CDI FY17 Request for Proposals

Open-source machine-learning toolkit for analysis and forecasting of environmental time series data

Submission Title: Open-source machine-learning toolkit for analysis and forecasting of environmental time series data

Lead PI: Kevin Schmidt

Mission Area: Core Science Systems

Region: Pacific Region

Organization: Geology, Minerals, Energy, and Geophysics Science Center

OrcId: 0000-0003-2365-8035

Phone: 6503295302

Email: kschmidt@usgs.gov

City: Menlo Park

State: CA

Co-PIs and Collaborators:

Type: CO-PI

Name: Collin Cronkite-Ratcliff

Mission Area: Natural Hazards

Region: Pacific Region

Organization: Geology, Minerals, Energy, and Geophysics Science Center

OrcId: 0000-0001-5485-3832

Phone: 6503295028

Email: ccronkite-ratcliff@usgs.gov

City: Menlo Park

State: CA

Type: Collaborator

Name: Aniruddha Basak

Mission Area: Not Applicable

Region: Pacific Region

Organization: Carnegie Mellon University Silicon Valley

OrcId: 0000-0001-7695-4087

Phone:

Email: aniruddha.basak@sv.cmu.edu

City: Moffett Field

State: CA

Type: Collaborator

Name: Ole Mengshoel

Mission Area: Not Applicable

Region: Pacific Region

Organization: Carnegie Mellon University Silicon Valley

OrcId: 0000-0003-2666-5310

Phone: 650-335-2887

Email: ole.mengshoel@sv.cmu.edu

City: Moffett Field

State: CA

Science Support Framework Element 1: Data

Science Support Framework Element 2: Science Data Lifecycle - Analysis

Science Support Framework Element 3: Science Data Lifecycle - Processing

In-Kind Match: $21,500.00

List of anticipated deliverables from the project: The required academic professor and student contributions, as collaborators from CMU Silicon Valley, are expected to be in the areas software development, data analysis, scripting tools, and mathematical statistical and modeling. The student will write and implement computer code to employ techniques such as Antecedent Water Index, Probabilistic Graphical Models, and Simulated Annealing optimization method. This may involve design and implementation of MySQL computer code. Machine learning optimization would be used to generally evaluate both rising and falling limb functions for any time-series data. The student would write and test computer code, perform experiments, and write a summary report including a description and distribution of the open-source software. Software products would be released as an R package through open-source software development, such as the USGS Bitbucket repository.

Lead Cost Center: GMEG- GGWSZT0000

Notes, Comments: We are excited by the possibilities of utilizing a large amount of existing time-series data in a unique collaboration blending earth science and machine learning expertise.

Project Description: Measured time-series data embody a wide range of geomorphic, hydrologic, and ecosystem processes and rates. For instance, spatial and temporal variability of time-series hydrologic data regulate the onset and duration of surface water flow, water availability for plants (e.g., ecosystem disturbance), and mass wasting. Although theoretical relationships exist to quantify the relation between rainfall and soil moisture (i.e., Richard’s equation), vegetation, variable infiltration, and soil stratigraphy preclude simple use. USGS researchers collect varying environmental soil moisture data, but no common means of evaluation exists to determine the magnitude, duration, and depth of moisture persistence. Inexpensive data loggers can collect large quantities of data (10’s of measurements per time step over 100,000’s of time steps) at a high temporal resolution, but USGS ability to easily reduce such large time-series data sets to representative metrics or models does not exist. We propose to derive a suite of machine-learning analysis tools for the associated rising and falling limbs of time-series response, bridging from the specific to the general. Our test bed data would be from post-wildfire monitoring sites. Parameter optimizations representing logarithmic wetting and drying functions would be used to evaluate response to individual rain storms, but also seasonal variations as a function of soil depth in comparison with vegetation regrowth and debris-flow producing storms.

Total Budget: $45,000.00

CDI16-SK5491

Open-source machine-learning toolkit for analysis and forecasting of environmental time-series data

PI: Kevin Schmidt, USGS, GMEG, Menlo Park, CA 94025, kschmidt@usgs.gov

Time-series data represent some of the most important forms of information chronicling the health and trajectory of earth and biological systems. In recognition of its importance in parallel with technologic advances for monitoring environmental data, the quantity and variety of time-series data is rapidly expanding. These data represent critical information for our work in each of the USGS mission areas. As a major provider of widely ranging time-series data to the public, the USGS plays a critical role in providing cutting edge tools for interpreting these complex data. We propose to develop and disseminate an open-source statistical computing package for decomposing and forecasting environmental time-series data, with emphasis on soil moisture as an analysis test bed. We would leverage existing soil-moisture data collected for a variety of USGS GMEG projects addressing post-fire erosion, shallow landsliding, and ecologic research questions.

In order to better analyze often diverse and complex data, recent USGS-Carnegie Mellon University collaborations have sought to introduce machine learning approaches to existing earth science analysis capabilities to readily identify: i) data outliers including instrumental error, ii) representative metrics of non-stationary time series, iii) means to parse functional packets (e.g., rising and falling limbs of response), with the goal of forecasting future behavior with greater accuracy over existing methods. Soil moisture data is one example of data that could benefit significantly from improved data analysis products. Although soil moisture is of critical importance across many fields, including a changing climate, landslide hazard assessments, landscape ecology, agriculture, water resources, and wildfire management, widespread in-situ monitoring of soil moisture has only begun in the last few years. The challenges of analyzing and forecasting soil moisture presents an ideal opportunity to take advantage of recent advances in machine learning. Software developed here would augment the science data lifecycle model by searching data for outliers and faulty instrumental measurements, preserving local maxima and minima through trend analyses, and decomposing to representative metrics available to the public.

Current work with partners at Scripps Institution of Oceanography (SIO) demonstrates a need for software with capabilities described above. In our work with SIO, we analyze the statistics of the antecedent seasonal rainfall when soil moisture exceeds given critical water contents (CWC). This analysis has significance for landslide hazard assessment because in some regions, landslide susceptibility increases once soil moisture exceeds threshold CWC. We are currently conducting this analysis using data from three of NOAA's Hydrometeorology Testbed (HMT) sites (http://hmt.noaa.gov/), where rain gauges and soil moisture sensors are co-located. However, because the soil moisture data are missing for some time periods, a reliable machine-learning model is needed to forecast the response of soil moisture to the available rain data.

Existing methods, however, such as moving averages, fail to smooth time series data while preserving peaks and valleys. We have developed novel method and software, HyperSTL, for extrema-preserving smoothing by optimizing the parameters of a decomposition technique, and Seasonal Decomposition of Time Series by Loess (STL). HyperSTL integrates machine learning and statistical algorithms and software, and optimizes an objective function over STL parameters using the decomposed components. HyperSTL successfully reduces noise, including instrumental variations, while preserving extrema and signal detail. So far, we have successfully demonstrated our novel method on post-fire soil moisture time-series data.

In this proposal, we aim for the development of HyperSTL Version 2.0. The key capabilities of the current HyperSTL (Version 1.0) will remain, but Version 2.0 will enable analysts, engineers, and scientists to perform a broader range of time series analysis functions. Specifically, the new software will enable users to explicitly state the goals of time series decomposition by formulating an objective function; decomposes data into components that capture, for example, long-term and daily variations separately; and reduces diurnal variations from data while maintaining the shape and amplitude of extrema (peaks and valleys). It will be easy for users not familiar with R, in which HyperSTL is implemented, to get started with the software and develop forecasting models for their datasets. To support the USGS mission of making our science publicly accessible, software products would be released as an R package through open-source software development, such as the USGS Bitbucket repository. R is an open-source programming language that is widely used throughout the scientific community and hosts one of the largest open-source libraries of scientific computing tools.

Funding requested here would provide support for graduate student (Basak) machine-learning capabilities at Carnegie Mellon Silicon Valley leveraged with USGS existing data streams and salary (Schmidt and Cronkite-Ratcliff). The resulting computational tools will offer data analysis capabilities relevant not just for soil moisture, but for a range of time-series data currently collected and distributed by USGS and our partners. These include, but are not limited to: seismic strong ground motion, streamflow; soil, water, and air temperature; snow and ground reflectance; and water quality indicators, vegetation indices, and ecological metrics. Distributing this product as an open source, web available product will make it readily available to the public and will facilitate efficient feedback, improvement, and extension from the broader community. It will also give us an opportunity to engage in the rapidly growing movement towards web-available science by providing an adaptable toolkit to enable discovery of trends and metrics in otherwise complicated datasets.

BUDGET

Budget Category	Federal Funding “Requested”	Matching Funds “Proposed”
1. PERSONNEL (SALARIES including benefits):	$42,000	
Federal Personnel Total:	$ 0	$19,000
Contract/Collaborator Personnel Total:	$42,000	$
Total Salaries:	$42,000	$19,000
2. TRAVEL EXPENSES:		
Travel Total (Per Diem, Airfare, Mileage/Shuttle) **x** # of Trips: CDI FY17 workshop and AGU FY17 Annual Meeting	$2500 (Basak &Mengshoel)	$2000 (Schmidt & Cronkite-Ratcliff)
Other Expenses (e.g. Registration Fees):	$500	$500
Total Travel Expenses:	$3000	$2500
Total Other Direct Costs:	$45,000	$0
Total Direct Costs:	$0	$0
Indirect Costs (%):	20%	25%
GRAND TOTAL:	$45,000	$21,500