Environmental Modelling & Software 101 (2018) 169–182

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Improving predictions of hydrological low-flow indices in ungaged basins using machine learning

Scott C. Worland [a, d, *], William H. Farmer [b], Julie E. Kiang [c]

[a] U.S. Geological Survey, Nashville, TN, USA
[b] U.S. Geological Survey, Denver, CO, USA
[c] U.S. Geological Survey, Reston, VA, USA
[d] Vanderbilt Institute for Energy and Environment, Vanderbilt University, Nashville, TN, USA

ARTICLE INFO

Article history:
Received 28 March 2017
Received in revised form
21 December 2017
Accepted 21 December 2017

Keywords:
Low streamflow
Ungaged basins
7Q10
Machine learning
Censored regression
Variable importance

ABSTRACT

We compare the ability of eight machine-learning models (elastic net, gradient boosting, kernel-k-nearest neighbors, two variants of support vector machines, M5-cubist, random forest, and a meta-learning ensemble M5-cubist model) and four baseline models (ordinary kriging, a unit area discharge model, and two variants of censored regression) to generate estimates of the annual minimum 7-day mean streamflow with an annual exceedance probability of 90% (7Q10) at 224 unregulated sites in South Carolina, Georgia, and Alabama, USA. The machine-learning models produced substantially lower cross validation errors compared to the baseline models. The meta-learning M5-cubist model had the lowest root-mean-squared-error of 26.72 cubic feet per second. Partial dependence plots show that 7Q10s are likely moderated by late summer and early fall precipitation and the infiltration capacity of basin soils.

Published by Elsevier Ltd.

1. Introduction

Water managers rely on streamflow data to allocate water resources, define the dilution potential of catchments, set ecological streamflow limits, and ensure sustainable watershed planning (Razavi and Coulibaly, 2012; Knight et al., 2014; Kapo et al., 2015). However, many streams do not have observed streamflow data and water managers must depend on the streamflow estimates from various prediction models (Mishra and Coulibaly, 2009; Razavi and Coulibaly, 2012; Luce, 2014). Improving the predictions of streamflow in ungaged basins has been a primary objective for hydrologists for decades and international initiatives have resulted in rapid advances in this field (Sivapalan et al., 2003; Hrachowitz et al., 2013; Blöschl, 2016). The two primary modeling strategies for predicting streamflow response in ungaged basins are: (1) deterministic physically based models—i.e. calculating streamflow based on distributed hydrologic parameters, and (2) statistical regionalization—i.e. using regression models to transfer hydrologic information from gaged to ungaged basins (Razavi and Coulibaly, 2012; Farmer and Vogel, 2016). This current paper focuses on the statistical regionalization of a low streamflow statistic: the annual minimum 7-day mean streamflow with an annual exceedance probability of 90% (7Q10).

A stream's "low flow" refers to the amount of water flowing in a stream during prolonged periods of little to no rainfall during an average non-drought year. The low-flow regime for a particular stream is controlled by the physical characteristics of its basin and the local climate (Smakhtin, 2001). The 7Q10 statistic describes a basin's expected low-flow and provides a way to compare directly the low-flow regimes of different basins. This statistic is commonly used to determine permitted point-source pollutant levels in streams (Ames, 2006). There are a number of other important low-flow metrics not discussed in this paper; several examples are the 7Q10 for a particular season or month, the annual minimum 7-day mean streamflow with an annual exceedance probability of 50% (7Q2), mean annual minimum, median September streamflow, and ecologically derived values (Knight et al., 2014; Kormos et al., 2016; Murphy et al., 2013; Raines and Asquith, 1997). The contribution of this research is the comparison of statistical estimation techniques; the choice of the specific response variable would not change the

* Corresponding author. U.S. Geological Survey Lower Mississippi Gulf Water Science Center, 640 Grassmere Park #100, Nashville, TN 37211, USA.
E-mail addresses: scworland@usgs.gov (S.C. Worland), wfarmer@usgs.gov (W.H. Farmer), jkiang@usgs.gov (J.E. Kiang).

https://doi.org/10.1016/j.envsoft.2017.12.021
1364-8152/Published by Elsevier Ltd.

structure of the analysis but we cannot conjecture how specific models would perform for a different target variable.

Low-flow regionalization methods attempt to predict low-flow metrics in ungaged basins by leveraging the correlation between basin characteristics and streamflow at gaged basins (Razavi and Coulibaly, 2012). The primary goal of 7Q10 regionalization is accurate predictions and not mechanistic explanations of what controls the 7Q10, and this distinction between prediction and explanation should guide the statistical analysis (Shmueli, 2010). Regardless of outcome goal or the type of model used, all hydrologic models require assumptions. Deterministic models, for example, assume that the physical relationships between parts of a hydrologic system are adequately captured by a set of static functions and decision rules, while stochastic models may depend on assumptions about the probabilistic constraints on parameters (i.e. "priors"), the choice of the likelihood and cost functions, the numerical methods used for parameter estimation (e.g., gradient descent, maximum likelihood, numerical integration, etc.), and choices about data preprocessing and transformation. Furthermore, hydrologic models often assume some level of stationarity (Lins and Cohn, 2011). These assumptions can have significant effects on the applicability of model results, and researchers must acknowledge how their model design choices propagate into conclusions drawn from the model.

This paper evaluates the predictive performance of various association-based models (e.g., linear regression models) that leverage the covariance structure between variables to make inferences and predictions. Association-based models have proved to be a useful engineering tool for predicting 7Q10s, and have become increasingly sophisticated in the last 30 years (Hrachowitz et al., 2013). Regression methods have evolved from simple ordinary least squares (Riggs, 1973; Thomas and Benson, 1970; Hardison, 1971) to time series weighted least squares (Tasker, 1980), generalized least squares (GLS) (Stedinger and Tasker, 1985), censored regression (Kroll and Stedinger, 1999), two step GLS-logistic regression (Funkhouser et al., 2008), truncated models, and catchment clustering methods (Law et al., 2009). There has also been an increased application of geostatistical low-flow regionalization methods—primarily ordinary kriging, top kriging, and physiographical spaced-based interpolation (Castiglioni et al., 2009, 2011).

Despite the recent methodological advances mentioned above, few studies have explored machine-learning methods to predict low-flow metrics in ungaged basins. Ouarda and Shu (2009) used an ensemble of artificial neural networks for predicting various low-flow metrics in Canada, Laaha and Blöschl (2006) used regression trees to predict Q95s in Austria, Schnier and Cai (2014) used model tree ensembles to predict a complete flow-duration curve (FDC) for streams in Illinois and Texas, and Booker and Woods (2014) used random forest models to predict several components of a FDC in New Zealand. These studies contributed valuable baseline assessments of the applicability of machine learning to streamflow-statistic estimation. Yet, however, they compare only 2–3 estimation techniques, each using a unique data set—a practice that confounds direct comparison of model performance between individual studies.

In this paper, twelve different modeling methods were applied to a publicly available data set (Falcone, 2011), and the multi-model comparison approach presented by Elshorbagy et al. (2010a, 2010b). and Shortridge et al. (2016) was used to determine the predictive performance of the models using multiple assessment criteria. Several machine-learning techniques were introduced—gradient boosting machines, kernel-k-nearest-neighbors, and elastic net-that, to our knowledge, have not yet been used to predict low-flow statistics. A meta-learning M5-Cubist model was also introduced that minimizes the overall generalization error by combining the cross-validated predictions of each machine-learning model. Finally, hydrologic insights to the physical controls of low streamflow were explored through a discussion of the relative importance of predictor variables and their corresponding partial-dependence functions for each model. The novelty of this contribution is the use multiple machine-learning models, the introduction of meta-modeling approaches for the regionalization of low-streamflow statistics, the comparison with models historically used to estimate 7Q10s, and the large gains in predictive accuracy over historical methods.

1.1. Research objectives and major findings

This paper provides the 7Q10 prediction performance estimates of twelve statistical estimation techniques—four "baseline" methods (type I Tobit regression, region of influence type I Tobit regression, ordinary kriging, and an average unit-area discharge null model) and eight machine-learning models: (1) M5-cubist regression trees, (2) gradient boosting machines, (3) kernel-K-nearest neighbors, (4) random forests, (5) elastic net, support vector machines with a (6) polynomial kernel and a (7) radial basis function kernel and an (8) ensemble meta-learning M5-cubist model is also explored. The specific research objectives are,

1. Use leave-one-out cross validation (LOO-CV) to simulate the prediction of 7Q10s at ungaged sites in three states in the southeast U.S. using eleven estimation techniques.
2. Compare the predictive accuracy of each model using root mean squared error (RMSE), unit area root mean squared area (UA-RMSE), median percentage error (MPE), and the Nash-Sutcliffe efficiency coefficient (NSE), and decompose the RMSE to examine what is controlling the error for each model.
3. Discuss the relative importance and partial dependence functions of predictor variables for each model.

We found that machine-learning methods can produce more accurate predictions of 7Q10s in ungaged basins than baseline models. Variable importance measures and partial dependence plots suggest that 7Q10s are partially driven by landcover, late summer and early fall precipitation, the infiltration rate of soils, and the variability of minimum and maximum monthly temperatures.

1.2. Background of machine learning in hydrology

Machine learning—also referred to as statistical learning, data-driven modeling, and computational intelligence—refers to a set of statistical methods that are optimized for predictive performance through a cross-validated parameter tuning process (Hastie et al., 2013; Kuhn and Johnson, 2013). These methods have been called black-box approaches and criticized for having little connection to the underlying physical processes being modeled (See references in Elshorbagy et al. (2010a) and See et al. (2007) for examples of these critiques in hydrology). Regardless, machine-learning techniques have become prevalent in the hydrology literature. Artificial neural networks have been used for predictions in hundreds of water-resource studies (Maier et al., 2010; Kasiviswanathan et al., 2016; Humphrey et al., 2016; Daliakopoulos and Tsanis, 2016). Random forest models have been used to predict natural and altered streamflow regimes in ungaged basins (Carlisle et al., 2010; Eng et al., 2013; Li et al., 2016); support vector machines have been used to forecast monthly streamflow (Kalteh, 2016; Guo et al., 2011) and to downscale low-flow indices (Joshi et al., 2013); genetic algorithms have been used to calibrate rainfall-runoff models (Goswami and O'Connor,