Microsoft

# AI for Earth Grantee Profile

## University of Massachusetts Boston

## Long-range flood prediction


## Summary

An accurate early warning system for severe floods in flood-prone regions would support response teams and help build resilience among vulnerable populations. A team of computer scientists and hydrologists/meteorologists is developing a flood prediction model that uses machine learning and an advanced AI data algorithm on Microsoft Azure to identify precursors to floods in specific at-risk regions, with the aim of accurately predicting floods with up to 15 days' lead time.


## An interdisciplinary approach to improving long-range forecasts for flood prediction

In most of the developing world, flooding is the most deadly and costly natural hazard. According to the World Resources Institute, nearly 80 percent of the total population exposed to river flood risk worldwide lives in just 15 countries—all considered least developed or developing. In Pakistan alone, 715,000 people were at risk in 2015. By 2030, river floods in Pakistan could affect an additional 2 million people.

> Nearly 80 percent of the total population exposed to river flood risk worldwide lives in just 15 developing countries.

In 2017, the World Bank reported that while average annual losses from natural disasters total more than $300 billion, the cost to people's well-being is the equivalent of a $520 billion drop in consumption—with poor people disproportionately affected. The report proposes that resilience-building interventions, including early warning systems, would help poor countries reduce the overall impact of disasters, for a potential gain in well-being equivalent to a $100 billion increase in annual consumption.

However, building resiliency in flood-prone parts of the developing world depends on increasing the quality of flood forecasts. Ideally, planners need 5 to 15 days of lead time to develop a response. The biggest impediment to providing such long-range flood forecasts is the poor quality of quantitative precipitation

forecasts. Numerous studies have shown that numerical weather forecast models systematically underestimate precipitation totals at forecast lead times as short as one day. At five days' lead time, forecast skill is limited to predicting the occurrence, but not magnitude, of precipitation. Accurate flood predictions also depend on accurate location forecasts—identifying the correct river basin to be affected.

One challenge is that the governing equations used by atmospheric models depend on having perfectly defined initial and boundary conditions—something which is simply not achievable. For the current generation of atmospheric models, the doubling time of small errors in the initial conditions is between 2 and 2.5 days.

Another challenge is model error. Atmospheric scientists have not yet developed a complete set of analytical equations for the atmosphere. Numerical models, therefore, require many important physical processes to be estimated rather than directly calculated from first principles, leading to model errors that can grow as the forecast horizon becomes longer than two to three days.

### An early warning system for vulnerable communities

The work of computer scientists Professor Wei Ding, water engineer Professor Shafiqul Islam, PhD student Yong Zhuang, and research hydrologist and meteorologist Dr. David Small may well represent a solution. Rather than using conventional differential-equation based atmospheric models, they are collaborating on developing machine learning models with the goal of accurately predicting floods up to 15 days in advance. The team's motivation is to build a system that can give early flood warnings to vulnerable populations around the world.

> The team is collaborating on developing machine learning models with the goal of accurately predicting floods up to 15 days in advance.

Their approach hinges on big data: accessing huge reanalysis datasets consisting of millions of variables from a range of sources, including the National Center for Atmospheric Research (NCAR), European Centre for Medium-Range Weather Forecasts (ECMRWF), and the National Oceanic and Atmospheric Administration (NOAA), as well as precipitation data from NASA satellites (e.g., TRMM). Big data and the machine learning approach enable them to ask the data—using an advanced AI data algorithm—to speak for itself and reveal any precursors for long-range flood prediction. Essentially, they are assessing whether historical data from previous floods can reveal patterns to enable future prediction.

### Collaborating at the speed of the cloud

Through a grant from the AI for Earth program, the team now has access to the Microsoft Azure platform, which provides both the computational power and the data platform for them to mine enormous datasets using their machine learning algorithm. As they refine their approach to use finer resolution (higher quality) data, the number of feature variables that they will be working with will increase from a few million to upward of a hundred million, requiring 100 TB of storage and real-time processing at regular intervals. The project will use an HDInsights Hadoop cluster running Spark to build a data pipeline to support this processing. They also plan to use the Data Science Virtual Machine software environment, including Microsoft Machine Learning Server with Python, Azure Machine Learning Workbench, and Jupyter Notebook.

> Azure provides the computational power and data platform to mine and share results from enormous finer resolution datasets.

Azure also facilitates an interdisciplinary research approach by enabling them to share results and collaborate directly on the data. The team credits the exciting potential of the project to their interdisciplinary structure. Rather than working in isolation as disparate workstreams, they view each other as true collaborators, wedding together their respective "faiths" in data and physics. These benefits extend beyond the team—they plan to share the solution more broadly with the people who need it the most.

### Going forward

The team's preliminary focus is to develop a forecasting model—including a precipitation prediction model, a streamflow prediction model, and an application programming interface (API)—for the Ganges basin, which they will make available to interested parties in Bangladesh and India. They are also working with case examples from Iowa (in partnership with the University of Iowa), Pakistan, and a few other regions that have experienced a significant flood in the past.

## About the project team

Wei Ding is an Associate Professor of Computer Science at the University of Massachusetts Boston. She has authored or co-authored over 105 referred research papers, one book, and holds two patents. Her current research interests include data mining, machine learning, artificial intelligence, and computational semantics with applications to astronomy, geosciences, and environmental sciences. Professor Ding and Professor Shafiqul Islam (Hydrology and Water Resources, Tufts University) are the primary investigators for this project.

In addition, Dr. David Small, a research hydrologist, meteorologist, and insurance industry data scientist, serves as an outside advisor. PhD student Yong Zhuang (Computer Science, University of Massachusetts Boston) is the primary research assistant to design and implement the machine learning algorithms. The group has extensive experience in flood and severe weather forecasting using machine learning and models of varying complexity. They have published five papers, including three in *ACM SIGKDD* and *IEEE BigData*.

# Resources

## Websites

Wei Ding publication page
AI for Earth

## Publications

Yahui Di; Wei Ding; Yang Mu; David L. Small; Shafiqul Islam; Ni-Bin Chang. "Developing machine learning tools for long-lead heavy precipitation prediction with multi-sensor data." 2015 IEEE 12th International Conference on Networking, Sensing and Control. *IEEE Xplore*. June 4, 2015

Dawei Wang; Wei Ding; Kui Yu; Xindong Wu; Ping Chen; David L. Small; Shafiqul Islam. "Towards Long-lead Forecasting of Extreme Flood Events: A Data Mining Framework for Precipitation Cluster Precursors Identification." 2013