ground water

Methods Note/

CrowdHydrology: Crowdsourcing Hydrologic Data and Engaging Citizen Scientists

by Christopher S. Lowry[1] and Michael N. Fienen[2]

Abstract

Spatially and temporally distributed measurements of processes, such as baseflow at the watershed scale, come at substantial equipment and personnel cost. Research presented here focuses on building a crowdsourced database of inexpensive distributed stream stage measurements. Signs on staff gauges encourage citizen scientists to voluntarily send hydrologic measurements (e.g., stream stage) via text message to a server that stores and displays the data on the web. Based on the crowdsourced stream stage, we evaluate the accuracy of citizen scientist measurements and measurement approach. The results show that crowdsourced data collection is a supplemental method for collecting hydrologic data and a promising method of public engagement.

Introduction

Advancements in technology and instrumentation in hydrogeology drive a perceived need for high-resolution spatially and temporally distributed measurements to quantify complex earth system processes. This increase in instrument resolution often comes at an increased cost in materials and personnel to install and manage these networks. To collect high-resolution data sets, many agencies and organizations have started to incorporate the use of citizen scientists (Graham et al. 2011). While there are some drawbacks relating to the accuracy of citizen scientist data (Engel and Voshell 2002), the potential for broader impact by engaging citizens and expanding monitoring networks is attractive. The expansion of mobile phone technology provides many opportunities to engage citizen scientists. As a result of these new developments, we developed and tested a method to monitor stream stage using crowdsource-based text messages of water levels. This network relies on untrained observers sending water level measurements at designated gauging staffs via mobile phone text messages to a server maintained by the researchers. On the basis of these data collected over one field season, we are able to investigate the accuracy of these data and density of field measurements.

Previous citizen scientist projects have involved classifying organisms/objects with fewer projects focusing on field measurements. A prime example is the Christmas Bird Count, which has been organized annually since 1900 by the Audubon Society (Wersma 2010; Audubon Society 2011). Others include identifying galaxies/stars (Galaxy Zoo; Land et al. 2008; Citizen Sky); data entry and measurements of fossils (OpenDinosaur), identifying birds (eBird; Wood et al. 2011); measuring precipitation (CoCoRaHs; Cifelli et al. 2005; RainLog; Weather Spotter Network); collecting images and qualitative stream data (Creek Watch); and documenting wildlife corridors by monitoring road kill (Wildlife Crossing). In the hydrologic sciences, the U.S. Environmental Protection Agency (EPA) runs a volunteer monitoring program for water quality.

In projects focusing on field measurements, such as the EPA monitoring program, maintaining quality control is accomplished with formal training of citizen scientists. Training may limit the number of observers involved and, as a result, reduce the temporal or spatial resolution. Training is necessary for projects

[1]Corresponding author: Department of Geology, University at Buffalo, 411 Cooke Hall, Buffalo, NY 14260; (716) 654-4266; fax: (716) 654-3999; cslowry@buffalo.edu

[2]U.S. Geological Survey, Wisconsin Water Science Center, Middleton, Wisconsin.

Received January 2012, accepted May 2012.

© 2012, The Author(s)
Ground Water © 2012, National Ground Water Association.

doi: 10.1111/j.1745-6584.2012.00956.x

that require multivariable measurements using calibrated instruments. In less involved projects, such as distributed measurements of precipitation, training may consist of a simple online slide show or video (CoCoRaHs; Cifelli et al. 2005). These single variable projects also require limited equipment and time commitment. Quality control issues are present in single variable projects but outliers may be identified using time-series analysis or other validation.

Predicting how the larger hydrologic scientific community will respond to the publishing of these data is uncharted territory. In astronomy, crowdsourced data analysis has become acceptable with more than 20 papers published using the Galaxy Zoo database (www.zooniverse.org; Raddick et al. 2010). In the case of classifying organisms or objects, an expert can later verify the classification. For crowdsourced hydrologic field measurements, experts usually cannot go back and verify a precipitation or water level measurement, which adds uncertainty to hydrologic measurements.

Crowdsourcing is a viable tool for collecting distributed measurements of stream stage based on both the ease with which stream stage can be measured and the ubiquity of mobile phones. Many rural and remote watersheds in the United States lack hydrologic data. The U.S. Geological Survey (USGS) has an extensive nationwide network of stream gauging stations (http://waterwatch.usgs.gov/new/). However, budget pressures impact the number of gauges, and the network resolution is typically limited to a single gauging station in a given watershed. As a result, assessing the impact of localized management practices and/or forecast future changes on water resources at a watershed or sub-watershed scale can be difficult.

The research presented here focuses on building Internet-based interface to harness the power of citizens to collect hydrologic data. Using mobile phones, citizen scientists can send hydrologic measurements (e.g., stream stage) via text message to a server that stores and displays the data on the web. A natural extension to using text messages is a smartphone application such as that used for CreekWatch (www.creekwatch.org). Text messages have the advantage of both simplicity and ubiquity. Virtually all mobile phones have text message capability. The development of a smartphone application adds a layer of design and implementation complexity. These data can be used by resource managers, researchers, teachers, and outdoor enthusiasts to monitor trends in stream stage. In some cases, recreation choices can be guided by nearly real-time observations in ungauged basins. The objectives of this research are (1) to build an inexpensive crowdsource-based hydrologic network to monitor stream stage and (2) to investigate the potential benefits and pitfalls for expanding such networks for use within the larger scientific community.

Methods

This section provides a broad overview of the project, starting with the staff gauges and progressing to data transmission and handling. The innovation of this project is the transmission of these data and distributing measurements to the larger community; not the method of measuring stream stage. To test this methodology, a network of stream gauging staffs (staff gauges) was installed in Summer 2011 at nine locations within the Buffalo, Cattaraugus, and Wiscoy Creek watersheds in Western New York, USA. These locations were chosen on the basis of popular fishing access points and hiking trails. Each staff gauge included a sign displaying a unique station number and a phone number for observers to send a text message of water level measurements. Text messages from local observers were used to generate a database of stream stages and populate an interactive web-based map (www.crowdhydrology.org). The web interface allowed users to freely view and download both current and historic data. The modular system allowed additional observations to be added to the hydrologic network by installing additional staff gauges in a watershed. This system was limited to stream stage, but the database could be easily adapted to include temperature, rainfall, and snow depth.

Signage

The system relied on untrained observers to send in measurements of stream stage. Instructional signage at each of the gauging locations was the only interaction between the scientists and the observers making measurements. Each gauging station had two signs: one located at the top of the gauging staff and a second sign posted on the shore. The sign on the top of the gauging staff gave the phone number and station number. For this study, the station number denoted by the two-letter state identification code along with a unique four character numeric code (e.g., NY1000). The on-shore sign detailed how to make a measurement and gave a short description of the project (Figure 1).

Data Retrieval and Automation

This section highlights the main aspects of Social.Water, the underlying software for the CrowdHydrology project. The Social.Water code is available at https://github.com/mnfienen-usgs/Social.Water. Social.Water relied solely on open-source programming languages (Python and JavaScript). Text messages were received at a GoogleVoice (google.com/voice) phone number and forwarded to a Gmail (mail.google.com) account. Social.Water read and parsed the messages, updated a simple database, and generated hydrographs and comma-separated variable (CSV) text files that were posted on the website crowdhydrology.org.

To interact with the Gmail account, we used IMAP (Internet Message Access Protocol) through a module built into Python. The code logged into the Gmail account, checked for new messages, downloaded the body of selected messages, and logged out of the account. IMAP

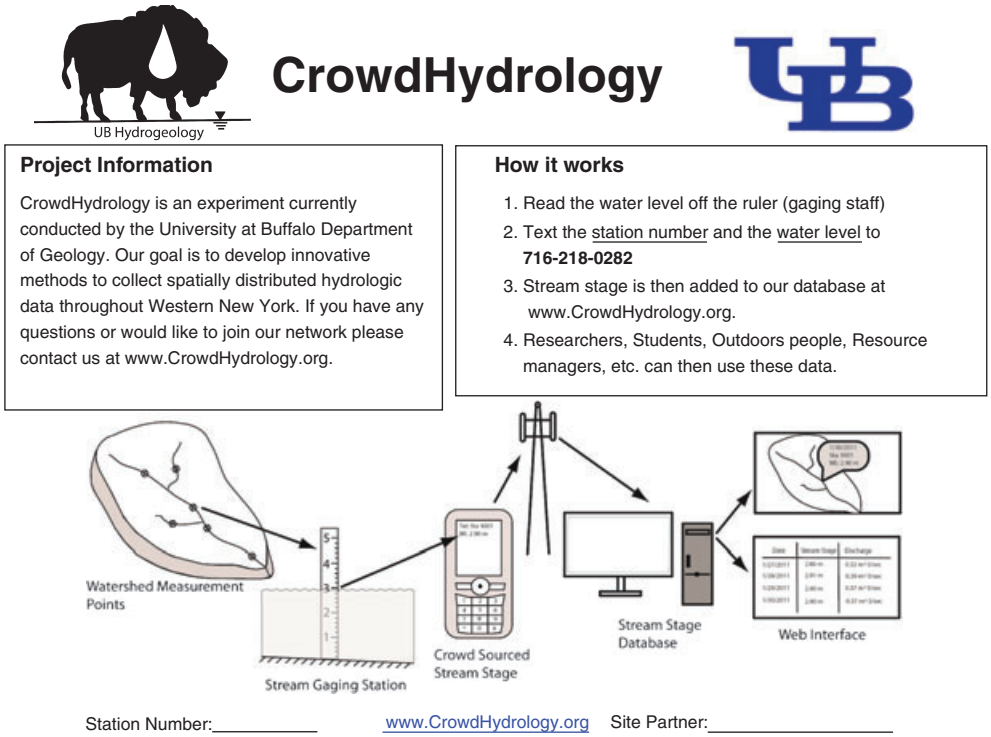

Figure 1. Streamside informational sign.

left the messages on the server, so a backup of the data was effectively maintained. Using Google Voice, all messages forwarded to the Gmail account included the text "SMS from" and a phone number in the subject line. Messages that did not contain data were avoided by only reading messages identified as "SMS from."

New messages were parsed to find the data point and to align the data with the proper gauge. The time of the measurement was taken from the time stamp of the e-mail message. To maintain simplicity of signage, little guidance regarding format of the text message was provided to the citizen scientists. In order to enable this level of flexibility, a fuzzy matching algorithm was required to properly align message content with gauges.

Social.Water used the Python package fuzzywuzzy (Cohen, 2011) to rank the level of agreement between two strings to match identifying text in the messages to the list of allowable gauge numbers (NY1000, NY1001, . . . , NY1008). The message was then assigned to the closest match. Heuristic evaluation of matching messages to the intended stations was conducted. The algorithm was adjusted until all messages that could be interpreted by scientists were correctly classified by the code.

The final step was to read the actual measurement. This was done using regular expressions that scanned the message for floating point values. The result was a code that interpreted messages similar to the way a human would do. Table 1 shows examples of the range of messages that were successfully interpreted by Social.Water. Some messages were vague, such as "water level 2.4" without indicating the gauge. Neither code nor a scientist could interpret such messages.

Table 1
Examples of Actual Messages That Were Successfully Interpreted by Social.Water

Date	Station Number	Text Message
October 9, 2011	NY1000	Station NY1000 water level 1.40
October 11, 2011	NY1000	Station#NY 1000 1.38
October 23, 2011	NY1007	1.80 Station NY 1007
November 25, 2011	NY1007	By 1007 2.15
November 25, 2011	NY1002	By 1002 ny 2.64

After parsing, a text file was updated and stored on the server containing time and value. This served as the database. Finally, a small JavaScript-enabled HTML page was written using the open-source Dygraph package. This generated an interactive hydrograph linked to a map on the main CrowdHydrology web site. The charts were also viewable on most smartphones. One goal of this implementation was to enable citizen scientists to see the results of measurements that they contributed to in near real time on the web. The Social.Water code was run on a web server inside a cron script every 5 min to facilitate this goal.

Results

Nine stream gauges were installed in two phases in Summer 2011. The initial test site was installed in May 2011 as a proof of concept. The result was over

Table 2
Distribution of Measurements at CrowdHydrology Sites

Station Number	Start Date	End Date	Number of Measurements
NY1000	April 19, 2011	December 15, 2011	133
NY1001	August 10, 2011	December 15, 2011	6
NY1002	August 10, 2011	December 15, 2011	6
NY1003	August 10, 2011	December 15, 2011	2
NY1004	August 10, 2011	December 15, 2011	3
NY1005	August 10, 2011	December 15, 2011	2
NY1006	August 10, 2011	December 15, 2011	1
NY1007	August 10, 2011	December 15, 2011	5
NY1008	August 10, 2011	December 15, 2011	5

one hundred recorded measurements between May and November 2011 (Table 2). Eight additional sites were installed in August 2011 and resulted in a combined total of 30 measurements (Table 2). An additional 17 messages collected were missing either the station number or the water level measurement. A total of 68 unique observers (phone numbers) contributed measurements. Eight-five percent of the citizen scientists sent in a single text message. The most prolific single contributor sent 60 measurements for a single station. At three of the nine stations, two or fewer measurements were recorded; no text messages were collected by observers outside our research group.

To validate measurement accuracy, a pressure transducer was installed at the NY1000 station from October 19 to November 25, 2011 at a frequency of 30 min (Figure 2). This gauging station was chosen because it was the most popular of the stations. The root mean square error (RMSE) between the pressure transducer water levels and crowdsourced water level measurements was 0.016 feet. These results were based on 19 water levels measurements from 9 observers over the 37-day period. The mean of the errors between the crowdsourced water levels and transducer measurements was not significantly different than zero using a Student's t-test ($p = 0.90$), indicating that the crowdsourced data were not biased high or low relative to the transducer data.

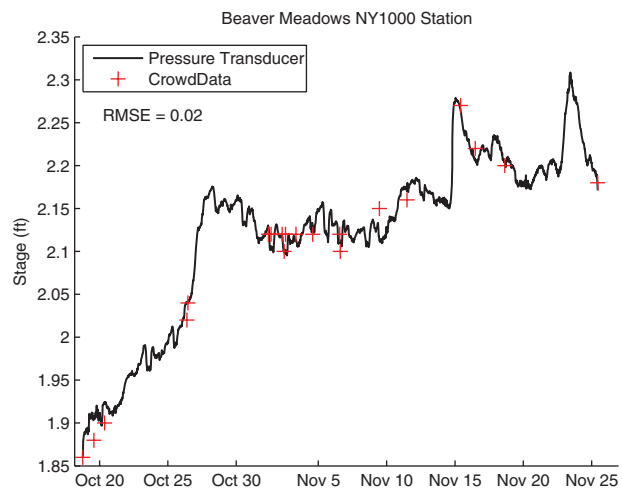

Figure 2. Beaver Meadows station NY1000 crowdsourced and pressure transducer water level measurements.

Discussion

Station Popularity

Location was the most important factor when assessing popularity of a station. The most popular station—NY1000—was located at the Buffalo Audubon Society's Beaver Meadows nature center. The nature center likely primes visitors, already in a conservation mindset, to participate. This site also experienced a unique transformation as a result of beavers returning to the habitat after being absent for over 10 years, which contributed to site popularity. The CrowdHydrology gauging staff was located only a few meters away from the first location that the beavers chose to build their new dam. As a result, this location received a substantial amount of traffic from visitors hoping to see the beavers or interested in the progress of the dam and lodge construction. The other eight locations were chosen based on popular trout-fishing locations. However, these gauging stations were not installed until August of 2011, which was near the end of the trout-fishing season in Western New York. More traffic is expected at these stations during a full trout-fishing season.

Dr. Smith Effect

Having an engaged citizen who is willing to send in a measurement is what we call the "Dr. Smith" effect. The success of any CrowdHydrology station location is based on both traffic passing by the station and the potential citizen scientist's motivation to send a text message. Choosing locations of stations based on easy access and other appealing associated factors (such as fishing locations) is within our control. However, it is more difficult to encourage observers to send text messages. A prime example is the comparison of NY1000 (Beaver Meadows) and NY1001 (Cattaraugus Creek). Both sites are high traffic locations: the Beaver Meadows site is on a

popular nature trail and the Cattaraugus Creek site is an excellent trout-fishing location at the edge of a local Farm store parking lot. Over the same period of time (August 10, 2011 to December 15, 2011), the Beaver Meadows site had almost 10 times the reported measurements taken as compared with the Cattaraugus Creek station (59 and 6, respectively). The owner of the Farm store indicated that his wife wanted to know if anyone ever sent messages because she walked past the Carraraugus Creek gauge every day—apparently she was not motivated to send messages herself. This exchange highlights issues relating to public engagement at the Cattaraugus Creek station. Public engagement was missing even though traffic was passing by the site. By comparison, the Beaver Meadows site experienced excellent public engagement. A single individual walked past the gauge three or four times a week and sent text messages from a single phone number. As a result, we have a fairly consistent record of water levels. The level of participation at any given station relies on placement of the staff gauge in high traffic areas and finding, or more likely hoping to find, a "Dr. Smith" who will become engaged in the project. These results suggest the need to find additional means to motivate participants in the future.

Accuracy

Validation of these data with a pressure transducer shows promising results; however, sampling period irregularities do exist. The RMSE as compared to the pressure transducer data (0.016 feet) is less than the highest resolution tick marks on the USGS Style A staff gauges used in the study (smallest numbered increment is 0.02 feet and, with training, resolution of 0.01 feet is possible). Engel and Voshell (2002) have questioned the accuracy of crowdsourced data, but data points at the reference location have confirmed high quality. This level of accuracy is encouraging since no training was given to the citizen scientists. However, the lack of training may explain the 17 messages that were sent in with insufficient data. There is an irregular sampling period and some hydrologic events, such as storm peaks, are unaccounted for in our crowdsourced data (see October 26 to November 3 and November 20 to 24 Figure 2). This irregular sampling period is an artifact of the method and cannot be removed. There was some anecdotal evidence for the external factors that control the regularity of measurements. During the summer of 2011 Western New York had record high temperatures that corresponded with a reduction in the number of text messages received. This may be a result of fewer people going outdoors because of the hot weather. Crowdsourced stage measurements may be better for baseflow conditions than for storm-related conditions.

Cost

The cost of setting up this crowdsourced network was modest because we were able to remove the most expensive budget items: personnel, telemetry, and pressure transducers. Each gauging station cost approximately $65. This included the gauging staff, mounting post, signage, and miscellaneous hardware. In terms of maintenance, we were fortunate to receive a text message that one of our gauges was damaged as a result of a storm. In the future, maintenance of the site may also be crowdsourced. The individual observer covers the telemetry costs and all text messages are handled by free available software (Google voice and Gmail). There was some associated cost for housing the web site on our servers but those may be minor as most institutions that have designated IT facilities and servers.

Conclusions

Spatially and temporally distributed measurements can come at a significant cost for equipment and personnel time. Such measurements often require sophisticated instrumentation and monitoring. These instruments are needed for measuring complex phenomena such as geochemical cycling. Less complex phenomena such as precipitation and stream stage can be monitored with simpler methods. Despite the limitations of such basic measurements, these data may be leveraged with additional information, such as stage/discharge relationships, which may enhance the utility of the information for hydrogeologic evaluation.

We have presented a crowdsourcing project monitoring stream stage in Western New York. Through the use of untrained observers who sent text messages of stream stage at designated locations throughout three watersheds, we were able to collect preliminary data showing the potential for crowdsourced hydrologic measurements as a tool for populating basic hydrologic networks. The network was set up at a minimal cost and the data were of high quality. Other examples, such as measurements of precipitation, have a long track record (e.g., CoCoRaHS) of successful application.

The number of measurements collected over our field season was most successful at a single gauging station (NY1000) and one user dominated data submission. A single observer can significantly add to the success of a given gauging station at an accessible location. It is important to maintain a balance between accessibility and the need for a geographically specific measurement location. Crowdsourcing may not work if geographic constraints limit the accessibility. However, we see potential for using these methods in areas such as national forests where hikers using remote trails may be the only source of data submission. Further research directions include both active recruitment of participants and collaboration with social scientists to better understand why some people participated while others did not, even in areas with substantial traffic.

The RMSE of crowdsourced data was on the same level as the gauging staff measurement gradations when compared with pressure transducer data. Crowdsourced data missed some high stage events, due to a lack of observers during rain events. This is characteristic of a method that uses passive engagement. Active recruitment might encourage more measurements in broader weather

conditions. Nonetheless these data are useful as long-term data sets that show weekly or monthly changes in water levels. These data would also be useful in investigating the characteristics of baseflow within a given watershed. The technology driving this project is not limited to stream stage or even to hydrologic data. The technology is transferable to any data that can be conveyed as a single value associated with a station number.

A key outcome of this project is the broader impact of engaging the general public in hydrologic research. Each of the 68 individuals who sent in a text message of water levels took the time to send in an observation. This is a starting point for discussions relating to water resources and human interactions within a local watershed. This public engagement cannot be achieved by putting a stilling well with a pressure transducer at the edge of a stream. Crowdsourced data supplements the collection of hydrologic data and shows promise for stimulating public engagement. Once engaged, the potential for "real-time" educational outreach to the same citizen scientists using associated websites or mobile applications may facilitate greater interaction between scientists and the public. With advances in mobile phone technology come the potential for significant gains in collecting spatially and temporally distributed measures in addition to interacting with engaged citizens, if we can only tap into our crowdsourcing potential.

Acknowledgments

Funding for this project came from the University at Buffalo. Access to the field sites was provided by the Buffalo Audubon Society and the New York State Department of Environment and Conservation. We thank Loren Smith and Charlotte Hsu for helping to publicize the project and Dave Yearke for computer support. The authors also thank Paul Juckem (USGS) and two anonymous reviewers for helpful comments and suggestions that greatly improved this paper.

Disclaimer

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

Audubon Society. 2011. Christmas Bird Count. http://birds.audubon.org/sites/default/files/documents/ChristmasBirdCount.pdf (accessed December 30, 2011).

Cifelli, R., N. Doesken, P. Kennedy, L.D. Carey, S.A. Rutledge, C. Gimmestad, and T. Depue. 2005. The community collaborative rain, hail and snow network: Informal education for scientist and citizens. *Bulletin of the American Meteorological Society* 86, no. 8: 1069–1077.

Citizen Sky. 2009. http://www.citizensky.org/ (accessed June 5, 2012).

Cohen, A. 2011. fuzzywuzzy. https://github.com/seatgeek/fuzzywuzzy (accessed January 3, 2012).

Creek Watch. 2010. http://creekwatch.org (accessed June 5, 2012).

CoCoRaHs. 1998. http://www.cocorahs.org/ (accessed June 5, 2012).

eBird. 2007. http://ebird.org/content/ebird/ (accessed June 5, 2012).

Engel, S.R., and J.R. Voshell Jr. 2002. Volunteer biological monitoring: can it accurately assess the ecological condition of streams? *American Entomologist* 48: 164–177.

Galazy Zoo. 2007. http://www.zooniverse.org (accessed June 5, 2012).

Graham, E.A., S. Henderson, and A. Schiloss. 2011. Using mobile phones to engage citizen scientists in research. *EOS Transactions of the American Geophysical Union* 92, no. 38: 313–315.

Land, K.A., Slosar, C. Lintott,D. Andreescu, S. Bamford, P. Murray, R. Nichol, M.J. Raddick, K. Schawinski, A. Szalay, D. Thomas, and J. Vandenberg. 2008. Galaxy Zoo: The large-scale spin statistics of spiral galaxies in the Sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society* 388, no. 4: 1686–1692.

OpenDinosaur. 2009. http://opendino.wordpress.com/ (accessed June 5, 2012).

Raddick, M.J., G. Bracey, P.L. Gay, C.J. Lintott, P. Murray, K. Schawinski, A.S. Szalay, and J. Vandenberg. 2010. Galaxy Zoo: Exploring the motivations of citizen science volunteers. *Astronomy Education Review* 9, 010103–1. DOI: 10.3847/AER2009036.

RainLog. 2005. http://rainlog.org/ (accessed June 5, 2012).

Wersma, Y.F. 2010. Birding 2.0: Citizen science and effective monitoring in the web 2.0 world. *Avian Conservation and Ecology* 5, no. 2: 13.

Weather Spotter Network. 1970. www.nws.noaa.gov/skywarn (accessed June 5, 2012).

Wildlife Crossing. 2010. http://www.wildlifecrossing.net/ (accessed June 5, 2012).

Wood, C., B. Sullivan, M. Iliff, D. Fink, and S. Kelling. 2011. eBird: engaging birders in science and conservation. *Plos Biology* 9, no. 12: 1–5.