Environmental Modelling & Software 26 (2011) 1583–1598

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

A data-driven approach for modeling post-fire debris-flow volumes and their uncertainty

Michael J. Friedel

Crustal Imaging and Characterization Team, United States Geological Survey, Denver Federal Center, Box 25046, MS 964, Lakewood, CO 80225, United States

ARTICLE INFO

Article history:
Received 30 December 2010
Received in revised form
20 July 2011
Accepted 21 July 2011
Available online 23 August 2011

Keywords:
Wildfire
Debris-flow volume
Self-organizing map
Genetic programming
Multivariate
Prediction
Nonlinear models
Nonlinear uncertainty

ABSTRACT

This study demonstrates the novel application of genetic programming to evolve nonlinear post-fire debris-flow volume equations from variables associated with a data-driven conceptual model of the western United States. The search space is constrained using a multi-component objective function that simultaneously minimizes root-mean squared and unit errors for the evolution of fittest equations. An optimization technique is then used to estimate the limits of nonlinear prediction uncertainty associated with the debris-flow equations. In contrast to a published multiple linear regression three-variable equation, linking basin area with slopes greater or equal to 30 percent, burn severity characterized as area burned moderate plus high, and total storm rainfall, the data-driven approach discovers many nonlinear and several dimensionally consistent equations that are unbiased and have less prediction uncertainty. Of the nonlinear equations, the best performance (lowest prediction uncertainty) is achieved when using three variables: average basin slope, total burned area, and total storm rainfall. Further reduction in uncertainty is possible for the nonlinear equations when dimensional consistency is not a priority and by subsequently applying a gradient solver to the fittest solutions. The data-driven modeling approach can be applied to nonlinear multivariate problems in all fields of study.

Published by Elsevier Ltd.

1. Introduction

The hazardous consequences of rainfall on basins motivate investigators to study its cause-and-effect on various hydrologic responses (Friedel, 2008; Gartner, 2008; Biswajeet and Lee, 2010; Fotopoulos et al., 2010; Scott et al., 2009; Stenson et al., 2011). Of these responses, debris flows have the most hazardous consequences and may be one reason why several modeling studies focus on this response. A review of traditional debris-flow modeling is provided by Bulmer et al. (2002). To date, the types of modeling include physical, empirical, and numerical approaches. Early physical models consider debris flows as a single phase Bingham (or Coulomb) continuum (Johnson, 1984). Takahashi (1980) considers particle–particle interactions but for homogenous mixtures without internal pressure on the fluid-matrix mixture. Later, these modeling assumptions are generalized by Iverson (1997) to include viscous pore fluid in a fluid-solid momentum transport approach. Because the debris-flow dynamics are nonlinear, time-dependent, and spatially varying, many researchers began digital investigations involving empirical and numerical approaches.

Empirically-based models are developed by fitting equations to field data for predicting post-fire debris-flow generation at the outlets of burned basins. While not explicitly describing the physics of debris flows, these models can provide a first-order prediction of debris-flow behavior. Some post-fire examples include equations devised using multiple linear regression (MLR) to predict debris-flow peak discharge as a function of variation in basin landform, burn severity, and rainfall (Cannon et al., 2003; Gartner, 2005). Although peak discharge is successfully used to model extreme flooding (Friedel et al., 2008) and extreme rainfall events (Friedel, 2008), many researchers consider it too uncertain for predicting post-fire debris flows (Pierson, 2004). In a recent study by Friedel (2010), the average range of prediction uncertainty in Colorado debris-flow peak discharge measurements is determined to span a factor of about six. For these reasons, there is a shift away from peak discharge in favor of alternative response variables that include the percent chance for debris-flow production (Cannon et al., 2004) and total volume of debris flows (Gartner et al., 2008).

Numerically-based models are also used to predict the timing and spatial movement of debris flows in response to rainfall on burned basins (Bunch et al., 2004; Elliott et al., 2005; Mikos et al., 2006; Rosso et al., 2007; Bathurst et al., 2007; Hsu et al., 2010). These models differ from standard basin hydrologic models with the addition of a friction slope. The friction slope term depends on

E-mail address: mfriedel@usgs.gov.

1364-8152/$ – see front matter Published by Elsevier Ltd.
doi:10.1016/j.envsoft.2011.07.014

which rheological model is chosen to represent the shear stress of a non-Newtonian fluid. One numerical modeling application is the creation of post-fire debris-flow inundation maps for basins burned during the 2002 Colorado wildfires (Elliott et al., 2005). In that study, the numerical problem is solved in two steps. First, the peak-discharge hydrographs associated with a 100-year storm event are estimated at the outlets of tributary basins burned as part of the Coal seam, Hayman, and Missionary Ridge wildfires. Second, the hydrographs are bulked and then used as input to an unsteady, unconfined, two-dimensional flow and transport model for predicting the timing and spatial extent of debris flows. The challenges in that study illustrate those common to other post-fire numerical modeling efforts: (1) poor spatial rainfall resolution, (2) poor spatial resolution of physical properties, (3) assumed homogeneity of the debris-flow field, (4) little or no streamflow and debris-flow information at basin outlets to calibrate and validate the model.

Despite the progress made in modeling post-fire debris flows, there remains a need for additional improvement (Han et al., 2007). This is particularly true with respect to the development of alternative nonlinear models and quantification of prediction uncertainty. Over the past decade, data-driven techniques have been introduced as alternative tools in hydrology (Dawson and Wilby, 2001; Han et al., 2007). One data-driven technique is the self-organizing map. The self-organizing map is a type of unsupervised neural network that maps nonlinear data vectors from a high- to low-dimensional model output space (Kohonen, 2001). Some applications include investigating the spatial and temporal trends in basin water quality data (Lischeid, 2003), estimating design hydrographs for ungauged basins (Lin and Wu, 2007), assessing the vulnerability of rainfall-induced debris flows (Lu et al., 2007), and developing post-fire landscape models at multi-state (regional) scales (Friedel, 2011). A more comprehensive review of applications in water-resources is provided by Kalteh and Berndtsson (2008). A second data-driven technique is symbolic regression. Symbolic regression is one type of genetic programming (GP) that searches for empirical relations using a specific form of the evolutionary algorithm (Koza, 1992, 1999). These algorithms share the common property of applying selection, variation, and reproduction to a population of structures that undergo evolution. Recent applications include the evolution of equations to estimate soil hydraulic properties (Parasuraman et al., 2007), estimate suspended sediment concentration (Aytek et al., 2008), forecast short-term streamflow with global climate change implications (Makkeasorn et al., 2009), and to project climate change impacts on landlocked salmon (Tung et al., 2009).

A common interest in the field of hydrology is the estimation of prediction uncertainty (Vecchia and Cooley (1987; Christensen and Cooley, 1999a,b; Cooley, 2004; Friedel, 2005; Friedel, 2006a,b; Gallager and Doherty, 2007; Friedel et al., 2008; Yu et al., 2008; Sreekanth and Datta, 2011). Part of the motivation for this analysis is the recognition that empirical (and numerical) models are non unique. That is, there are many alternate combinations of model coefficients (or parameter values) that can satisfy the same best-fit criteria. Because predictions made using a given model represent one set of many, there is range over which they vary. In this study, the following objectives focus on burned basins in the western United States: (1) evolve a set of nonlinear multivariate debris-flow volume equations; and (2) quantify and compare model statistics and prediction uncertainty among these nonlinear equations to a published linear equation. This study extends the work of Gartner (2008) who sought to devise debris-flow equations based on the traditional multiple linear regression approach. We demonstrate the novel application of genetic programming to evolve nonlinear post-fire debris-flow volume equations from variables associated with a data-driven conceptual model of the western United States (Friedel, 2011). In addition to providing new equations, this study illustrates the applicability of an inverse technique for estimating nonlinear post-fire debris-flow prediction uncertainty. The general nonlinear modeling approach can be applied to multivariate problems in all fields of study.

2. Conceptual models and data

The selection of variables for use in this study is based on a conceptual post-fire landscape model provided by Friedel (2011). In that study, conceptual models are delineated at the multi-state (regional) scale using data from six hundred burned basins in nine western states (Gartner et al., 2005), the self-organizing map technique (Kohonen, 2001), the partitive cluster technique (Vesanto and Alhoniemi, 2000), and the Davies-Bouldin criteria (Davies and Bouldin, 1979). Given that the conceptual model variables are delineated in terms of probability (low, moderate, high), the model selection process proceeds by successive screening for variables with high (or moderate) likelihood in three categories.

1. Runoff-initiated debris-flow volume discharge (the response variable). Models characterized by a high likelihood for flooding or nonevents are dropped from consideration.
2. Occurrence in California, Colorado, and Utah. Models characterized by debris flows occurring in an alternate combination of states are dropped from consideration.
3. Susceptibility (explanatory) variables in common with a linear debris-flow volume equation (Gartner et al., 2008) for western United States. Models that are not characterized by a high likelihood for basin slopes greater or equal to 30 percent, burn severity characterized as area burned moderate plus high, and total storm rainfall are dropped from consideration.

Of the eight conceptual regional landscape models in Friedel (2011), only number five (RLM-5) depicts basin debris-flow response and susceptibility variables that are similar to those used in the linear debris-flow volume equation presented by Gartner et al. (2008):

$$\ln V = 0.59{*}\ln(G30/1e6) + 0.65(\mathrm{BMH}/1e6)^{1/2}{*}(\mathrm{TSR}{*}1000)^{1/2}{+}7.21 \quad (1)$$

where V is the debris flow total volume in m^3; G30 is the basin area with slopes greater or equal to 30 percent in m^2; burn severity characterized as area burned moderate plus high, in m^2; and TSR is the total storm rainfall in m. A summary of the conceptual model is given below.

RLM-5: This conceptual regional landscape model includes mostly Montana basins but also some basins from Arizona, California, Colorado, and Utah. These basins typically are underlain by metamorphic and sometimes sedimentary rocks that have a high likelihood for post-fire runoff-initiated debris-flow events. The independent *land surface* features that best characterize this region are small basins with medium gradients and relief ratio that are highly ruggedized; and medium values of basin gradients exceeding 30 and 50 percent. The *geologic texture* is best characterized by high variable values of organic matter, permeability, soil thickness, and hydrologic group; and high values of clay content, erodibility, and soil thickness. The *rainfall* variables include high values of total storm amount, storm duration, average storm intensity, and recurrent rainfall; and a medium likelihood for *post-fire* medium to high and high burn severity areas.