**Impact statement for paper: Machine learning in drought prediction**

Yujie Cai, Michael Fein, Thee Ngamsangrat, Jim Zhang,

Institute of Applied Computer Science, Harvard, Cambridge, United States of America

May 4, 2022

**1. Data**

This study introduces new datasets to the field of drought study: multi-channel satellite imagery as well as telemetry data: Surface Evaporation and Run-off data from ERA5-Land, Surface (soil) Moisture data from Sentinel-1, and SentinelHub satellite imagery. Although there have been sporadic studies on the application of satellite images for drought prediction, most of the studies in drought research community still use meteorological information collected from ground-based weather stations mainly. The use of satellite imagery and remote sensing dataset essentially introduces a new chunk of datasets to the field.

Satellite telemetry data inevitably requires manual aggregation, calculation, and conversion, similar to other machine learning datasets that require crowd workers' annotation. We have carefully searched the public ratings for data providers and we believe that the employees of our data providers are treated reasonably in accordance with local laws and social practices. In addition, although satellite maps are newly introduced in the field of drought research, they have been used for quite some years in other research areas, and there are no significant ethical or legal risks associated with satellite acquisition, labeling, analysis, supply, and secondary distribution.

In the process of applying the new datasets, we ensured that no intellectual property infringement occurred. First of all, all datasets were obtained from open-source service providers that release their datasets freely to the public. The research and publication of our research may involve the redistribution of processed data. So, we also ensured that the distribution of model comply with all relevant laws and original source of the data is legally and compliantly labeled.

We also consider the human element in the data collection and processing. We believe that all satellite and surface weather data that we use do not contain measurement of human individuals. We recognize that ultra-high-resolution satellite imagery may constitute issues such as privacy infringement if they capture discernable humans in their imagery, but the resolution of the dataset we use is up to 60 x 60 miter and more on the scale of 9 x 9 or 36 x 36 kilometer. We believe, at this resolution, our dataset poses little risk of violating the privacy and rights of human individuals.

In summary, our use of satellite map datasets and meteorological datasets is justified.

**2. Experiments**

We have conducted a large number of experiments on satellite maps and meteorological datasets. We have presented the most important properties of each dataset as much as possible. However, due to length limitations, we cannot present all the properties of the datasets in their entirety. We have included the source of the data in the appendix of the article and fully disclosed how to access data, the official website of the data provider that contains full description of the datasets; we also provide in our GitHub page the full codes of our research. We believe that we have ensured a high enough level of transparency in our experiments on the data.

Our experiments are consistent with what we claim in our paper. To ensure the generalizability of our experiments, and to ensure that our drought prediction models have good prediction quality over as wide a range of scenarios as possible, we deliberately targeted our research at the continental United States, a vast geographic area with 9 climate sub-regions, and stratified our sampling across the nine climate regions to ensure that our models were trained to see drought patterns from a variety of climate types.

Finally, we need to disclose that we are unable to verify the quality of each datasets in detail due to the time and financial constraints of the study, except for the basic sanity check.

**3. Application Scenarios**

As an applied research project that aspires to practice AI for social good, we anticipate that our model will be used in real-world drought prediction scenarios. We are quite confident in the availability of data and the accessibility of our model. All of our data are from open-source service providers and the data are freely available. Our models can be trained and run on any mid-end computer or on a free version of Google Colab.

We would like to remind the users that, our model can only predict the occurrence of droughts and cannot give a higher level prediction of how much water reserves are needed as drought aid or relief, thus our model may lead to mis-matching of water resources in drought areas. In practice, our model also inevitably gives wrong predictions - it is only 80% accurate in the half-year-long run, which may lead to wasted relief supplies and water reserves if agencies and local governments rely solely on our model.

When tested on the continental U.S. dataset, the model was more accurate in drought-prone areas, which coincidentally are mostly more economically disadvantaged than the east and west coastlines. Since the model is instead more accurate in these economically weak areas, harms identified are not likely to fall disproportionately on populations that already experience marginalization or vulnerable.