Microsoft

AI for Earth Grantee Profile

CoCoRaHS, Colorado State University

Enhancing a citizen science dataset with AI

Summary

Precipitation can vary a lot over surprisingly small distances, as demonstrated by the Spring Creek flood in Fort Collins, Colorado, in 1997, when 14.5 inches of rain fell in a highly concentrated area and caused a deadly flash flood in nearby neighborhoods that had significantly less rain. From that disaster was born the Collaborative Community Rain, Snow, and Hail Network—CoCoRaHS for short—which works with thousands of volunteers to gather daily data on precipitation. CoCoRaHS provides small-scale coverage that helps weather services issue timely alerts on severe weather conditions that can save lives, and its accumulated records also help other organizations, from climatology to agriculture, engineering, and insurance, with long-term planning. Now thanks to a Microsoft AI for Earth grant, CoCoRaHS is improving the quality of its reports through AI, pulling more information out of the reports with natural language processing, and making that data more available through Azure Notebooks and Power BI.

Enhancing climate data and research with AI

In popular opinion, weather forecasts are notoriously inaccurate. The classic example is rainfall prediction: it never seems to rain as much (or as little) as "they" said it would. Some of this is due to the inherent challenges of predicting the behavior of vast chaotic atmospheric systems, but some is also due to differences in location between where the weather data is gathered and where people are experiencing it.

> "Since precipitation is so variable across small areas, it makes it difficult to make too many assumptions about what's going on." – Julian Turner, CoCoRaHS

Airports are a primary location for gathering weather data, partly because the safety of air travel requires highly accurate reporting for those locations and partly because the wide-open flat areas mean the data won't be influenced by local factors such as the clusters of large buildings (and heat output) of an urban downtown. But as it's said, nobody lives at the airport—weather systems may act over wide areas, but weather is still a local phenomenon.

The degree to which weather can be highly localized has only come to be appreciated in the past couple decades. Unexpected and tragic events have provided the impetus for this deeper understanding. For instance, in late July 1997, the city of Fort Collins, Colorado, had a major storm dump heavy rains across town—but not evenly. Spring Creek, which runs through the center of town and passes close by Colorado State University, rose in a flash flood that killed five people and caused over $200 million in damage to the university and surrounding neighborhoods. The flooding was a big surprise, as neither the predicted nor reported rainfall in that area indicated the danger. However, a bucket survey carried out afterwards showed that while the area around the university was reporting six to ten inches of rain, a very small area nearby which happened to include the source of the creek had as much as 14.5 inches of rain. The heavier amount of rain wasn't visible on radar and wasn't being measured on the ground in that locality, and so no one was prepared for it.

At the time of the Spring Creek flood, Nolan Doesken was an assistant state climatologist for Colorado. Doesken and his team conducted the bucket survey to determine why the flooding had occurred. Through this process, Doesken learned three things. One, the variability of rainfall was much greater over much smaller distances than expected. Within just a couple miles across Fort Collins, the rainfall ranged from as little as two inches to over a foot. Two, that variability meant a lot more local data-gathering was needed to track and warn of these kinds of dangers. Three, many people—ordinary citizens—were eager and able to help with gathering this data. These discoveries inspired Doesken to start the Community Collaborative Rain, Hail, and Snow Network—CoCoRaHS for short.

Using citizen scientist observers to improve weather forecasting

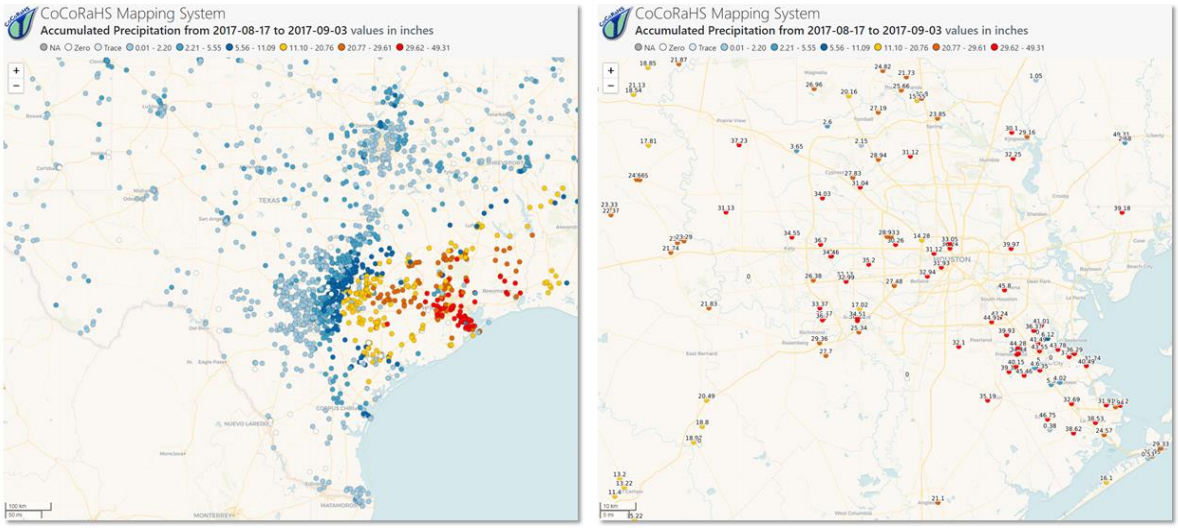
CoCoRaHS works with thousands of trained volunteer observers throughout the United States, Canada, and the Bahamas to gather daily data on precipitation. Using simple and inexpensive tools such as plastic rain gauges, the volunteers measure rain or snowfall each morning—or more often during severe weather—and submit their readings to the CoCoRaHS database. Reports of hail are also submitted as it occurs. This data is immediately available to the US National Weather Service (NWS) and similar organizations concerned with short-term weather prediction, providing detailed on-the-ground confirmation of what's actually happening weather-wise. With over 19,000 active observers in the network, CoCoRaHS provides the small-scale coverage that helps the NWS issue timely alerts on severe weather conditions that can save lives. Additionally, with over twenty years of still-

Simple tools like this plastic rain gauge help over 19,000 active observers provide small-scale coverage on severe weather conditions that can help save lives.

CoCoRaHS, CSU

accumulating data, CoCoRaHS also helps other organizations from climatology to agriculture, engineering, and insurance with long-term planning.

Although the Internet of Things has made automated local sensors more widespread, on-the-ground manual measurements by humans continue to provide vitally reliable and accurate information that automated systems cannot. For example, automated rain or stream-flow gauges can be overwhelmed or knocked out by high winds and flooding. People can make reasonable decisions for collecting data that fixed automated systems cannot, such as choosing where to measure snowfall to account for wind drifts. Also, weather variations can be so localized that meteorologists can doubt long-range automated sensors like radar even when they are providing accurate information. Human observers can verify that radar signs of very localized heavy rainfall, for instance, are indeed correct. Beyond direct numerical measurements, the volunteers can also submit observation notes providing additional useful context, such as hail mixed in with rain, or the relative dryness of plant life for monitoring drought conditions.



CoCoRaHS observations of Hurricane Harvey.

Comparative studies have shown that the data collected by CoCoRaHS's trained volunteers is sufficiently high quality to be valuable for scientific research and practical use. Still, with 4.4 million records being created every year, quality control is an important and necessary, but time-consuming, part of the process. Large errors, such as missing a decimal point and submitting 22 inches of rain instead of 0.22, are easy to catch, but smaller ones such as submitting 0 instead of 0.22 are much harder. Also, sometimes new volunteers have been measuring precipitation on their own for years and submit large batches of historical data all at once, putting a strain on CoCoRaHS's existing quality control systems and people. This is where automated systems can shine, and AI and machine learning offer opportunities to improve. And those opportunities are coming to CoCoRaHS thanks to a Microsoft AI for Earth grant.

Improving weather data quality with machine learning

"It's really hard to quality check weather data because anomalies happen," says Julian Turner, the chief software developer for CoCoRaHS. "Since precipitation is so variable across small areas, it makes it difficult to make too many assumptions about what's going on." Just as human observers provide a check on the weather forecasts produced with computer calculations and remote sensor data, so can machine learning help identify suspect data for humans to review. Using both the CoCoRaHS daily precipitation records (whose data include quantitative as well as textual observation notes) as well as the ticketing system already in use for reporting possible errors, CoCoRaHS in collaboration with Solliance, a Microsoft Gold Cloud Partner and Preferred Partner for data and AI, will develop machine learning and deep learning models using Microsoft Azure Machine Learning and Microsoft Azure Databricks to look for patterns within datasets and anomalies across datasets that should be flagged for a second look. Further analysis of external datasets from NOAA will help refine the models and make CoCoRaHS's long-term datasets more valuable for climatological data users.

> "It's really hard to quality check weather data because anomalies happen." – Julian Turner

Machine learning also offers opportunities to draw more quantifiable data out of the condition monitoring reports. These reports now feature sliding indicators for some conditions, such as relative dryness to help monitor drought, but whether those conditions are good or bad depends on the situation and how the observers describe it. With natural language processing, Turner hopes to use topic detection, named entity recognition, key phrase extraction, text sentiment analysis and text classification models to evaluate these observation notes and make it easier to incorporate that information into weather and climate research and analysis.

The AI for Earth grant offers further benefits by allowing CoCoRaHS to easily share their data and analyses through Microsoft Azure cloud services with the public. Turner says, "We're all about teaching people climate and weather, and I'm also really interested in teaching people, especially kids, the basics of understanding data visualizations and how decision makers make decisions based on those charts and graphs. And that's where using Power BI and the Azure Notebooks come in." Azure Notebooks can combine the dataset, the data processing, and the data output all in one place with a text summary for distribution. Power BI will let Turner create interactive dashboards and data visualizations without having to build custom APIs and a UI to support them. With these tools, CoCoRaHS network coordinators could easily visualize statistics on the weather in their region, and observers could see how their reports contribute to forecasting and research—or even simply enjoy the satisfaction of seeing for themselves that yes, it really was raining harder in their neighborhood than the surrounding area.

Looking ahead

The main role of CoCoRaHS is to collect high-quality data that otherwise couldn't be gathered and distribute that to other organizations for use and research. By increasing and improving the quantity, quality, and availability of those datasets, this project will greatly benefit many others in the community. If the machine learning models prove effective, more datasets may also be used, such as the hail, significant weather, and evapotranspiration observation datasets.

About CoCoRaHS

CoCoRaHS is an acronym for the Community Collaborative Rain, Hail and Snow Network—a unique, non-profit, community-based network of volunteers of all ages and backgrounds working together to measure and map precipitation. By using low-cost measurement tools, stressing training and education, and utilizing an interactive website, its aim is to provide the highest quality data for natural resource, education, and research applications. The network originated with the Colorado Climate Center at Colorado State University in 1998, and now has over 19,000 active volunteers across the US, Canada, and the Bahamas. CoCoRaHS is active in all fifty US states as well as the territories of Puerto Rico and the US Virgin Islands.

Resources

Websites

CoCoRaHS home site
CoCoRaHS introductory videos on YouTube: 2-minute introduction and 4-minute animated synopsis of the CoCoRaHS story

Press

"Even Big Data Starts Small, Ep. 1." *The Crowd & The Cloud*. PBS. April 6, 2017. http://www.pbs.org/video/crowd-cloud-even-big-data-starts-small/ (This episode of the series features the CoCoRaHS project, starting at 05:00 until about 17:20.)