Ecological Informatics 33 (2016) 17–23

Contents lists available at ScienceDirect

Ecological Informatics

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

MATLAB Hydrological Index Tool (MHIT): A high performance library to calculate 171 ecologically relevant hydrological indices

Mohammad Abouali, Fariborz Daneshvar, A. Pouyan Nejadhashemi *

Biosystems and Agricultural Engineering, 216 Farrall Hall, Michigan State University, United States

ARTICLE INFO

Article history:
Received 14 November 2015
Received in revised form 27 February 2016
Accepted 6 March 2016
Available online 17 March 2016

Keywords:
Ecologically relevant hydrological index
High performance computing
Big data
Ecohydrology

ABSTRACT

Ecologically relevant hydrologic indices are being used for developing environmental flow baselines. EflowStats is an R-package developed by the US Geological Survey that calculates 171 ecologically relevant hydrologic indices. Although EflowStats provides a set of easy to use functions, once it is applied to a large area with thousands of streams, it requires a long processing time. This particularly limits its application in problems that are dealing with optimizations, requiring evaluation of thousands of scenarios. As a result, a new package called MATLAB Hydrological Index Tool (MHIT) was developed, incorporating different technologies, such as vectorization, memoization, and parallelism to considerably reduce the processing time. Both EflowStats and MHIT were tested against a long-term (20 years) daily streamflow data set from the Saginaw River Watershed in Michigan. In summary, MHIT reduces the processing time by 227 times for the study area. Therefore, MHIT is much more suitable for dealing with Big Data.

1. Introduction

The concept of "Environmental Flow", which deals with the required timing, quantity and quality of the streamflow, was introduced in order to sustain the wellbeing of both human and natural ecosystems (Moore, 2004; Arthington et al., 2006; Hirji and Davis, 2009). Indices quantifying environmental flow are usually divided into five sets that characterize various aspects of streamflow (Poff et al., 1997; Olden and Poff, 2003; Henriksen et al., 2006; Kennen et al., 2009). These five categories are Magnitude (M), Frequency (F), Duration (D), Timing (T), and Rate of change (R). Each category is divided into further sub-groups: High (H), Average (A), and Low (L). Overall, 171 different ecologically relevant indices were introduced (Poff et al., 1997). Table S1 in the supplementary material summarizes how these 171 indices are distributed among the different categories.

To assist in the calculation and the analysis of these indices, the US Geological Survey (USGS) originally developed two software packages: (1) the Hydrologic Index Tool (HIT), which calculates these 171 indices, and (2) the National Hydrologic Assessment Tool (NATHAT), which uses the calculated indices to evaluate stream conditions. These software programs are commonly used to determine environmental flow standards and to evaluate past and proposed future hydrological adaptation strategies (Kennen et al., 2007; Snelder et al., 2009; Buchanan et al., 2013). In addition to these two packages, the USGS developed an R package, called EflowStats that can be used to calculate these indices (EflowStats, 2015). EflowStats is an open source software that provides a separate function for each of these 171 indices and is hosted on GitHub. Users could perform the calculation interactively; or automate the calculation by wrapping EflowStats within some extra R codes. Therefore, EflowStats is considered to be more flexible than previous software (HIT and NATHAT); hence, it was used more often in recent publications (Caldwell et al., 2015; Herman et al., 2015; Vis et al., 2015).

Although EflowStats provides great flexibility to calculate the aforementioned indices, it lacks the required coding mechanism and design to be scalable when applied to a large data set. An initial test done for this study showed that the current R implementation of EflowStats takes on average about a quarter of a minute to calculate all the 171

Abbreviations: CPU, Central Processing Unit; EPA, Environmental Protection Agency; FH1, high flood pulse count; HIT, Hydrological Index Tool; HUC, Hydrological Unit Code; I/O, input/output; INF, infinity; MA10, spread in daily flow (the ratio of the difference between 80th and 20th percentiles of the logs of mean daily flows to median); MA11, Spread in daily flow (the ratio of the difference between 75th and 25th percentiles of the logs of mean daily flows to median); MA1, mean of the daily mean flow; MA2, median of the daily mean flow; MA5, skewness; MA9, SPREAD in daily flow (the ratio of the difference between 90th and 10th percentile of the logs of mean daily flows to median); MH15, high flow discharge index corresponding to 1% exceedance value; MH16, high flow discharge index corresponding to 10% exceedance value; MH17, high flow discharge index corresponding to 25% exceedance value; MH21, high flow volume index (threshold equal to median); MH22, high flow volume (threshold equal to three times of median); MH23, high flow volume (threshold equal to seven times of median); MH24, high peak flow (threshold equal to median); MH25, high peak flow (threshold equal to three times of median); MH26, high peak flow (threshold equal to seven times of median); MHIT, MATLAB Hydrological Index Tool; NaN, Not a Number; NATHAT, NATional Hydrological Assessment Tool; NSE, Nash–Sutcliffe efficiency; PBIAS, Percent Bias; RSR, ratio of the root mean square error to the standard deviation; SWAT, Soil and Water Assessment Tool; TA1, constancy; TA2, predictability; USDA, United State Department of Agriculture; USGS, United State Geological Survey.

* Corresponding author.
E-mail address: pouyan@msu.edu (A.P. Nejadhashemi).

http://dx.doi.org/10.1016/j.ecoinf.2016.03.004
1574-9541/

indices for one stream containing 20 years of daily data. This can be acceptable when dealing with single streams or a small number of streams. However, in cases where there are thousands of streams, which can be qualified as "Big Data", the required time could become days, months, or even years. This severely limits the applicability and usability of EflowStats in watershed management that involves evaluation of thousands of management scenarios to achieve the optimum configuration. For examples, the recent applications of EflowStats in: optimization of conservation practice implementation strategies (Herman et al., 2015), large-scale environmental impact assessment (Woznicki et al., 2016), and landscape design for selection and placement of bioenergy crops (Herman et al., 2016) required hundreds of hours to perform the analysis, which can limit the real world applications. Therefore, it is very crucial to develop a high performing code that significantly reduces the computational time, which is the goal of this study. In order to achieve this goal, a new software called MATLAB Hydrological Index Tool (MHIT) is introduced.

2. Materials and methods

2.1. Study area

To test the performance of MHIT, long-term (1993–2012) daily streamflow data from the Saginaw River Watershed located in Michigan, was used. The Saginaw watershed contains 13,831 reaches, draining the total area of 16,120 km^2 to Lake Huron (Fig. 1). The Saginaw River Watershed was selected for this study because it is the largest six-digit Hydrological Unit Code (HUC-6) in Michigan and is identified as an area of concern by the US Environmental Protection Agency (EPA, 2013). In order to obtain long-term daily streamflow data, the Soil and Water Assessment Tool (SWAT) was used. SWAT is a conceptual model developed by the U.S. Department of Agriculture (USDA) for water resources management (Arnold et al., 1998). SWAT uses elevation, climate, soil, and land management data to model both water quantity and water quality at a daily time-step (Gassman et al., 2007). Three criteria including: Nash–Sutcliffe efficiency (NSE), Percent Bias (PBIAS), and ratio of the root mean square error to the standard deviation of measured data (RSR) were used for model calibration and validation. According to Moriasi et al. (2007) the model should meet the following criteria on the monthly basis to be considered as satisfactory: NSE > 0.5, −0.25% < PBIAS < 0.25% and RSR < 0.7. The model was calibrated for the period of 2001–2005; and validated for the period of 2006–2010 against daily streamflow data from nine monitoring stations operated by the USGS (Figs. S1 to S3 in the supplementary material).

2.2. Performance improvement in MHIT

2.2.1. Vectorization and avoiding unnecessary function calls

MHIT was implemented in MATLAB® and vectorized in order to achieve a high performing code while appealing to a large group of researchers with variety of programming skills. To vectorize a code, loop-based and scalar-oriented codes were eliminated and replaced by matrix and vector operations (MATLAB® Online Documentation, 2015). For example, here the calculation of one of the 171 indices that is called the high flood pulse count (FH1) is explained in more detail.

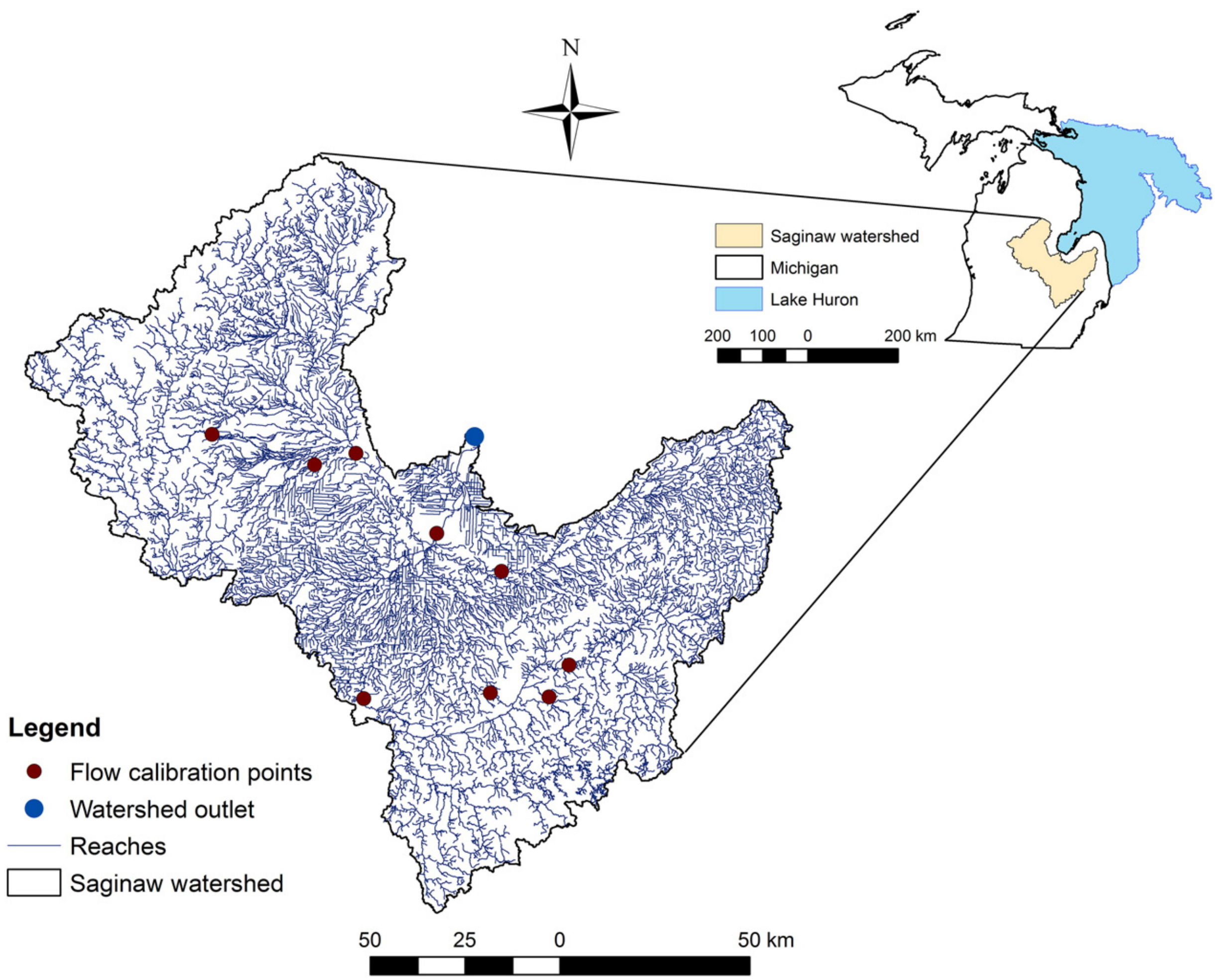

Fig. 1. The Saginaw River Watershed.