1. Project Details

1.1 Name of project: Developing a Cloud-based Open-Source Platform for an Automatic High-throughput Monitoring System to Safeguard Stream Water Quality

1.2 Project lead and contact details: Tao Wen (Syracuse University, twen08@syr.edu)

1.3 Project partners and contact details:

- Lingzhou Xue (Penn State University, lingzhou@psu.edu)
- Amal Agarwal (eBay, amalag.19@gmail.com)
- Josh Woda (USGS, jwoda@usgs.gov)
- Anthony Castronova (CUAHSI, acastronova@cuahsi.org)
- Shuang Zhang (Texas A&M University, shuang-zhang@tamu.edu)

1.4 Proposed start and end date: September 2021 – March 2022

1.5 Budget Requested: $10,000

Deliverable	Requested Funding
Developing R-based data preprocessing pipeline	$2,000
Developing R Shiny app for online demonstration and application	$5,000
Publishing R package for offline application	$3,000

2. Project Outline

2.1 Project description: Streams and rivers integrate the products of natural processes and anthropogenic activities within their corresponding watershed. The investigation of stream water quality data (e.g., chloride and sodium concentrations, specific conductivity) can shed light on the source, mixing, and transportation/migration of the released material from human activities, among which unconventional oil and gas (UOG) production and road salting has caused public concerns related to stream water quality. Wastewater leaked from well pads and salts spread for road deicing occasionally causes surface water quality impairments, e.g., increased level of salinity and toxic element in downstream waters. Detecting these impairments using the existing surface water quality data for multiple locations in a large region is computationally challenging as it requires the integration and analysis of datasets of various types including water quality data, potential polluter locations, and stream flowlines in usually complex stream networks.

The ongoing development of automatic sensor devices in U.S. streams provides a much larger and denser water quality dataset (Figure 1). Combined with the publicly accessible geoscience database (e.g., Water Quality Portal), the advancement of cloud computing (e.g., Amazon AWS) and open-source resources (e.g., R Shiny, GitHub) for web application development makes it possible to develop computer algorithms to ***automatically*** detect stream water quality impairments for the community [e.g., GeoNet model from Agarwal et al. (2020)[1] that is written in R]. However, before being integrated to the stream network for subsequent data analysis, water quality datasets are usually untidy, hampering the application of Artificial Intelligence (AI) and machine learning tools to help detect water impairments from human activities, e.g., UOG development and road salting. Current research efforts (i.e., GeoNet) also lack the consideration of additional hydrological controls, e.g., precipitation. In addition, applying the GeoNet model (in its current form) to a larger stream network is largely hampered due to its relatively low-efficiency coding. Members in the community of hydrology, environmental science, and beyond might find it hard to apply this GeoNet model to other regions due to the lack of a "plug-and-go" R package for the GeoNet. Furthermore, to facilitate the community to learn and use the GeoNet model, an interactive R Shiny application deployed to the cloud server is also in need.

[1] Agarwal, A., et al., 2020. Assessing Contamination of Stream Networks near Shale Gas Development Using a New Geospatial Tool. Environ. Sci. Technol. https://doi.org/10.1021/acs.est.9b06761

In this proposal, we seek funding to support student(s) and faculty to work on revising and improving a previously published R algorithm – GeoNet with respect to code efficiency by utilizing the power of the Rcpp package and parallel computing. In addition, we also plan to improve the hydrologic robustness of the GeoNet algorithm by integrating the precipitation data into the stream-network-based model as precipitation can dilute the stream water that might ultimately affect the statistical inference to determine whether a tested location is a pollution source. Throughout the project, the student will be supported by a team of geoscientists, statisticians, and data scientists. In the development phase, we will utilize the Syracuse University cluster server – HTCondor for algorithm development and testing. In addition, to maximize the applicability of the project outcome to the broad community, we also plan to develop a streamlined workflow for the data acquisition, data preprocessing, and data cleaning procedures so that the protocol is reproducible. To validate the developed GeoNet algorithm, we will first compare the output of the improved GeoNet model with that of the original GeoNet (with relatively low efficiency) on the Pennsylvania stream network (Agarwal et al., 2020), and then potentially apply it to a broader geographic region across the U.S. to exemplify the improved GeoNet (Figure 1). To disseminate results from this project, we plan to publish a new R package and a publicly accessible R Shiny web application so that the community can explore the improved GeoNet algorithm both offline and online. A variety of open-source resources will be used in this project, including but not limited to GitHub, R Notebook, Singularity, Docker, and Binder.

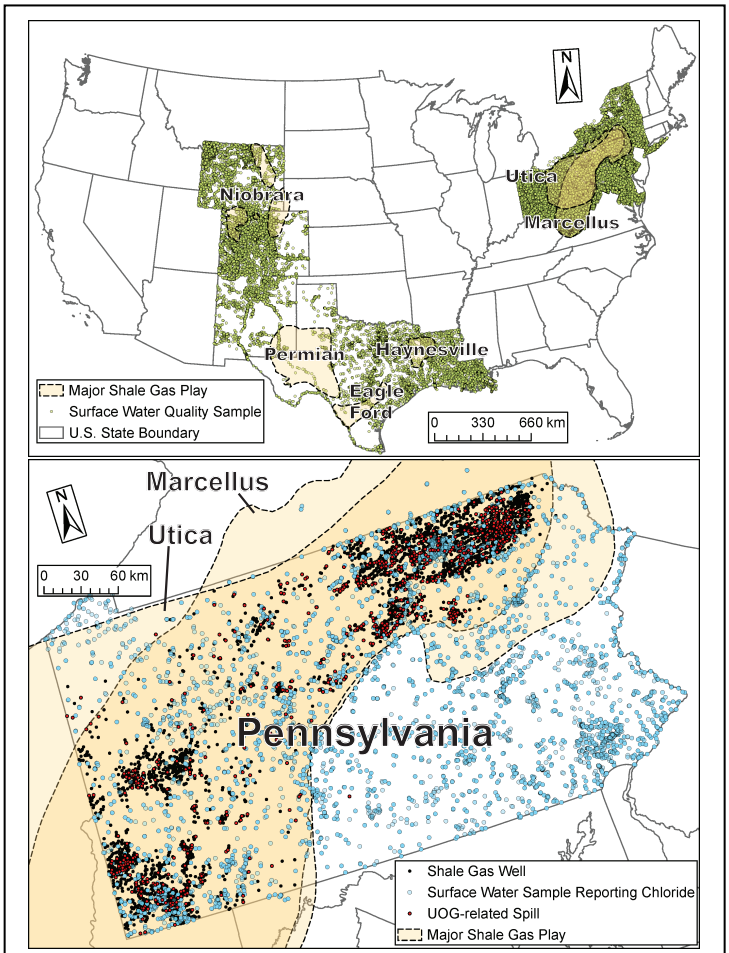

Figure 1. Top: >55,000 surface water monitoring sites available in ten U.S. major fossil-fuel production states. **Bottom:** location of unconventional oil and gas wells (black circle), surface water quality samples (blue circle), and UOG-related fluid release events (red circle) mapped in PA.

2.2 Project objectives (learning & technical): In this proposal, we seek funding to continue to work on improving a published stream-network-based geospatial-analysis tool – GeoNet by addressing three issues: i) lack of scalability of the original GeoNet when dealing with large sensor-based data (e.g., conductivity), ii) relatively low code efficiency, and iii) lack of consideration of the impact of precipitation on stream quality (i.e., via dilution). After improvements, the revised GeoNet will resolve the hydrological relationship among stream monitoring stations as well as between these stations and potential polluters (e.g., Figure 1). Our specific aims are described below:

- Aim 1: Gather, compile, clean, and pre-process stream water quality data to test the impact of UOG and road salting in ten major fossil-fuel production states in the U.S. The algorithms and workflow for accomplishing this will be shared as an R Notebook that can be discovered in both the CUAHSI HydroShare and GitHub repository, and executed in online computing environments.

- Aim 2: Improve computational efficiency of the prototype version of GeoNet by using Rcpp and parallel computing. Source codes of the revised GeoNet will be disseminated via GitHub and CUAHSI HydroShare. Our preliminary test of using Rcpp for one chunk of the GeoNet source codes shows a 100-times improvement in the running time.
- Aim 3: Evaluate the improved GeoNet implementation in Pennsylvania and compare it against Agarwal et al. (2020) to ensure model validity. Results will be disseminated during the ESIP meeting
- Aim 4: Develop an R Shiny web application for the revised GeoNet algorithm to demonstrate the GeoNet capability, engage the broader community and disseminate our scientific findings.
- Aim 5: Package the GeoNet algorithm as an R package and submit it to the CRAN repository so it can be used by the greater community to advance water science research.

2.3 Project significance and impact: The improved GeoNet algorithm to be developed in this project will be the first hydrology-informed stream-network-based geospatial-analysis tool that is capable of ***automatically*** analyzing multiple datasets including stream network, water quality, and precipitation to detect stream water quality impairments potentially related to human activities including UOG and road salting. Earth science-major undergraduate and graduate student(s) will be trained in this project, and they will be well suited with data science knowledge for future research projects. This project will develop, implement, and test community-recommended best practices to streamline the development, validation, application, and dissemination of the GeoNet algorithm by making use of multiple open-source and cloud computing resources. The outcome from this project can be used as a showcase of the best practice of performing data-driven studies in the field of geoscience. In addition, the improved Geonet algorithm can be used to detect stream water quality pollution in other geographic regions and to guide the design of more efficient stream monitoring networks. To evaluate the impact of this project, we will first compare its testing result on Pennsylvania stream networks with published results. In the long term, we can quantify the impact of this project by monitoring the usage of the to-be-or published R Shiny application and R package for the improved GeoNet algorithm.

2.4 Description of key project steps and timeline:

	Sept.	Oct.	Nov.	Dec.	Jan.	Feb.	March
Data acquisition and preprocessing	x	x					
Improving GeoNet efficiency		x	x	x	x		
Developing an R Shiny app					x	x	
Preparing GeoNet codes for submission to CRAN as an R package						x	x

2.5 Description of additional funding currently supporting this work: Additional funding includes start-up funds from PI Wen and Syracuse University research computing accessible to PI Wen for free of charge.

3. Outreach

3.1 What groups/audiences will be engaged in the project?

In this project, we plan to engage researchers from universities, relevant federal agencies (e.g., USGS), and community-based institutions (e.g., CUAHSI) to in refining the hydrological and technological aspects of the GeoNet algorithm. If applicable, we will also reach out to state-level government agencies, e.g., the Pennsylvania Department of Environmental Protection (PA DEP), as the outcome of this project is of high value to environmental regulators. In addition, environmental justice groups, environmental organizations, and the broader public will be engaged via both social media and personal contact. ESIP members from the collaboration areas including Cloud Computing and Machine Learning will also be contacted if applicable.

3.2 How will you share the knowledge generated by the project?

The outcome from this project will be shared with the ESIP community and beyond via multiple channels: (1) ESIP 2022 Summer Meeting, (2) A R Shiny web application to demonstrate the capability of the improved GeoNet algorithm, and (3) A R package to be prepared and submitted to CRAN during the duration of this project. Although it is out of the scope of this project, one peer-review journal article describing the improved GeoNet algorithm and its application in the proposed study area will be planned and submitted after the closure of this project.

3.3 Who (agencies/individuals) should be aware of this project, i.e., potential outreach targets?

- ESIP members in the collaboration areas: Cloud Computing, Machine Learning, and Data Stewardship Committee
- Environmental regulators at various levels, e.g., EPA and PA DEP. We will use existing contacts to reach out to them.
- Environmental data (particularly stream water quality data) provider, e.g., USGS. We have team members and additional contacts from USGS.

4. Project Partners

4.1 Description of project partners (agencies/individuals) and their involvement

Tao Wen is an Assistant Professor in the Department of Earth and Environmental Sciences at Syracuse University. As Project Lead, Wen will plan and monitor the progress of this project in collaboration with the other five Project Partners. Josh Woda is a trained hydrologist from USGS, and he will contribute to this project by supervising the student(s) on hydrology and water quality. Lingzhou Xue, an Associate Professor in the Department of Statistics at Penn State, and Amal Agarwal, a Data Scientist at eBay both have expertise in statistical inference and network-based analysis. The team of Xue, Agarwal, Wen, and other researchers co-developed the earlier version of GeoNet. In this project, Xue and Agarwal will guide the student(s) on the algorithm development and improvement. Shuang Zhang is an Assistant Professor in the Department of Oceanography at TAMU. Together with Wen, Zhang will supervise the student(s) on stream water quality and data acquisition. Anthony Castronova is a hydrologist and data scientist at the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI). Castronova will contribute his expertise in software development and data science to this project, specifically guidance for disseminating the results of this project using CUAHSI services such as the HydroShare platform and hosting R Shiny web application in CUAHSI's community app gallery. Among the project partners, Woda and Castronova are from USGS and CUAHSI, respectively, both of which are ESIP sponsors.

4.2 How will this project engage members of the ESIP community?

In this project, we plan to first establish the engagement with the ESIP community by reaching out to the ESIP members of collaboration areas of the Data Stewardship Committee as one of the aims of this project is to compile water quality data for the subsequent model development and testing. Input from the members of the Data Stewardship Committee will be sought. Members of other ESIP collaboration areas including Cloud Computing and Machine Learning will also be engaged as the corresponding stage of this project is ongoing. The results from this project will be disseminated to all participants in the ESIP 2022 Summer Meeting. Feedback from the audience will be collected to further improve the revised algorithm. In addition, the developed R Shiny application and R package will also be shared with all members of the ESIP community. User feedback from the ESIP community will help guide the post-project development of the to-be-improved GeoNet algorithm. As reviewing the existing collaboration areas listed on the ESIP website (https://www.esipfed.org/get-involved/collaborate), we conclude that there might be a broad community-level interest to establish a new cluster particularly targeting at either the reporting standard of stream water quality data or the data-driven studies of the interaction of energy-water-soil, which we deem might be helpful to broaden ESIP member engagement.