanwhile, we will write about the results analysis and wrap up the project.

We have received extensive support from our teaching fellow, Zona, and our head instructor, Dr. Tanner. They provided tremendous help in finding proper directions for exploration. We are looking forward to presenting you with our updated modeling results this Friday as well as next week. The following two sections walk through our current progress and immediate future plans step by step.

**Current progress:**

* *Data Sanity Check.* Since we switch to predict on categorical targets, we need to specify a threshold of drought/no drought. We did it by looking at the “None” column in the drought dataset, making None=100 drought and any other thing to be no drought. Per Francesca and Vidur’s suggestion, we have explored the balance of the dataset given this threshold for every year and every climate region. The slight imbalance across the year is not too concerning, since it fluctuates around 40%-60% between one of two categories. However, the regional imbalance is expected to occur.
* *Modeling and Improvements.* We implemented three types of models for predicting our categorical target. We considered logistic regression as our first set of models, and we tried both predicting binary targets and six-level targets (including “no drought”). The best result (accuracy: 67%) is for using “8 weeks before” to predict “26 weeks ahead”, training from 2012-01-01 onwards , using binary targets. The second type is the tree-based model, namely decision tree and random forest with bagging. We did not have better accuracy (much greater than the logistic regression outputs). As for the deep learning methods, we get significant improvement with an accuracy of 95% for predicting 1 weeks ahead. The detailed results are shown in the research paper.

**Next steps:**

* *Modeling.* We will finalize our LSTM pipeline and implement one additional RNN model. Moreover, regional models are worth implementing. In the remaining days we want to fine-tune all the models we have, and do careful analysis for the results.
* *Documentation and Paper.* It is a good merit to take notes while we are developing the project. We will keep a log and update the documentation regularly. Also, we shall find a way to better discuss our results, combining high-level ideas and detailed data as support.
* *Github.* Clean up github code, make sure we include some pedagogical examples for future users.

Again, we want to thank you for your enthusiasm and strong support for our project! Please let us know if you have any additional suggestions or comments regarding the summary above, and we look forward to keeping in contact with you throughout the remainder of the semester.

Sincerely yours,

Yujie, Jim, Michael, Thee