IMPROVING THE PREDICTABILITY OF HYDROLOGIC INDICES IN ECOHYDROLOGICAL APPLICATIONS

By

Juan Sebastian Hernandez Suarez

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Biosystems Engineering – Doctor of Philosophy

2021

ABSTRACT

IMPROVING THE PREDICTABILITY OF HYDROLOGIC INDICES IN ECOHYDROLOGICAL APPLICATIONS

By

Juan Sebastian Hernandez Suarez

Monitoring freshwater ecosystems allow us to better understand their overall ecohydrological condition within large and diverse watersheds. Due to the significant costs associated with biological monitoring, hydrological modeling is widely used to calculate ecologically relevant hydrologic indices (ERHIs) for stream health characterization in locations with lacking data. However, the reliability and applicability of these models within ecohydrological frameworks are major concerns. Particularly, hydrologic modeling's ability to predict ERHIs is limited, especially when calibrating models by optimizing a single objective function or selecting a single optimal solution. The goal of this research was to develop model calibration strategies based on multi-objective optimization and Bayesian parameter estimation to improve the predictability of ERHIs and the overall representation of the streamflow regime. The research objectives were to (1) evaluate the predictions of ERHIs using different calibration techniques based on widely used performance metrics, (2) develop performance and signature-based calibration strategies explicitly constraining or targeting ERHIs, and (3) quantify the modeling uncertainty of ERHIs using the results from multi-objective model calibration and Bayesian inference. The developed strategies were tested in an agriculture-dominated watershed in Michigan, US, using the Unified Non-dominated Sorting Algorithm III (U-NSGA-III) for multi-objective calibration and the Soil and Water Assessment Tool (SWAT) for hydrological modeling. Performance-based calibration used objective functions based on metrics calculated on streamflow time series, whereas signature-based calibration used ERHIs values for objective

functions' formulation. For uncertainty quantification purposes, a lumped error model accounting for heteroscedasticity and autocorrelation was considered and the multiple-try Differential Evolution Adaptive Metropolis (ZS) (MT-DREAM$_{(ZS)}$) algorithm was implemented for Markov Chain Monte Carlo (MCMC) sampling. In relation to the first objective, the results showed that using different sets of solutions instead of a single optimal introduces more flexibility in the predictability of various ERHIs. Regarding the second objective, both performance-based and signature-based model calibration strategies were successful in representing most of the selected ERHIs within a ±30% relative error acceptability threshold while yielding consistent runoff predictions. The performance-based strategy was preferred since it showed a lower dispersion of near-optimal Pareto solutions when representing the selected indices and other hydrologic signatures based on water balance and Flow Duration Curve characteristics. Finally, regarding the third objective, using near-optimal Pareto parameter distributions as prior knowledge in Bayesian calibration generally reduced both the bias and variability ranges in ERHIs prediction. In addition, there was no significant loss in the reliability of streamflow predictions when targeting ERHIs, while improving precision and reducing the bias. Moreover, parametric uncertainty drastically shrank when linking multi-objective calibration and Bayesian parameter estimation. Still, the representation of low flow magnitude and timing, rate of change, and duration and frequency of extreme flows were limited. These limitations, expressed in terms of bias and interannual variability, were mainly attributed to the hydrological model's structural inadequacies. Therefore, future research should involve revising hydrological models to better describe the ecohydrological characteristics of riverine systems.

ACKNOWLEDGMENTS

First, I want to thank God. Without His grace, love, inspiration, and blessings, this research would not have been possible.

I would like to express my sincere appreciation to my advisor Dr. A. Pouyan Nejadhashemi for his relentless support, encouragement, and guidance during my PhD. Beyond being an excellent mentor, I consider him a friend. I would also like to thank my committee members: Dr. Kalyanmoy Deb, Dr. Timothy Harrigan, and Dr. Mohsen Zayernouri for their guidance and advice throughout my research.

I am very grateful to the Colombian Ministry of Science, Technology and Innovation (Minciencias), Fulbright Colombia, and the Office of International Students and Scholars at Michigan State University (MSU) for the financial support granted for my doctoral studies. I am also grateful to the College of Agriculture and Natural Resources, the College of Engineering, the Department of Biosystems and Agricultural Engineering (BAE), and the Graduate School at MSU for providing fellowships to encourage and disseminate my research and finalize my dissertation. I would like to extend my sincere thanks to the BAE administrators and staff for their help with all the (tedious) administrative side of the program. I would also like to thank my lab mates and friends for all their help and fun moments that made my time in the US very enjoyable.

Many thanks to my father, mother, and brother for all their love, patience, motivation, and support from the distance. Dennise, this journey would have been unbearable without you. Thank you very much for sharing your days with me and making me infinitely happy. Finally, I dedicate this work to the memory of my Grandma Mariela. We miss you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

ABC: Approximate Bayesian Computation

AIC: Akaike Information Criterion

ALPHA_BF: Baseflow alpha factor ($days^{-1}$)

ANCOVA: Analysis of Covariance

ANFIS: Adaptive Neuro-Fuzzy Inference Systems

ANN: Artificial Neural Network

AR (1): Autoregressive Model with Lag 1

AUC: Area Under Receiver Operating Characteristic Curve

BBN: Bayesian Belief Networks

BIC: Bayesian Information Criterion

BIOMIX: Biological mixing efficiency

BRT: Boosted Regression Trees

CA: Correspondence Analysis

CANMX: Maximum canopy storage (mm H_2O)

CAR (1): Continuous Autoregressive Model with Lag 1

CART: Classification and Regression Trees

CCA: Canonical Correspondence Analysis

CDF: Cumulative distribution function

CH_K (2): Effective hydraulic conductivity in main channel alluvium (mm hr^{-1})

CH_N (2): Manning's "n" value for the main channel

CN2: Curve Number for moisture condition II

COIN: Computational Optimization and Innovation Laboratory at Michigan State University

d: Index of Agreement

DCA: Detrended Correspondence Analysis

DCCA: Detrended Canonical Correspondence Analysis

DH: High flows duration indices

DH1: Annual maximum daily flow ($m^3 s^{-1}$)

DH2: Annual maximum of 3-day moving average flow ($m^3 s^{-1}$)

DH3: Annual maxima of 7-day means of daily discharge ($m^3 s^{-1}$)

DH4: Annual maxima of 30-day means of daily discharge ($m^3 s^{-1}$)

DH5: Annual maxima of 90-day means of daily discharge ($m^3 s^{-1}$)

DH6: Variability of annual maximum daily average flow

DH7: Variability of annual maximum of 3-day moving average flow

DH8: Variability of annual maximum of 7-day moving average flow

DH9: Variability of annual maximum of 30-day moving average flow

DH10: Variability of annual maximum of 90-day moving average flow

DH15: High flow pulse duration with a threshold equal to the 75th percentile of the entire flow record (days)

DH16: Variability in high flow pulse duration with a threshold equal to the 75th percentile of the entire flow record

DH17: High flow duration with a threshold equal to the median flow (days)

DH19: High flow duration with a threshold equal to 7 times the median flow (days)

DH20: High flow duration with a threshold equal to the 75th percentile value for the median annual flows (days)

DH21: High flow duration with a threshold equal to the 25th percentile value for the median annual flows (days)

DH23: Flood duration with a threshold equal to the flow equivalent for a flood recurrence of 1.67 years (days)

DL: Low flows duration indices

DL1: Annual minimum daily flow ($m^3 s^{-1}$)

DL2: Annual minimum of 3-day moving average flow ($m^3 s^{-1}$)

DL3: Annual minima of 7-day means of daily discharge ($m^3 s^{-1}$)

DL4: Annual minima of 30-day means of daily discharge ($m^3 s^{-1}$)

DL5: Annual minima of 90-day means of daily discharge ($m^3 s^{-1}$)

DL6: Variability of annual minimum daily average flow

DL7: Variability of annual minimum of 3-day moving average flow

DL8: Variability of annual minimum of 7-day moving average flow

DL9: Variability of annual minimum of 30-day moving average flow

DL10: Variability of annual minimum of 90-day moving average flow

DL11: Mean of 1-day minima of daily discharge

DL12: Mean of 3-day minima of daily discharge

DL16: Low flow pulse duration (days)

DREAM: Differential Evolution Adaptive Metropolis

EPCO: Plant uptake compensation factor

ERHI: Ecologically Relevant Hydrologic Index

ESCO: Soil evaporation compensation factor

FDC: Flow Duration Curve

FH: High flows frequency indices

FH1: High flood pulse count with a threshold equal to the 25th percentile of the entire flow record ($year^{-1}$)

FH2: Variability in high flood pulse count with a threshold equal to the 25th percentile of the entire flow record

FH4: High flood pulse count with a threshold equal to 7 times median daily flow ($year^{-1}$)

FH5: Flood frequency with a threshold equal to the median flow ($year^{-1}$)

FH8: Flood frequency with a threshold equal to the 25th percentile of the entire flow record ($year^{-1}$)

FH9: Flood frequency with a threshold equal to the 75th percentile of the entire flow record ($year^{-1}$)

FHV: FDC very-high-segment volume

FL: Low flows frequency indices

FL1: Low flood pulse count ($year^{-1}$)

FL2: Variability in low flood pulse count

FLV: FDC low-segment volume

FMS: FDC midsegment slope

FMV: FDC high-segment volume

GA: Genetic Algorithm

GAM: Generalized Additive Model

GARP: Genetic Algorithm for Rule set Production

GLM: Generalized Linear Model

GLS: Generalized Least-Squares

GRNN: Generalized Regression Neural Network

GW_DELAY: Groundwater delay time (days)

GW_REVAP: Groundwater "revap" coefficient

GWQMN: Threshold depth of water in the shallow aquifer required for return flow to occur (mm H_2O)

HIT: Hydrologic Index Tool

HRU: Hydrologic Response Unit

I-IBI: Macroinvertebrate Index of Biotic Integrity

IBI: Fish Index of Biotic Integrity

ICI: Invertebrate Community Index

IHA: Indices of Hydrologic Alteration

IoA: Index of Agreement

IQR: Interquartile Range

KGE: Kling-Gupta Efficiency

MA: Average flows magnitude indices

MA12: Mean monthly flow for January ($m^3 s^{-1}$)

MA13: Mean monthly flow for February ($m^3 s^{-1}$)

MA14: Mean monthly flow for March ($m^3 s^{-1}$)

MA15: Mean monthly flow for April ($m^3 s^{-1}$)

MA16: Mean monthly flow for May ($m^3 s^{-1}$)

MA17: Mean monthly flow for June ($m^3 s^{-1}$)

MA18: Mean monthly flow for July ($m^3 s^{-1}$)

MA19: Mean monthly flow for August ($m^3 s^{-1}$)

MA20: Mean monthly flow for September ($m^3 s^{-1}$)

MA21: Mean monthly flow for October ($m^3 s^{-1}$)

MA22: Mean monthly flow for November ($m^3 s^{-1}$)

MA23: Mean monthly flow for December ($m^3 s^{-1}$)

MA24: Variability in January flows

MA25: Variability in February flows

MA26: Variability in March flows

MA27: Variability in April flows

MA28: Variability in May flows

MA29: Variability in June flows

MA30: Variability in July flows

MA31: Variability in August flows

MA32: Variability in September flows

MA33: Variability in October flows

MA34: Variability in November flows

MA35: Variability in December flows

MA42: Variability across annual flows

MA44: Variability across annual flows

MA45: Skewness in annual flows

MAG: Magnificent Seven indices

MAG1: First L-moment

MAG2: Second L-moment

MAG3: Third L-moment

MAG4: Fourth L-moment

MAG5: Autoregressive lag-one AR(1) correlation coefficient

MAG6: Amplitude of the seasonal signal

MAG7: Phase of the seasonal signal

MARS: Multivariate Adaptive Regression Splines

MCDM: Multicriteria Decision-Making

MCMC: Markov Chain Monte Carlo

MH: High flows magnitude indices

MH10: Mean maximum October monthly flow ($m^3 s^{-1}$)

MH11: Mean maximum November monthly flow ($m^3 s^{-1}$)

MH21: High flow volume (days)

MH22: High flow volume (days)

MH23: High flow volume (days)

MH6: Mean maximum June monthly flow ($m^3 s^{-1}$)

MH7: Mean maximum July monthly flow ($m^3 s^{-1}$)

MHIT: MATLAB Hydrologic Index Tool

ML: Low flows magnitude indices

ML7: Mean minimum July monthly flow ($m^3 s^{-1}$)

ML8: Mean minimum August monthly flow ($m^3 s^{-1}$)

ML9: Mean minimum September monthly flow ($m^3 s^{-1}$)

ML14: Mean of annual minimum flows

ML15: Low flow index

ML16: Median of annual minimum flows

ML17: Baseflow index based on the seven-day minimum flow

ML18: Baseflow index based on the seven-day minimum flow

ML19: Variability of baseflow index based on the lowest annual daily flow

ML21: Variability across annual minimum flows

ML22: Specific mean annual minimum flows ($m^3 s^{-1} km^{-2}$)

MLP: Multilayer Perceptron

MLR: Multiple Linear Regression

MOEA: Multi-objective evolutionary algorithm

MT-DREAM$_{(ZS)}$: Multiple-try Differential Evolution Adaptive Metropolis (ZS)

NASS: National Agricultural Statistics Service

NCDC: National Climatic Data Center

NCEI: National Centers for Environmental Information

NED: National Elevation Dataset

NMDS: Nonmetric Multidimensional Scaling

NOAA: National Oceanic and Atmospheric Administration

NRCS: Natural Resources Conservation Service

NSE: Nash-Sutcliffe Efficiency

NSE$_{rel}$: Relative NSE

NSE_{sqrt}: Root-squared-transformed NSE

NSGA-II: Non-dominated Sorting Genetic Algorithm II

NSGA-III: Nondominated Sorted Genetic Algorithm III

OF: Objective function

PBIAS: Percent Bias

PCA: Principal Component Analysis

PCoA: Principal Coordinates Analysis

PLSR: Partial Least Squares Regression

PO: Polar Ordination

PQQ: Predictive quantile-quantile plot

Q25: flow exceeded 25% of the time

Q75: flow exceeded 75% of the time

r: Correlation coefficient

R^2: Coefficient of Determination

R4MS4E: Fourth Root Mean Quadrupled Error

RA: Rate of change indices

RA1: Rise rate ($m^3\ s^{-1}\ d^{-1}$)

RA2: Rise rate variability

RA3: Fall rate ($m^3\ s^{-1}\ d^{-1}$)

RA4: Fall rate variability

RA6: Change of flow- increasing ($m^3\ s^{-1}$)

RA7: Change of flow - decreasing ($m^3\ s^{-1}$)

RA8: Reversals ($year^{-1}$)

RA9: Reversals variability

RCHRG_DP: Deep aquifer percolation fraction

RDA: Redundancy Analysis

REVAPMN: Threshold depth of water in the shallow aquifer for "revap" or percolation to the deep aquifer to occur (mm H_2O)

RF: Random Forests

RIVPACS: River Invertebrate Prediction and Classification System

RMSE: Root Mean Squared Error

RR: Runoff Ratio

SBX: Simulated Binary Crossover

SCS: Soil Conservation Service

SDM: Species Distribution Model

SEM: Structural Equation Modeling

SOL_AWC: Available water capacity of the soil layer (mm H_2O mm^{-1} soil)

SSURGO: Soil Survey Geographic Database

SURLAG: Surface runoff lag coefficient

SVM: Support Vector Machine

SWAT: Soil and Water Assessment Tool

TA: Average flows timing indices

TH: High flows timing indices

TH1: Julian date of annual maximum

TH2: Julian date of annual maximum variability

TL: Low flows timing indices

TL1: Julian date of annual minimum

TL2: Julian date of annual minimum variability

TL4: Seasonal predictability of non-low flow

U-NSGA-III: Unified Non-dominated Sorting Algorithm III

US: United States

USDA: US Department of Agriculture

USEPA: US Environmental Protection Agency

USGS: US Geological Survey

WFG: Walking Fish Group

WWAP-UN: United Nations World Water Assessment Programme

WXGEN: SWAT Stochastic Weather Generator

1 INTRODUCTION

One of the major concerns in the twenty-first century is the increasing pressure on water resources worldwide. Nearly 80% of the global population are exposed to high levels of threat to water security (Vörösmarty et al., 2010). In addition, freshwater ecosystems are deeply fragmented by built infrastructure, with only 23% of rivers longer than 1000 km arriving uninterrupted to the ocean (Grill et al., 2019). Unfortunately, this crisis is not only limited to water quantity but also is expanded to water quality. According to the United Nations World Water Assessment Programme (WWAP-UN), over 80% of global wastewater is discharged to waterbodies without any treatment (WWAP-UN, 2017). Moreover, most agriculture and urban runoff are delivered to freshwater and marine ecosystems without any water quantity and quality control (Eckart et al., 2017; Mateo-Sagasta et al., 2018). Summed to climate change, all these factors have increased the occurrence of waterborne disease. In addition, the resulting biodiversity and ecosystem losses imperil valuable ecosystem services necessary to sustain human societies (Hipsey et al., 2015; Pham et al., 2019).

In the United States, the Clean Water Act was enacted in 1972 to restore and maintain US waters' chemical, physical, and biological integrity. In river systems, chemical integrity can be associated with instream water quality. Meanwhile, physical integrity can be described in terms of water quantity, physical habitat, and stream’s geomorphology. Likewise, biological integrity is expressed in terms of abundance, composition, and diversity of freshwater organisms. Since these three components support biotic systems necessary for human and environment well-being, the paradigm comprising these concepts is known as stream or river health (Karr, 1999; Maddock, 1999). Stream health is generally measured using bioassessments, which have gained popularity for supporting water quality management, complementing chemical and

microbiological criteria (US EPA, 2011). Particularly, fish and benthic macroinvertebrates are commonly used as biological indicators. Fish are suitable for monitoring broad habitat conditions, streams connectivity, and long-term effects, whereas benthic macroinvertebrates are preferred when assessing local conditions and short-term effects (Herman and Nejadhashemi, 2015).

Stream health monitoring is usually done sparsely in time and space (Einheuser et al., 2012). Knowing the stream health condition of every single stream within a watershed is desirable for environmental management and policymaking. However, extensive biological monitoring is costly, time-consuming, and impractical for large areas. Therefore, modeling techniques have been developed to extend available information to locations with lacking biological data (Woznicki et al., 2016a). Stream health models generally use landscape attributes (e.g., land use/cover, slope, soils, geology) and instream physical and chemical characteristics (e.g., temperature, streamflow, nutrients, sediments, substrate) as explanatory variables to predict instream biological responses (Einheuser et al., 2012; Sowa et al., 2016). Streamflow is considered a master variable that dictates patterns and processes occurring in rivers and streams, including water quality, physical habitat formation, and life cycles of living organisms (Walker et al., 1995). Therefore, by studying the streamflow behavior over time (i.e., streamflow regime), it is possible to approximate the overall stream health condition (Poff et al., 1997; Richter et al., 2003).

The streamflow regime is generally described using metrics or indices related to five major facets: magnitude, duration, frequency, timing, and rate of change of flows (Sofi et al., 2020). *Magnitude* refers to the volume of water passing through a fixed location per unit of time. Based on this facet, streamflow can be classified as high, average, or low flow. Streamflow plays

different roles depending on its magnitude (Sofi et al., 2020); for instance, low flows maintain instream water quality conditions, define the longitudinal stream connectivity, enable fish and nutrients to move, and allow natural selection by purging invasive species (Poff et al., 1997). Meanwhile, high flows give shape to streams, flush away pollutants, and maintain lateral connectivity, favoring floodplains, wetlands, and riparian vegetation (Poff et al., 1997). *Duration* is the length of time associated with a flow event being read horizontally in a hydrograph. This facet influences the persistence of aquatic and riparian species and controls fish growth potential and development under flooding events (Bunn and Arthington, 2002). *Frequency* refers to how often a streamflow magnitude occurs over a specific period of time, and it is generally described using Flow Duration Curves (FDC). This facet is important for controlling aquatic and riparian species' life cycles and productivity (Bunn and Arthington, 2002). For example, frequency regulates how often fish can move upstream or to floodplains for migration or reproduction (Poff et al., 1997). *Timing* is the degree to which flow events are temporally autocorrelated, indicating that the system has memory. For instance, certain rivers always experience high flows in spring and low flows in summer. This facet works as a trigger the system needs to start a process (e.g., fish spawning). Additionally, timing helps to maintain species diversity (Bunn and Arthington, 2002). Finally, the *rate of change* describes how fast the system goes up and down. This facet influences species persistence and coexistence and controls the establishment of nonnative species (Poff et al., 1997; Sofi et al., 2020). In summary, there are many indices describing the aforementioned streamflow regime facets (Olden and Poff, 2003), and they are known as ecologically relevant hydrologic indices (ERHIs).

Calibrated hydrological models are used to predict streamflows beyond monitoring stations. Consequently, ERHIs can be estimated in ungauged locations using results obtained

from hydrological modeling. When developing models for predicting ERHIs, three fundamental questions emerge in the process: (1) how much the predictability of ERHIs is affected by the choice of model calibration techniques? (2) how to calibrate a hydrologic model to improve the prediction of multiple ERHIs simultaneously? and (3) how reliable are the hydrological modeling results when predicting EHRIs?

To address these questions, the goals of this research were to (1) evaluate the predictions of ERHIs using a hydrological model when it is calibrated using single- and multi-objective techniques based on widely used performance metrics, (2) develop calibration strategies for improving the predictability of ERHIs and the overall streamflow regime using a hydrologic model, and (3) quantify the modeling uncertainty of ERHIs using the results obtained from the developed calibration strategies under the previous section.

The outcome of this research is a framework for linking hydrologic model calibration and uncertainty quantification when predicting ERHIs. This framework includes the development of novel calibration strategies aimed to improve the accuracy of ERHIs predictions while maintaining a balanced representation of different streamflow regime facets. Ultimately, it is expected that the overall performance of ERHIs' uncertainty quantification is improved. This can help policymakers with decision-making in the context of water and natural resources management.

2 LITERATURE REVIEW

2.1 OVERVIEW

During the last three decades, explaining cause-effect relationships between natural and anthropogenic disturbances with measures of stream health have motivated the growing application of statistical, machine learning, and soft computing methods. The aim of this review is to provide insight into the most widely used methods for predicting biological variables based on macroinvertebrate and fish species in riverine ecosystems. Therefore, we describe several methods including multiple linear regression, generalized linear models, generalized additive models, boosted regression trees, random forests, artificial neural networks, fuzzy logic-based, and Bayesian belief networks along with recent applications of these. Moreover, issues regarding variable selection, model interpretability, ensemble modeling, and model evaluation and overfitting are discussed. Recent advances have suggested the need for integrated modeling systems to enhance predictive ability and improve interpretability. However, trade-offs between model complexity and accuracy demand research efforts in uncertainty quantification/propagation in model ensembles. Additionally, models should be perceived as complementary tools that require further validation with field measurements. Therefore, a consensus regarding monitoring and modeling practices for stream health applications is recommended.

2.2 INTRODUCTION

Current and future threats to freshwater ecosystems due to changes in environmental conditions and impacts of anthropogenic activities require urgent and well-informed actions (Strayer and Dudgeon, 2010; Vörösmarty et al., 2010; Waldron et al., 2017). Therefore, health

assessments of these ecosystems are critical to promoting their protection and restoration (Beechie et al., 2010). During the last decades, many environmental legislations have been increasingly supporting the introduction of biological assessments in local, regional and national monitoring programs (Hering et al., 2010; Hill et al., 2017). Typically, biotic indices are derived from biological assessments to represent the stream health condition. The stream health concept comprises the physical, chemical and biological capacity to maintain the structure and functioning of freshwater ecosystems, required for supporting living systems (Karr, 1999; Maddock, 1999). These indices are based on one or multiple metrics describing abundance, richness, diversity or composition of biological assemblages (Herman and Nejadhashemi, 2015). Furthermore, biomass, probability of occurrence, and incidence (presence/absence) data provide information regarding the level of impairment, which is also useful for stream health evaluation (Hill et al., 2017; Smucker et al., 2013). Biological measurements in aquatic ecosystems can be obtained from several biological assemblages and their selection is usually subjected to the type of study to perform. Benthic macroinvertebrates are preferred when studying localized effects of habitat and water quality alterations, due to their limited movement within a water body (Kerans and Karr, 1994). Meanwhile, fish communities are preferred when evaluating changes in flow regime and spatial connectivity (Karr, 1981). Benefits of stream health evaluation include the possibility to explore the environmental mechanisms driving ecosystem alterations (Herman and Nejadhashemi, 2015). Likewise, indicators of stream health can help with the identification of degraded areas and the provision of necessary inputs to design protection and restoration projects (Walters et al., 2009).

Stream health models have been introduced to relate observed biological data with environmental and landscape variables with the goal of establishing reference conditions (Feio

and Poquet, 2011; Hawkins et al., 2000), predicting biological variables and indicators in unsampled locations (Merriam et al., 2015; Waite et al., 2010), classifying streams by impairment condition (Brown et al., 2012; Maloney et al., 2009), and predicting biological variables and indicators given the implementation of conservation practices (Hall et al., 2017; Herman et al., 2015; Sowa et al., 2016) and changes in environmental and landscape stressors (Einheuser et al., 2013b, 2013a, 2012). These models have been enhanced by the advances in landscape methods for studying freshwater ecosystems (Johnson and Host, 2010; Steel et al., 2010), species distribution models (SDMs) (Elith and Graham, 2009; Li and Wang, 2013; Van Echelpoel et al., 2015) and habitat suitability models (Ahmadi-Nedushan et al., 2006; Yi et al., 2017). Still, due to the complexity and nature of the problem, stream health models are mainly empirically-based rather than mechanistic or process-based. However, hierarchical approaches incorporating climate, hydrologic, hydraulic, water quality and/or physical habitat models have been suggested to improve the current models' predictability, interpretability and accuracy (Daneshvar et al., 2017a; Einheuser et al., 2013a, 2013b, 2012; Guse et al., 2015; Herman et al., 2015; Holguin-Gonzalez et al., 2014, 2013a, 2013b; Jähnig et al., 2012; Kail et al., 2015; Kennen et al., 2008; Woznicki et al., 2016a, 2016b; Yi et al., 2017). Modeling approaches include traditional regression models (e.g. multiple linear regression, generalized linear models), ordination and classification methods (e.g. principal component analysis, redundancy analysis), clustering methods (e.g. self-organizing maps, *k*-nearest neighbors), structural equation modeling (SEM), machine learning and soft computing techniques (e.g. fuzzy logic, neural networks, evolutionary computation). In this paper, we review the most widely used methods able to model both continuous and categorical stream health data based on macroinvertebrate and fish assemblages. These methods comprise of traditional statistical approaches, machine learning,

and soft computing methods. The specific objectives of this study are to (1) summarize the main characteristics of the selected modeling methods and their applications, and (2) identify features requiring further research for improving stream health modeling practices.

2.3 MODELING METHODS

In this section, we describe the most widely-used methods for stream health modeling. In general, the approaches presented herein are data-driven since we are dealing with natural systems; however, soft computing methods are more suitable for incorporating expert elicitation. In addition, both soft computing and machine learning methods are more flexible regarding statistical assumptions than traditional statistical modeling approaches.

2.3.1 Statistical Methods

Statistical methods considered herein are mainly focused on modeling approaches based on linear regression. However, a general overview of multivariate methods for ordination is also presented. A statistical model is a specification of probability distributions reproducing observed data, establishing mathematical relationships between explanatory and response variables (Nelder and Baker, 2006). Multivariate methods can be used for either ordination or classification purposes. Ordination is the arrangement of biological data samples along one or more gradients (Austin, 1976), whereas classification is the assignment of biological data samples into groups based on a measure of similarity (or dissimilarity) (Mitteroecker and Bookstein, 2011).

2.3.1.1 Linear statistical methods

Linear statistical methods have been applied mainly to elucidate relationships between landscape, habitat, and water quality factors and biological variables. In this category, there are methods with different complexity level like Multiple Linear Regression (MLR), Generalized Linear Models (GLMs) and Generalized Additive Models (GAMs), which have been widely

implemented in ecological applications. MLR fits a linear equation using the observed data to model the relationship between a set of explanatory variables and a response variable, assuming independent and identically normal distributed errors. Meanwhile, GLMs introduce some flexibility allowing different error distributions, selected from the exponential family of sampling models (e.g. normal, binomial, Poisson, gamma), and relate the response variable Y with the explanatory variables X using a pre-specified link function g (McCullagh and Nelder, 1989). The link function provides the relationship between the linear predictor η and the expected value (i.e. mean) of the response variable $E(Y|X)$:

$$g[E(Y|X)] = \eta = \alpha + X\beta \quad (1)$$

where, α and β are the intercept and the vector of linear weights, respectively. When selecting different link functions, GLMs comprise linear, logistic, and Poisson regressions, among others. For instance, logistic regression is preferred when modeling species presence/absence, while Poisson regression is more suitable when modeling the count data (Li and Wang, 2013). In order to further account for nonlinearities, GAMs extend GLMs, expressing η as the sum of unspecified nonparametric linear or nonlinear smoothing functions f_i, applied over the set of p explanatory variables (Hastie et al., 2009):

$$g[E(Y|X)] = \alpha + f_1(X_1) + \cdots + f_p(X_p) \quad (2)$$

To ensure that the smoothing functions are identifiable, they are restricted to have zero mean (Maloney et al., 2012). These functions are commonly estimated using a scatterplot smoother (e.g. cubic spline) as the basic building block (Hastie et al., 2009; Zuur et al., 2009).

Linear statistical models were initially introduced to build empirical associations between landscape and stream health attributes. Early efforts were mostly concerned in identifying the main stressors and landscape components (i.e. riparian buffer, watershed) affecting water quality,

physical habitat, and/or freshwater biological communities (Van Sickle et al., 2004). Many of these works are reviewed by Ahmadi-Nedushan et al. (2006), Johnson and Host (2010) and Steel et al. (2010). Linear statistical models have also been continuously applied as benchmarks for comparison with other statistical methods and machine learning approaches. For macroinvertebrate communities, traditional regression models are still used to relate abiotic stressors with species occurrence in order to explore habitat and water quality preferences, especially in headwaters (Pond et al., 2017) and tropical regions (Damanik-Ambarita et al., 2016; Everaert et al., 2014; Jerves-Cobo et al., 2017). However, prediction of the biological condition under different spatial and temporal domains and scales have been also addressed in some studies for both macroinvertebrate and fish assemblages (Frimpong et al., 2005; Johnson and Host, 2010; Van Sickle and Burch Johnson, 2008).

Representative studies employing MLR for stream health prediction include models developed by Waite et al. (2010) to predict several macroinvertebrate metrics using watershed- and riparian-scale variables. Results showed that the best models explained 41-74% of the variation requiring only two or three explanatory variables after stepwise selection using the Akaike Information Criterion (AIC) estimator. Likewise, Merriam et al. (2015, 2013) employed linear regression and deletion tests to predict two indices based on benthic macroinvertebrate abundance as a function of surface mining, underground mine permit density, residential development, and location attributes. The results suggested that the interactions between different land uses are more important than a single land use effect.

On the other hand, GLMs have been very popular for predicting species occurrence and distribution in freshwater ecosystems (Ahmadi-Nedushan et al., 2006). Van Sickle et al. (2004) implemented a linear regression and a negative binomial GLM for projecting fish and

macroinvertebrate biological indices as a function of landscape and streamflow variables under different timeline scenarios, including reference conditions. Donohue et al. (2006) employed a stepwise binary logistic regression and logarithmic and quadratic regressions to obtain national-wide relationships between catchment and water quality attributes and stream ecological status based on an index representing the structure of benthic macroinvertebrate communities. Other studies have implemented GLMs in an integrated ecological modeling framework involving hydrodynamic, water quality and stream habitat suitability models to predict macroinvertebrate-based stream health at a reach scale (Holguin-Gonzalez et al., 2013b, 2013a; Kuemmerlen et al., 2014). Sui et al. (2014) developed a predictive model employing a geomorphology-based hydrological model to determine ten flow indices. Then, a GLM was implemented in order to relate those indices to the occurrence probabilities of 50 fish species at a watershed scale. In a recent study, Gieswein et al. (2017) used GLM to quantify the pairwise stressor interactions (strength and significance). This was performed following implementation of a decision tree-based model for identifying stressor hierarchy. A Boosted Regression Trees (BRT) model was used when analyzing the relationships between several factors (i.e. riparian land use, physical habitat quality, nutrients, natural variables) and fish, macroinvertebrates and macrophytes assemblages.

With respect to GAMs, Maloney et al. (2012) compared the standard and boosted version of this method for macroinvertebrate and fish metrics prediction, using watershed, stream, and site attributes as explanatory variables. Results indicated that gradient boosting applied to GAMs avoids overfitting and provides interpretable relationships, which is an advantage in comparison with traditional machine learning techniques. Additionally, regular GAM has been also compared with a GAM based on principal component analysis (PCA) for fish richness and

diversity prediction (Zhao et al., 2014). Results showed different selected explanatory variables for each approach, generating different outcomes for the response variables. However, PCA-based GAM performed better during cross-validation tests and was found to be more suitable when predictors are highly correlated (Zhao et al., 2014). More recently, Almeida et al. (2017) evaluated the effect of sampling effort in terms of transect length on fish metrics. This was performed for a large Mediterranean watershed using a GAM with sampling area as a predictor. Results indicated that fish indices that are obtained using predictive models are more sensitive to sampling strategies than simpler biotic metrics that are model-independent, showing a decrease in their values with increasing sampling area, despite observed higher richness.

Regarding spatial-scale effects, Johnson and Host (2010) listed representative studies from 2000 to 2008 involving invertebrates and fish assemblages. For each study, the authors reported the scales (e.g. habitat, reach, local, ecosystem, watershed, regional, ecoregion) at which explanatory variables explained the instream biological response. Johnson and Host (2010) showed meaningful differences in the scale's importance among the reviewed studies, due to the different region sizes and disturbance levels for each case. A study by Frimpong et al. (2005) used linear and piecewise regression to compare the performance of stream habitat indices obtained at the watershed-scale and observations at the reach-scale for fish metrics prediction. Results indicated that watershed-scale variables provided better predictions for stream health than reach-scale variables. Additionally, predictive ability decreased with the spatial extent, which might be attributed to the increase of the attributes' heterogeneity. In another study, Van Sickle and Burch Johnson (2008) developed a distance weighting model based on linear regression for estimating specific land use areas within watersheds that best explain fish index of biotic integrity (IBI). With this approach, it is possible to compare different scales of landscape

influence on stream health. Furthermore, linear models have been also used for multimetric indices formulation. Pont et al. (2009) implemented a procedure involving stepwise, multilinear, logistic, and Poisson regressions to build a predictive IBI for aquatic vertebrates (fish and aquatic amphibians). Particularly, this procedure was developed to discriminate natural and anthropogenic effects over the biotic metrics computation. Implementing the same approach, Moya et al. (2011) developed a multimetric index for macroinvertebrate assemblages. This index based on predictive models successfully discriminated between the reference and disturbed sites.

2.3.1.2 Ordination methods

Ordination refers to multivariate statistical methods commonly classified into indirect (unconstrained ordination) and direct gradient analysis (constrained ordination) (De'ath, 1999; Guo et al., 2015b). Ordination methods are generally preferred when analyzing multiple species at multiple sites (Ahmadi-Nedushan et al., 2006). The main objective of ordination is to reduce dimensionality to identify patterns in the data while describing relationships with explanatory variables (e.g. environmental gradients). As a result, data samples are ordered in such a way that similar points are placed together (Ahmadi-Nedushan et al., 2006).

Indirect gradient analysis only uses samples collected in one data matrix, extracting dominant or orthogonal axes of variation. Any additional information regarding explanatory variables is used afterwards to enhance results' interpretation. These methods can be classified into distance-based techniques (e.g. polar ordination – PO, principal coordinates analysis – PCoA, nonmetric multidimensional scaling – NMDS) and Eigen analysis-based techniques, which can be derived from linear models (e.g. PCA) or from unimodal (nonlinear) models (e.g. correspondence analysis – CA, detrended correspondence analysis – DCA). Contrariwise, in direct gradient analysis, variables of interest are directly related to explanatory variables. Therefore, these techniques are preferred for habitat modeling (Ahmadi-Nedushan et al., 2006).

Direct gradient methods can also be based on linear models (e.g. redundancy analysis – RDA) or unimodal models (e.g. canonical correspondence analysis – CCA, detrended canonical correspondence analysis – DCCA). It is worth noting that linear models are preferred for short gradients, whereas unimodal models are, in general, more suitable for aquatic habitat modeling (Ahmadi-Nedushan et al., 2006). Further details regarding ordination techniques can be found elsewhere (Borcard et al., 2011; Kent, 2006; Zuur et al., 2007).

Multivariate methods for biological assessments, like the River Invertebrate Prediction and Classification System (RIVPACS) and its variants (Abbasi and Abbasi, 2012; Feio and Poquet, 2011), are mainly based on ordination methods. These multivariate methods attempt to predict ratios of taxa observed vs. expected – O/E, carefully choosing reference sites while using categorical (e.g. presence/absence) rather than continuous biological data. Some recent applications of ordination methods within stream health modeling were reported by Gazendam et al. (2016), indicating that the integration with other techniques is needed for identifying relationships between environmental variables and stream health using PCA or CCA. For instance, D'Ambrosio et al. (2014), used CCA and variance partitioning to evaluate the relationships between instream habitat, spatial location, and geomorphic characteristics on fish and macroinvertebrate-based stream health indices. This study was conducted considering highly modified drainage channels as a consequence of agricultural activities. Results provided key ecological drivers for each biological community, under different stream geomorphic condition and location. Additionally, it is worth noting that ordination methods have been mainly implemented to assess the influence of explanatory variables on response variables before implementing more complex approaches for predicting stream health indicators (Lin et al., 2016).

2.3.2 Machine Learning

Machine learning is a form of artificial intelligence employing statistical, probabilistic and optimization algorithms to identify relationships and patterns from datasets. The resulting outcomes can be used for data analysis, visualization and prediction (Mitchell, 1999).

2.3.2.1 Decision tree-based methods

Decision tree family of models are hierarchical structures also known as Classification and Regression Trees (CART) (Breiman et al., 1984). These models divide the predictor space into regions with a homogenous response, then fitting a constant to each region. A decision tree grows using binary splits, resulting in a dendrogram with varying numbers of branches (De'ath, 2007). Each split is defined by threshold values accompanying the explanatory variables. Regression trees fit the mean response to observations, while classification trees, which are used for categorical data, fit the most frequent class as the constant. Usually, CART are grown to a maximum and then pruned using cross-validation approaches to prevent overfitting (Hastie et al., 2009). CART have been applied in several stream health applications, and nowadays are commonly used as a benchmark approach for comparison with other methods (Ambelu et al., 2010; He et al., 2010; Holguin-Gonzalez et al., 2014, 2013a; Maloney et al., 2009; Ocampo-Duque et al., 2007; Waite et al., 2012; Wang et al., 2007). Known drawbacks of CART are the difficulty in modeling smooth functions and producing very different results when making small changes to the training data (Elith et al., 2008). However, ensemble methods based on computational intensive procedures as boosting and bootstrap aggregation (a.k.a. bagging) have shown better and promissory results in ecological applications (De'ath, 2007). Thus, two ensemble methods, Boosted Regression Trees (BRT) and Random forests (RF), are further described in the next sections.

2.3.2.1.1 Boosted regression trees

Boosted regression trees (BRT) method is an advanced form of regression that combines a large number of regression trees using the boosting technique to increase predictive performance (De'ath, 2007; Friedman, 2001; Hastie et al., 2009). Boosting is a forward sequential procedure that aims to find and merge results from multiple models (e.g. decision trees), emphasizing on observations poorly represented by an existing combination of models (Brown et al., 2012). For BRT, boosting works as an optimization technique, minimizing the difference between predicted and observed values, adding at each step a new tree that best reduces this difference (Elith et al., 2008). The technique updates the residuals in each iteration, preserving the existing trees unchanged while extending the overall model. Hence, the final BRT model is a linear combination of several decision trees. Like other regression methods, it is possible to define the error distribution in BRT models in order to consider different response types (e.g. Gaussian, Poisson, binomial).

BRT models are controlled by two important parameters: the learning rate (*lr*), which determines the contribution of each tree, and the tree complexity (*tc*), which controls the number of terminal nodes and interactions. Both *lr* and *tc* define the required number of trees (*nt*) for an optimal prediction (Elith et al., 2008). In addition, the stochasticity of BRT models is controlled by the "bag fraction", which refers to the observations that are randomly drawn to train each new tree, with optimal values between 0.5 and 0.75 (Elith et al., 2008). In ecological applications, small values for *lr* (<0.001), and therefore high *nt* (>1000), are preferred in order to avoid overfitting, reduce the contribution of each tree, and increase predictive reliability (Elith et al., 2008). Values for *tc* are defined depending on the data availability and are restricted to the desired computing time. High values for *tc* implies a slower *lr* to keep a similar optimal *nt* (Elith et al., 2008).

Works conducted by Moisen et al. (2006), Elith et al. (2006) and Leathwick et al. (2006a), are among the first studies employing BRT models for ecological applications. Particularly, they showed that BRT models are more flexible and outperform regression models like GLM or GAM in variable selection, higher variance explanation, and lower prediction error. Moreover, BRT models are suitable for handling nonlinear relationships and can model smooth functions and interactions (Elith et al., 2008). Applications in stream health modeling include the determination of quantitative relationships between landscape variables and instream biological response (Chee and Elith, 2012; Gieswein et al., 2017; Golden et al., 2016; Pilière et al., 2014; Steel et al., 2017; Tonkin et al., 2014; Waite and Van Metre, 2017), prediction (Brown et al., 2012; Clapcott et al., 2017; Elias et al., 2016; Leclere et al., 2011; May et al., 2015; Waite et al., 2014, 2012), multimetric indices formulation (Clapcott et al., 2014; Esselman et al., 2013), and setting of instream water quality/ecological objectives and disturbance thresholds (Clapcott et al., 2012, 2010; Wagenhoff et al., 2016). Moreover, the most of recent studies involving BRTs implementation have dealt with regional and national scales.

Nonlinear relationships using BRTs have been explored using different explanatory and response variables, and distinct objectives. For instance, Chee and Elith (2012) analyzed the patterns of occurrence for 17 native and alien riverine fish species. In that study, the explanatory variables comprised of 20 environmental predictors including physiographic, bioclimatic, edaphic and land cover attributes. The survey method was also considered as a categorical explanatory variable. Results showed that several, but not all, of the developed models are transferable to adjacent regions. Meanwhile, Tonkin et al. (2014) explored the effects of distance and barriers on the occurrence of macroinvertebrate assemblages. Hence, four different BRT models (i.e. considering different factors driving invertebrate colonization) were evaluated. In

another study, Golden et al. (2016) addressed the relationship between landscape variables at the watershed and riparian buffer scales with instream nutrient concentration and fish IBI under low flow conditions. Landscape variables included temporal and geographic position attributes, and indicators of runoff and point and non-point nutrient sources. Similarly, Pilière et al. (2014) explored the relationships between environmental stressors and freshwater invertebrates represented by the Invertebrate Community Index (ICI). Predictors included geography, water quality, physical-habitat quality, and toxic pressure variables. The results suggested that it is necessary to fit explanatory variables interactions to increase predictive ability and model interpretability. In a most recent work, Waite and Van Metre (2017) used BRTs to identify the most important stressors explaining the macroinvertebrate condition in streams. The final model was determined sequentially eliminating variables according to cross-validation performance. Three macroinvertebrate metrics and a multimetric index were used as response variables. Results indicated that watershed-scale stressors acted as surrogate variables for instream stressors. However, given the performance metrics for model validation, the authors did not recommend using the fitted BRT models for prediction in unsampled sites. On the other hand, there are several studies that implemented BRTs for understanding instream processes that affect species distribution. For instance, a study by Steel et al. (2017) explored the relationship between streamflow regime and water temperature metrics with the Shannon-Wiener diversity index, total richness, and total density per square meter of benthic macroinvertebrate assemblages. Results indicated that macroinvertebrate diversity and total richness showed the best predictive performances, with metrics related to spring snowmelt recession and variability in summer water temperature having the greatest relative influence. Meanwhile, the total density per square meter

of benthic macroinvertebrates showed the poorest fitness, limiting the interpretability of the modeling results (Steel et al., 2017).

In general, studies using BRTs concerned with predicting the stream health condition using fish and macroinvertebrate communities are recent. Leclere et al. (2011) attempted to select an appropriate statistical method for predicting nine fish species occurrence in large river systems at a reach-scale. Compared methods were CART, GLM and BRT, where the latter showed the best performance. Specified predictors comprised qualitative (occurrence of shallow waters and shelters), semi-quantitative (range of coverage/magnitude of bottom subtract, current velocity, shade, macrophytes, complexity structures), and quantitative (value for depth, stream width, subtract diversity, cover of bed sediment) variables. The results showed that the BRT method selected a greater number of variables, giving more importance to continuous variables. Furthermore, this method provided a better ecological interpretability and consistency in the obtained response curves. Other studies have used BRT models to predict benthic macroinvertebrate metrics at a regional scale. For instance, Waite et al. (2012) used land use and land cover explanatory variables to obtain richness and O/E ratios. In addition, the study compared MLR, CART, Random forests (RF) and BRT predictive performances. Results indicated that BRT outperformed the other methods and provided additional information regarding potential interaction among explanatory variables. Meanwhile, Brown et al. (2012) used benthic macroinvertebrate index of biotic integrity (I-IBI) as the response variable. Landscape variables at the watershed and riparian-buffer scales were selected as explanatory variables. In the study, population density and agricultural and urban land use were the predictors with the highest influence on the response. However, the final BRT model was not able to capture the minimum and maximum values of the observed data. Therefore, the results suggested

that the outcomes from BRT models should not be used to predict index values at specific sites. Instead, the model is recommended to be used to predict impairment condition due to watershed disturbance (Brown et al., 2012).

In a more recent study, Elias et al. (2016) implemented a two-level nested model to predict macroinvertebrate occurrence. The first level attempted to predict four water quality variables (dissolved oxygen, phosphates, ammonium and nitrates) with a BRT model. Then, these variables were used to predict reference conditions of stream health employing a different modeling approach. However, results showed that the BRT model was only successful in predicting nitrates. Other studies have addressed related fields to stream health modeling like surface water and groundwater interactions and water quality modeling (Poor and Ullman, 2010; Smucker et al., 2013). For instance, Johnson et al. (2017) successfully estimated the effects of groundwater seepage on stream temperature in unsampled sites at headwaters. Hence, landform and precipitation covariates, representing spatial and temporal variables, respectively, were selected as explanatory variables. During the modeling process, these variables were sequentially eliminated using ranking, clustering and model simplification approaches.

The impacts of the scale of study on model predictability have been also addressed with BRT models in which depending on the size and location of study, the importance of natural gradients and anthropogenic stressors are different (May et al., 2015; Waite et al., 2014). However, it was suggested that using BRT models for small spatial scales provide more accurate predictions compare to large-scale studies (Waite et al., 2014). A study by Clapcott et al. (2017) attempted to estimate site-specific contemporary and reference values for a macroinvertebrate index, evaluating spatial-scale effects on the predictions. Models at national and regional scales were tested for comparison. In general, the proportion of native vegetation in upstream

catchments was the primary predictor, while the remaining relevant predictors varied regionally. Main environmental predictor at large scales also included flow variability, habitat category, substrate composition, summer temperature and average upstream slope. Regional models showed that low flow remaining downstream after daily water allocation, and calcium concentration of rocks in the catchments were relevant. Other methods (i.e. ANCOVA and RF) were also evaluated. Results indicated that models at different scales were equally informative. However, the authors recommended using finer-scale predictors to improve the model accuracy. These variables include substrate size, nutrient concentrations, streamflow, and temperature. Additionally, it was shown that regression tree-based methods did not overestimate biological condition scores because the methods do not extrapolate beyond the range of observations. Meanwhile, these methods are able to predict stream classes that are unrepresented in the available observations (Clapcott et al., 2017).

BRT model interpretability has been found to be suitable for defining disturbance thresholds and setting instream objectives. Particularly, functional indicators for ecosystem processes are often used as response variables for the task mentioned above (Clapcott et al., 2012, 2010; Wagenhoff et al., 2016). Functional indicators include variables related to primary production, ecosystem respiration, organic matter breakdown, cellulose decomposition potential, among others (Clapcott et al., 2010). Furthermore, it has been suggested that statistical analysis based on single-stressor models have the tendency to provide spurious thresholds for management purposes (Wagenhoff et al., 2016). Meanwhile, the use of sediment-specific macroinvertebrate metrics has been encouraged for improving prediction and threshold definition (Wagenhoff et al., 2016). Other findings include the importance of spatial variation of the predictors for increasing predictive power (Clapcott et al., 2012). Finally, additional applications

of BRT models at national and continental scales comprise the definition of multimetric indices. Examples include the formulation of fish community indicators in large regions (Esselman et al., 2013) and multimetric indices development based on water quality-based predictive modeling, measurements of macroinvertebrate and fish assemblages, and indicators for ecosystem processes (Clapcott et al., 2014).

2.3.2.1.2 Random forests

Similar to BRT, Random forests (RF) are collections of individual CART (Breiman, 2001). This method fits several trees using bootstrap samples of the training data while employing a small number of randomly selected predictors from the explanatory variables (Snelder and J. Booker, 2013). The bootstrapping process attempts to reduce the variance of estimated outputs (Hastie et al., 2009), which is typically high for single large regression trees. For each bootstrapped sample, the largest tree is grown but not pruned, and aggregation is made by averaging or majority voting the trees (Carlisle et al., 2009b; Cutler et al., 2007). The size of randomly selected predictors is usually $\sqrt{p}$ or $\log(p)$, being p the number of explanatory variables (De'ath, 2007). This method requires a large number of trees to ensure convergence (Booker et al., 2015). Additionally, because of the bootstrap sampling, RF excludes over 37% of observed data for growing the regression or classification trees. This non-drawn portion is called out-of-bag samples (Prasad et al., 2006), and error estimates are computed using these samples. Then, these error estimates are used for regression trees aggregation (Carlisle et al., 2009b). When used for classification, RF determines the most frequent class across all trees for each observation within the out-of-bag portion. Estimating the error using the out-of-bag samples is almost equivalent to perform *k*-fold cross-validation in which once the error stabilizes, the

training is terminated (Hastie et al., 2009). Therefore, because a large number of trees provides limited generalization errors, RF method prevents overfitting (Prasad et al., 2006).

RF applications include predicting the instream biological condition and simulating ecologically-relevant hydrologic indices in undisturbed sites and ungauged locations. Most of the studies have been developed at regional and national scales. Some RF models have attempted to evaluate the effects of human activities and relevant environmental factors on natural aquatic ecosystems (He et al., 2010), while others have addressed the prediction of instream biological condition in ungauged locations (Carlisle et al., 2009a; Hill et al., 2017). Many of these studies have been focused on predicting macroinvertebrate taxa richness and composition (Álvarez-Cabria et al., 2017; Booker et al., 2015; Carlisle et al., 2009a; Chinnayakanahalli et al., 2011; Patrick and Yuan, 2017; Vander Laan et al., 2013; Waite et al., 2014). Other studies have addressed the RIVPACS-type approach for determining biological condition (Carlisle et al., 2009a; Chinnayakanahalli et al., 2011). Moreover, fish assemblages composition and richness (He et al., 2010; Patrick and Yuan, 2017), and fish biomass (Álvarez-Cabria et al., 2017) have been also modeled. Other applications include determining reference conditions (Clapcott et al., 2017). For biological condition prediction, common explanatory variables comprise of geospatial datasets including land cover, land use, topography, climate, soils, societal infrastructure, and hydrologic modification. Some of these studies have addressed natural flow regime and water quality and temperature roles on biological condition predictability (Booker et al., 2015; Chinnayakanahalli et al., 2011; Patrick and Yuan, 2017; Vander Laan et al., 2013). For instance, Patrick and Yuan (2017) used the 171 Hydrologic Index Tool (HIT) indices as predictors, which were obtained with statistical modeling. Other studies have identified variables describing natural and human activities as important predictors (Carlisle et al., 2009a; He et al., 2010),

especially land use changes and flow modification. Regarding ecologically-relevant hydrologic indices, common explanatory variables are related to geospatial data describing natural watershed characteristics. However, RF models usually include a reduced number of indices (between 1 and 36), and prediction errors range from 15% to 40% (Carlisle et al., 2009b). Other applications include river classification (Dhungel et al., 2016; Snelder and J. Booker, 2013), where classification success has ranged from 34% to 75%. Additionally, there are studies that have analyzed potential effects of climate change (Dhungel et al., 2016), and environmental flow, and flow-ecology relationships, for environmental impact assessment (Buchanan et al., 2017).

2.3.2.2 Artificial neural networks

Artificial neural networks (ANNs) are nonlinear models with many parameters flexible enough to approximate any smooth function. ANN is a learning method based on the idea of building linear combinations of the specified explanatory variables and then modeling the response variables as nonlinear functions of these linear combinations (Hastie et al., 2009). ANNs are labeled as "black box" models and are known for being useful for prediction but not very useful for producing understandable models (i.e. provide limited insight into the relative influence of explanatory variables). However, multiple methods for understanding ANNs results are available, including sensitivity analysis and randomization tests (Gevrey et al., 2003; Olden and Jackson, 2002).

A typical two-stage regression or classification network model is known as a feed-forward neural network. Under this approach, linear combinations of explanatory variables X are transformed into Z elements called "hidden units" using a nonlinear function σ, known as the "activation function". Usually, σ is the sigmoid function, which is a smooth version of the step

function. However, Gaussian radial basis functions are also commonly used (Hastie et al., 2009; Mathon et al., 2013). For ANN setting, more than one layer of hidden units can be used (Lek and Guégan, 1999). Afterwards, these *Z* elements are linearly combined. Then, the linear combinations are transformed using an output function to provide the response variables. The constants involved in the linear combinations are known as "weights". The aforementioned ANNs are called multilayer perceptrons (MLPs) and are very popular among ecological applications. In order to fit ANNs with observed data, a two-pass procedure known as back-propagation, which is a gradient descent algorithm, is typically employed in order to determine the weights (Hastie et al., 2009). Hence, key MLP parameters include the number of times that the training data is used to update the weights of the hidden units (i.e. epochs) and the number of hidden layers and units. Alternatives to MLP, such as the Generalized Regression Neural Network (GRNN), have been recently applied for stream health modeling (Mathon et al., 2013; Sutela et al., 2010).

ANNs were first introduced in ecological applications during the 1990s (Lek et al., 1996; Lek and Guégan, 1999); however, since then the method has been widely used. Goethals et al. (Goethals et al., 2007) presented a review of 26 representative studies addressing macroinvertebrate prediction, covering a period from 1998 to 2006. In those studies, the response variables were usually presence/absence, abundance and derived indicators (e.g. richness, average score per taxon, exergy), using landscape, instream and water quality attributes as explanatory variables (i.e. inputs). Studies using fish biological data have been also implemented for evaluating restoration projects and understanding the effects of changes in physical habitat and water quality variables (Olaya-Marín et al., 2013, 2012; Olden et al., 2008).

Other studies have reported an increased use of self-organizing maps (SOMs) for predicting macroinvertebrate and fish distributions and exploring relationships with landscape and environmental variables (Chon, 2011; Tsai et al., 2016). SOMs, also known as the Kohonen networks, are unsupervised ANNs that are implemented for pattern classification, clustering, and ordering purposes (Kalteh et al., 2008), and are also known for approximating probability density functions of the input data (Chon, 2011). Recent works have addressed variable selection and uncertainty analysis to gain insight into the results interpretability, which is one of the main ANN's drawbacks. For instance, Mouton et al. (2010) compared six different methods reviewed by Gevrey et al. (2003) for evaluating explanatory variables' individual contribution in predicting macroinvertebrates abundance. The results indicated that some techniques are more sensitive and less stable than others are. However, it was shown that the different methods were able to provide consistent results regarding the order of importance for the explanatory variables. Likewise, Grenouillet et al. (2011) and Guo et al. (2015a) evaluated the variability and uncertainty of an ensemble of several models including GLM, GAM, CART, RF, ANNs, among others, in the prediction of fish distributions in streams and lakes at a national level, respectively. This evaluation included the comparison between the individual predictions provided by every single model and an average of ensemble models. Results indicated that ensemble modeling results improve accuracy when modeling different species and revealed uncertainty dependence upon geographical extent. In addition, Gazendam et al. (2016) developed an ANN model to predict two macroinvertebrate indices (Hilsenhoff's Biotic Index and richness) while testing combinations of physically-based input variables (geomorphic, riparian, hydrology and watershed-level). Results showed that considering both watershed- and reach-scale (geomorphic

and riparian) inputs can improve model performance and are consistent with findings using ordination techniques (D'Ambrosio et al., 2014, 2009).

2.3.2.3 Other methods

Additional statistical and machine learning methods that have been implemented for SDM and habitat suitability include Multivariate adaptive regression splines (MARS), Support Vector Machines (SVMs), and Partial Least Squares Regression (PLSR). Other methods such as Maximum Entropy (Phillips et al., 2006) and Genetic algorithm for rule set production (GARP) (Stockwell and Peters, 1999) are not further described herein because they are intended for using presence-only data. It is worth noting that these two methods are especially suitable when working with small sample sizes and incomplete datasets (Li and Wang, 2013; Yi et al., 2017).

MARS (Friedman, 1991) are multidimensional extensions of GAMs that build up from basis functions fitting separate splines for different intervals (with extremes known as *knots*) of the predictor variables. MARS algorithm finds the location and number knots using an exhaustive search procedure in a forward/backward stepwise fashion (Prasad et al., 2006). MARS models are better suited than CART for continuous variables, can handle large numbers of explanatory variables with low order interactions, and automatically quantify interaction effects (Li and Wang, 2013; Prasad et al., 2006). However its interpretability is limited when analyzing species-environmental relationships, its parameter identification is not straightforward, and it is highly sensitive to extrapolation (Prasad et al., 2006). MARS method is mainly implemented for multi-species modeling (Guo et al., 2015a) and is considered as an alternative to other regression methods such as GLM and GAM (Barry and Elith, 2006; Li and Wang, 2013; Yi et al., 2017).

On the other hand, SVMs (Vapnik, 2000) is a machine learning algorithm that also uses basis functions known as *kernels* to map data into a new nonlinear feature hyperspace attempting

to simplify data patterns. Then, the data is classified while the margins between hyperplanes that are used to define classes are maximized. These hyperplanes are determined by a set of support vectors using quadratic programming. An advantage of SVMs is the reduced number of tuning parameters. Also, SVM models are not prone to overfitting. However, because this method does not provide a simple representation or a pictorial graph, the interpretation of modeling results is difficult (Cutler et al., 2007), the algorithm is computationally demanding, and the tuning parameters are poorly identifiable when data is not linearly separable (Guo et al., 2015b). SVM have been applied mainly for modeling species-habitat relationships for fish communities (Fukuda et al., 2013; Fukuda and De Baets, 2016; Muñoz-Mas et al., 2018, 2016) and for predicting the occurrence and macroinvertebrate-based stream health indices using landscape and water quality attributes (Ambelu et al., 2010; Fan et al., 2017; Hoang et al., 2010; Lin et al., 2016; Sor et al., 2017). In general, the studies have indicated similar performances between ANN and SVM when predicting continuous variables, although SVM usually performs slightly better than ANN. Meanwhile, methods such as RF have shown better results than SVM for classification purposes.

PLSR (Wold et al., 2001) is a method that projects explanatory and response variables into a new space where MLR is performed. PLSR is known for handling multicollinearity and strong correlation of predictors while allowing a high interpretability of the resulting regression coefficients (Villeneuve et al., 2015). Moreover, this method is suitable when the number of explanatory variables is greater than the number of observations (Abouali et al., 2016b). PLSR method has been used as an explicative method for quantifying relationships between several stressors and stream health indicators. This method is especially suitable for identifying the most

relevant input variables before any predictive model training (Abouali et al., 2016b; Einheuser et al., 2013a; Villeneuve et al., 2015).

In an attempt to facilitate the integration with expert elicitation regarding causal relationships, different approaches are being integrated within the SEM framework for stream health modeling (Riseng et al., 2011; Surridge et al., 2014). For instance, a recent study by Villeneuve et al. (2018) shows an application of the aforementioned framework using PLSR for evaluating the direct and indirect effects of multiple stressors on macroinvertebrate indices in nested spatial scales (watershed, reach, and site). Results showed that the direct effects of instream water quality conditions decrease when indirect effects from land use and hydro-morphological alterations are considered for macroinvertebrate assemblages.

2.3.3 Soft Computing Methods

Soft computing is a collection of paradigms that attempt to represent complex systems in an environment of imprecision, uncertainty and partial truth, resembling human mind and biological systems' learning. Soft computing approaches include fuzzy logic, neurocomputing, evolutionary computation and probabilistic reasoning (Zadeh, 1994, 1998).

2.3.3.1 Fuzzy logic-based methods

Fuzzy logic-based methods are used to model nonlinear relationships employing linguistic terms instead of numeric values. These methods incorporate membership functions, fuzzy set operations, and *if-then* rules for mapping from a given input to an output. This process is also called "fuzzy inference" (Ocampo-Duque et al., 2006). The membership functions are curves with values between 0 and 1 that represent the degree of membership of an input variable's value (element) to a certain fuzzy set, where 0 and 1 represent non- and full membership, respectively (Adriaenssens et al., 2006). Given that the same element may belong to several sets at the same time, expert knowledge is necessary to define the overlap between

different membership functions for the same variable. Then, the obtained membership values are combined for different variables using fuzzy *if-then* rules incorporating fuzzy set operations. Outcomes from different fuzzy rules are then aggregated into a final fuzzy score. Fuzzy set operations include union, intersection and additive complement. Following Adriaenssens et al. (2004a), fuzzy *if-then* rules consist of antecedent and consequent parts, which mainly rely on expert knowledge. The antecedent part states conditions for the explanatory variables (i.e. input), while the consequent describes the corresponding values of the response variables (i.e. output). When both parts are concerned with statements that define the value of the variables without considering any explicit function relating the explanatory and response variables, the model belongs to the Mamdani-Assilian type. On the other hand, when the consequent part establishes a linear or nonlinear relationship between the explanatory and response variables, the model belongs to the Takagi-Sugeno type. It would be necessary to implement a weighting operation using a specified decision-making method (e.g. Analytic Hierarchy Process) to define the influence of each input variable in the final fuzzy score (Ocampo-Duque et al., 2006). Once the fuzzy rules are implemented, the defuzzification operation is performed, if necessary, in order to obtain a numerical output under no-fuzzy contexts (Ocampo-Duque et al., 2006). Fuzzy logic models are generally trained employing optimization routines based on the Shannon-Waver entropy (tuning the fuzzy sets parameters looking for entropies values greater than 0.85) or Genetic Algorithms to generate uniformly distributed fuzzy sets (Yi et al., 2017). Then, the nearest ascent hill-climbing algorithm can be used for optimizing the consequent part of each fuzzy rule (Yi et al., 2017).

Compared to ANNs, fuzzy logic-based methods provide a more transparent insight into the influence and interactions of input variables. Moreover, the analysis of the model is more

intuitive and can be performed in a semi-qualitative manner while uncertainty quantification can be easily integrated. However, setting the membership functions and fuzzy rules, which are jointly referred as the knowledge database, is not an easy task (Adriaenssens et al., 2004a). In order to address this issue, several techniques such as Adaptive neuro-fuzzy inference systems (ANFIS) (Jang, 1993) were introduced. ANFIS is a hybrid method that combines ANNs and fuzzy logic. An adaptive network is a multilayer feed-forward neural network where the hidden units may or may not have parameters. These parameters are determined during the training process using observed data by minimizing the predictive error. Therefore, the membership functions are introduced in the first hidden layer of parameterized units. Successive layers, which contain non-parameterized units, represent the *if-then* fuzzy rules while incorporating the fuzzy set operations. Each successive layer attempts to represent the different specified fuzzy rules. The last layer before computing the overall fuzzy score comprises parameterized units. In this layer, the linear and non-linear relationships for Takagi-Sugeno type models are introduced. These relationships are referred as "output membership functions" (Jang, 1993). The number of parameters in the ANFIS model depends on the number of explanatory variables, the number of membership functions per variable, the membership functions shape, and the output membership function type. Thus, the number of parameters should not be greater than the number of available observations for model training in order to avoid overfitting (Woznicki et al., 2016a).

Early fuzzy logic applications for predictive modeling and ecosystem management were reviewed by Adriaenssens et al. (2004a). The integration of expert knowledge, the use of qualitative data, and the capacity to incorporate uncertainty assessments have motivated the growing use of fuzzy logic-based methods in ecological applications (Adriaenssens et al., 2004a; Ocampo-Duque et al., 2006). Recent studies include the prediction of macroinvertebrate

abundance (Adriaenssens et al., 2006; Mouton et al., 2009; Van Broekhoven et al., 2006), ecological status classification (Ocampo-Duque et al., 2007), the development of multimetric stream health indices (Marchini et al., 2009), and the prediction of fish species occurrence, distribution or derived indices (Boavida et al., 2014; Muñoz-Mas et al., 2012). When applied for prediction, fuzzy logic-based models typically use water quality, hydrologic and hydro-morphological attributes and knowledge based on species preferences and tolerances. In an attempt to implement fuzzy logic in a more data-driven fashion, ANFIS has been recently implemented using process-based models to evaluate the impacts of conservation practices and environmental changes on stream health indices at a watershed scale (Einheuser et al., 2013a, 2013b, 2012). For instance, the Soil and Water Assessment Tool (SWAT) has been employed to simulate streamflow series in order to obtain ecologically-relevant hydrologic indices. Those hydrologic indices are later used to predict stream health indices using ANFIS (Herman et al., 2016, 2015). Others studies have also incorporated water quality simulations (sediment and nutrients) for predicting stream health indicators (Woznicki et al., 2016a, 2016b). Moreover, Abouali et al. (Abouali et al., 2016b), employed the same integrated framework to develop a two-phase approach coupling Partial Least Square Regression (PLSR) and ANFIS for predicting macroinvertebrate and fish-based indices. Results showed a significant improvement in the prediction power while omitting the need for variable selection. It is worth noting that ensemble modeling frameworks integrating process-based models are especially suitable for evaluating climate change effects on biological assemblages (Daneshvar et al., 2017a). Recently, Yi et al. (2017) presented a detailed revision of fuzzy logic integration with process-based models intended for habitat modeling. Remarkable applications include micro-habitat

selection/evaluation for fish and macroinvertebrate assemblages and dam operation/removal assessments (e.g. CASiMiR model).

2.3.3.2 Bayesian belief networks

Bayesian belief networks (BBN) are directed acyclic graphs having nodes linked by probabilities (Pearl, 1986). Each node represents constants, discrete or continuous variables, or continuous functions in the model, whereas the arrowed links indicate direct correlation or causal relationships between nodes (McCann et al., 2006). There are two types of nodes: parent or independent nodes (nodes that don't have arrows incoming or outgoing), and child nodes (with one or both incoming and outgoing arrows). Each node has an associated probability distribution, which is unconditional for parent nodes and conditional for child nodes. The outcomes of each node are known as *states* (McDonald et al., 2015). Probability distributions are usually defined for each node in terms of the *states* (i.e. conditional probability tables, CPTs) (McCann et al., 2006). Consequently, a BBN is comprised of a qualitative component referring to the network structure and a quantitative component given by the CPTs within each node (Phan et al., 2016).

The structure of the network is iteratively determined using expert knowledge and/or prior data, where metrics such as correlation coefficients are often used to define causal links. It is important to note that causal links should be carefully defined in order to prevent high levels of uncertainty in the model's outputs (McDonald et al., 2015). Conditional independence tests are often used for learning the BBN structure. However, when multiple BBN are tested, model selection using the Bayesian Information Criterion (BIC), or optimization routines are also employed (Aguilera et al., 2011). Determining relevant variables is usually addressed by knowing that the nodes within a network can be unconditionally separated, or conditionally separated/connected if prior knowledge is given in other nodes (Aguilera et al., 2011). On the other hand, determining the probability values populating CPTs is usually done using prior data

and the Bayes' theorem for probabilities propagation from parent nodes. Training algorithms have been developed in accordance with the BBN description of the joint probability distribution of the nodes within the network (Pérez-Miñana, 2016). CPTs, which are marginal probability distributions, are often parameterized and calculated using approaches based on Monte Carlo simulation (Phan et al., 2016), Gibbs sampling or dynamic discretization (Nojavan A. et al., 2017; Pérez-Miñana, 2016), maximum likelihood or the Laplace correction (Aguilera et al., 2011), or the Expectation maximization algorithm for small and incomplete datasets, and the gradient learning algorithm for large incomplete datasets and continuous data (McDonald et al., 2015). In general, populating CPTs requires either expert knowledge, the use of several data-based methods, or both (Phan et al., 2016).

There are recent reviews covering BBN applications in areas such as environmental modeling (Aguilera et al., 2011), ecosystem services modeling (Landuyt et al., 2013; Pérez-Miñana, 2016), ecological risk assessment (McDonald et al., 2015), and water resources management (Phan et al., 2016), which also include some early studies related to stream health modeling (Adriaenssens et al., 2004b; Marcot et al., 2001). The aforementioned reviews addressed different aspects of BBN model development/application such as data pre-processing, complexity, training, optimization, validation methods, variations and extensions (e.g. dynamic Bayesian networks for time series modeling and Hidden Markov Models for higher order relationships, see Tucker and Duplisea (2012)), integration with other modeling techniques, and software comparison. Moreover, given the extensive application of BBN in ecology and environmental science, there are guidelines addressing good modeling practices (Chen and Pollino, 2012; Marcot et al., 2006), overfitting, uncertainty quantification, and salient issues (Marcot, 2017, 2012).

Representative BBN applications in stream health modeling include the evaluation of cumulative environmental impacts of multiple stressors on ecosystem health using traditional and scientific knowledge (Mantyka-Pringle et al., 2017). The study incorporated biotic factors describing wildlife health, food webs, wildlife populations, fish health, macroinvertebrate metrics (density, richness and diversity), among others. Other recent studies have also addressed the estimation of the interactive effect of land use and climate change (considering its influence on water quality, physical factors and habitat characteristics) on fish population success (Turschwell et al., 2017), EPT macroinvertebrates indices integrating SEM (Li et al., 2018), and both fish and macroinvertebrate richness (Mantyka-Pringle et al., 2014). Likewise, BBN models have been developed for predicting macroinvertebrate indices using land use, physicochemical, and hydro-morphological factors (Allan et al., 2012; Forio et al., 2015; McLaughlin and Reckhow, 2017). However, McLaughlin and Reckhow (2017) could not find strong causal relationships or a high predictive power when relating water quality parameters (e.g. nutrients, chlorophyll, dissolved oxygen) and habitat attributes to benthic macroinvertebrates in streams. The results may indicate that the BNN model is reflecting associations rather than strict causal relationships between variables. In addition, low predictive power might be a consequence of not considering all relevant factors affecting the response variable. However, other studies have been able to identify relevant stressors. For instance, Forio et al. (2015) indicated that flow velocity is a major variable determining stream health, which is highly sensitive to natural streamflow alterations by dams and water abstractions. In addition, Dyer et al. (2014) ended up with a similar conclusion when analyzing the effects of climate change and stream regulation on ecologically-relevant hydrologic indices and water quality attributes using BBN. Moreover, there are studies implementing BBN for environmental flows decision-making processes (Leigh et al.,

2012; Shenton et al., 2014) and water management for fish species conservation (Peterson et al., 2013; Vilizzi et al., 2013). On the other hand, Death et al. (2015) compared BNN with logistic regression, artificial neural networks, classification trees, and random forests while predicting a macroinvertebrate-based stream health index at a national scale. Results indicated that BNN moderately outperformed the other methods; however, the model preparation is more time-consuming. BBN have been also implemented for evaluating the habitat suitability of invasive macroinvertebrate species using purely data-driven and expert knowledge-based approaches (Boets et al., 2015).

2.4 KNOWLEDGE GAP ANALYSIS

Explaining cause-effect relationships between environmental and anthropogenic factors with measures of ecological integrity or health in freshwater ecosystems is not a straightforward task because of the complex, nonlinear and uncertain nature of these systems (Niemi and McDonald, 2004). However, growing applications of statistical and soft computing methods have been employed to address the aforementioned challenges. When selecting a modeling approach, aspects regarding variable selection, interpretability of modeling results, modeling ensembles for increasing predictive ability, and model evaluation and overfitting should be considered. Particularly, we have identified the need for developing guidelines regarding three main aspects of stream health modeling practice: variable selection, model evaluation, and data acquisition and uncertainty analysis for modeling ensembles. In this section, we briefly describe the aspects mentioned above and indicate the corresponding research priorities.

Strategies for variable selection include trial and error, expert knowledge, statistical analysis, heuristics, or combinations of these methods (Falcone et al., 2010; May et al., 2008). However, there is not an agreement about how to proceed with variable selection when

developing stream health models (Woznicki et al., 2015). Variable selection is a critical step in any empirical modeling exercise, compromising performance, efficiency and interpretability (Li et al., 2015). Moreover, the number of selected input variables defines the number of model parameters to be calibrated, the computational effort, and is critical for overfitting prevention (Galelli et al., 2014). According to the studies reviewed in this paper, ordination and classification methods, nonparametric rank correlations and Bayesian methods are usually implemented. For instance, Woznicki et al. (Woznicki et al., 2015) compared PCA, Spearman's rank correlation and Bayesian variable selection when modeling macroinvertebrates and fish based indices with ANFIS. Results showed that Bayesian variable selection provided the best final models (Woznicki et al., 2015). Other variable selection approaches are based on stepwise procedures using relative quality (e.g. using AIC) (Clapcott et al., 2017), backward elimination (Fox et al., 2017), clustering employing SOMs (Bowden et al., 2005), partial mutual information (Fernando et al., 2009) and using interpretability tools based on sensitivity and perturbation analysis and the relative importance of the predictors (Elith et al., 2008; Gevrey et al., 2003). Nevertheless, there is a lack of studies comparing different variable selection methods with different stream health modeling approaches. Thus, further research in this area is encouraged to gain insight into the influence of different ensembles of variable selection approaches and modeling methods over predictive ability and model parsimony.

Traditional statistical methods based on linear regression (e.g. MLR, GLM) are generally transparent and their coefficients can be well interpreted (Li and Wang, 2013). However, those methods usually show the lowest predictive power. For more complex models, there have been initiatives introducing expert elicitation into the model formulation and training. In those cases, fuzzy logic and Bayesian belief networks approaches have played an important role (Mantyka-

Pringle et al., 2017; Mouton et al., 2009). On the other hand, when working with data-driven approaches, some alternatives have been formulated depending on the implemented modeling method. For instance, in decision tree-based methods (e.g. CART, BRT, RF), results are interpreted estimating the relative influence of predictor variables and are visualized and examined using partial dependence plots (Hastie et al., 2009). Relative influence in single trees can be measured based on the number of times a variable is selected for splitting. The number of times is weighted by a squared sum measure of model improvement as a result of adding each split to the individual tree. The final relative influence is obtained by averaging the corresponding results for the individual trees and then standardizing the final values (Elith et al., 2008). A similar option for obtaining relative influence measures can be applied using the out-of-bag samples, assuming that relative decrease in prediction accuracy is related to the variable influence (Carlisle et al., 2009a). The out-of-bag samples observations are used to obtain a decrease in prediction accuracy when the explanatory variables are randomly permuted in each tree. Then, the decrease is averaged and standardized across all trees (Carlisle et al., 2009a). Partial dependence plots attempt to represent the effect of a variable when assigning average values for all other variables in the model. The plots can reveal strong interactions when using multiple variables and are especially suitable for detecting disturbance thresholds or ranges. However, these plots are limited to low-dimensional views (Hastie et al., 2009). Other methods for analyzing the contribution of input variables are based on partial derivatives, perturbation of the input variables, and successive variation in a certain input variable while the remaining are kept constant, among others that are specifically designed for ANNs (e.g. neural interpretation diagram, Garson's algorithm, randomization test, stepwise methods) (Gevrey et al., 2003; Olden and Jackson, 2002). However, to significantly improve the interpretability of modeling results, it

is necessary to advance towards frameworks which incorporate process-based models to describe disturbance factors (Araújo and New, 2007).

Ensemble modeling frameworks have been introduced to improve predictive ability while understanding cause-effect and multiscale dynamics driving instream changes due to alterations in landscape and environmental factors. The inclusion of process-based modeling approaches into integrated stream health modeling frameworks has been developed for different scales (i.e. macro-, meso- and micro-scale). For instance, at a macroscale (e.g. watershed, ecoregion), main advancements are related to the representation of bioclimatic and hydrologic factors using climatic, hydrologic, hydraulic, and water quality models. At meso- (e.g. river segments, hydromorphologic units: pool, riffle, run,…) and micro-scales (e.g. point locations, substrate), computational fluid dynamics, hydraulic, water quality and physical habitat models predicting local velocities, depths and physic-chemical constituents are often implemented (Daneshvar et al., 2017b; Yi et al., 2017). For instance, Jähnig et al. (2012) proposed a framework following the driver-pressure-state-impact concept. Drivers include main watershed and instream stressors (e.g. climate, land use, river alteration). Pressures on freshwater ecosystems comprise hydrologic and hydraulic stress, sediment intake, substrate stability, among others, which can be represented by process-based models developed in several areas such as ecohydrology and ecohydraulics. State refers to the outputs driven by the aforementioned pressures (e.g. extreme events, sediment, velocity, depth, substrate, nutrients). Finally, to evaluate the impact of the different state variables, species distribution, aquatic biodiversity and stream health measures can be obtained (Jähnig et al., 2012). A similar comprehensive framework was recently proposed by Kail et al. (2015), introducing a module for simulating stream channel geomorphological evolution. Other authors have put more attention on states describing the flow regime, using ecologically-relevant

hydrologic indices (Woznicki et al., 2016a) and linkages with climate change scenarios (Daneshvar et al., 2017a; Guse et al., 2015). It is also worth noting that there is a trade-off between model complexity and uncertainty. In ensemble stream health modeling, gaining model predictive power and interpretability implies multiple information sources, an elevated number of model parameters and the persistence of epistemic errors in model formulation. Hence, higher outcome uncertainties might be expected. Therefore, increased efforts studying uncertainty propagation and shrinking in ensemble modeling must be addressed to promote more transparent decision-making processes. On the other hand, with the advent of big data sources and applications, including image processing and long-term monitoring data (Kuemmerlen et al., 2016), recent advances in deep learning are encouraged to be implemented and integrated within existing modeling frameworks (Babbar-Sebens et al., 2015; Lecun et al., 2015). Furthermore, it is necessary to evaluate how the current strategies for biological data acquisition are compatible with data derived from remote sensing products and traditional environmental information systems and networks, and how these strategies can help us to better understand the health of a stream. In summary, stream health modeling should be seen as complementary tools that require continuous validation with field measurements (Kuehne et al., 2017). Therefore, clear guidelines regarding monitoring stream health for modeling purposes should be considered in future studies.

With respect to *model evaluation and overfitting*, the commonly used performance measures depend on the type of response variable. When categorical variables are predicted (e.g. presence/absence, impairment condition), the percentage of correctly classified observations, true statistical skills, sensitivity, specificity, Cohen's kappa, and the area under receiver operating characteristic curve (AUC) are commonly calculated (Manel et al., 2001; Sor et al., 2017).

Meanwhile, using several performance measures when conducting modeling exercises is recommended (Maloney et al., 2009). When predicting continuous response variables (e.g. a stream health index, biomass) commonly used performance measures are the correlation coefficient (r), the coefficient of determination (R^2), the Nash-Sutcliffe efficiency, the root mean squared error (RMSE) and the deviance between observed and predicted values (Goethals et al., 2007). On the other hand, the application of the algorithm's formulation and the k-fold cross validation techniques can be used to minimize model's overfitting as they were widely used in studies reviewed here. However, to the best of our knowledge, there are no standard guidelines for evaluating the stream health model performance. Even though, for other aspects of environmental modeling, several guidelines have been developed including Bennett et al. (2013) and Moriasi et al.(2007, 2015). Therefore, for stream health modeling, a combination of the aforementioned criteria or new criteria should be further evaluated with respect to their applicability/usefulness.

2.5 SUMMARY AND CONCLUSION

In this study, we provided an overview of different statistical, machine learning, and soft computing methods widely used in ecological applications and stream health modeling based on data describing macroinvertebrates and fish assemblages. The main advantages and disadvantages for the reviewed methods are summarized in Table 1. It is worth noting that statistical methods are simpler and more interpretable than other methods, while their prediction power is generally low. On the other hand, models based on machine learning techniques provide a better accuracy in reproducing observed stream health indices, and are more suitable for representing complex, nonlinear systems. Nevertheless, these methods can be difficult to interpret and hardly provide insight into model parameters' meaning and relative importance.

Thus, soft computing methods, which can be integrated with machine learning techniques, are favorable because they allow the insertion of expert elicitation and partial information, enhancing interpretability of ecological models. However, model formulation is usually time consuming; especially for very complex models. Meanwhile, soft computing models that are structured based on expert knowledge are susceptible to misrepresenting causal relationships, and consequently are likely to provide higher structural uncertainties.

Therefore, frameworks supporting the integration of process-based models for driving multi-scale stressors and employing ensembles of different empirical modeling techniques, are being recommended. Meanwhile, these types of modeling techniques are vulnerable to uncertainty propagation resulted from the modeling process and components and data sources, which can affect the consistency and reliability of the modeling results. Meanwhile, there is a growing amount of literature providing better practices for data preparation, optimal model design, model interpretation, performance evaluation, variables relative importance, among others, for specific methods such as decision trees, ANN, fuzzy logic and BBN. Therefore, it is crucial to develop guidelines addressing the aforementioned aspects for stream health modeling practice.

Table 1 Summary of advantages, disadvantages and applications for the methods described in this study

Method	Advantages	Disadvantages	Applications	
			Macroinvertebrates	Fish
Multiple Linear Regression	• Straightforward implementation and interpretation • Computational effort is low	• Low predictive power • Method assumptions (i.e. normality, homoscedasticity) are usually violated • Parameter estimation is unstable under multicollinearity and strong correlated variables	(Merriam et al., 2015, 2013; Pond et al., 2017; Waite et al., 2012, 2010)	(Frimpong et al., 2005; Van Sickle and Burch Johnson, 2008)
Generalized Linear Models	• Straightforward implementation and interpretation • Flexible with the selection of error distributions • Computational effort is low	• Low predictive power • Model structure (distributions selection) must be defined a priori	(Damanik-Ambarita et al., 2016; Death et al., 2015; Donohue et al., 2006; Everaert et al., 2014; Gieswein et al., 2017; Holguin-Gonzalez et al., 2013a, 2013b; Jerves-Cobo et al., 2017; Kuemmerlen et al., 2014; Moya et al., 2011; Pont et al., 2009; Sauer et al., 2011; Van Sickle et al., 2004)	(Fukuda et al., 2013; Gieswein et al., 2017; Grenouillet et al., 2011; Guo et al., 2015a; Hermoso et al., 2011; Kwon et al., 2015; Leclere et al., 2011; Patrick and Yuan, 2017; Sui et al., 2014)
Generalized Additive Models	• Suitable for modeling nonlinear relationships • Uses non-parametric basis functions	• Prone to overfitting • Reduced interpretability of modeling results	(Maloney et al., 2012; Sauer et al., 2011)	(Almeida et al., 2017; Fukuda et al., 2013; Grenouillet et al., 2011; Guo et al., 2015a; Maloney et al., 2012; Zhao et al., 2014)

Table 1 (cont'd).

Ordination methods	• Suitable when analyzing multiple species in multiple sites • Straightforward interpretation • Computational effort is low • Suitable for variable selection and exploratory analysis	• Methods' assumptions (e.g. linearity, unimodality) are likely to be violated • Interpretability is compromised when high correlations are present without clear causal relationships • Some methods are sensitive to the relative scaling and noise of the explanatory variables	(D'Ambrosio et al., 2014; Lin et al., 2016; Pond et al., 2017)	(D'Ambrosio et al., 2014, 2009; Kwon et al., 2015; Patrick and Yuan, 2017)
Classification and Regression Trees	• Do not require assumptions about data distribution • Interactions between predictors are modeled and can be easily visualized	• Smooth functions are poorly modeled • Provide very different results when making small changes to the training data • Large trees are poorly interpretable • Not suitable for modeling continuous datasets (e.g. temporal dynamics)	(Ambelu et al., 2010; Death et al., 2015; Holguin-Gonzalez et al., 2014, 2013a; Maloney et al., 2009; Ocampo-Duque et al., 2007; Sauer et al., 2011; Waite et al., 2012; Wang et al., 2007)	(Grenouillet et al., 2011; Guo et al., 2015a; He et al., 2010; Kwon et al., 2015; Leclere et al., 2011; Wang et al., 2007)
Boosted Regression Trees	• Suitable for modeling smooth functions and interactions between predictors • Insensitive to outliers • Exclude irrelevant predictor variables • Do not extrapolate beyond the range of observations	• Time consuming for large number of trees or low learning rates • Prone to overfitting • Maximum and minimum values for response variables are poorly reproduced • Interactions between multiple (more than three) explanatory variables are difficult to visualize and interpret • Model interpretability is limited	(Brown et al., 2012; Clapcott et al., 2017, 2014, 2012; May et al., 2015; Pilière et al., 2014; Steel et al., 2017; Tonkin et al., 2014; Wagenhoff et al., 2016; Waite et al., 2014, 2012; Waite and Van Metre, 2017)	(Chee and Elith, 2012; Clapcott et al., 2014; Esselman et al., 2013; Golden et al., 2016; Leclere et al., 2011)

Table 1 (cont'd).

Random Forests	• Resistant to overfitting • Cross-validation is not necessary because a similar approach is automatically performed during model training • Good for classification purposes	• Time consuming for large number of trees • Less accurate than BRT • Interactions between multiple (more than three) explanatory variables are difficult to visualize and interpret • Model interpretability is limited • Cannot be extrapolated beyond the range of observations	(Álvarez-Cabria et al., 2017; Booker et al., 2015; Carlisle et al., 2009a; Chinnayakanahalli et al., 2011; Clapcott et al., 2017; Death et al., 2015; Fox et al., 2017; Hill et al., 2017; Patrick and Yuan, 2017; Vander Laan et al., 2013; Waite et al., 2012)	(Álvarez-Cabria et al., 2017; Fukuda et al., 2013; Fukuda and De Baets, 2016; Grenouillet et al., 2011; Guo et al., 2015a; He et al., 2010; Kwon et al., 2015; Olaya-Marín et al., 2013; Patrick and Yuan, 2017; Tuulaikhuu et al., 2017; Vezza et al., 2015)
Artificial Neural Networks	• Suitable for modeling nonlinear relationships • Vast literature addressing aspects such as variable selection, sensitivity analysis, model ensembles, and optimal design • Good performance when modeling continuous data	• Model interpretability is limited • Relative importance of predictor variables is more difficult to determine than other approaches	(Chon, 2011; Gazendam et al., 2016; Goethals et al., 2007; Mathon et al., 2013; Mouton et al., 2010; Sauer et al., 2011)	(Chon, 2011; Fukuda et al., 2013; Grenouillet et al., 2011; Guo et al., 2015a; Mathon et al., 2013; Olaya-Marín et al., 2013, 2012; Olden et al., 2008; Sutela et al., 2010; Tsai et al., 2016)
Multivariate Adaptive Regression Splines	• Suitable for modeling smooth functions • Can handle a large number of explanatory variables with low order interactions • Automatically quantify interaction effects	• Model interpretability is limited • Highly sensitive to extrapolation (prone to under and overestimation) • Model parameters are difficult to identify	(Sauer et al., 2011)	(Hermoso et al., 2011; Kwon et al., 2015; Leathwick et al., 2006b)
Support Vector Machines	• Suitable for modeling nonlinear relationships • Reduced number of algorithm parameters. • Overfitting is unlikely • Good performance when modeling continuous data	• Model interpretability is limited • Algorithm is computationally expensive • Model parameters are difficult to identify when data is not linearly separable	(Ambelu et al., 2010; Fan et al., 2017; Hoang et al., 2010; Lin et al., 2016; Sor et al., 2017)	(Fukuda et al., 2013; Fukuda and De Baets, 2016; Kwon et al., 2015; Muñoz-Mas et al., 2018)

Table 1 (cont'd).

Partial Least Squares Regression	• Handles multicollinearity and strong correlation of predictors • Straightforward interpretation • Suitable when the number of explanatory variables is greater than the number of observations	• Interpretability is compromised when high correlations are present without clear causal relationships • Sensitive to the relative scaling and noise of the predictor variables	(Abouali et al., 2016b; Riseng et al., 2011; Surridge et al., 2014; Villeneuve et al., 2018, 2015)	(Abouali et al., 2016b; Einheuser et al., 2013a; Villeneuve et al., 2015)
Fuzzy Logic-based	• Provides insight into the influence and interactions of explanatory variables. • Uncertainty quantification is easily integrated into the models • Suitable for including expert elicitation and partial information along data	• Computational effort rapidly increases with the number of predictors • Introduction of expert elicitation into models can be time consuming	(Adriaenssens et al., 2006; Herman et al., 2016; Herman and Nejadhashemi, 2015; Marchini et al., 2009; Mouton et al., 2009; Ocampo-Duque et al., 2007; Van Broekhoven et al., 2006; Woznicki et al., 2016b, 2016a)	(Abouali et al., 2016b; Boavida et al., 2014; Einheuser et al., 2013a, 2013b, 2012; Fukuda et al., 2013; Fukuda and De Baets, 2016; Herman et al., 2016, 2015; Muñoz-Mas et al., 2012; Woznicki et al., 2016b, 2016a; Yi et al., 2017)
Bayesian Belief Networks	• Uncertainty quantification is easily integrated into the models • Suitable for including expert elicitation and partial information with data • Able to handle missing values in input dataset • Can be extended to account for feedback loops and time series modeling	• Computational effort and data demand rapidly increases with the number of variables • Loss of accuracy and information because of variable discretization • Model performance greatly depends on the qualitative network definition (i.e. network formulation itself is an important source of error) • Time series modeling is computationally demanding	(Allan et al., 2012; Boets et al., 2015; Death et al., 2015; Forio et al., 2015; Li et al., 2018; Mantyka-Pringle et al., 2014; McLaughlin and Reckhow, 2017)	(Mantyka-Pringle et al., 2017, 2014; Peterson et al., 2013; Turschwell et al., 2017; Vilizzi et al., 2013)

3 INTRODUCTION TO METHODOLOGY AND RESULTS

This dissertation is comprised of three studies developing a framework for linking multi-objective calibration and uncertainty quantification for ecohydrological models. The first study evaluates the impacts of two multi-objective calibration strategies in the replication of a comprehensive list of ecologically relevant hydrologic indices. The second study builds upon the first study by introducing an optimization constraint for improving the representation of a subset of hydrologic indices. Furthermore, different categories of hydrologic indices targeting distinct streamflow regime facets are explicitly considered during the objective functions' formulation. The third study links the advances from the previous two studies to quantify the uncertainty of ecologically relevant hydrologic indices using Bayesian parameter estimation.

The first study, titled "Evaluation of the Impacts of Hydrologic Model Calibration Methods on Predictability of Ecologically-relevant Hydrologic Indices", evaluated the performance of multi-objective model calibration in replicating 167 hydrologic indices of ecohydrological interest using the median values of near-optimal Pareto solutions. Two calibration strategies were compared. The first strategy consisted of three objective functions based on the Nash-Sutcliffe Efficiency (NSE), each one accentuating different flow conditions. The second strategy explicitly divided the streamflow time-series into three segments representing low, moderate, and high flows using the 25% and 75% flow quantiles as thresholds. Then, an objective function based on the root-mean-square error (RMSE) was formulated for each portion. The Non-dominated Sorting Genetic Algorithm III (NSGA-III) was implemented to obtain near-optimal Pareto solutions under each strategy. SWAT was used to simulate daily streamflows at the outlet of the Honeyoey Creek – Pine Creek Watershed, located in east-central Michigan, US. Then, the MATLAB Hydrologic Index Tool (MHIT) was used to compute the

167 hydrologic indices for each near-optimal Pareto solution. Pareto solutions were clustered into three groups using the *k*-means method. Generalized Least-Squares (GLS) was used to analyze the difference among the different clusters in their prediction of average streamflows. Meanwhile, the replication of hydrologic indices was evaluated using a ±30% relative error range as reference. Finally, the performance of multi-objective calibration was compared against traditional single-objective model calibration using different NSE versions targeting different flow conditions.

The second study, titled "A Novel Multi-Objective Model Calibration Method for Ecohydrological Applications", developed calibration strategies for generating a balanced representation of the overall streamflow regime in terms of magnitude, frequency, duration, timing, and rate of change. The second study used the same hydrological model and study area as the first study. Two multi-objective calibration strategies were evaluated based on the findings of the first study. On one side, the first strategy selected six objective functions representing as many hydrologic indices as possible within a ±30% relative error range. Moreover, an optimization constraint was devised targeting a subset of indices of ecohydrological interest to be within the error range. This subset was comprised of 32 Indices of Hydrologic Alteration (IHA) describing the central tendency of streamflow attributes and seven indices (a.k.a. Magnificent seven) representing fundamental stochastic properties of streamflow time series. On the other side, the second strategy consisted in the formulation of six objective functions, each one explicitly targeting groups of hydrologic indices representing a particular streamflow regime facet. These hydrologic indices were part of the same subset of 39 indices. The Unified Non-dominated Sorting Genetic Algorithm III (U-NSGA-III) was applied to generate near-optimal Pareto solutions under each strategy. Additionally, preferred tradeoff solutions were identified

using various multicriteria decision-making methods. Results for both strategies were compared in terms of performance of the near-optimal Pareto solutions and preferred tradeoff solutions, accuracy in the replication of the subset of hydrologic indices, the representation of water balance and flow duration curve characteristics, and accuracy in the replication of hydrologic indices' variability.

The final study, titled "Probabilistic Predictions of Ecologically Relevant Hydrologic Indices Using a Hydrological Model", evaluated the effects of prior knowledge obtained from multi-objective optimization on the uncertainty analysis of simulated hydrologic indices of ecohydrological interest. For this purpose, two experiments were formulated. In the first experiment, non-informative priors were considered when calibrating model and error parameters using Bayesian parameter estimation. In the second experiment, near-optimal Pareto solutions from multi-objective calibrations were used to build a multivariate prior distribution for calibrating model and error parameters using Bayesian inference under an independent time period from the one used for multi-objective calibration. In both experiments, the same likelihood function was employed, considering heteroscedasticity and autocorrelation effects. The multi-objective strategy used here was the same as the one used for the second study's first strategy. The U-NSGA-III algorithm was used for multi-objective calibration, the multiple-try Differential Evolution Adaptive Metropolis$_{(ZS)}$ (MT-DREAM$_{(ZS)}$) algorithm was implemented as the Markov Chain Monte Carlo method for sampling the posterior distributions, and the hydrological model, study area, and subset of hydrologic indices were the same as the second study. The reliability, precision, and bias in streamflow and hydrologic indices predictions were evaluated and compared for each experiment.

4 EVALUATION OF THE IMPACTS OF HYDROLOGIC MODEL CALIBRATION METHODS ON PREDICTABILITY OF ECOLOGICALLY-RELEVANT HYDROLOGIC INDICES

4.1 INTRODUCTION

Alterations driven by human interventions and changing environmental conditions are threatening water security and freshwater biodiversity around the world (Bunn and Arthington, 2002; Carpenter et al., 2011; Dudgeon et al., 2006; Hipsey et al., 2015; Vörösmarty et al., 2010). Traditionally, stream condition evaluation has used chemical and microbiological constituents as criteria (Karr and Yoder, 2004). However, the lack of holistic approaches resulted in further degradation of aquatic ecosystems (Hering et al., 2010; Jelks et al., 2008). To overcome this issue, biological assessments were introduced to provide additional insight into the overall ecological integrity of streams (US EPA, 2011; Woznicki et al., 2016a), and therefore, can be used for environmental management and decision-making.

Biological assessments measure the biota (e.g. fish, benthic macroinvertebrates, periphyton, amphibians) within a stream to obtain information regarding its biological integrity (US EPA, 2011). In this context, biological integrity is the capacity to support and maintain a "balanced, integrated, and adaptive" biological system within the expected structure and function of the natural habitat of a particular region (Karr, 1996; Karr and Dudley, 1981). Stream health integrates the physical, chemical and biological integrity of a stream, which supports living systems that are necessary for human well-being (Karr, 1999; Maddock, 1999).

Stream health indices are generally classified into biotic indices, multi-metric indices, and multivariate methods (Herman and Nejadhashemi, 2015). Biotic indices use only one metric, and multi-metric indices use multiple metrics to evaluate stream health (Herman and

Nejadhashemi, 2015). Biotic metrics are individual characteristics comprised of species abundance and condition, species richness and composition, or trophic composition (Herman and Nejadhashemi, 2015). Multivariate methods use the reference condition approach to predict ratios of taxa observed vs. expected – O/E, and implement statistical and modeling tools that relate environmental features with observed organisms. These tools include cluster analysis, ordination techniques, discriminant analysis, Artificial Neural Networks, self-organizing maps, evolutionary algorithms, Bayesian Belief Networks, and others (Abbasi and Abbasi, 2012; Feio and Poquet, 2011). However, due to limited economic resources, it is not possible to obtain biotic metrics or O/E ratios for all streams within a watershed. Therefore, stream health evaluation based on field data is very limited. To address this difficulty, several modeling approaches have been introduced to extend the available information to ungauged locations (Woznicki et al., 2015).

Streamflow regime has been recognized as a key determinant for sustaining biodiversity and ecological integrity. Thus, ecologically-relevant hydrologic indices are often used as predictors for stream health models besides landscape factors and water quality indicators (Poff and Zimmerman, 2010; Woznicki et al., 2016a). Prediction of ecologically-relevant hydrologic indices include the use of regional statistic approaches (Carlisle et al., 2009b; Dhungel et al., 2016; Knight et al., 2012; Patrick and Yuan, 2017; Sanborn and Bledsoe, 2006; Yang et al., 2016), and hydrological modeling (Caldwell et al., 2015; Kennen et al., 2008; Kiesel et al., 2017; Olsen et al., 2013; Vis et al., 2015; Wenger et al., 2010; You et al., 2014). The use of hydrological models is especially preferred when it is necessary to evaluate the change of stream health driven by modifications in land use, environmental conditions or management practices (Poff et al., 2010; Shrestha et al., 2016; Woznicki et al., 2016b). However, hydrologic models'

ability to replicate ecologically-relevant indices is limited. Some studies have shown that the use of typical calibration approaches (i.e. single-objective based on widely used performance metrics) produces poor representations of some streamflow regime characteristics (Murphy et al., 2013; Vis et al., 2015). For instance, while average conditions are generally well-predicted, low- and high-flow indices are frequently over or under predicted (Wenger et al., 2010). Moreover, no model has been found to provide all selected ecologically-relevant hydrologic indices within ± 30% of the observed values (Caldwell et al., 2015; Vis et al., 2015). Therefore, several studies have proposed to explicitly include ecologically-relevant hydrologic indices into the objective functions for model calibration to improve the overall performance of streamflow regime simulations (Murphy et al., 2013; Shrestha et al., 2014; Vis et al., 2015). For instance, Kiesel et al. (2017) and Zhang et al. (2016) used multi-metric (i.e. aggregated) objective functions based on a reduced number of ecologically-relevant hydrological indices (12 and 16 indices, respectively). They found that it is possible to obtain better overall representations of streamflow regime compared to objective functions based only on conventional performance metrics (e.g. coefficient of efficiency, mean squared errors, correlation coefficient). However, optimal solutions were still unable to effectively represent all hydrological indices individually, especially when they are not explicitly included in the objective function formulation (Kiesel et al., 2017). In addition, optimal solutions are sensitive to the weights assigned to each ecological-relevant hydrological index in the multi-metric objective function (Zhang et al., 2016). On the other hand, different authors have suggested the use of typical performance metrics with proper transformations (Garcia et al., 2017; Oudin et al., 2006; Pushpalatha et al., 2012) or explicit hydrograph partitioning (Pfannerstill et al., 2014) for model calibration. However, these approaches have mainly shown improvements in the representation of target flow conditions

rather than the overall streamflow regime, or have been evaluated using traditional performance metrics instead of ecologically-relevant hydrological indices.

Furthermore, the aforementioned studies have mainly implemented single-objective algorithms for model calibration. Therefore, these previous studies present no integrated perspectives on relationships between different performance metrics or hydrological indices involved in the model calibration process. On the other hand, multi-objective optimization algorithms are very useful for evaluating the tradeoffs between different metrics and objective functions involved in hydrologic model calibration (Price et al., 2012), and can provide sets of solutions able to represent different flow conditions (Efstratiadis and Koutsoyiannis, 2010; Reed et al., 2013). However, Pareto-optimal solutions are not usually evaluated employing ecologically-relevant hydrological indices, but instead pure hydrological signatures based on, for example, Flow Duration Curve (FDC) segments or runoff ratios (Shafii and Tolson, 2015; van Werkhoven et al., 2009). Moreover, many studies have been more concerned about selecting a best single solution than analyzing the whole set of Pareto-optimal solutions, which could provide better results for overall streamflow regime representation. For instance, Vis et al. (2015) attributed high model efficiencies when using the median of several optimum solutions as a "more robust prediction" for ecologically-relevant hydrologic indicators. Therefore, the objective of this study is to identify an objective function best suited for stream health model applications using a multi-objective optimization algorithm for model calibration. Typical performance metrics that represent different flow conditions and explicit hydrograph partitioning are considered in this study. For this purpose, the Soil and Water Assessment Tool (SWAT) watershed model and the NSGA-III multi-objective optimization algorithm are jointly implemented. Then, a set of 167 ecologically-relevant hydrologic indices is used for evaluating

the ability of the resulting Pareto-optimal solutions in representing the overall streamflow regime.

4.2 MATERIALS AND METHODS

Two strategies based on a many-objective optimization technique for model calibration were compared to evaluate their abilities to predict ecologically-relevant hydrologic indices (Figure 1). The first multi-objective strategy calibrates the model based on three different forms of Nash-Sutcliffe Efficiency (NSE) described by Krause et al. (2005) and Pushpalatha et al. (2012) that are suitable for evaluating high, medium, and low flows. In the second strategy, observed daily flow time series were divided into three categories (high, medium, and low flows) using explicit thresholds for low and high flow (flows exceeded 75% and 25% of the time, respectively). For each category, an objective function based on the root-mean-square error (RMSE) was computed. Pareto-optimal solutions for calibration model parameters were obtained for each multi-objective strategy employing the NSGA-III algorithm (Deb and Jain, 2014; Jain and Deb, 2014).

For both strategies, the Soil and Water Assessment Tool (SWAT) (Arnold et al., 2012; Neitsch et al., 2011) was used to simulate daily streamflow discharge time series for every stream segment, and the MATLAB Hydrological Index Tool (Abouali et al., 2016a) was employed to calculate 171 hydrologic indices intended to characterize streamflow regime (Olden and Poff, 2003). Hydrologic indices were computed for each Pareto-optimal point obtained from both multi-objective strategies and for the observed flow dataset. Then, model outputs and indices were evaluated with respect to the observed values using statistical analysis. For this purpose, Pareto-optimal points for each multi-objective strategy were clustered into three groups using the *k*-means method. Generalized Least-Square (GLS) estimation, considering

autocorrelation for the residue, was implemented for streamflow time-series. Meanwhile, the median errors between the hydrologic indices based on Pareto-optimal solutions and observed time series were evaluated with respect to the ±30% uncertainty bound for the observed values, as reported by previous studies (Caldwell et al., 2015; Kennard et al., 2010a; Vis et al., 2015). Finally, results were compared with the optimal solution obtained using a single-objective approach with an objective function based on the standard NSE.

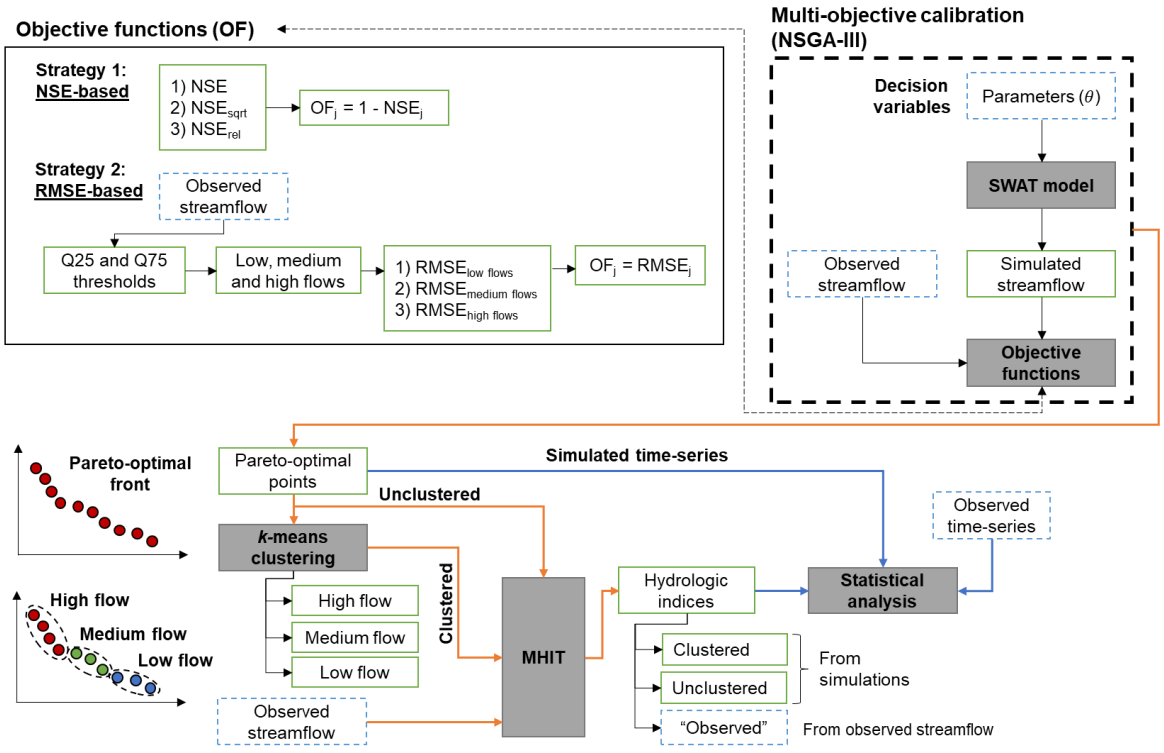


Figure 1 A schematic diagram presenting the overall multi-objective model calibration and evaluation process. Q25 and Q75 are the flows exceeded 25% and 75% of the time, respectively, NSE is the standard Nash-Sutcliffe Efficiency, NSE_{sqrt} is the root-squared-transformed NSE, NSE_{rel} is the relative NSE, RMSE is the Root-Mean-Square Error, and MHIT is the MATLAB Hydrological Index Tool

4.2.1 Study area

In order to perform environmental flow analysis, it is desirable to identify areas where urbanization is limited, streamflow is not regulated or its alteration is negligible, and observed discharge records are available for almost all the studied period (Olden and Poff, 2003). The

Honeyoey Creek–Pine Creek Watershed, with a drainage area of 1,010 km^2, meets all the criteria, because urbanization is less than 4%, streamflow regulation is limited, and observed streamflow data for the period is complete. The study area is located in the Saginaw Bay Watershed in east-central Michigan (Figure 2), which is the largest watershed in the state and is identified as an area of concern by the US Environmental Protection Agency (USEPA, 2015). The watershed has an average slope of 1.9% ranging from 12-39% in the headwaters to 0-1.4% in the lowlands (USGS, 2018). The region has a temperate climate with distinct seasons (Andresen and Winkler, 2009). The average annual rainfall is about 840 mm for the period 1981-2010 (NOAA-NCEI, 2020). However, the precipitation regime is bimodal, with maxima in May and September, and minima in February and July. Mean annual air temperature in the watershed is 9 °C, with a minimum monthly temperature of -9 °C in January and a maximum monthly temperature of 28 °C in July. The dominant land use is agriculture, covering about 50% of the watershed, followed by forests (24%), wetlands (16%) and pasturelands (7%) (USDA-NASS, 2012). Over 60% of the river network's riparian vegetation has not been altered by human activities. The dominant soil textures are loamy sand, sandy loam, loam/clay loam, and sand, which cover about 30, 26, 20 and 11% of the study area, respectively (USDA-NRCS, 2020). The average flow is about 103 m^3/s at the outlet of the watershed. High flows occur between March and May as a result of snow melting and high precipitation, while low flows occur between July and October, during summer and fall seasons (USGS, 2020). High flows, considered in this study as those that are exceeded at most 25% of the time (Q25), have magnitudes above 11.3 m^3/s while low flows are defined as those with values below 3.9 m^3/s, which is the discharge exceeded 75% of the time (Q75).

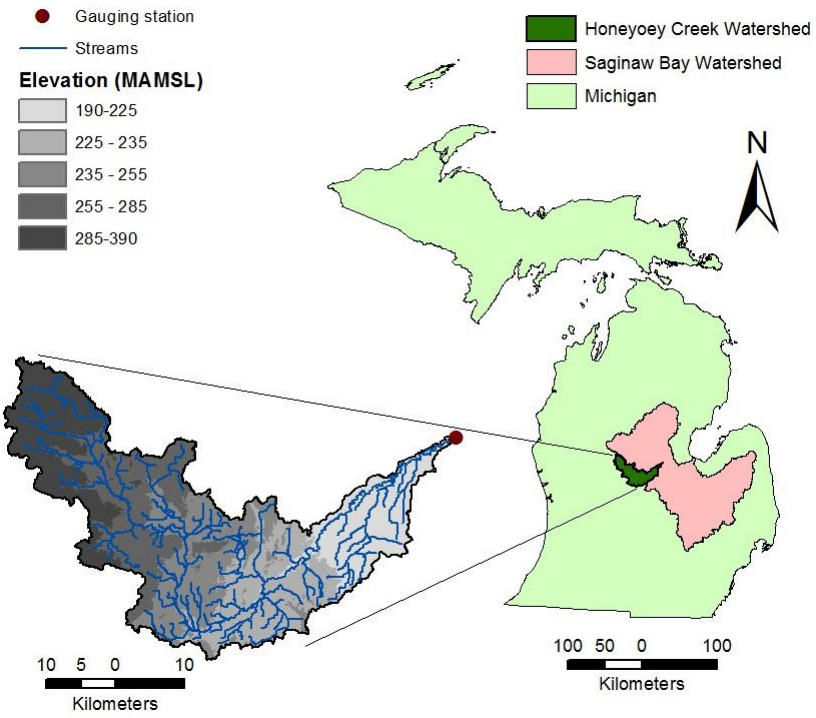


Figure 2 Location and topography of the study area

4.2.2 Data Collection

Datasets required for the hydrologic modeling comprise topography, land use, soil properties, climate, and observed streamflow discharge. The National Elevation Dataset from the US Geological Survey (USGS) with 30 m spatial resolution was used to represent the topography of the watershed (USGS, 2018). The land use characteristics were obtained from the Cropland Data Layer developed by the National Agricultural Statistics Service of the US Department of Agriculture (USDA-NASS) with 30 m spatial resolution (USDA-NASS, 2012). The soil properties were compiled from the Natural Resources Conservation Service's (NRCS) Soil Survey Geographic (SSURGO) Database, at a scale of 1: 250,000 (USDA-NRCS, 2020). Daily time series for precipitation and temperature from 2001 through 2014 were obtained from two weather stations that belong to the National Climatic Data Center (NOAA-NCEI, 2020). Relative

humidity, solar radiation and wind speed time series for the same time span were provided by the stochastic weather generator WXGEN (Neitsch et al., 2011) included in SWAT. Daily streamflow discharges between 2003 and 2014 were obtained from the Pine River Near Midland gauging station (ID 04155500) (USGS, 2020).

4.2.3 SWAT Model description

The Soil and Water Assessment Tool (SWAT version 2012, rev. 614) is a semi-distributed, continuous-time, process-based hydrological model, which simulates water flow, sediment transport, and water quality processes in watersheds (Arnold et al., 1998). SWAT divides a watershed into subwatersheds that are further discretized into multiple units with homogeneous land use, slope, and soil characteristics known as hydrologic response units (HRU). The main processes in SWAT include snow accumulation and melting, evapotranspiration, infiltration, percolation losses, surface runoff, channel routing, and groundwater flows (Neitsch et al., 2011).

SWAT is used in this study for daily flow simulation between 2003 to 2014 for all 749 defined stream segments in the Honeyoey Creek–Pine Creek watershed. Fifteen parameters were selected for model calibration whose description and ranges of variation are presented in Table 2. The calibration period was defined between 2003 and 2008, while the validation period spans between 2009 to 2014. Meanwhile, two years of warm-up period (2001-2002) were considered to stabilize initial conditions of soil water (Cibin et al., 2010).

4.2.4 Hydrologic indices

The 171 hydrologic indices reported by Olden and Poff (2003) are evaluated for all Pareto-optimal solutions after completing each multi-objective model calibration. Then, these indices are compared with the respective indices for the observed dataset, including the calibration and validation periods. The evaluated hydrologic indices characterize the streamflow

regime in terms of magnitude, frequency, duration, timing and rate of change of flows (Poff et al., 1997; Richter et al., 1996) for a given daily time-series. These indices are classified into eleven groups: magnitude for low (ML), average (MA), and high (MH) flow conditions; frequency for low (FL), and high (FH) flow conditions; duration for low (DL), and high (DH) flow conditions; timing for low (TL), average (TA), and high (TH) flow conditions; and rate of change for average (RA) flow conditions. The hydrologic indices are computed using the MATLAB Hydrological Index Tool (MHIT), which has shown better computing performances in comparison to other available packages when handling high number of datasets (Abouali et al., 2016a).

4.2.5 Objective functions

In this study, we contrast the ability of two commonly used procedures in representing a wide number of the streamflow metrics related to environmental flows indicated in section 4.2.4. Each procedure refers to a specific three-dimensional objective space (Figure 3). In summary, the first strategy utilizes three different NSE-based efficiency criteria to evaluate the efficiency of high, medium (overall), or low flow conditions. In the second strategy, efficiency computation is done after flow time series are explicitly partitioned into high, medium, and low flows using statistical thresholds for flow separation. Further details are presented next.

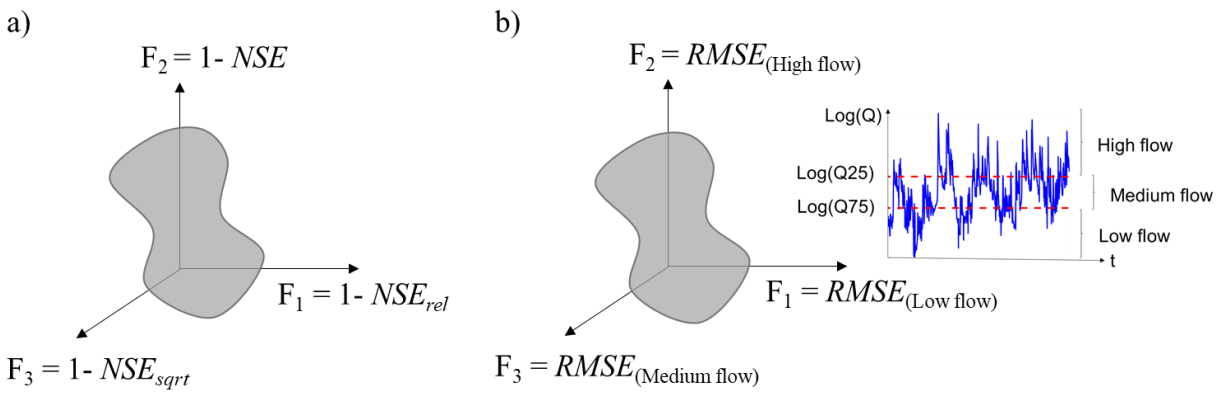


Figure 3 Objective spaces for the SWAT model calibration: a) using different forms of Nash-Sutcliffe efficiency; b) after hydrograph partitioning using Q25 and Q75 thresholds

4.2.5.1 Nash-Sutcliffe efficiency-based objective functions

In this strategy, NSE-based objective functions are formulated to represent different parts of the observed hydrograph. Krause et al. (2005) and Pushpalatha et al. (2012) indicated that standard NSE, Eq. 1 (Nash and Sutcliffe, 1970) is very sensitive to high flows on continuous simulations, given that the differences between simulated and observed values are squared. On the other hand, NSE calculated on root squared transformed flows (Eq. 2, NSE_{sqrt}) has been found to provide a more balanced performance because the errors are more equally distributed on high and low flow portions of the hydrograph (Oudin et al., 2006; Pushpalatha et al., 2012). Additionally, relative NSE (Eq. 3, NSE_{rel}), described by Krause et al. (2005), suppresses the influence of peak flows on the efficiency computation, making it more sensitive to low flows. Pushpalatha et al. (2012) showed that NSE computed on the reciprocal of flow values (inverse transformed flows) is better suited for low flow conditions, focusing on the 20% lowest flows on average. However, in this study we decided to use the NSE_{rel} given that some of the ecologically-relevant hydrologic indices (e.g., low flow index, base flow indices, indices based on moving averages) computed by MHIT for low flows are based also on overall flow values. Therefore, NSE, NSE_{sqrt}, and NSE_{rel} were used to represent high, medium, and low flows, respectively.

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}\left(O_i - P_i\right)^2}{\sum_{i=1}^{n}\left(O_i - \overline{O}\right)^2} \quad (1)$$

$$\mathrm{NSE}_{\mathrm{sqrt}} = 1 - \frac{\sum_{i=1}^{n}\left(\sqrt{O_i} - \sqrt{P_i}\right)^2}{\sum_{i=1}^{n}\left(\sqrt{O_i} - \overline{\sqrt{O}}\right)^2} \quad (2)$$

$$\mathrm{NSE}_{\mathrm{rel}} = 1 - \frac{\sum_{i=1}^{n}\left(\frac{O_i - P_i}{O_i}\right)^2}{\sum_{i=1}^{n}\left(\frac{O_i - \overline{O}}{\overline{O}}\right)^2} \quad (3)$$

where, *O* and *P* are the observed and predicted values, respectively. For all NSE-based criteria, the objective functions (OF) were minimized by computing 1 – NSE, which have a range between zero and infinite. Values for any form of NSE range from minus infinite to one, while the corresponding OFs range from zero to infinite. A perfect fit between simulated and observed values is achieved when all NSE are equal to one and corresponding OFs are equal to zero.

4.2.5.2 Root-Mean-Square Error-based objective functions

In this strategy, RMSE-based objective functions are formulated for streamflow calibration. The time series for the entire study period (2003-2014) was divided into three categories representing high, medium, and low flows using the Q25 and Q75 thresholds obtained from the observed data. To have the same amount of observed and simulated points in each category, the simulated time series are divided following the observed time series partitioning. Then, the RMSE is computed for each flow category:

$$\mathrm{RMSE}_{\mathrm{j}} = \sqrt{\frac{1}{n_j}\sum_{i=1}^{n_j}\left(O_i - P_i\right)^2} \quad (4)$$

where, j refers to the flow category and n_j is the number of observations for each category. Each minimization OF is equal to the computed RMSE for each category. A perfect fit between observed and simulated values yields an RMSE equal to zero.

4.2.6 Many-objective optimization algorithm

Multi-objective evolutionary algorithms (MOEAs) are population-based heuristic search methods that use randomly generated points that move towards a Pareto-optimal front using evolutionary operators (Coello Coello et al., 2007). MOEAs have been widely implemented during the last two decades for water resources applications (Efstratiadis and Koutsoyiannis, 2010; Maier et al., 2014; Reed et al., 2013). For instance, the Non-dominated Sorting Genetic Algorithm II (NSGA-II) (Deb et al., 2002a) has been widely-used for hydrologic model calibration (e.g. Bekele and Nicklow, 2007; Confesor and Whittaker, 2007; Lu et al., 2014; Shafii and De Smedt, 2009). The popularity of NSGA-II is mainly given by its simplicity, modularity, parameter-less property, and good performance for difficult two-objective problems (Deb and Gupta, 2006). However, without any extensions or combinations with other approaches, the NSGA-II by itself has shown shortcomings for solving problems with three or more objectives (Deb and Jain, 2014; Reed et al., 2013; Sindhya et al., 2013).

NSGA-III, which is based on the NSGA-II framework, is the evolutionary many-objective optimization algorithm used to implement the two multi-objective calibration strategies. NSGA-III is an elitist reference-point-based procedure that uses non-domination sorting to solve problems with four or more objectives. This procedure has also shown good performance solving cases with three objectives (Seada and Deb, 2016). The main difference between NSGA-II and NSGA-III is the niching method which is a procedure to maintain diversity among solutions (Deb, 2001). NSGA-II uses crowding distances, while NSGA-III is reference-directions-based (Deb and Jain, 2014; Jain and Deb, 2014). A reference direction is a

line that crosses both the origin and a supplied reference point in the objective space. Selection operation is not explicit in NSGA-III given that for each reference direction, only one population individual is expected (Seada and Deb, 2016). The general outline of the algorithm is as follows:

1. The algorithm begins by generating a population of size *N* and a number of *H* reference points distributed in the objective space with *M* dimensions (i.e., number of objectives is *M*). The number of reference points is $H = \binom{M+p-1}{p}$, where *p* is the number of divisions, along each objective, used to distribute reference directions on the front. The parameter *p* is chosen suitably so as to create a population size adequate to hold a number of trade-off solutions.
2. Next, NSGA-III proceeds in a similar fashion as NSGA-II. Using recombination and mutation, the current parent population P_t is used to generate an offspring population Q_t. The parent and offspring populations are combined into $R_t = P_t \cup Q_t$ (of size 2*N*), and then the R_t members are sorted using non-domination ranking. A new intermediate set S_t is generated selecting the first Pareto front until the size of S_t is equal or greater than *N*. The rank of the last selected individual in S_t is obtained, corresponding to the last front F_l. The population members included in S_t but not included in F_l (expressed as $S_t \setminus F_l$) are directly selected for the next generation P_{t+1}.
3. The new population P_{t+1} is completed by selecting population members from F_l based on the NSGA-III niching method. For this purpose, objective values and supplied reference points are normalized to have a commensurate range. After the normalization, the ideal point coincides with the origin of the objective space. Next, reference directions are constructed by joining the ideal point with each reference point. Then, each member of $S_t \setminus F_l$ is associated with a reference point according to its proximity (i.e. perpendicular distance) to the corresponding reference direction. Reference points that have the least number of related

population members in $S_t \setminus F_l$ are considered to be associated with a member of F_l. Each member of F_l set is therefore selected one-at-a-time making the latter association to fill the remaining slots for P_{t+1}.

4. The whole evolutionary process is repeated until a predefined termination criterion is reached (e.g., number of generations/function evaluations, negligible improvement of Pareto-optimal solutions, small change in performance metrics).

NSGA-III's parameters are the population size (equal to the number of reference points), stopping-criteria (in this case, the number of generations), crossover and mutation probabilities, and distribution indices for each genetic operation (i.e., Simulated Binary Crossover – SBX, and polynomial mutation). The NSGA-III implementation used in this study was programmed in Java and was provided by the Computational Optimization and Innovation (COIN) Laboratory at Michigan State University. The Java code was adapted for this study to have a connection with SWAT to perform the automatic calibration process.

4.2.7 Model evaluation

In order to perform the statistical analysis for model evaluation, the Pareto-optimal points are clustered into three groups representing high, medium, and low flow conditions. Hence, the *k*-mean clustering method (Arthur and Vassilvitskii, 2007) is employed for each calibration strategy in order to identify separate sets of solutions that show better performances for each flow condition. The three clusters are identified for each calibration method, using the corresponding objective functions presented in section 4.2.5. Thus, Pareto-optimal solutions with the highest NSE and NSE_{rel} values are going to be collected in the high and low flow clusters, respectively. Solutions with balanced NSE and NSE_{rel} values will comprise the medium flow cluster. Then, the simulated streamflow time series from each group are compared with the

observed dataset, to evaluate whether the estimated mean differences between the simulations and observations are significant. Therefore, the difference of each simulation with respect to the observed dataset is used as response to fit a simple regression with intercept using GLS estimation with Autoregressive model with lag 1, or AR (1), to account for the serial correlation for the time series. The differences are considered significant when the reported *p*-value is less than 0.05 (i.e. 95% confidence interval for the estimated mean of the distribution does not span zero), with positive (or negative) values for the difference indicating over/ under-estimation of the actual observation. The process was repeated comparing high, medium, and low streamflow categories using the Q25 and Q75 thresholds defined for the RMSE-based calibration strategy. However, because the extracted time series for each flow category are irregularly spaced temporally, we modeled the difference from the corresponding samples between the observed series and each simulated series. Three different methods were used: Normal distribution, Student's t-distribution, and GLS with a Continuous Autoregressive model with lag 1, or CAR (1). Then, we determine the most appropriated test based on the smallest Akaike Information Criterion (AIC) value.

On the other hand, the hydrologic indices obtained for each Pareto-optimal solution are also grouped following the same three clusters determined above (high, medium, and low flow conditions). For each cluster and hydrologic index, the difference between the simulated and observed values are computed and divided between the respective observed values to obtain the relative error. Then, it is determined whether the median relative errors for each group are within or outside the ±30% uncertainty bound. The comparison described above is also performed with no clustering of the Pareto-optimal solutions, in order to evaluate the effect of accounting for all solutions in the median values of the predicted hydrologic indices.

4.3 RESULTS AND DISCUSSION

4.3.1 Convergence and spread of Pareto-optimal fronts obtained with multi-objective calibration strategies

The NSGA-III algorithm was implemented for both NSE- and RMSE-based strategies using a population size of 100 points, a maximum number of 500 generations, a crossover probability of 0.9, a mutation probability of 1/15 (i.e., the reciprocal of the number of calibration parameters), and distribution indices of 10 and 20 for SBX and polynomial mutation, respectively. The convergence to the Pareto-optimal front was evaluated using the hypervolume indicator, which is a measure of the volume enclosed by a Pareto front with respect to a specified reference point (Auger et al., 2009). The Pareto-optimal front was selected when a steady behavior of the hypervolume indicator was observed across the preceding generations. In this study, the reference points for each strategy were selected with $NSE = NSE_{sqrt} = NSE_{rel} = 0$, and $RMSE_H = RMSE_M = 30$ m^3/s and $RMSE_L = 10$ m^3/s, which approximately correspond to the extreme objective function values visited by the optimization algorithm. The hypervolume indicator was computed for the non-dominated front obtained at the end of each generation using the Walking Fish Group (WFG) algorithm (While et al., 2012, 2016), and the resulting values for each strategy were normalized to range between 0 and 1 (Figure 4). Figure 5 shows the final non-dominated fronts obtained after 349 and 484 generations using the NSE- and RMSE-based calibration strategies, respectively. In the same figure, clusters corresponding to high, medium and low flow conditions, obtained with the *k*-means method, are also shown.

The solutions along the NSE-based Pareto-optimal front range from 0.22 to 0.76 for NSE, from 0.37 to 0.73 for NSE_{sqrt}, and from 0.57 to 0.81 for NSE_{rel}. The Pareto-optimal front is characterized by two distinct regions with significant tradeoffs. One of those regions, the low flow cluster, shows NSE_{rel} above 0.77 (Figure 5a) with lower values for NSE and NSE_{sqrt}

spanning from 0.22 to 0.35, and from 0.37 to 0.45, respectively. The other region, the high and medium flow clusters, shows acceptable NSE and NSE_{sqrt} values (i.e. from 0.73 to 0.76 and from 0.58 to 0.71, respectively) while NSE_{rel} ranges from 0.57 to 0.71. These results indicate that solutions with a very good representation of low discharges provide a poor representation of peak and medium flows (Guo et al., 2014; Shafii and De Smedt, 2009). However, the results also suggest that low discharges still exhibit acceptable efficiencies for the best representations of high and medium flows. Additionally, a strong linear correlation (R^2 = 0.91) between the NSE and NSE_{sqrt} objective functions is observed, though it is weaker (R^2 = 0.002) for smaller values for the NSE_{rel} objective function (i.e., low flow cluster). Hence, it is likely that NSE and NSE_{sqrt} are providing similar information for model calibration and therefore similar results can be achieved discarding one of these two objective functions. With respect to the RMSE-based Pareto-optimal front, the performance measures range from 8.3 to 20.4 m^3/s for $RMSE_H$, from 2.1 to 4.9 m^3/s for $RMSE_M$, and from 0.81 to 3.6 m^3/s for $RMSE_L$ (Figure 5b). These values indicate that, in general, the RMSE-based values range from 30% to 130% of the observed average discharges for each flow category (high, medium, and low) defined using the observed Q25 and Q75 thresholds. Moreover, the RMSE-based Pareto-optimal clusters are layered along the $RMSE_H$ direction. This behavior suggests that the high flow cluster can represent some medium and low discharges that are well represented by medium and low flow clusters.

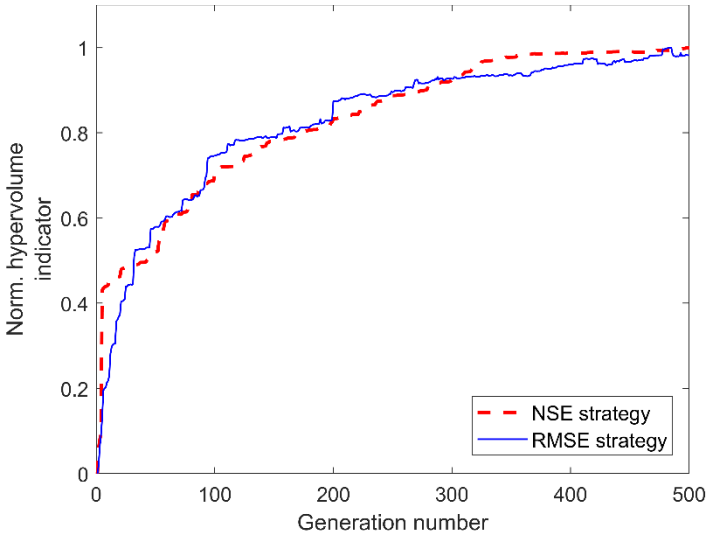
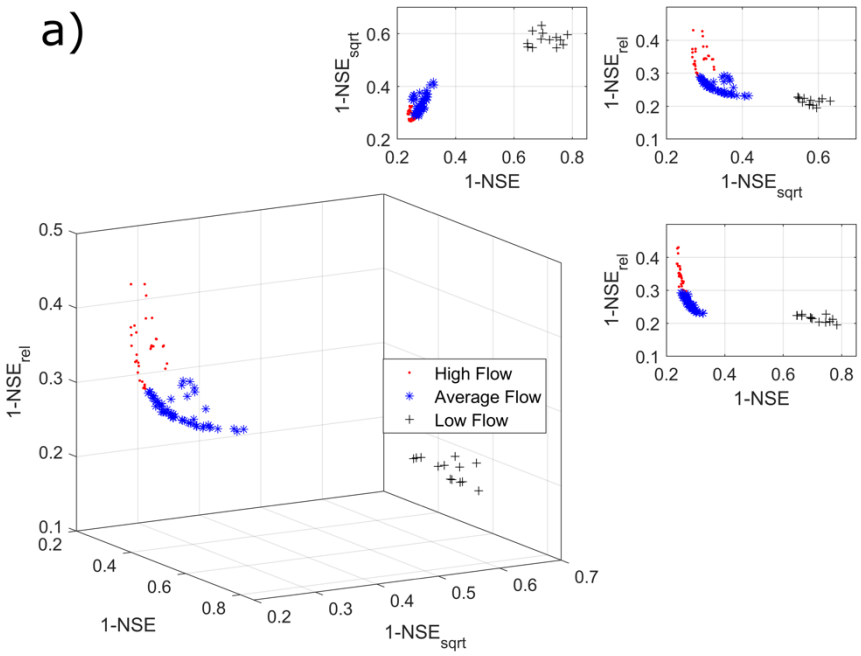


Figure 4 Normalized hypervolume indicator behavior over the NSGA-III search process for each calibration strategy



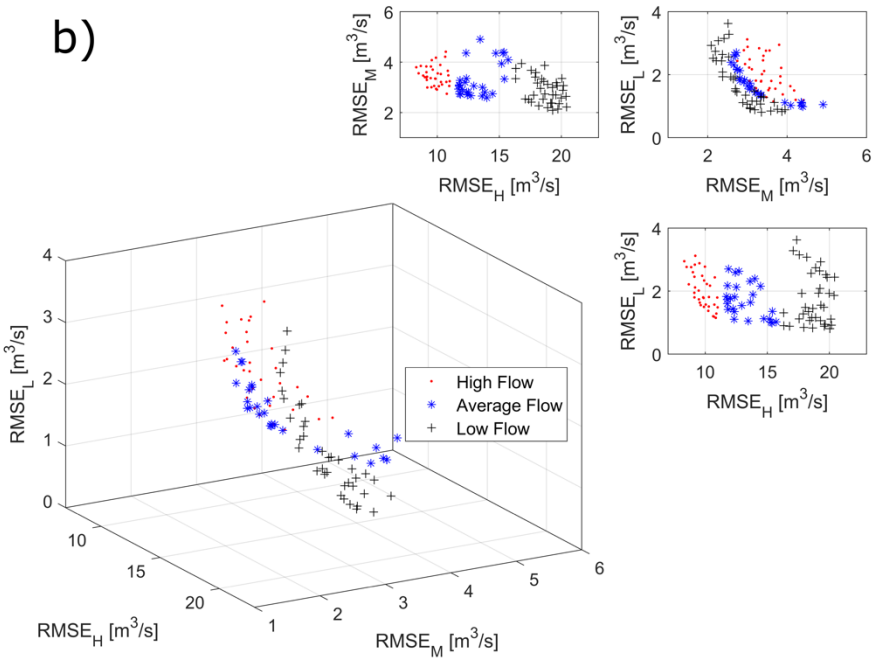


Figure 5 Clustered Pareto-optimal solutions obtained for each multi-objective calibration strategy employing NSGA-III algorithm and *k*-means clustering method a) NSE-based and b) RMSE-based

4.3.2 Reduction of initial parameter ranges by multi-objective calibration strategies

The Pareto-optimal calibrated SWAT parameter ranges varied according to the multi-objective calibration strategy (see Table 2). In general, results suggest that NSE-based calibration strategy was able to provide narrower calibrated ranges for model parameters than RMSE-based strategy. Moreover, for some parameters, the implementation of the *k*-means method to the Pareto-optimal fronts allowed the identification of different ranges depending on the role of the objective functions in each cluster (e.g., importance of NSE and $RMSE_H$ for high flow, and importance of NSE_{rel} and $RMSE_L$ for low flow). In order to compare the results for calibration parameters, we considered whether or not they showed a significant reduction in their ranges after the calibration process (i.e., narrower calibration range with respect to initial calibration range), and whether or not they showed similar final ranges in each multi-objective calibration strategy. Regarding the significant reduction in calibration ranges, we found that a group of

parameters describing mainly HRU and groundwater components showed very similar initial and final ranges for both calibration strategies. These parameters were BIOMIX (biological mixing), CANMX (max. canopy storage), EPCO (plant uptake), GW_REVAP (groundwater "revap" coefficient) and REVAPMN (groundwater threshold depth for "revap"). On the other hand, the groundwater parameter GWQMN (threshold depth for flow return) reduced in range for both multi-objective strategies. However, the ranges varied depending on the calibration strategy. Likewise, some parameters mainly related to groundwater and routing components also reduced in calibration range. On the other hand, the following ranges were very similar for both calibration strategies: ESCO (soil evaporation), GW_DELAY (groundwater delay time), ALPHA_BF (baseflow factor), RCHRG_DP (percolation factor), CH_N (2) (Manning coefficient), and CH_K (2) (alluvium hydraulic conductivity). It is worth noting that GW_DELAY, ALPHA_BF and CH_N (2) were within different ranges depending on contrasting flow conditions (i.e., high and low flow clusters). For example, GW_DELAY resulted in a range of 0 to 0.1 days for high flow conditions, and a range of 237 to 309 days for low flow conditions using the NSE-based strategy. Meanwhile, CN2 (curve number for moisture condition II) and SOL_AWC (soil water capacity) showed contrasting ranges in high and low flows for NSE-based strategy. For instance, low flow conditions favored positive multiplicative factors for CN2, increasing runoff potential, while providing negative multiplicative factors for SOL_AWC, reducing the available water capacity of soils. For high flow conditions, CN2 and SOL_AWC results were the opposite. However, RMSE-based strategy did not provide reduced calibrated ranges for these two parameters. Finally, SURLAG (surface runoff lag) showed very similar reduced ranges for all flow conditions in the NSE-based strategy, while providing a reduced range only for high flow condition in the RMSE-based strategy.

Table 2 Calibrated ranges obtained with Pareto-optimal solutions. Values without brackets correspond to NSE-based strategy results while values within brackets correspond to RMSE-based strategy results

Parameter**	Initial range	Calibrated ranges per cluster			
		All solutions	High Flow	Medium Flow	Low flow
BIOMIX	0-1	0-0.97 [0.02-0.99]	0-0.97 [0.02-0.96]	0-0.63 [0.03-0.99]	0.04-0.87 [0.07-0.99]
CN2*	(-0.25)-0.25	(-0.25)-0.25 [(-0.25)-0.25]	(-0.25) -(-0.22) [(-0.25)-0.23]	(-0.25) -(-0.21) [(-0.25)-0.25]	0.246-0.249 [(-0.24)-0.25]
CANMX	0-100	10-100 [1.4-91]	10-68 [4.7-78]	11-95 [20-91]	36-100 [1.4-70]
ESCO	0.01-1	0.74-1 [0.6-1]	0.85-0.96 [0.6-1]	0.74-0.91 [0.9-1]	0.91-1 [0.9-1]
EPCO	0.01-1	0.01-0.9 [0.01-0.9]	0.01-0.79 [0.01-0.8]	0.01-0.9 [0.01-0.9]	0.09-0.47 [0.1-0.9]
GW_DELAY	0-500	0-309 [0-499]	0-0.1 [0.01-0.34]	0-0 [0-415]	237-309 [141-499]
ALPHA_BF	0-1	0.05-0.29 [0.05-0.33]	0.23-0.29 [0.13-0.32]	0.19-0.28 [0.1-0.33]	0.05-0.11 [0.05-0.24]
GWQMN	0-5000	0-4861 [38-2018]	2-154 [38-636]	0-614 [479-2018]	1221-4861 [331-649]
GW_REVAP	0.02-0.2	0.02-0.2 [0.03-0.2]	0.02-0.2 [0.03-0.2]	0.03-0.19 [0.1-0.16]	0.11-0.2 [0.12-0.17]
REVAPMN	0-1000	48-959 [0.15-840]	50-953 [0.15-840]	48-959 [0.88-550]	60-378 [16.41-450]
RCHRG_DP	0-1	0.28-0.63 [0.28-0.75]	0.52-0.63 [0.28-0.64]	0.43-0.55 [0.3-0.45]	0.28-0.41 [0.3-0.75]
CH_N (2)	0.001-0.3	0.03-0.23 [0.02-0.3]	0.03-0.04 [0.02-0.04]	0.03-0.04 [0.02-0.08]	0.13-0.23 [0.05-0.3]
CH_K (2)	0-500	10-34 [12-57]	23-31 [21-51]	22-34 [28-57]	10-21 [12-52]
SOL_AWC*	(-0.25)-0.25	(-0.25)-0.25 [(-0.25)-0.23]	0.02-0.25 [(-0.19)-0.23]	0.11-0.25 [(-0.25)-0.16]	(-0.25) -(-0.13) [(-0.22)-0.23]
SURLAG	1-24	1-1.4 [1-19.2]	1-1.1 [1-1.2]	1-1.2 [1-13.4]	1.1-1.4 [1-19.2]

* These parameters are treated as global multiplying factors that modify the assigned values for each HRU depending on soil type and land use

** ALPHA_BF, Baseflow alpha factor ($days^{-1}$); BIOMIX, Biological mixing efficiency; CANMX, Maximum canopy storage (mm H_2O); CH_K (2), Effective hydraulic conductivity in main channel alluvium (mm hr^{-1}); CH_N (2), Manning's "n" value for the main channel; CN2, Initial SCS runoff number for moisture condition II; EPCO, Plant uptake compensation factor; ESCO, Soil evaporation compensation factor; GW_DELAY, Groundwater delay time (days); GWQMN, Threshold depth of water in the shallow aquifer required for return flow to occur (mm H_2O); GW_REVAP, Groundwater "revap" coefficient; REVAPMN, Threshold depth of water in the shallow aquifer for "revap" or percolation to the deep aquifer to occur (mm H_2O); RCHRG_DP, Deep aquifer percolation fraction; SOL_AWC, Available water capacity of the soil layer (mm H_2O mm^{-1} soil); SURLAG, Surface runoff lag coefficient.

4.3.3 Flow duration curves and streamflow time series representation

Figure 6 presents FDC and hydrographs for the simulation period. Visual inspection of the simulated and observed curves reveals that NSE-based strategy provides less variability than RMSE-based strategy, represented by the width of light gray bound of solutions. For instance, Q25 and Q75 for NSE-based strategy ranged from 7.4 to 11.8 m^3/s and from 2.6 to 4.3 m^3/s, respectively. Meanwhile, Q25 and Q75 for RMSE-based strategy ranged from 4.6 to 12.2 m^3/s and from 2.6 to 5.9 m^3/s, respectively. This means that the higher uncertainty level for streamflow simulations given by the RMSE-based calibration strategy is consistent with the wide CN2 and SOL_AWC calibrated ranges (Table 2). Note that some extreme discharges, especially low flow events, lie outside all the simulation bounds provided by Pareto-optimal solutions considered here. Also, some descending limbs and subsequent low flow pulses are poorly simulated in both calibration and validation periods. Therefore, limitations in the representation of extreme low and high flow related indices are expected. However, different sources of error may play a role here including input data uncertainties and structural inadequacies (Price et al., 2012). In both calibration strategies, high flow cluster bounds include most of the observed FDC, while shrinking the dispersion of simulated FDCs. It is worth noting that the group of simulated FDCs obtained from the NSE-based calibration strategy splits into two branches at the portion representing discharges exceeded 25% of the time. For the aforementioned calibration strategy, only the low flow cluster does not have any simulated FDC representing the corresponding branch for high discharges. Additionally, medium flow cluster shows the largest variability in the NSE-based calibration strategy, while lower flow cluster does in the RMSE-based strategy.

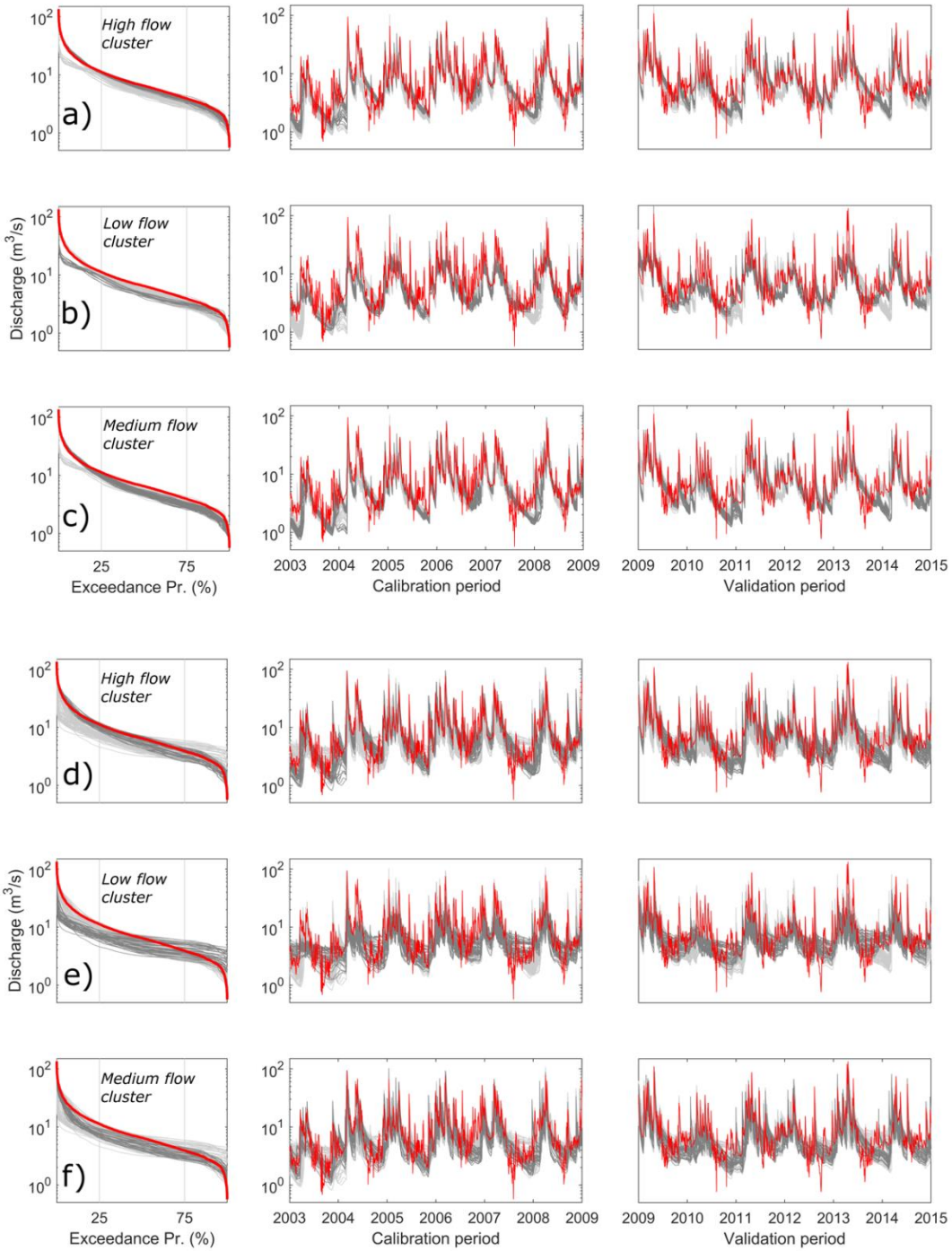


Figure 6 Flow duration curves and time series obtained from Pareto-optimal solutions (light gray) and clustered (high, medium, and low flow) solutions (dark gray) for NSE-based (a, b, and c), and RMSE-based (d, e, and f) multi-objective calibration strategies. Red lines correspond to observed streamflow values

4.3.4 Statistical analysis for predicted streamflow time-series

We performed the statistical analysis of the mean difference between observed and simulated streamflow time series for the simulation period comprised from 2003 to 2014. Most of results showed that GLS with CAR (1) correlation is substantially better than the other methods that ignore the serial correlation for the time series, as indicated by much smaller AIC (results not showed here). In a few rare cases, t-student model is better than GLS-CAR (1), meaning it is even more important to model the heavy-tail distribution rather than model the serial correlation. The percentage of Pareto-optimal solutions in each cluster without enough evidence of significant difference with a confidence level of 95% (Table 3), only account for the results obtained with GLS with AR (1) or CAR (1) correlations. Results for the statistical analysis indicated that, in general, the NSE-based strategy provides many more Pareto-optimal solutions without significant differences, compared to RMSE-based strategy. Table 3 also shows that most solutions that belong to the high flow cluster of the Pareto front yield good mean representations of overall, high, and medium streamflow time series values in both multi-objective calibration strategies. However, the percentages for high flow category do not surpass 47%, because of the recurrent under-estimation of the mean of time series comprising high discharges. Surprisingly, medium flow cluster resulted in more solutions with good mean representation of discharges below Q25 threshold (i.e., low flow values) in both calibration strategies. Therefore, it is possible to infer that solutions with simultaneous good performances based on both NSE and NSE_{rel} (Pareto front region conformed by high and medium flow clusters, see Figure 5) lead to sound representation of overall streamflow time series and specific flow conditions, which is consistent with the graphical results obtained for FDCs and hydrographs presented in Figure 6.

Table 3 Percentage of Pareto-optimal solutions without evidence of significant mean difference (α=0.05) between simulated and observed time-series, considering different time series categories and clusters for both calibration strategies. Cluster with highest percentage for each flow category are in bold

Category	Cluster	Percentage	
		NSE-based	RMSE-based
Complete time series	**High flow**	**88%**	50%
	Low flow	0%	0%
	Medium flow	27%	0%
High flow extracted time series	**High flow**	**47%**	28%
	Low flow	0%	0%
	Medium flow	13%	0%
Medium flow extracted time series	**High flow**	**94%**	56%
	Low flow	0%	5%
	Medium flow	22%	11%
Low flow extracted time series	High flow	0%	17%
	Low flow	62%	11%
	Medium flow	**75%**	26%

4.3.5 The level of predictability of ecologically-relevant hydrologic indices using multi and single-objective strategies

4.3.5.1 Multi-objective calibration strategies

For the period from 2003 to 2014, we computed the relative errors between 171 ecologically-relevant hydrologic indices obtained from the Pareto-optimal solutions and those obtained from the observed hydrograph. Results were organized according to the eleven hydrologic index groups defined by Olden and Poff (2003), which are described in section 4.2.4. For each group and multi-objective calibration strategy, we determined the indices with median relative errors within ±30%, using different sets of Pareto-optimal solutions: complete Pareto front (i.e. all points) and high, medium, and low flow clusters obtained with the *k*-means method. Indices whose median relative errors were outside the ±30% bound for all different collections of Pareto-optimal solutions, are reported in Table 4 and described in Table A1. Hence, we considered that these indices were not well represented by the calibration strategies and the

model structure employed in this study. Note that we discarded four hydrologic indices related to the frequency and duration of zero-flow days (DL18-DL20) and low flow spells (FL3), all equal to zero for this case of study. Therefore, the NSE-based calibration strategy was able to provide acceptable representation of 128 indices (77%), while the RMSE-based calibration strategy did the same for 123 indices (74%) out of 167 indices. In general, the RMSE-based strategy provided more dispersion for indices values than the NSE-based strategy (see Table 5).

Regarding the magnitude of flow events, both multi-objective calibration strategies provided acceptable results for 76 out of 94 indices (81%). These strategies were not able to fully represent the variability of flows across months and years, and the magnitude (mean and median) of annual extreme flows. For instance, under average flow conditions, results showed poor representation of the variability of some summer and fall monthly flows (i.e., MA29-MA33, which are expressed in terms of the coefficient of variation) and the skewness in annual flows (i.e., MA45, represented in terms of the difference between the mean and median annual flows,). Additionally, the NSE-based calibration strategy generated high relative errors for the variability across annual flows, expressed in terms of the range or $90^{th} – 10^{th}$ percentiles (i.e., MA42 and MA44, which include extreme flow values). For low flow conditions, the mean and median of annual minimum flows were not well replicated (i.e., ML14 and ML16, respectively), also affecting the results for some indices depending on these values (e.g., low flow index, ML15; baseflow index, ML19; and variability across annual minimum flows, ML21). For high flow conditions, both calibration strategies had limitations in representing high flow volumes (e.g., MH21) and the mean maximum monthly flows for some summer and fall months (e.g., MH6, June; MH7, July; MH10, October). Both multi-objective calibration strategies generated acceptable median values for 10 out of 13 indices (77%) describing the frequency of flow events.

The indices that were not well represented include the low flow pulse count (i.e., FL1) and some flood frequency indices that use the median and 75th percentile of flows as upper thresholds (i.e., FH5 and FH9). Moreover, the NSE-based strategy yielded poor representations of a high flood pulse count index based on a very high upper threshold (i.e., FH4, which uses 7 times median flow). Meanwhile, the RMSE-based strategy produced limited results for high flood pulse count (i.e. FH1) and flood frequency using percentile 25th as threshold (i.e. FH8). For the duration of flow events, both calibration strategies resulted in acceptable values for 32 out of 41 indices (78%). The results for this group of hydrologic indices were consistent with the poor representation of some indices describing the magnitude and frequency of flow events. For instance, duration indices related to magnitude and variability of daily and annual minima (i.e., DL1 and DL11, and DL6-DL8, respectively) yielded elevated relative errors for both strategies. Similarly, high flow indices with the median and 75th percentile of flows used as thresholds (i.e., DH17 and DH15, respectively), produced high relative errors too. Likewise, the NSE-based strategy presented difficulties representing high flow duration using seven times the median as an upper threshold (i.e., DH19). Moreover, the RMSE-based strategy yielded large deviations for the annual minima of 3-day means of daily discharge (i.e., DL2), the mean annual 3-day minimum of daily discharge (i.e. DL12), and indices related to flood duration (i.e. DH20, DH23) because of poor results for pulse count. With respect to the timing of flow events, all the four indices were well reproduced using both multi-objective calibration strategies. However, the NSE-based strategy was not able to produce median relative errors within ±30% for the seasonal predictability of non-low flow (i.e. TL4). Finally, regarding the rate of change in flow events, both multi-objective calibration strategies reproduced 6 out of 9 indices (67%) with median

relative errors outside ±30% for the fall rate (i.e., RA3) and change of flow for increasing and decreasing discharges (i.e., RA6 and RA7, respectively).

In general, hydrologic indices presented in Table 4 are mainly influenced by extreme low and high flows and attained a poor performance due to the model's limited depiction of descending limbs and low flow pulses that take place in the transition from summer to fall seasons. Additionally, between June (summer) and October (fall) is when the lowest annual flows are expected to occur in the study area. For instance, ML14, ML16, ML21, MH10, FL1, DL11, DL16, DH15, RA6 and RA7 indices (see Table A1 for description) are all directly related to discharges occurring in the period indicated above. Moreover, the aforementioned indices are key indicators for the description of the flow regime of perennial streams, as indicated by Olden and Poff (2003). Additionally, it is important to mention that extreme high flows are also being under-predicted, as indicated by the statistical analysis performed in section 4.3.4. Therefore, it is reasonable to assume that high flow indices with large upper thresholds values produced more deviated results.

Table 4 List of ecologically-relevant hydrologic indices with all, high, medium, and low flow Pareto-optimal solutions having median relative errors outside the ±30% bound, for each multi-objective calibration strategy

Hydrologic index group	No. of indicators	Median values outside ±30% relative error**		
		Both NSE- and RMSE-based	Only NSE-based	Only RMSE-based
Magnitude of flow events				
Average flow conditions	45	MA29, MA30, MA31, MA32, MA33, MA45	MA42, MA44	MA34
Low flow conditions	22	ML7, ML8, ML14, ML15, ML16, ML19, ML21, ML22		ML9, ML17
High flow conditions	27	MH6, MH7, MH10, MH21	MH22	MH11, MH23
Frequency of flow events				
Low flow conditions	3	FL1		
High flow conditions	11	FH5, FH9	FH4	FH1, FH8

Table 4 (cont'd).

Duration of flow events				
Low flow conditions	20	DL1, DL6, DL7, DL8, DL11, DL16		DL2, DL12
High flow conditions	24	DH15, DH17, DH21	DH19	DH20, DH23
Timing of flow events				
Average flow conditions	3			
Low flow conditions	4		TL4	
High flow conditions	3			
Rate of change in flow events				
Average flow conditions	9	RA3, RA6, RA7		

In comparison to previous studies where hydrological modeling was employed to predict ecological-relevant hydrologic indices (Caldwell et al., 2015; Kiesel et al., 2017; Murphy et al., 2013; Shrestha et al., 2014; Vis et al., 2015), the use of the median of different optimal-Pareto sets improved the representation of some indicators (e.g. Julian day of annual minimum, TL1; high flood pulse count, FH1; rise rate, RA1; reversals, RA8). However, key indices related to the frequency and duration of high and low flow pulses (e.g., FL1, DL16, and DH15, see below) were consistently poorly simulated. For instance, Table 5 presents the lowest relative error obtained for a suite of 32 indices included in the software Indicators of Hydrologic Alteration (IHA) (The Nature Conservancy, 2009) that were evaluated by Shrestha et al. (2014). For this group of indices, while calibrated solutions can properly reproduce the magnitude, duration and timing of different flow conditions (with difficulties for DL1, the annual minima of daily flows), the frequency and duration of low flood pulses (i.e., FL1 and DL16, respectively), the duration of high flow pulses (i.e. DH15), and the fall rate (i.e. RA3) still showed high deviances. These outcomes might be related to the limited model reproduction of some descending limbs and low

flood pulses that occur at the beginning of the fall season as observed in Figure 6. The latter is also confirmed with the high relative errors obtained for the average maximum monthly flows for June, July and October (i.e. MH6, MH7 and MH10) and the variability of mean flows among the same months (i.e. MA29 to MA33). It is important to note that most of the indices in Table 5 were well represented by high flow clusters, which are dominated by good performances for NSE or $RMSE_H$. However, the misrepresented low flow indices FL1 and DL16 obtained the lowest relative errors using optimal-Pareto solutions from low flow clusters. These clusters have a better description of low flow discharges, particularly in the seasonal transition from summer to fall.

Table 5 The lowest median relative error and corresponding interquartile range (IQR) and flow cluster for each multi-objective calibration strategy for the Indicators of Hydrologic Alteration (IHA). Values that exceed ±30% bound of relative error are highlighted

IHA group	Hydrologic index**	NSE-based			RMSE-based		
		Relative error	IQR	Cluster	Relative error	IQR	Cluster
Magnitude of monthly water conditions	MA12	-3.4%	10.5%	High	-6.9%	11.4%	High
	MA13	0.2%	6.1%	Low	-7.9%	9.8%	High
	MA14	12.9%	7.3%	High	9.3%	10.4%	High
	MA15	-7.8%	2.5%	High	-5.9%	6.2%	High
	MA16	1.2%	3.9%	High	-3.0%	8.5%	High
	MA17	-0.5%	7.4%	High	-4.5%	14.9%	High
	MA18	-1.0%	24.2%	All	-1.8%	32.6%	All
	MA19	7.5%	10.2%	Medium	-1.0%	30.6%	Medium
	MA20	1.8%	7.8%	High	0.7%	22.9%	High
	MA21	-21.9%	6.5%	High	-23.7%	38.6%	Low
	MA22	-17.9%	13.7%	High	-25.0%	20.7%	High
	MA23	-7.1%	8.1%	High	-13.9%	13.7%	High

Table 5 (cont'd)

Magnitude and duration of annual extreme water conditions (mean daily flow)	DL1	**36.8%**	10.5%	Medium	**71.8%**	48.2%	High
	DL2	10.3%	8.5%	Medium	**38.5%**	38.7%	High
	DL3	2.1%	16.8%	All	19.8%	33.4%	High
	DL4	-8.7%	6.8%	High	2.2%	30.6%	All
	DL5	-17.6%	6.1%	High	-14.8%	39.2%	Low
	DH1	-17.1%	3.7%	High	-14.6%	11.9%	High
	DH2	-8.9%	3.9%	Medium	-6.5%	10.3%	High
	DH3	-2.2%	3.8%	High	0.5%	9.8%	High
	DH4	-0.3%	4.2%	All	1.5%	10.1%	High
	DH5	-0.4%	5.1%	All	0.2%	6.3%	High
	ML17	13.1%	4.1%	Medium	**32.4%**	25.1%	High
Timing of annual extreme water conditions	TL1	6.7%	4.0%	Low	12.2%	12.8%	Low
	TH1	-0.4%	10.2%	Medium	0.4%	32.8%	All
Frequency and duration of high and low pulses	FL1	**-65.5%**	25.7%	Low	**-75.7%**	16.2%	Low
	DL16	**213.5%**	167.2%	Low	**144.9%**	181.2%	Low
	FH1	-28.4%	3.4%	High	**-44.9%**	16.1%	High
	DH15	**60.0%**	34.4%	Low	**100.0%**	53.4%	High
Rate and frequency of water condition changes	RA1	18.9%	11.6%	Medium	-16.9%	27.3%	Medium
	RA3	**-50.0%**	2.8%	High	**-53.0%**	13.2%	High
	RA8	5.5%	7.3%	Low	0.9%	25.2%	Low

4.3.5.2 Single-objective calibration

We obtained the individual model simulations that minimized each NSE-based objective function from the optimized Pareto front obtained after the NSGA-III algorithm implementation, with their corresponding results for the ecologically-relevant hydrologic indices. Maximum attained values for NSE, NSE_{sqrt} and NSE_{rel} were 0.76, 0.73 and 0.81, respectively. Optimal NSE and NSE_{sqrt} models were able to simulate 119 out of 167 indices (71%) within ±30% of relative error each, while optimal NSE_{rel} model did the same for 78 indices (47%). Compared to NSE-based multi-objective calibration strategy, some of the indices reported in Table 4 were represented within the acceptability threshold of ±30% using any of the single-objective calibrated models. As expected, the optimal NSE model provided acceptable results for high

flow related indices: mean maximum monthly flows for June and July (MH6 and MH7, respectively), high flow volume using as threshold three times the median annual flow (MH22), high flow duration with seven times the media flow as the upper threshold (DH19) and the seasonal predictability of non-low flows (TL4). However, key indices as MA19 (mean monthly flow for August), DL2 (annual minimum of 3-day average flow), TL1 (Julian date of annual minimum) and RA8 (reversals), included in Table 5, felt out the acceptability range defined in this study. On the other hand, the optimal NSE_{sqrt} model provided acceptable results for the mean maximum October flow (MH10) which is a key index for perennial streams (related to low flows during the fall season). Similarly, NSE_{sqrt} produced acceptable outcomes for high flood pulse count with seven times the median daily flow as the upper threshold (FH4), in addition to MH22 (high flow volume), DH19 (high flow duration) and TL4 (seasonal predictability), also given by the optimal NSE model. However, optimal NSE_{sqrt} model provided poor representation for MA19 (August mean flow), DL2 (3-day annual minimum), RA1 (rise rate) and R8 (reversals). Finally, the optimal NSE_{rel} model, which is insensitive to peak flows and biased towards low flows, surprisingly improved the representation of the high flow pulse duration (DH15), a key indicator for perennial streams. This occurs because NSE_{rel} significantly reduces the influence of absolute differences during high flow events (Krause et al., 2005). Therefore, the NSE_{rel} objective function has the property of benefiting simulations that better describe the overall shape of the hydrograph, which can be graphically evinced in Figure 6 for the NSE-based low flow cluster simulations. Key indices that cannot be represented by the NSE_{rel} optimal model within ±30% relative error include MA14-MA17 and MA21-MA22 (mean monthly flows for March-June and October-November, respectively), ML17 (seven-day minimum flow divided by mean annual daily flows averaged across all years), DL5 (seasonal magnitude of minimum

annual flow), and DH2-DH5 (magnitude of maximum annual flow from 3-day duration to seasonal). As expected, aforementioned indices are mainly related with high flow events.

4.4 CONCLUSIONS

This study evaluated the predictability of 167 ecologically-relevant hydrologic indices using different approaches for model calibration. We compared the performance of two multi-objective and three single-objective formulations employing the NSGA-III multi-objective optimization algorithm and the SWAT model structure. In general, the two multi-objective formulations performed better than the single-objective formulations in calculating the hydrologic indices, within a range of acceptability given by ±30% relative error. However, no specific approach was able to outperform the others for all the same set of hydrologic indicators. In this sense, all the evaluated formulations can be used to represent different targeted ecologically-relevant hydrologic indices. An advantage of a multi-objective calibration approach over a single objective alternative is the direct provision of a non-subjective range of variation for the quantity of interest after the optimization process, given by the diversity of the set of Pareto-optimal solutions.

Among the multi-objective formulations tested herein, the NSE-based strategy provided the highest number of well-predicted indices and the smallest dispersion (i.e., uncertainty) over the different sets (all points, low, medium, and high flow clusters) of Pareto-optimal simulations. The results indicated that low flows show acceptable efficiencies for the best representations of high and medium flows. Consequently, the Pareto front region comprised of high and medium flow clusters contained the highest percentages of Pareto-optimal solutions with no evidence of significant mean difference between simulated and observed time-series. Likewise, this Pareto-optimal region provided the highest number of hydrologic indicators with the lowest median

relative error and dispersion measure (i.e., interquartile range). Furthermore, this method provided groups of solutions able to simultaneously describe different streamflow regime components for distinct flow conditions, which has been proven to be a very difficult task for a single optimal solution found by the current single-objective calibration strategies (including multi-metric approximations) and available model structures.

The multi-objective strategies were able to explain up to 77% of the ecologically-relevant hydrologic indices. Important indices related to the frequency and duration of high and low flow pulses were consistently poorly simulated. Limited model depiction of descending limbs and low flow pulses that take place in the transition from summer to fall seasons resulted in weak predictability of low flow indices. This issue has been clearly identified in previous studies and is subject of current hydrology research (Garcia et al., 2017; Murphy et al., 2013; Pfannerstill et al., 2014; Shrestha et al., 2014).

In this study, we proposed the use of NSE_{rel} performance measure in a multi-objective framework in order to improve the representation of low and extreme low flow events while maintaining a good overall representation of other flow conditions. However, we showed that an NSE_{rel} objective function improves low flows while highly sacrificing the representation of other flow conditions, as opposed to the standard NSE for high flows which improves high flows, maintaining acceptable representation of other flow conditions. Therefore, during the calibration process we observed that NSE_{rel} affects the overall central tendency values (e.g. median and mean flows) for different temporal scales (e.g. monthly, annual) negatively impacting the representation of low flow indicators (e.g. baseflow index).

We also demonstrated that the use of different set of solutions, instead of a single optimal solution, introduces more flexibility in the predictability of different hydrologic indices of

ecological interest. Moreover, we were able to identify a reduced group of poorly represented indices that are closely related (Table 4). This systematic identification would facilitate the formulation of additional objective functions intended to improve model performance or to detect model inadequacies that can be addressed to reduce structural uncertainties in future research efforts.

5 A NOVEL MULTI-OBJECTIVE MODEL CALIBRATION METHOD FOR ECOHYDROLOGICAL APPLICATIONS

5.1 INTRODUCTION

The streamflow regime is widely acknowledged as a key determinant of the ecological integrity of riverine ecosystems (Poff et al., 1997; Sofi et al., 2020). Both climate and human-driven alterations to natural streamflow fluctuations affect the structure and functioning of these ecosystems, threatening biodiversity and restricting the provision of ecosystem services (Palmer and Ruhi, 2019; Vörösmarty et al., 2010). Therefore, understanding and evaluating the impacts of climate change and human interventions on the streamflow regime is critical to inform and prioritize environmental management alternatives (Hassanzadeh et al., 2017; Mittal et al., 2016).

A broadly accepted approach to characterizing streamflow regimes is to compute flow statistics from streamflow hydrographs. These statistics, also known as hydrologic signature metrics, streamflow characteristics (SFCs), or ecologically relevant hydrologic indices (ERHIs), generally represent five fundamental facets: magnitude, frequency, duration, timing, and rate of change of flows (Poff and Zimmerman, 2010). Currently, there are over 200 flow statistics relevant to stream ecology (Archfield et al., 2014; Olden and Poff, 2003; Vogel et al., 2007). These indices are usually employed in ecohydrological applications such as stream classification (Kennard et al., 2010b; Mcmanamay et al., 2014), prediction of stream health or distribution of riverine species (Hernandez-Suarez and Nejadhashemi, 2018; Kakouei et al., 2017), and environmental flow determination (Mathews and Richter, 2007; Poff et al., 2010). Since these applications generally cover large spatial scales, statistical and hydrological models have been increasingly used, especially to predict regional changes in ERHIs due to climate and anthropogenic factors (Caldwell et al., 2015; Mittal et al., 2016; Yang et al., 2016).

Hydrological models are usually preferred over regional statistical approaches because they can explicitly consider modifications in land use, environmental conditions, and management practices (Hall et al., 2017; Shrestha et al., 2016). Moreover, some environmental flow frameworks recommend using hydrological models for predicting streamflow in poorly gauged or ungauged locations (Peters et al., 2012; Poff et al., 2010). However, there is a growing number of studies revealing important limitations of hydrological models in representing ERHIs, especially when these models are calibrated based on traditional performance metrics such as the Nash-Sutcliffe efficiency (NSE) (Murphy et al., 2013; Shrestha et al., 2014; Vigiak et al., 2018; Vis et al., 2015). These limitations include over or underprediction of low- and high-flow indices (Wenger et al., 2010), high errors/uncertainties when predicting ERHIs related to timing, duration, frequency, and/or rate of change of flows (Murphy et al., 2013; Shrestha et al., 2014; Vigiak et al., 2018), and different sets of equally well-performing model parameters (in terms of traditional metrics) yielding very different performances in terms of ERHIs (Vis et al., 2015).

Current model calibration approaches for addressing limitations in ERHIs' representation can be classified into two major categories. In the first category (hereafter referred to as performance-based), objective functions are formulated based on traditional performance metrics with different streamflow transformations (e.g., square root, logarithm, inverse) to stress or balance the importance of different flow conditions. On the other hand, calibration approaches in the second category (hereafter referred to as signature-based) explicitly incorporate SFCs of interest into the objective functions (Hallouin et al., 2020; Kiesel et al., 2020, 2017; Pool et al., 2017; Vis et al., 2015; Zhang et al., 2016). In ecohydrological applications, the choice of SFCs of interest has been mainly based on riverine species preferences (Hallouin et al., 2020; Kiesel et al., 2020, 2017; Pool et al., 2017), whereas hydrological applications usually target Flow

Duration Curve (FDC) features, runoff ratios, and basic discharge statistics (Chilkoti et al., 2018; Euser et al., 2013; Fernandez-Palomino et al., 2020; Pfannerstill et al., 2017, 2014; Sahraei et al., 2020; Shafii and Tolson, 2015; Yilmaz et al., 2008). Some applications using performance-based approaches target specific flow conditions (Garcia et al., 2017; Mizukami et al., 2019), whereas others use one or multiple objective functions to attain an acceptable overall representation of the streamflow regime (Hallouin et al., 2020). When combining multiple objective functions, studies either use aggregated single-objective functions (Vis et al., 2015) or pure multi-objective approaches (Chilkoti et al., 2018; Hernandez-Suarez et al., 2018; Sahraei et al., 2020). In general, signature-based approaches provide better predictions of pre-selected SFCs compared to performance-based approaches (Hallouin et al., 2020). However, those SFCs that are not included in the original objective function formulation are not necessarily well-represented or better-performing than traditional approaches using streamflow transformations (Hallouin et al., 2020).

During the last decade, researchers have obtained a better understanding of the implications of model calibration into EHRIs replication. For instance, several studies have demonstrated that the objective function choice or formulation influences the prediction of flow statistics (Kiesel et al., 2020; Pool et al., 2017; Shafii and Tolson, 2015; Vis et al., 2015). Also, these studies showed that optimality in terms of traditional performance metrics does not necessarily result in optimal solutions for ecohydrological purposes (Hallouin et al., 2020; Kiesel et al., 2020). In ecohydrological applications, regardless of the optimization scheme for model calibration, it is uncommon to find solutions yielding acceptable results for all ERHIs of interest. Also, finding an individual simulation with acceptable results for both low- and high-flow conditions is unusual. Therefore, simulation ensembles such as median or averages of optimal

results, or their clusters, are recommended (Hernandez-Suarez et al., 2018; Vis et al., 2015). It is worth noting that most of the calibration approaches used in previous ecohydrological studies have run on single-objective mode (i.e., multi-metric, aggregated functions). Hence, those results depend on the weight assigned to each ERHI or performance metric considered within the objective function (Zhang et al., 2016), and tradeoffs among different indices, performance metrics, or regime facets are not fully explored.

The goal of this study was to develop calibration strategies providing a balanced streamflow regime representation among the different regime facets (i.e., magnitude, frequency, duration, timing, and rate of change). Two strategies were developed to compare both performance- and signature-based calibration approaches. The strategy using a performance-based approach was improved by incorporating a novel constraint formulation to obtain simulations with targeted ERHIs within pre-defined acceptability thresholds. For the signature-based strategy, tradeoffs between different streamflow regime facets were explicitly considered. These calibration strategies were implemented in an agriculture-dominated watershed in Michigan, US, using the recently developed evolutionary multi-objective optimization algorithm called Unified Non-dominated Sorting Genetic Algorithm III (U-NSGA-III) and the Soil and Water Assessment Tool (SWAT). To the best of our knowledge, previous multi-objective calibration approaches for ecohydrological applications have not explicitly considered optimization routines constraining the performance of ERHIs of interest. Likewise, this is the first time that a multi-objective calibration approach is applied to targeted ERHIs, pursuing a balanced representation of the overall streamflow regime while explicitly considering different regime facets.

5.2 MATERIALS AND METHODS

5.2.1 Overview

Two different strategies for multi-objective calibration were evaluated to improve the representation of the overall streamflow regime in a watershed model. *Strategy 1* employed a constrained performance-based approach, whereas *Strategy 2* used a constraint-free signature-based approach (Figure 7). *Strategy 1* consisted of three major steps. In the first step, the goal was to identify a reduced set of performance metrics that jointly represented a wide list of ERHI. Then, in the second step, a tailored constraint was formulated to generate individual simulations with an acceptable replication of a reduced set of ERHIs of interest. This formulation was based on pre-defined acceptability criteria for ERHI replication. Moreover, the selection of ERHI of interest was performed by targeting a balanced representation of different flow regime facets. In the third step, the outputs of the previous steps were used as inputs to formulate a multi-objective optimization problem for model calibration. Meanwhile, *Strategy 2* consisted of two major steps. In the first step, a reduced set of ERHI was defined to provide a balanced representation of different regime facets. Then, several objective functions representing different regime facets were formulated. These objective functions were considered as inputs of the problem formulation in step 2. This formulation was intended to explore tradeoffs in the simulation of different regime facets. For each strategy, near-optimal Pareto solutions were obtained using an evolutionary multi-objective optimization algorithm. Finally, preferred tradeoff solutions were identified and compared using multicriteria decision-making (MCDM) methods.

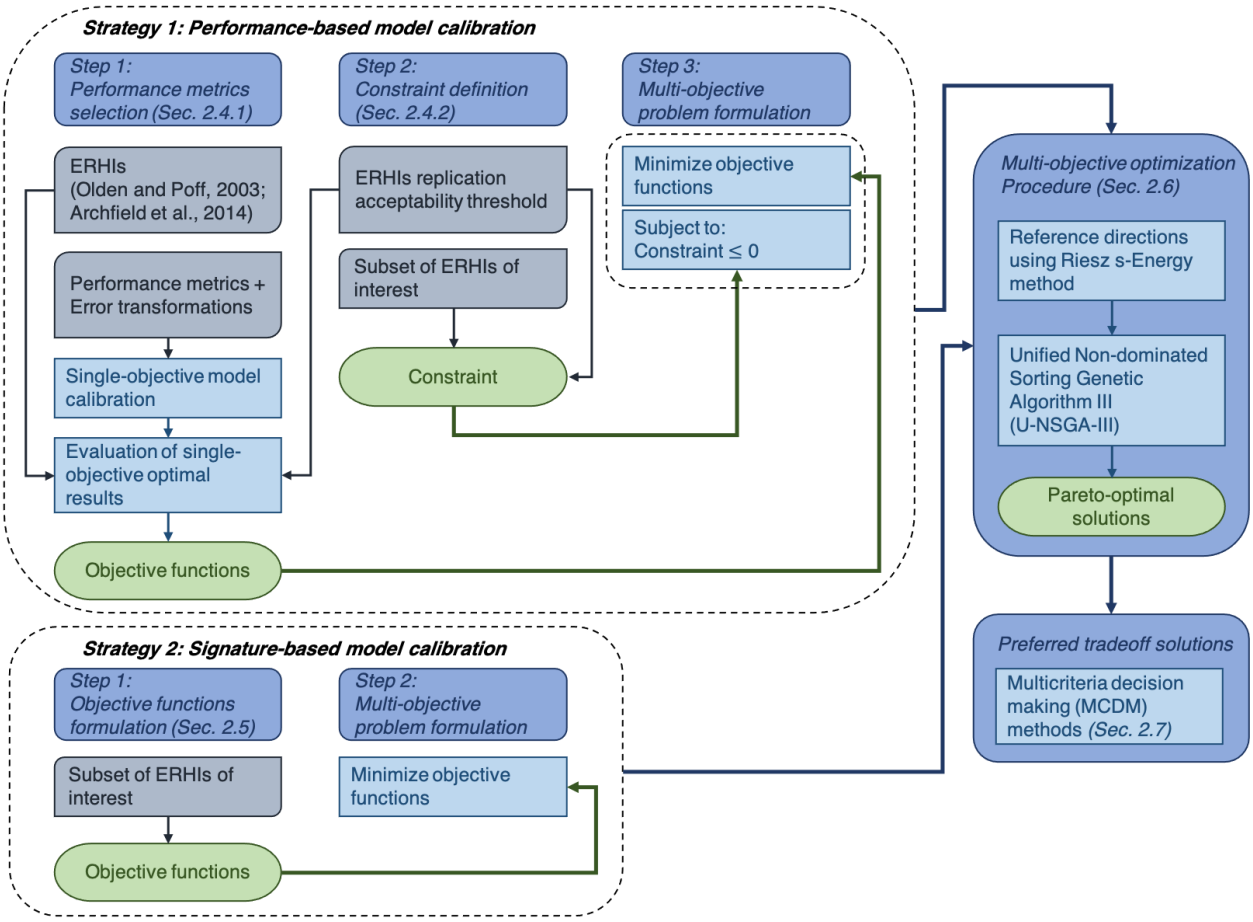


Figure 7 Overview of the two multi-objective strategies for model calibration evaluated in this study

5.2.2 Study Area

The proposed strategies were evaluated in the Honeyoey Creek-Pine Creek Watershed (Hydrologic Unit Code 0408020203), located in east-central Michigan, US (Figure 8). This watershed has a drainage area of 1010 km^2 and is situated within the Saginaw River Watershed, which drains into Lake Huron. The Saginaw River Watershed is identified as an area of concern by the US Environmental Protection Agency (USEPA) due to water pollution, wildlife habitat degradation, loss of recreational values, among others (USEPA, 2015). According to data from the National Agricultural Statistics Service (NASS) of the US Department of Agriculture (USDA), agriculture is the dominant land use (~50% of the area),

followed by forests (~24%), wetlands (~16%), pasturelands (~7%), and urban development (~3%) (USDA-NASS, 2012).

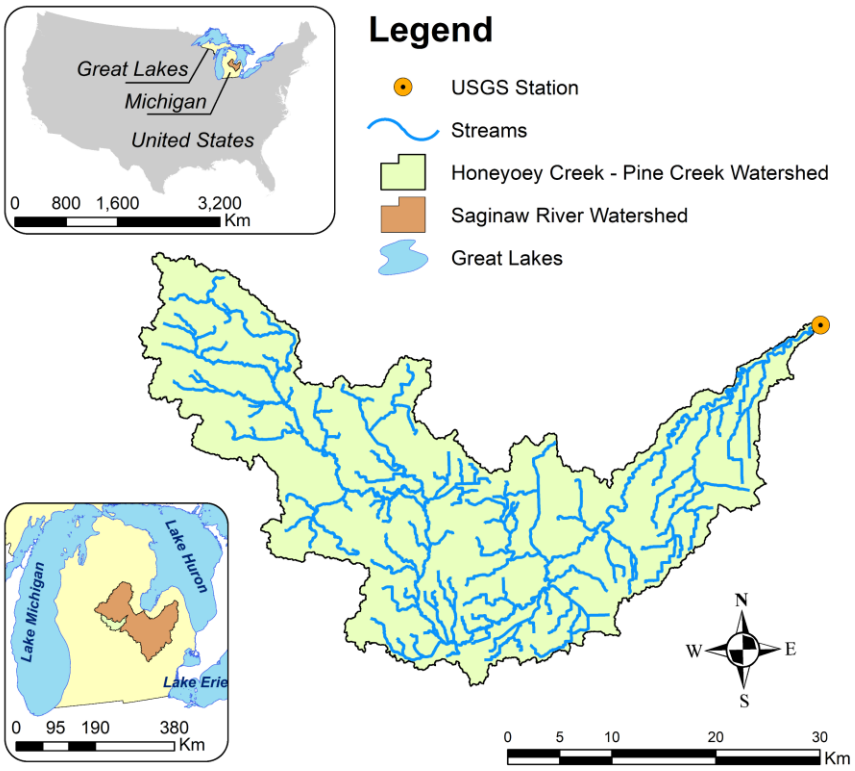


Figure 8 Location of the Honeyoey Creek - Pine Creek Watershed

5.2.3 Watershed Model

The Soil and Water Assessment Tool (SWAT 2012, Rev. 622) was used to simulate the streamflow regime in the study area. SWAT is a semi-distributed, process-based, continuous-time watershed model that can operate on a daily or sub-daily time step. SWAT is mainly used to evaluate the impact of land use and management practices on water, sediments, nutrients, pesticides, and bacteria yields at the watershed scale (Arnold et al., 2012). When using SWAT, a watershed is divided into subwatersheds, which are further discretized into Hydrologic Response Units (HRUs). HRUs are geographical units with homogeneous land use, soil, and topographical characteristics. SWAT inputs controlling the water balance include daily or sub-daily precipitation, maximum and minimum air temperatures, solar radiation, wind speed, and relative

humidity. SWAT simulates the watershed hydrology in two phases: land (loading) and water network (routing). Simulated hydrological processes include snow accumulation and melting, canopy storage, plant growth, evapotranspiration, infiltration, surface runoff, soil water redistribution, lateral flow, groundwater flows, and channel routing (Neitsch et al., 2011).

In this study, SWAT was used to obtain daily streamflow from 2003 to 2014 (calibration period) and from 1983 to 1994 (validation period) at the outlet of the Honeyoey Creek-Pine Creek Watershed (Figure 2). A warm-up period of two years was used to minimize the effect of initial conditions on the simulations. Simulated streamflow values were compared against daily observations obtained from the Pine River Near Midland US Geological Survey (USGS) gauging station (ID 04155500) (USGS, 2020). Input daily precipitation and max/min temperature data from 1981 to 2014 were collected from two weather stations provided by the National Centers for Environmental Information (NCEI) of the National Oceanic and Atmospheric Administration (NOAA) (NOAA-NCEI, 2020). The missing weather input data were estimated using SWAT's stochastic weather generator WXGEN (Neitsch et al., 2011). The watershed was divided into 250 subwatersheds, each consisting of a unique HRU obtained from dominant land use, soil, and slope characteristics. These subwatersheds were delineated using stream network data from the National Hydrography Dataset (NHD) and pre-defined units obtained from the Michigan Institute for Fisheries Research (Einheuser et al., 2012). Elevation data with a 30-m resolution was obtained from the National Elevation Dataset provided by the USGS National Map (USGS, 2018). Land use was extracted from the 30-m resolution Cropland Data Layer (CDL), which was obtained from USDA-NASS (2012). Soil characteristics were extracted from the Soil Survey Geographic Database (SSURGO) provided by the USDA Natural Resources Conservation Service (NRCS) (USDA-NRCS, 2020). Potential evapotranspiration was calculated using the

Penman-Monteith equation (Monteith, 1965), whereas surface runoff was computed using the Soil Conservation Service (SCS) curve number method (USDA-SCS, 1972). Streamflow was routed through the channel network using the variable storage coefficient method (Williams, 1969). The model was calibrated by adjusting 15 parameters whose description and calibration ranges are reported in Table 6.

Table 6 Calibration parameters and ranges

Parameter	Description	Calibration range
BIOMIX[a]	Biological mixing efficiency	[0, 1]
CANMX[a]	Maximum canopy storage (mm H_2O)	[-0.25, 0.25]
CN2[b]	Initial Soil Conservation Service (SCS) runoff number for moisture condition II	[0, 100]
ESCO[a]	Plant uptake compensation factor	[0, 1]
EPCO[a]	Soil evaporation compensation factor	[0, 1]
ALPHA_BF[a]	Baseflow alpha factor ($days^{-1}$)	[0, 1]
GW_DELAY[a]	Groundwater delay time (days)	[0, 500]
GWQMN[a]	Threshold depth of water in the shallow aquifer required for return flow to occur (mm H_2O)	[0, 5000]
GW_REVAP[a]	Groundwater "revap" coefficient	[0.02, 0.2]
REVAPMN[a]	Threshold depth of water in the shallow aquifer for "revap" or percolation to the deep aquifer to occur (mm H_2O)	[0, 1000]
RCHRG_DP[a]	Deep aquifer percolation fraction	[0, 1]
CH_N2[a]	Manning's *n* value for the main channel	[0, 0.3]
CH_K2[a]	Effective hydraulic conductivity in main channel alluvium (mm h^{-1})	[0, 500]
SOL_AWC[b]	Available water capacity of the soil layer (mm H_2O mm^{-1} soil)	[-0.25, 0.25]
SURLAG[a]	Surface runoff lag coefficient	[1, 24]

Notes:
a Values are replaced in the SWAT input files by a drawn value from the calibration range.
b Values are replaced in the SWAT input files by the existing parameter value (defined during the model set up) multiplied by 1 plus a drawn value from the calibration range.

5.2.4 *Strategy 1*: Constrained Performance-Based Model Calibration

5.2.4.1 Performance Metrics Selection

A reduced set of performance metrics were used for objective functions' formulation from a list of widely used measures (see Table 7). These measures included NSE (Nash and Sutcliffe, 1970), original and modified versions of the Kling-Gupta Efficiency (KGE, Gupta et al., 2009; Kling et al., 2012), the index of agreement (IoA, Willmott, 1981), and the coefficient

of determination (R^2). The Fourth Root Mean Quadrupled Error (R4MS4E) was also considered in order to emphasize the largest residuals expected under high flow conditions.

Since both NSE and the Root Mean Square Error (RMSE) vary only with the sum of squared model residuals, just the former was contemplated in this study. Following Gupta et al. (2009), NSE and KGE can be expressed in terms of three components representing correlation, bias, and variability. Correlation relates to timing and hydrograph shape. Meanwhile, bias and variability are aimed to reproduce the first and second moments of the distribution of observations, which mainly affect magnitude-related SFCs. These three components interact differently under each performance measure. For instance, bias is scaled by the standard deviation of observations in NSE. Thus, in presence of high variability, the bias component might be less important when obtaining optimal values. In addition, correlation and variability components interact with each other in NSE, which generally results in underestimation of the latter (Gupta et al., 2009). As an alternative, KGE provides a more balanced representation of correlation, bias, and variability, while avoiding interactions among these components (Gupta et al., 2009). By considering R^2 as an additional measure, we aimed to evaluate the role of the correlation component in ERHIs replication. Meanwhile, IoA was included to consider a different way of normalizing the sum of square errors and its effects on ERHIs replication.

To accentuate different flow conditions (i.e., low, moderate, and high), relative errors and error transformations were considered. Except for R4MS4E, all measures included their standard versions along with logarithmic, inverse, and square root transformations. Relative error versions were only used for NSE and IoA. It is worth mentioning that, in general, the standard versions favor high flows representation, square root transform is used for highlighting moderate or

average flow conditions, whereas logarithmic, inverse, and relative error versions accentuate low flows (Bennett et al., 2013; Krause et al., 2005).

Table 7 Performance metrics and transformations considered for the selection process

Metric	**Range**	**Formula**
Nash-Sutcliffe Efficiency (NSE)	(-∞, 1]	$1 - \frac{\sum_{i=1}^{n}(S_i - O_i)^2}{\sum_{i=1}^{n}(O_i - \mu_o)^2}$
Kling-Gupta Efficiency (KGE)	(-∞, 1]	*Original (Gupta et al., 2009):* $1 - \sqrt{(1-r)^2 + (1-\alpha)^2 + (1-\beta)^2}$ *Modified (Kling et al., 2012):* $1 - \sqrt{(1-r)^2 + (1-\gamma)^2 + (1-\beta)^2}$ $r = \frac{Cov_{so}}{\sigma_s \sigma_o}; \alpha = \frac{\sigma_s}{\sigma_o}; \beta = \frac{\mu_s}{\mu_o}; \gamma = \frac{\frac{\sigma_s}{\mu_s}}{\frac{\sigma_o}{\mu_o}}$
Index of Agreement (IoA)	[0, 1]	$1 - \frac{\sum_{i=1}^{n}(S_i - O_i)^2}{\sum_{i=1}^{n}(\lvert S_i - \mu_s \rvert + \lvert O_i - \mu_o \rvert)^2}$
Coefficient of Determination (R^2)	[0, 1]	$r^2 = \left(\frac{Cov_{so}}{\sigma_s \sigma_o}\right)^2$
Fourth Root Mean Quadrupled Error (R4MS4E)	[0, ∞)	$\sqrt[4]{\frac{1}{n}\sum_{i=1}^{n}(S_i - O_i)^4}$
Transformations		
Standard		$S_i = y_i; O_i = \hat{y}_i$
Square root		$S_i = \sqrt{y_i}; O_i = \sqrt{\hat{y}_i}$
Logarithmic		$S_i = \ln y_i; O_i = \ln \hat{y}_i$
Inverse		$S_i = y_i^{-1}; O_i = \hat{y}_i^{-1}$
Relative		For NSE: $S_i - O_i = \frac{y_i - \hat{y}_i}{\mu_o}; O_i - \mu_o = \frac{\hat{y}_i - \mu_o}{\mu_o}$ For IoA: $\lvert S_i - \mu_s \rvert + \lvert O_i - \mu_o \rvert = \frac{\lvert y_i - \mu_s \rvert + \lvert \hat{y}_i - \mu_o \rvert}{\mu_o}$

Notes:

$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i; \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2}; Cov_{so} = \frac{1}{n}\sum_{i=1}^{n}(O_i - \mu_o)(S_i - \mu_s)$

If $\mu = \mu_o$ *or* $\sigma = \sigma_o \rightarrow X_i = O_i$; if $\mu = \mu_s$ *or* $\sigma = \sigma_s \rightarrow X_i = S_i$

If transformation is 'Relative': if $\mu = \mu_o \rightarrow X_i = \hat{y}_i$; if $\mu = \mu_s \rightarrow X_i = y_i \hat{y}$ = observed values; y= simulated values; n = number of observations

Single-objective model calibration was executed for each performance metric and transformation indicated above, resulting in 23 individual optimization problems. Each minimization objective function *f* was defined as $1 - P_m$, where P_m is the transformed performance metric to be maximized; for R4MS4E, $f = P_m$ (as this metric has to be minimized). As a next step, 171 ERHIs reported by Henriksen *et al.* (2006), and seven ERHIs proposed by Archfield *et al.* (2014) were computed for each of the 23 optimal solutions. Simulated ERHIs were compared against those obtained from streamflow observations by calculating relative errors for each index e_{rel} for a vector of model parameters θ, as follows:

$$e_{rel_i}(\theta) = \left(\frac{I_i(y(\theta)) - I_i(\hat{y})}{I_i(\hat{y})}\right) \cdot 100 \quad (1)$$

where I_i is the *i*-th hydrologic index evaluated for simulations $y(\theta)$ and observations $\hat{y}$. Then, ERHIs within a pre-defined relative error threshold were identified for each optimal solution. It was expected that optimal results from different performance metrics and transformations yield different well-replicated ERHIs. Therefore, the final choice of performance metrics was determined by selecting up to six transformed measures that jointly represented the maximum number of ERHIs within the pre-defined relative error threshold. For the selection procedure, the transformed measure with the highest number of ERHIs within the acceptability threshold was selected. Then, another transformed measure was identified based on the remaining ERHIs and added to the list of selected measures. The previous step was repeated until either attaining the maximum number of well-replicated ERHIs or the pre-defined maximum number of objective functions. It is worth noting that the fraction of non-dominated solutions with respect to the total population increases with the number of objective functions, slowing down the search process (Deb and Jain, 2014). Likewise, a higher population size is required to maintain a good exploration of large dimensional spaces, which increases the number of function

evaluations and the overall computational time. For these reasons, we decided to limit the number of objective functions to six.

In this study, a real-parameter Genetic Algorithm (GA) (Goldberg, 1991) was used for single-objective optimization. Particularly, tournament selection, simulated binary crossover (SBX, Deb & Agrawal, 1994), and polynomial mutation (Deb, 2001) were designated as GA operators. The optimization algorithm ran for 250 generations with a population size of 100, resulting in a total of 25,000 model evaluations for each problem. The crossover probability and distribution index for the SBX operator were defined as 0.9 and 10, respectively. Likewise, the mutation probability and distribution index for the polynomial mutation operator were defined as 1/15 (i.e., the reciprocal of the number of calibration parameters) and 20, respectively. On the other hand, the relative error threshold for EHRI replication was defined as $\pm$30%, following uncertainty in the estimation of hydrologic indices reported by Kennard *et al.* (2010a) when using 15-year time series. This threshold has also been used in previous ecohydrological studies to evaluate the performance of ERHIs predictions (Caldwell et al., 2015; Hernandez-Suarez et al., 2018; Vis et al., 2015).

5.2.4.2 Constraint Definition

Traditionally, hydrologic signatures have been used in model calibration either as objective functions or post-calibration evaluation criteria (Shafii and Tolson, 2015). Here, we used a set of relevant signatures as constraints given a pre-defined acceptability threshold. This set can be identified by the modeler depending on the ecohydrological application needs. In this study, we used 32 Indicators of Hydrologic Alteration (IHA) (The Nature Conservancy, 2009), divided into five categories (Table 8), each representing specific streamflow regime facets. In addition, seven indices presented by Archfield *et al.* (2014), which describe fundamental stochastic properties of streamflow time series, were included in the constraint definition. The

aforementioned 39 indices are described in Table 8. For consistency, an acceptability threshold of $\pm$30% relative error was used for constraining ERHIs prediction.

Table 8 List of 39 Ecologically Relevant Hydrologic Indices of interest used for multi-objective model calibration

Category	Index*	Description	Associated variability index*
IHA Group 1: magnitude of monthly water conditions (IHA_1)	MA12 – MA23	Mean monthly flows from January to December ($m^3 s^{-1}$)	MA24 – MA35
IHA Group 2: magnitude and duration of annual extreme water conditions (IHA_2)	DL1 – DL5	Annual minimum with 1-, 3-, 7-, 30-, and 90-day moving average flow ($m^3 s^{-1}$)	DL6 – DL10
	DH1 – DH5	Annual maximum with 1-, 3-, 7-, 30-, and 90-day moving average flow ($m^3 s^{-1}$)	DH6 – DH10
	ML17	Baseflow index based on the 7-day minimum flow	ML18
IHA Group 3: timing of annual extreme water conditions (IHA_3)	TL1	Julian day of annual minimum	TL2
	TH1	Julian day of annual maximum	TH2
IHA Group 4: frequency and duration of high and low pulses (IHA_4)	FL1	Mean low flow pulse count per water year ($year^{-1}$)	FL2
	DL16	Mean low flow pulse duration (days)	DL17
	FH1	Mean high flow pulse count per water year with a threshold equal to the 75th percentile of the entire flow record ($year^{-1}$)	FH2
	DH15	Mean high flow pulse duration with a threshold equal to the 75th percentile of the entire flow record (days)	DH16
IHA Group 5: rate and frequency of water condition changes (IHA_5)	RA1	Rise rate ($m^3 s^{-1} d^{-1}$)	RA2
	RA3	Fall rate ($m^3 s^{-1} d^{-1}$)	RA4
	RA8	Reversals ($year^{-1}$)	RA9
Magnificent seven (MAG) (Archfield et al., 2014)	MAG1 – MAG4	First four L–moments (mean, coefficient of variation, skewness, and kurtosis)	
	MAG5	Autoregressive lag-one AR(1) correlation coefficient	
	MAG6 – MAG7	Amplitude and phase of the seasonal signal	

* Index abbreviations for Indicators of Hydrologic Alteration (IHA) as presented by Olden and Poff (2003).

The constraint, which was formulated as the sum of two components, aggregates the performance of all EHRIs of interest into a single measure. The first component is the number of indices with relative errors outside the pre-defined acceptability threshold for EHRIs replication. The second component is a weighted sum of relative violations by each index with respect to the pre-defined acceptability threshold. The constraint can be expressed as follows:

$$CV(\theta) = \sum_{i=1}^{m} k_i(\theta)\left[1 + w_i\left(\frac{1}{\tau}\frac{|I_i(y(\theta)) - I_i(\hat{y})|}{I_i(\hat{y})} - 1\right)\right] \quad (2)$$

$$k_i(\theta) = \begin{cases} 0, & \text{if } \frac{1}{\tau}\frac{|I_i(y(\theta)) - I_i(\hat{y})|}{I_i(\hat{y})} - 1 \leq 0 \\ 1, & \text{Otherwise} \end{cases}$$

$$w_i = \frac{1}{g \cdot h_i}$$

where *CV(θ)* is the constraint violation for simulations *y(θ)*, *m* is the total number of indices (i.e., 39 in this study), τ is the acceptability threshold expressed as the absolute value of a fraction between 0 and 1 (0.30 is used for this study), w_i is the weighting factor for the *i*-th index, *g* is the number of index categories (i.e., 6 in this study), and h_i is the total number of indices in the category that contains the *i*-th index. The weighing factor was explicitly incorporated to provide a balanced contribution from different streamflow regime facets. A solution is considered feasible when *CV(θ)* attains a value of zero, but for ease of handling the constraint with an optimization algorithm, we convert it to an inequality constraint as $CV(\theta) \leq 0$. By introducing the constraint formulation presented above, the optimization algorithm is forced to find streamflow simulations in which all ERHIs of interest are estimated within the acceptable range (i.e., the relative error is within ±30%). It is worth noting that the constraint definition is flexible enough to designate different acceptability thresholds τ_i for each index. This might be necessary when it is needed to iteratively relax certain acceptability conditions to find feasible solutions.

5.2.5 *Strategy 2*: Unconstrained Signature-Based Model Calibration

Under this strategy, an objective function was formulated for each index category presented in Table 8, as follows:

$$f_j(\theta) = \sum_{i \in G_j} \frac{|I_i(y(\theta)) - I_i(\hat{y})|}{I_i(\hat{y})} \quad (3)$$

where $f_j(\theta)$ is the objective function for the *j*-th category, and G_j is the set of indices belonging to the *j*-th category. Each objective function represents the total error obtained under each index category. Relative errors were used to normalize the contribution from different indices. No constraints were formulated for this calibration strategy. Therefore, opposite to *Strategy 1*, no pre-defined acceptability thresholds for ERHIs replication and no weighting factors were required in *Strategy 2*.

5.2.6 Evolutionary Multi-Objective Optimization Algorithm

In both calibration strategies, the goal was to determine the values for the vector of model parameters θ (i.e., decision variables) that minimize the objective functions formulated for each strategy. Each decision variable θ_p, p = 1, 2, …, 15, could take a value within the ranges defined in Table 6. In *Strategy 1*, those model simulations with $CV(\theta) \leq 0$ were considered as feasible (see Eq. 2), the remaining were infeasible. An evolutionary multi-objective optimization algorithm, U-NSGA-III (Seada and Deb, 2016), was implemented to address the optimization problems resulting from each strategy. U-NSGA-III is a population-based algorithm that employs crossover and mutation operators along with non-dominated sorting and reference directions to move towards near-optimum Pareto solutions. Reference directions are vectors evenly filling the objective space. This algorithm can be used for single-, multi- (i.e., 2 or 3 objective functions), and many-objective (i.e., >3 objective functions) optimization problems, and stems from the NSGA-III algorithm (Deb and Jain, 2014). It is worth mentioning that U-NSGA-III can handle both unconstrained and constrained problems. For unconstrained problems, during the non-domination sorting, any two solutions are compared using just the objective function values. A solution x^1 dominates a solution x^2 when 1) x^1 is no worse than x^2 in all objective functions, and 2) x^1 is better than x^2 in at least one objective function (Deb, 2001). In

constrained problems, the concept of constraint-domination is used instead. A solution x^1 constraint-dominates a solution x^2 when 1) x^1 is feasible and x^2 is infeasible, 2) both x^1 and x^2 are infeasible but x^1 has a lower constraint violation *CV*, or 3) both x^1 and x^2 are feasible and x^1 dominates x^2 using the traditional domination principle (Jain and Deb, 2014). In non-domination sorting, feasible solutions will always be on top of infeasible solutions. Likewise, the selection operation when creating the offspring population is modified for constrained problems (Jain and Deb, 2014).

NSGA-III and U-NSGA-III have been implemented in previous water resources applications, such as multivariate model calibration using streamflow and evapotranspiration data (Herman et al., 2020), multi-objective calibration targeting different flow conditions (Hernandez-Suarez et al., 2018), irrigation scheduling (Kropp et al., 2019; Mwiya et al., 2020), reservoir design and operation (Chen et al., 2020; Pourshahabi et al., 2020), and optimization of land use practices (Raschke et al., 2021). In this study, an interface for modifying SWAT input files and executing the model was developed in Python 3.7. This interface also included the computation of the ERHIs reported by Henriksen *et al.* (2006) and Archfield *et al.* (2014), and was coupled with the Python library *pymoo* (Blank and Deb, 2020) to implement the U-NSGA-III algorithm. The stopping criterion was set as a maximum of 1000 generations for the multi-objective optimization, with a number of reference directions assigned equal to 100. Well-spaced reference directions were generated using the recently developed Riesz s-Energy method (Blank et al., 2021) included in the *pymoo* library. The operators and parameters chosen for crossover and mutation were the same as the ones presented in section 5.2.4.1 for the GA, which are standard and recommended (Deb et al., 2002b). Convergence to a near-optimal solution was

analyzed using the Hypervolume indicator (Auger et al., 2009), which is a measure of the collective volume of the region dominated by the Pareto-optimal solutions in the objective space.

5.2.7 Selection of Preferred Tradeoff Solutions

Since we were interested in obtaining solutions providing balanced representations of different streamflow regime facets, we compared a set of preferred solutions from different MCDM methods. Particularly, two approaches were implemented: compromise programming (Zeleny, 2011), and the pseudo-weight method (Deb, 2001). The compromise programming approach identifies the closest Pareto-optimal solution to a reference point using a user-defined distance metric. Usually, the reference point is the ideal point, representing the best-expected objective function values. In this study, the ideal point was the origin of the objective space. As distance metrics, we used the ℓ_p norm with $p = 2$ (Euclidian distance) and $p \to \infty$ (Chebyshev distance). The latter is preferred for non-convex Pareto-optimal solutions. The metrics for a Pareto-optimal solution were computed as follows (Branke et al., 2008):

$$\ell_p = \left(\sum_{m=1}^{M} |f_m - z_m|^p\right)^{\frac{1}{p}} \quad (4)$$

$$\ell_{p\to\infty} = \max_m(|f_m - z_m|) \quad (5)$$

where M is the number of objective functions, f_m is the value for the m-th objective function, and z_m is the value of the m-th component of the reference point. Before applying any distance metrics, the objective functions were normalized to values between 0 and 1. Meanwhile, the *pseudo-weight* method generates a vector for each Pareto-optimal solution representing the relative importance (or weight) of each objective function. The sum of the different weights in each vector is forced to one. The pseudo-weight w_i for the i-th component in a Pareto-optimal solution was computed as follows (Deb, 2001):

$$w_i = \frac{\left(f_i^{\max} - f_i\right) / \left(f_i^{\max} - f_i^{\min}\right)}{\sum_{m=1}^{M} \left(f_m^{\max} - f_m\right) / \left(f_m^{\max} - f_m^{\min}\right)} \quad (6)$$

where $f_i^{\max}$ and $f_i^{\min}$ are the maximum and minimum values for the *i*-th objective function among all Pareto-optimal solutions, respectively. The denominator in Eq. 6 guarantees that the sum of all pseudo-weight vector components for a Pareto-solution is equal to one. Pseudo-weights are proportional to the difference between the maximum objective function value and the solution's value for a particular component. Thus, a higher pseudo-weight indicates that the point is closer to the minimum objective function value for that component. In other words, a higher pseudo-weight value indicates a higher preference for the corresponding objective function. In this study, we selected the most balanced Pareto-optimal solution as the one with the closest pseudo-weight vector to the *M*-dimension target vector $[1/M \quad \cdots \quad 1/M]$. Different target vectors can be used to explore how a Pareto solution changes when giving more relevance to a particular objective function.

5.2.8 Evaluation of Calibration Results Using Water Balance, Flow Duration Curve Characteristics, and Additional Hydrologic Indices

The Flow Duration Curve (FDC) is the complement of the streamflow cumulative distribution function (Vogel and Fennessey, 1994). FDCs are signatures of runoff variability and summarize a watershed's ability to generate streamflow values of different magnitude (Yilmaz et al., 2008). FDCs have been widely used for model evaluation and calibration (Fenicia et al., 2018). Since a FDC is a frequency-domain representation of a hydrograph, information concerning to streamflow timing is lost, limiting its utility to diagnose the overall streamflow regime. However, some characteristics extracted from FDCs are useful for understanding key hydrological processes and their ecohydrological significance (McMillan, 2020a, 2020b). In this study, we computed the percent bias (PBIAS) of four indices extracted from FDCs to evaluate

the consistency between calibration results and SFCs that have been typically used in signature-based model calibration. The characteristics derived from FDCs were the very-high-segment volume (FHV), high-segment volume (FMV), midsegment slope (FMS), and low-segment volume (FLV) (Ley et al., 2016; Yilmaz et al., 2008). The aforementioned segments were subjectively defined by Yilmaz et al. (2008) using the 2%, 20%, and 70% flow exceedance probabilities. FMS is a signature of the vertical soil moisture redistribution and streamflow flashiness. Likewise, FHV provides additional information regarding streamflow flashiness and quantifies watershed reactions to large precipitation events. Meanwhile, FMV quantifies the watershed response to heavy rainfall. Finally, FLV, which is related to long-term baseflow, was computed using the modification reported by Casper et al. (2012) to reduce the effect of the difference in lowest simulated and observed flows on the PBIAS computation. Long-term water balance was also considered by computing the PBIAS in the overall runoff ratio (RR) (Yilmaz et al., 2008).

The IHA indices that were selected in this study are computed from metrics obtained on an annual basis and represent the central tendency (i.e., mean) of annual metrics (Table 8). When setting environmental flows or evaluating streamflow regime alteration, widely used methods such as the Range of Variability Approach (Richter et al., 1997, 1996) also consider the associated interannual variability in those metrics. These methods define streamflow alteration targets as a function of central tendency and variability metrics. These targets are defined for each streamflow regime facet using meaningful indices and are further customized depending on the available ecological information of the study area (Poff et al., 2010; Richter et al., 1997). Given the relevance of streamflow variability in ecohydrological applications, especially in the definition of limits of streamflow alteration, we evaluated the impact of the two calibration

strategies defined in this study (which use only central tendency indices) in the replication of associated variability indices. These variability indices are expressed here in terms of coefficients of variation following Henriksen et al. (2006).

5.3 RESULTS AND DISCUSSION

5.3.1 Performance of Single-objective Model Calibration Using Transformed Metrics

The relative errors for EHRIs replication under each optimal solution using the transformed measures indicated in section 5.2.4.1 are presented in Figure 9. These results were obtained as part of the objective functions' selection routine under *Strategy 1*. Using hierarchical clustering with Euclidean distances and Ward's method, five groups of performance metrics were identified based on their similarity in replicating ERHIs. These groups are presented in Table 9 and can be visualized in Figure 9 for the different categories of hydrologic indices (performance metrics were arranged by similarity in the *y*-axis). These groups revealed that optimal solutions using R^2 and relative-transformed metrics as objective functions behaved drastically different, compared to the other evaluated metrics. Generally, optimal simulations using the former metrics were able to represent those ERHIs that did not fall within the ±30% relative error threshold using KGE- and sum-of-square-errors-based metrics. For example, in Figure 3b indices ML9 and ML10 are better represented by R^2 and relative-transformed metrics than any other metrics. Similar examples can be observed for DL6 and DL11 in Figure 3d, for FH5 in Figure 3g, or for RA3 and RA6 in Figure 3k. Still, the overall performance in ERHIs replication was very poor for R^2 and relative-transformed metrics (having less than 51% of ERHIs within the threshold according to Table 9). Other poor-performing metrics included inverse- and log-transformed NSE. These results suggest that those measures should be used as

complementary criteria rather than objective functions in single-objective model calibration when targeting the overall streamflow regime representation.

Different performance metrics or groups of metrics are more suitable in replicating specific index categories or streamflow regime facets (see Table 9). In terms of *magnitude*, standard or square-root-transformed metrics are preferred when targeting average and high flows (MA and MH, respectively), whereas low flows (ML) were best represented by optimal solutions when using R^2 for model calibration. Regarding *duration*, KGE and KGE' provided the best performing solutions for low flows (DL), whereas standard and square-root-transformed metrics were better suited for high flows (DH). For *frequency*, most of the standard, square-root-, and inverse-transformed metrics were better suited for low flows (FL), whereas KGE_{sqrt} yielded the highest proportion of well-replicated high flow (FH) indices. With respect to *timing*, standard square-root-, or most of the log-transformed metrics are preferred when targeting average (TA) and high flows (TH). Meanwhile, IoA_{rel}, IoA_{log}, KGE'_{inv}, and R^2 were the best performing metrics when looking for optimal solutions in replicating low flows timing (TL). Those indices representing the change of flow and reversals showed an acceptable replication under some R^2-based or relative-transformed metrics. However, in general, an acceptable representation of *rate of change* indices (RA) was quite difficult, and none of the performance metrics that were employed in this study resulted in an outstanding performance. Finally, all of the Magnificent Seven indices (MAG) were well-replicated by optimal results using standard, square-root- or log-transformed metrics.

It is worth noting that none of the 23 identified optimal solutions were able to represent five indices within the pre-defined acceptability threshold of ±30% relative error. These indices were the mean duration of flows exceeded 25% of the time (DH21), mean low flow pulse

duration (DL16), mean low flow pulse count (FL1), mean number of high flow events using the flow exceeded 25% of the time as a threshold (FH9), and mean high flow volume using the median annual flow as a threshold (MH21).

5.3.2 Selected Metrics for Constrained Performance-Based Model Calibration

The selection process of objective functions under *Strategy 1* resulted in three different lists of six transformed measures jointly representing 168 out of 178 ERHIs within the ±30% relative error range. These lists had in common the first five measures: standard and inverse KGE, and standard, inverse, and square root R^2. The sixth measure was either R^2_{log}, KGE'_{inv}, or IoA_{rel}. We decided to proceed with the list containing IoA_{rel} because, opposite to the other two lists, this one represented all rate of change indices within the ±30% acceptability threshold. The optimal solution using standard KGE was able to provide the highest number of indices within the error threshold (i.e., 128 indices or 72% of all ERHIs, see Table 9). Note that the selected list of metrics includes most of the best performing measures for each group reported in Table 9 (i.e., metrics in bold). However, this list did not well-represent five indices related to flow variability and high flow magnitude: variability in annual minima of daily flows (DL6 and ML21), variability in February and August flows (MA25 and MA31, respectively), and mean peak flows using the median annual flow as a threshold (MH24). These indices were added to the five indices that were not represented by an optimal solution (see Section 5.3.1).

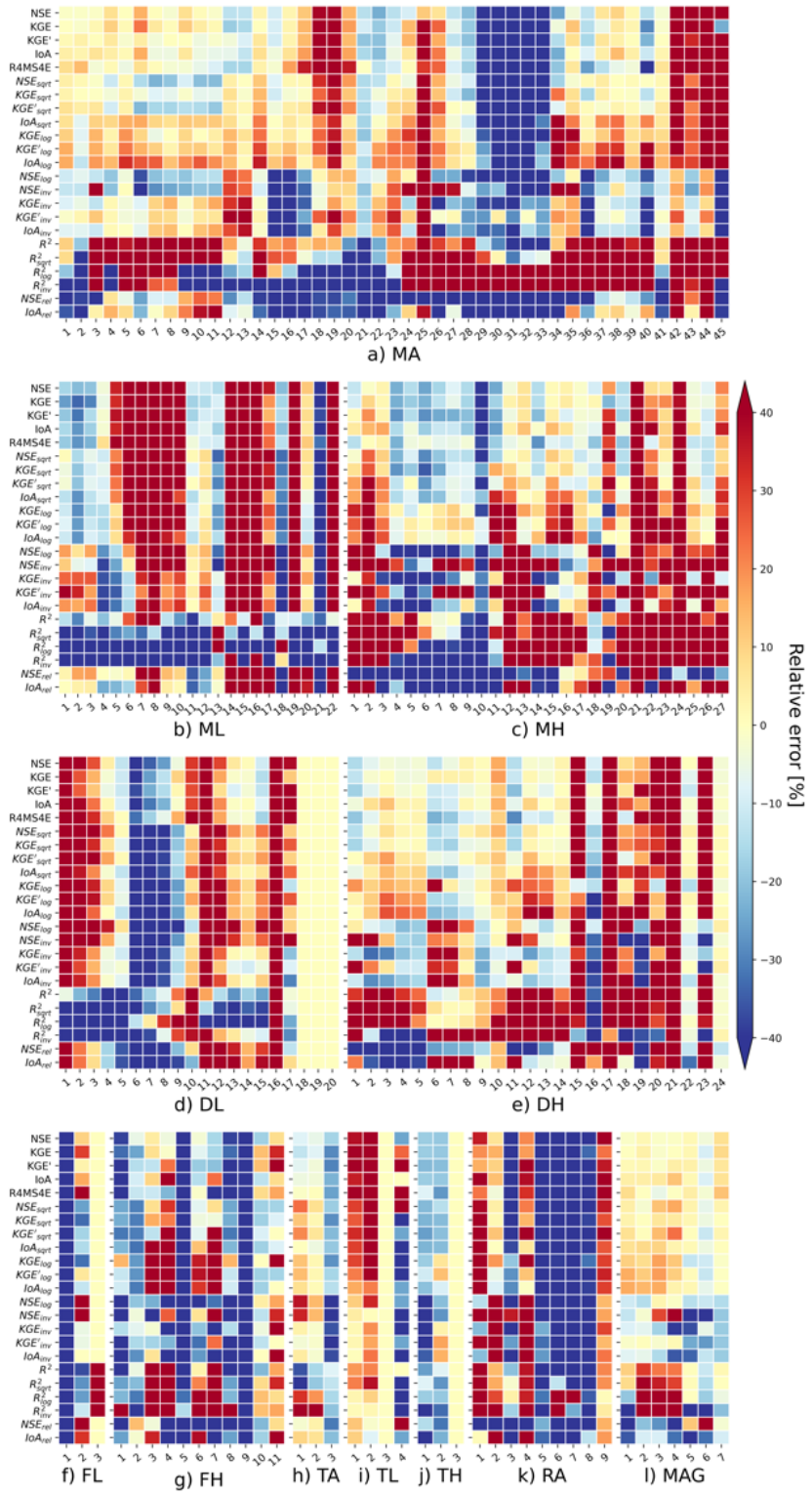


Figure 9 Heatmaps with relative errors for 178 ecologically relevant hydrologic indices when optimizing different transformed measures. Panels a) to l) represent an individual category of hydrological indices as presented in Table 9

Table 9 Proportion of indices falling within the ±30% relative error threshold under different categories of hydrologic indices. Proportions are reported for each performance metric considered in the single-objective calibration process. Performance metrics were grouped following proportions similarity. The best performing metric overall is in bold within each group. Proportions are color-coded as follows: 100% are dark green (excellent), 70-99% are light green (good), 55-69% are dark yellow (fair), 40-54% are light yellow (poor), and 0-39% are red (very poor)

Group	Performance metric		Hydrologic indices category*												Overall
			MA	ML	MH	DL	DH	FL	FH	TA	TL	TH	RA	MAG	
			Number of indices												
		Near-optimal value	45	22	27	20	24	3	11	3	4	3	9	7	178
1	*NSE*	0.78	76%	41%	89%	60%	79%	67%	64%	100%	50%	100%	22%	100%	70%
	KGE	0.88	76%	41%	89%	75%	79%	67%	55%	100%	50%	100%	44%	100%	**72%**
	KGE'	0.86	73%	41%	85%	75%	75%	67%	55%	100%	25%	100%	22%	100%	69%
	IoA	0.93	73%	45%	78%	65%	79%	67%	55%	100%	50%	100%	11%	100%	67%
	R4MS4E (m^3/s)	10.7	67%	45%	89%	60%	75%	33%	55%	100%	25%	100%	22%	100%	66%
	NSE_{sqrt}	0.72	76%	36%	85%	55%	79%	67%	64%	100%	50%	100%	22%	100%	68%
	KGE_{sqrt}	0.87	76%	41%	89%	55%	79%	33%	73%	100%	50%	100%	33%	100%	70%
	KGE'_{sqrt}	0.85	76%	45%	78%	55%	79%	67%	36%	100%	50%	100%	11%	100%	66%
2	$\boldsymbol{IoA_{sqrt}}$	0.93	76%	55%	78%	65%	75%	67%	45%	100%	50%	100%	33%	100%	**69%**
	KGE_{log}	0.83	71%	45%	63%	55%	79%	33%	36%	100%	50%	100%	33%	100%	63%
	KGE'_{log}	0.84	73%	45%	63%	55%	75%	67%	45%	100%	50%	100%	33%	100%	64%
	IoA_{log}	0.92	69%	55%	63%	60%	63%	67%	36%	100%	100%	100%	44%	100%	64%
3	NSE_{log}	0.55	71%	41%	44%	35%	71%	33%	36%	33%	75%	67%	22%	100%	54%
	NSE_{inv}	0.45	64%	36%	22%	45%	54%	33%	36%	33%	75%	67%	11%	57%	46%
	$\boldsymbol{KGE_{inv}}$	0.67	71%	50%	41%	65%	67%	67%	55%	67%	75%	67%	33%	71%	**60%**
	KGE'_{inv}	0.67	78%	45%	30%	70%	54%	67%	55%	67%	100%	67%	22%	100%	59%
	IoA_{inv}	0.81	73%	45%	44%	65%	58%	67%	36%	67%	75%	67%	33%	71%	58%
4	$\boldsymbol{R^2}$	0.80	42%	73%	37%	70%	38%	0%	36%	67%	100%	100%	33%	86%	**51%**
	R^2_{sqrt}	0.78	33%	23%	19%	40%	29%	33%	45%	67%	50%	100%	44%	71%	35%
	R^2_{log}	0.77	16%	5%	15%	30%	33%	33%	36%	67%	75%	100%	33%	57%	26%
	R^2_{inv}	0.59	4%	9%	11%	50%	21%	33%	27%	33%	75%	33%	33%	14%	20%
5	NSE_{rel}	0.79	33%	45%	22%	50%	38%	33%	36%	67%	75%	67%	22%	57%	38%
	$\boldsymbol{IoA_{rel}}$	0.93	42%	50%	19%	45%	38%	33%	36%	100%	100%	67%	33%	43%	**41%**

* MA = magnitude – average flows, ML = magnitude – low flows, MH = magnitude – high flows, DL = duration – low flows, DH = duration – high flows, FL = frequency – low flows, FH = frequency – high flows, TA = timing – average flows, TL = timing – low flows, TH = timing – high flows, RA = rate of change, MAG = magnificent seven indices.

5.3.3 Overall Performance of Pareto-Optimal Solutions

Each multi-objective calibration strategy was executed using 20 threads in parallel on a machine equipped with two Intel® Xeon® CPU E5-2640 Processor at 2.5 GHz with 64 GB RAM running Ubuntu 16.04.7 LTS. Total computation time for *Strategies 1* and *2* were 32.43 and 30.86 hours, respectively. *Strategy 1* successfully identified Pareto solutions satisfying the defined constraint for all 39 ERHIs of interest. The first feasible solution was found at generation 48, and convergence to a near-optimal Pareto front was achieved after 800 generations once the hypervolume indicator started to show a steady behavior (Figure 10a). Pareto front sizes at the end of each generation over the U-NSGA-III search process did not exceed 35 solutions, with 25 near-optimal Pareto solutions for the 1000th generation. Similarly, *Strategy 2* converged to a near-optimal Pareto front after 800 generations. In this case, Pareto front sizes at the end of each generation mostly varied between 20 and 40 solutions, with 29 near-optimal Pareto solutions for the 1000th generation.

Performance of near-optimal Pareto solutions for both strategies improved with respect to the initial random population sampled from uniform distributions of model calibration parameters (Figures 10b and c). Near-optimal solutions from *Strategy 1* resulted in linear correlations *r* between 0.80 and 0.85, whereas *Strategy 2* provided results with a broader range for *r* between 0.70 and 0.85. All Pareto solutions from *Strategy 1* overestimated up to 1.3 times the standard deviation in observations while showing a ratio between simulated and observed means between 0.95 and 1.05. Meanwhile, *Strategy 2* resulted in a more balanced and wider set of near-optimal Pareto solutions in terms of both simulated/observed standard deviation and mean ratios (α and β, respectively). Under both strategies, the standard deviation of model residuals was around 60-70% of the standard deviation of observations.

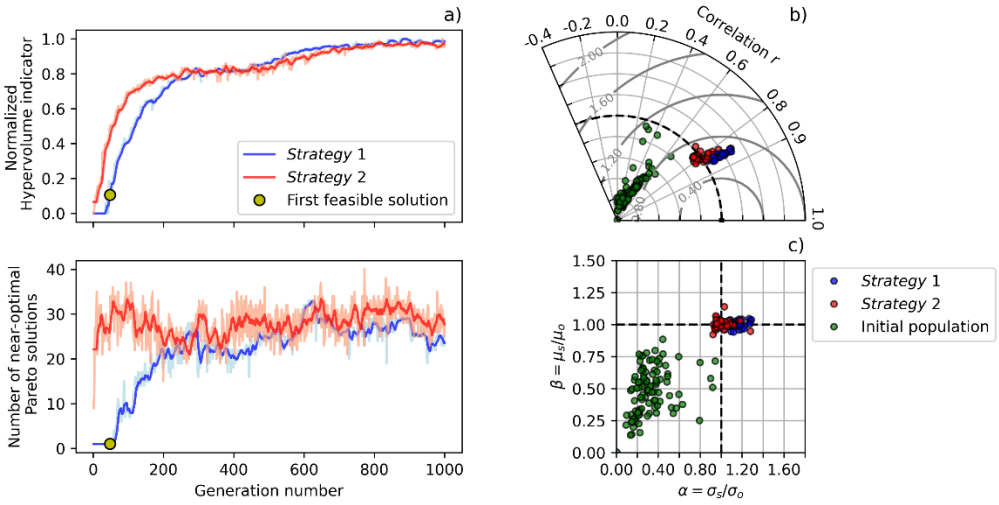


Figure 10 Overall performance of the two model calibration strategies: a) 10-generations moving average of normalized hypervolume indicator and number of Pareto solutions over the U-NSGA-III search process, lighter colors represent values for each generation; b) Taylor diagram for the initial population and Pareto solutions at the last generation, contour lines represent the ratio of the standard deviation of residuals and standard deviation of observations, α is the ratio of simulated and observed standard deviations, and r is the linear correlation coefficient; c) behavior of the ratio of simulated and observed means (β) obtained for the initial population and Pareto solutions at the last generation

A summary of the metrics and objectives (median, interquartile range (IQR), maximum, and minimum) that were used to obtain the near-optimal Pareto and preferred tradeoffs solutions are presented in Table 10. In both strategies, the performance metric showing the highest variability, as presented by IQR, was $KGE_{inv,}$ which emphasizes low flow conditions. In general, near-optimal Pareto solutions from *Strategy 2* showed a higher variability of objective function values compared to *Strategy 1*. In terms of $KGE_{inv,}$ *Strategy 2* provided an overall better performance, but also had solutions with very low values (minimum was -1.43). It is worth noting that none of the maximum values for the performance metrics chosen under *Strategy 1* were as high as those found when executing single-objective model calibration. For example, the maximum KGE of 0.83 obtained from the near-optimal Pareto set from *Strategy 1* reported in Table 10 was below the near-optimum value of 0.88 reported for KGE in Table 9. These results

indicate that simulations with ERHIs of interest within ±30% relative error are not necessarily close to an optimum in terms of a particular performance metric.

Table 10 Overall performance of near-optimal Pareto and preferred tradeoffs solutions under each model calibration strategy. Values in parenthesis correspond to the validation period

Metric*	Near-optimal Pareto solutions *Strategy 1*				Near-optimal Pareto solutions *Strategy 2*				*Preferred tradeoffs*		
	Median	IQR	Max	Min	Median	IQR	Max	Min	Compromise prog. (*Strategy 1*)	Pseudo-weights (*Strategy 1*)	Compromise prog. (*Strategy 2*)
Strategy 1 performance metrics											
KGE	0.77	0.06	0.83	0.67	0.77	0.04	0.83	0.68	0.75 (0.81)	0.76 (0.81)	0.77 (0.74)
KGE_{inv}	0.46	0.25	0.60	0.22	0.56	0.61	0.65	-1.43	0.40 (0.58)	0.41 (0.57)	0.47 (0.53)
R^2	0.72	0.01	0.74	0.70	0.64	0.07	0.71	0.53	0.73 (0.71)	0.71 (0.69)	0.66 (0.58)
R^2_{sqrt}	0.72	0.02	0.74	0.70	0.66	0.05	0.73	0.62	0.74 (0.71)	0.73 (0.70)	0.69 (0.64)
R^2_{inv}	0.40	0.02	0.43	0.37	0.44	0.04	0.52	0.37	0.40 (0.38)	0.41 (0.38)	0.41 (0.35)
IoA_{rel}	0.92	0.01	0.92	0.90	0.88	0.04	0.92	0.82	0.91 (0.92)	0.91 (0.92)	0.90 (0.91)
Strategy 2 objectives											
f_1	11.6%	1.0%	14.5%	10.5%	10.7%	2.1%	22.8%	8.8%	11.4% (9.1%)	11.6% (9.2%)	11.3% (13.6%)
f_2	11.8%	3.1%	18.1%	7.8%	15.5%	9.7%	40.4%	8.2%	13.4% (5.5%)	11.7% (6.3%)	8.6% (7.3%)
f_3	26.4%	1.3%	27.7%	20.9%	12.6%	11.6%	26.0%	3.3%	26.0% (21.9%)	26.7% (19.6%)	15.8% (13.6%)
f_4	11.9%	5.1%	21.2%	2.2%	13.5%	37.9%	107.9%	4.1%	15.0% (29.8%)	10.5% (23.5%)	18.9% (40.6%)
f_5	16.9%	1.5%	19.9%	14.6%	14.2%	11.6%	27.9%	6.8%	18.6% (17.1%)	15.1% (25.1%)	10.5% (23.0%)
f_6	7.9%	3.6%	12.1%	4.2%	6.9%	7.6%	15.0%	3.8%	7.1% (9.4%)	7.2% (8.1%)	6.3% (10.0%)

*f_1 = objective function for IHA Group 1 (magnitude of monthly water conditions); f_2 = objective function for IHA Group 2 (magnitude and duration of annual extreme water conditions); f_3 = objective function for IHA Group 3 (timing of annual extreme water conditions); f_4 = objective function for IHA Group 4 (frequency and duration of high and low pulses); f_5 = objective function for IHA Group 5 (rate and frequency of water condition changes); f_6 = objective function for Magnificent Seven indices. See Equation 3.

5.3.4 Replication of Ecologically Relevant Hydrologic Indices of Interest

Figure 11 shows the distribution of relative errors for each ERHI of interest and model calibration strategy during both calibration and validation periods. During the calibration period, which was defined between 2003 and 2014, all the indices computed from the near-optimal

Pareto set from *Strategy 1* fell within the ±30% relative error range. Meanwhile, *Strategy 2* provided median values for almost all ERHI of interest (MA19, i.e., August mean flow, was the exception) within the same range, with some near-optimal Pareto solutions generating index values outside this range. In general, *Strategy 1* resulted in a lower variability of EHRIs values compared to *Strategy 2*. Additionally, median relative errors had a similar behavior among both calibration strategies. Some exceptions included DH1 to DH5 (i.e., duration of annual maxima) in Figure 11b, which exhibited opposite trends under both strategies (i.e., overestimation for *Strategy 1* and underestimation for *Strategy 2*).

Indices that showed the highest variability during the calibration period in both strategies were mostly related to low flow conditions. These indices include DL1 to DL5 (i.e., duration of annual minima) and ML17 (i.e., baseflow index) in Figure 5b, row 1; DL16 and FL1 (i.e., low flow pulse duration and frequency), and RA3 (i.e., fall rate) in Figure 5c, row 1; and MAG3 and MAG4 (i.e., skewness and kurtosis) in Figure 5d, row 1. *Strategy 1* presented the most biased results for indices in the IHA_3 category representing the *timing* of annual extremes. This is consistent with the f_3 median value of 26.4% for this strategy, reported in Table 10, which is the highest median value among both strategies and objectives used in *Strategy 2*. It is worth noting that median values for objectives f_2 and f_4 (related to *duration* and *frequency*, respectively) were lower and better for *Strategy 1* than *Strategy 2*.

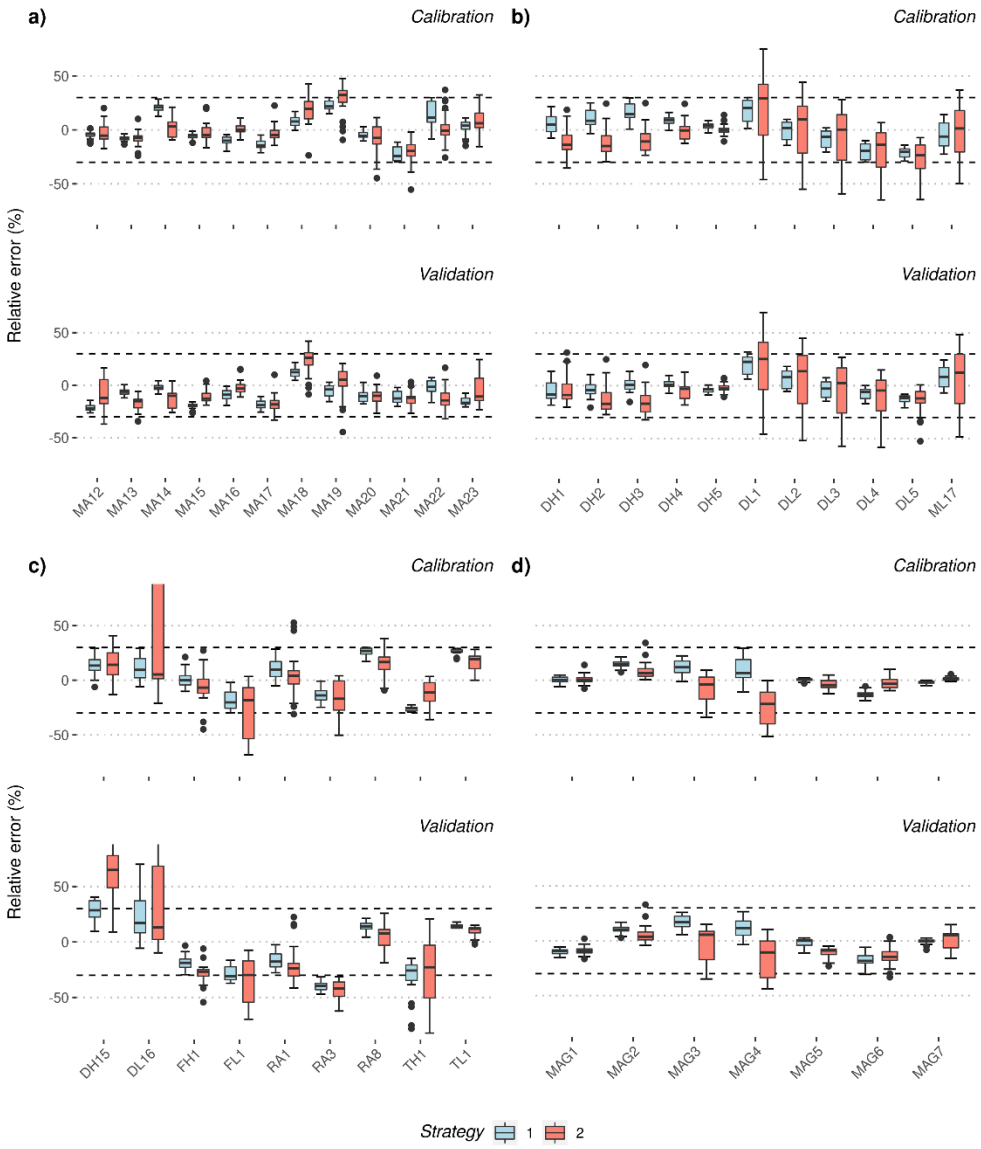


Figure 11 Boxplots representing the distribution of relative errors for each Ecologically Relevant Hydrologic Index of interest for the near-optimal Pareto solutions obtained under each model calibration strategy, horizontal dashed lines represent the ±30% interval: a) magnitude of monthly water conditions; b) magnitude and duration of annual extreme water conditions; c) duration and frequency of high and low pulses, rate and frequency of water condition changes, and timing of annual extreme water conditions; d) Magnificent seven indices. Index abbreviations are listed in Table 8

Near-optimal Pareto sets of model parameters obtained during model calibration were validated for a 12-year period between 1983 and 1994. Validation results for the replication of ERHIs of interest are also presented in Figure 11. From the list of 39 ERHI of interest, the

median relative errors for three indices fell outside the acceptability range of ±30%. These indices were RA3 for both strategies, FL1 for *Strategy 1*, and DH15 for *Strategy 2* (Figure 5c, row 2). RA3 relative error values were mostly distributed within -50% and -40%, and median relative error values for FL1 and DH15 were very close to -30% and 30% limits, respectively (excepting DH15 for *Strategy 2*). These results are indicative of the robustness of both calibration strategies. Variability of ERHIs of interest behaved similarly during the calibration and validation periods. However, DH15 and TH1, related to duration and timing of high flow events, drastically increased their variability during the validation period.

5.3.5 Performance of Preferred Tradeoff Solutions

We obtained three different solutions (two from *Strategy 1* and one from *Strategy 2*) targeting a balanced representation of the different streamflow regime facets. Both Euclidean and Chebyshev distances used for the compromise programming method selected the same preferred solution from the near-optimal Pareto sets under each calibration strategy. For *Strategy 2*, the pseudo-weight method's preferred solution was the same as the compromise programming method. The overall calibration performance for these preferred solutions is reported on the right side of Table 10.

There are no major differences between the three preferred solutions in terms of performance during the calibration period. *Strategy 2* provided slightly better results for KGE and KGE_{inv}, whereas *Strategy 1* solutions presented better R^2 and R^2_{sqrt} values. It is worth noting that NSE values were 0.61 and 0.60 for compromise programming and pseudo-weight solutions in *Strategy 1*, respectively. The preferred solution under *Strategy 2* attained a lower NSE of 0.56. Regarding the replication of ERHIs of interest, the *Strategy 2* preferred solution provided, in average, lower absolute relative errors for five of six categories of hydrologic indices during the

calibration period. Meanwhile, *Strategy 1* preferred solutions attained better replication results for the IHA_4 category, which is related to frequency and duration of high and low pulses.

During the validation period, preferred solutions from *Strategy 1* improved in terms of KGE, KGE_{inv} and IoA_{rel} (~7%, ~42%, and ~1%, respectively) while slightly worsening in terms of R^2, R^2_{sqrt} and R^2_{inv} (~3%, ~4%, and ~6%, respectively). NSE improved to 0.65 and 0.63 (~6%) for the compromise programming and pseudo-weight method solutions, respectively. Average absolute relative errors improved for the three first ERHI categories (i.e., *magnitude* of monthly flows, *duration* and *timing* of extremes) and deteriorated for the remaining categories, especially for the IHA_4 category. Regarding *Strategy 2*, the preferred solution generally worsened in terms of both performance metrics and ERHI replication, especially for the IHA_4 category, which exceeded the 30% threshold on average. Likewise, the validation NSE was reduced to 0.48 (14% reduction).

5.3.6 Representation of Water Balance and Flow Duration Curve Characteristics

Percent bias for long-term water balance and FDC characteristics are presented in Figure 12. In the same figure, FDCs for the preferred tradeoff solutions and near-optimal Pareto sets under each strategy are compared against the observed FDC during the calibration period. Generally, absolute biases of FDC characteristics for the validation period were lower than the calibration period. Most of the near-optimal Pareto solutions over-estimated FMV (high-segment volume) and FMS (midsegment slope), whereas FLV (low-segment volume) was mostly under-estimated. Opposite to *Strategy 1*, FHV (very-high-segment volume) was mostly under-estimated in *Strategy 2*. The maximum absolute bias for *Strategy 1* was below 30%. For *Strategy 2*, the maximum bias was just below 100%. Meanwhile, the largest variability resulted from FMS under both *Strategies*, whereas the highest variability for FLV occurred under *Strategy 2*. The overall RR (runoff ratio) showed a lower variability compared to the FDC characteristics.

Moreover, this index was mostly under-estimated during both calibration and validation periods. Concerning the preferred tradeoff solutions, none of them exceed the 30% absolute threshold for any water balance or FDC characteristic. For these solutions, the most biased FDC characteristic was FMS and the least biased was FHV. It is worth noting that, during validation, the minimum observed flow was at least 0.5 m^3/s lower than the minimum simulated flow for any preferred tradeoff solution.

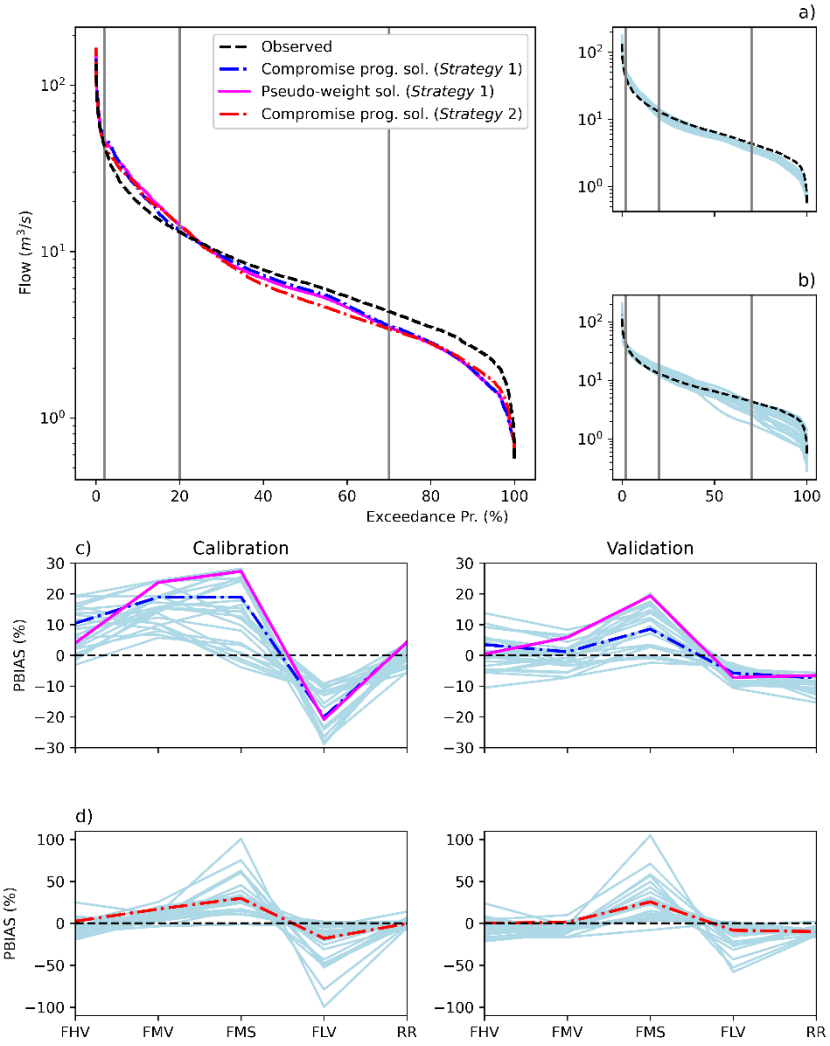


Figure 12 Flow duration curves (FDCs) for the preferred tradeoff solutions identified for each calibration strategy compared against the observed FDC from 2003 to 2014: a) FDCs from near-optimal Pareto solutions for *Strategy 1*; b) FDCs from near-optimal Pareto solutions for *Strategy 2*; c) and d) represent the bias for FDC and water balance measures for *Strategy 1* and *Strategy 2*, respectively, under calibration and validation periods

5.3.7 Relationship between Water Balance, Flow Duration Curve Characteristics, and Ecologically Relevant Hydrologic Indices of Interest

The results obtained in the previous section indicated that constraining or targeting ERHIs of interest during model calibration did not drastically worsen long-term water balance and FDC representation. Instead, calibration and validation results for ERHIs were relatively consistent with the behavior of the five SFCs addressed above. For instance, FLV under-estimation is related to the observed under-estimation of most of the monthly mean flows (indices MA12-23 showed in Figure 11a). Likewise, baseflow index behavior (see ML17 in Figure 11b) under both calibration and validation periods was consistent with RR. Lower values in the latter (i.e., lower simulated mean flow) resulted in an increase of ML17, which is computed as the ratio between the minimum 7-day flow and the overall mean flow (assuming the minimum 7-day flow does not change drastically). Another example is the under-estimation of FHV, which is related to the under-estimation of DH indices (Figure 11b), which can be caused by missing high flow events or low volume events. The same logic applies to FHV over-estimation.

On the other hand, FMS interpretation posed a different and remarkable case. In this study, simulations under both strategies were prone to yield steeper midsegment slopes. An initial explanation for this behavior was that the chosen model structure and calibrated parameters favored flashiness (i.e., abrupt ascendant and descendant streamflow changes after the occurrence of rainfall events). However, this explanation contradicted the observed underestimation of the fall rate (see RA3 in Figure 11c). When explicitly considering the *timing* facet (neglected when constructing FDCs), we obtained a more consistent interpretation. For this purpose, it is worth noting that end-of-summer monthly flows (i.e., MA18 and MA19, linked to July and August months, respectively) were drastically over-predicted, whereas September and

October monthly flows were under-predicted. Also, the timing of flow minima (TL1), which usually occurs during the summer season, was generally over-estimated. The latter followed the lower simulated fall rate for the spring-to-summer transition. The lagged timing prediction in annual minima resulted in the over-estimation of summer flows. Likewise, there was an additional delay in the transition towards the fall season. This delay was one of the reasons for the observed under-prediction in monthly flows for the fall season. Adding up this behavior across all the simulated years mainly explained the FMS results. Given the consistency between ERHIs constraining/targeting and FDC/water balance characteristics, model structure inadequacies in representing intermediate and baseflows were likely the main factors contributing to the previous inaccuracies.

5.3.8 Replication of Variability in Ecologically Relevant Hydrologic Indices

Figure 13 shows the distribution of relative errors for IHA variability indices under each model calibration strategy. Opposite to *Strategy 2*, many of the interannual variabilities of monthly flows were not captured by *Strategy 1* for the calibration period using the ±30% relative error threshold. However, most of these indices were well represented during the validation period under both strategies. According to Figure 13a, the variabilities of winter flows (i.e., MA24-25, MA34-35) were over-predicted, with median relative errors as high as ~110%. Variabilities of summer flows (i.e., MA29-31) were generally under-predicted, with absolute median relative errors as high as ~50%. Indices representing variabilities in magnitude and duration of annual extreme water conditions were mostly well represented under both strategies. Compared to *Strategy 1*, *Strategy 2* resulted in more over-predicted indices under this category outside the acceptability threshold, especially those representing the duration of high flows (Figure 13b). It is worth noting that median relative errors for the variability in the duration of

annual 1-day minimum flows (i.e., DL6) were slightly below -30% under both strategies and for both calibration and validation periods. Regarding other streamflow facets (i.e., frequency, rate of change, and timing), some calibration and validation results showed a contrasting behavior (Figure 13c). For the calibration period, most of variability indices median relative errors fell within the acceptability threshold regardless of the strategy. The most problematic index was the coefficient of variation of the Julian day of annual minimum (i.e., TL2), which was over-predicted with median relative errors around ~100%. Meanwhile, most indices of variability in frequency/duration of flow pulses and variability in rate of change of flows were largely over-predicted during the validation period, with median relative errors as high as ~120%. Therefore, our results suggest that water resources managers must be particularly cautious when defining streamflow regime alteration limits based on simulated low flow timing, rate of change, and extreme events duration and frequency given the observed bias in both associated central tendency and variability indices during model validation.

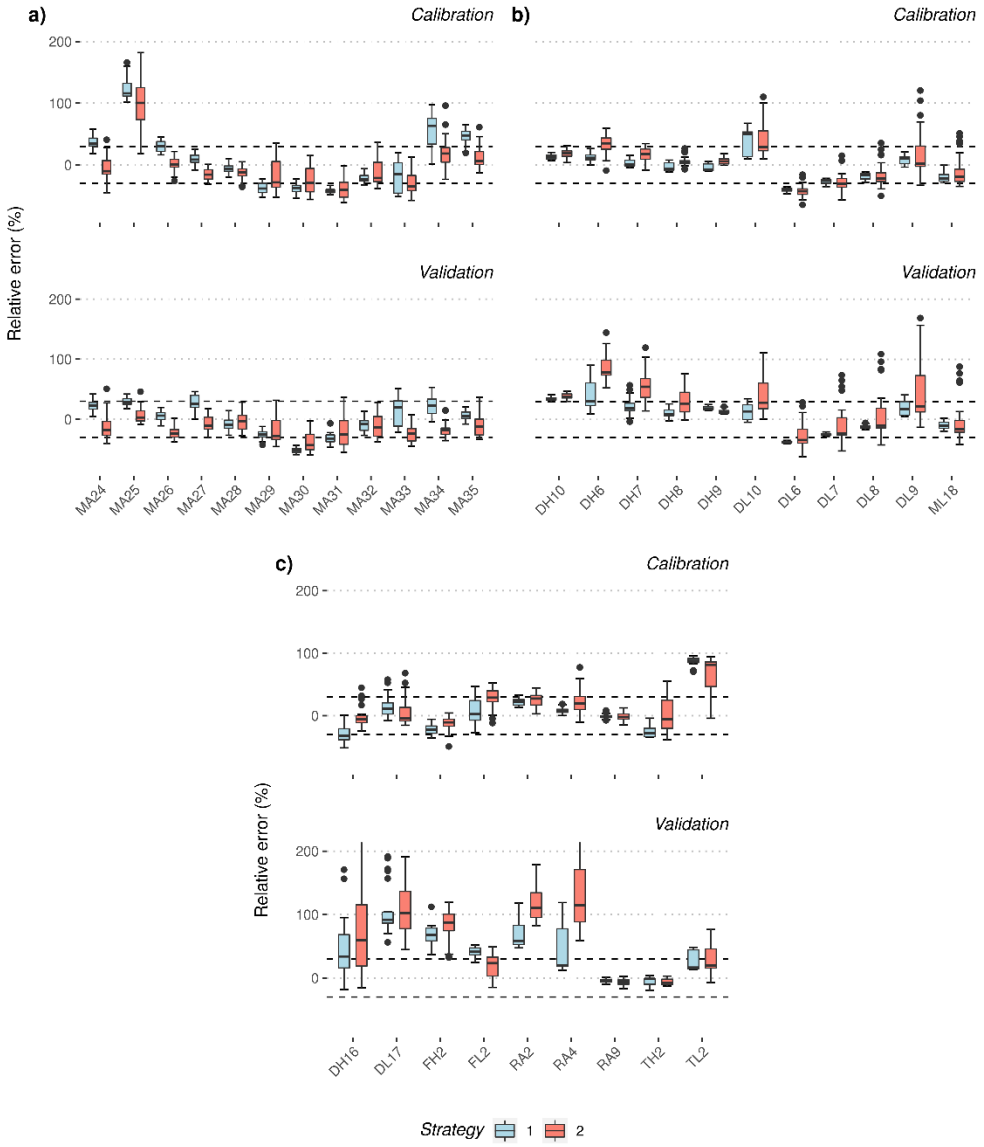


Figure 13 Boxplots representing the distribution of relative errors for variability hydrologic indices under each model calibration strategy, horizontal dashed lines represent the ±30% interval: a) variability in the magnitude of monthly water conditions; b) variability in the magnitude and duration of annual extreme water conditions; c) variability in the duration and frequency of high and low pulses, rate and frequency of water condition changes, and the timing of annual extreme water conditions. Index abbreviations are presented in Table 3

5.4 CONCLUSIONS

Implementing the performance-based calibration strategy confirmed that various performance metrics and transformations are better suited for particular streamflow regime facets. Also, it was revealed that R^2 and relative-transformed metrics behaved drastically

different comparing to KGE- and sum-of-square-errors-based metrics when replicating hydrologic indices. Moreover, results showed that a balanced representation of the streamflow regime is not directly related to the improvement of a particular performance metric. Instead, it responded to tradeoffs among different performance-based objective functions stressing different regime facets (i.e., magnitude, duration, frequency, rate of change, and timing) and flow conditions (i.e., high, moderate, and low flows).

The successful implementation of the signature-based calibration strategy demonstrated that it is possible to obtain consistent hydrological responses by simultaneously targeting multiple streamflow regime facets. More importantly, this was achieved without using any performance-based objective function. However, compared to the latter, the signature-based strategy resulted in higher variability in the near-optimal Pareto solutions, many of them with simulated indices falling outside the acceptability threshold (±30% relative error). Similarly, this strategy resulted in a highly variable representation of water balance and FDC characteristics compared to the performance-based strategy. Therefore, performance-based calibration is preferable. It is worth noting that the variability in the near-optimal Pareto solutions obtained under the two calibration strategies was driven mostly by the representation of low flows, as revealed by the highly variable inverse-transformed KGE values and low-flow related FDC characteristics among these solutions.

The model calibration framework developed here can also be used as a diagnosis tool. For instance, results revealed limitations of the SWAT model structure when representing the vertical redistribution of soil moisture, fall rate, and timing of annual extremes. Likewise, the representation of low flow magnitude and timing, rate of change of flows (especially rise and fall rates), and duration and frequency of extreme flows was limited in terms of interannual

variability. These limitations impact the definition of limits to hydrologic alteration, which are relevant when defining environmental flows and managing social-hydrological systems. Thus, water managers and modelers must account for limitations in hydrologic indices replication when defining or selecting streamflow regime targets as part of broader ecohydrological frameworks and applications in ungauged or poorly gauged watersheds.

In this study, we focused on analyzing the objective space and output variables of interest. Analyzing the near-optimal decision variables (i.e., model parameters) and intermediate variables representing other water cycle components (e.g., evapotranspiration, soil moisture, groundwater) was out of the scope of this study. Our framework detected modeling limitations when representing various streamflow regime facets, which is useful to address structural inadequacies and improving the overall modeling process. Future research should involve redesigning hydrological models and tailoring modeling practices (e.g., input data processing, model parameters selection, choosing calibration/validation time periods/lengths) to better represent ecologically-relevant characteristics of riverine ecosystems. Likewise, we recommend future studies to analyze model parameter behavior and other water cycle components when using any of the proposed calibration methods. In this regard, the proposed performance-based method is flexible enough to implement multi-variable and multi-site model calibration.

6 PROBABILISTIC PREDICTIONS OF ECOLOGICALLY RELEVANT HYDROLOGIC INDICES USING A HYDROLOGICAL MODEL

6.1 INTRODUCTION

Hydrologic signatures (a.k.a., hydrologic indices) are quantitative features that characterize the statistical properties of hydrologic time series (McMillan, 2020b). These signatures, which are most likely obtained from streamflow data, have received increasing attention due to their significance in understanding hydrologic and ecological processes (Carlisle et al., 2017; McMillan, 2020a). In hydrological modeling, streamflow signatures are typically used for model evaluation (Euser et al., 2013; Gupta et al., 2008; Jehn et al., 2019), model calibration (Shafii and Tolson, 2015), and for informing watershed management in ungauged and poorly gauged watersheds (Guo et al., 2021). Ecologically relevant hydrologic indices (ERHIs) are a subset of hydrologic signatures that can be obtained from hydrologic simulations to predict the biological condition of freshwater ecosystems in sites lacking streamflow or biological data (Hernandez-Suarez and Nejadhashemi, 2018; Mazor et al., 2018). Likewise, simulated ERHIs can be used to evaluate hydrologic and ecological alterations due to anthropogenic interventions or changes in climate and land use (Bejarano et al., 2019; McKay et al., 2019; Sengupta et al., 2018).

Using hydrologic models to simulate ERHIs introduces uncertainty. However, uncertainty sources are not only limited to modeling uncertainties (i.e., inputs, structure, parameters, initial/boundary conditions, and measurement errors), but also include the signature computation method (Westerberg et al., 2016; Westerberg and McMillan, 2015), the time series length, and non-stationarity effects (Kennard et al., 2010a). For this reason, simulated ERHIs can result in large prediction errors, especially when hydrologic models do not explicitly target those

ERHIs during model calibration (Hallouin et al., 2020; Vigiak et al., 2018). Whether a calibration method accounts for uncertainties or not, it can be broadly classified into probabilistic and deterministic methods (Tasdighi et al., 2018).

On one side, probabilistic methods for model calibration consider different uncertainty sources and can be classified into informal and formal approaches (Schoups and Vrugt, 2010). The most important difference in these two approaches resides in the likelihood function formulation (Beven and Binley, 2014). Informal methods use subjective measures (i.e., Limits of Acceptability) to identify those simulations that are a good fit to the observations (i.e., behavioral solutions), and then generate predictive distributions of the model parameters using sampling algorithms (Vrugt and Beven, 2018). Generally, the characteristics of model residual errors are treated implicitly and mapped onto the resulting parameter distributions (Beven and Smith, 2015). In this context, hydrologic signatures are typically used to further constrain the identification of behavioral solutions (Blazkova and Beven, 2009). Examples using informal methods and EHRIs can be found in Kiesel et al. (2020, 2017). Meanwhile, formal methods explicitly consider a model of residual errors to formulate the likelihood function, providing uncertainty estimates for the parameters of both hydrological and error models (McInerney et al., 2017; Smith et al., 2015). Under a Bayesian framework, different sources beyond parameter uncertainty can be explicitly addressed (Moges et al., 2021). However, the most common practice, which is also conceptually problematic, uses a lumped error model to account for those sources (Ammann et al., 2019). One of the major difficulties with hydrologic signatures has been the formulation of closed-form and tractable likelihood functions (Sadegh et al., 2015). Thus, these methods have been mainly implemented when the modeling objective is to predict the streamflow time series. However, in recent years, the application of signature-based formal

probabilistic calibration was introduced by implementing Approximate Bayesian Computation (ABC) methods. ABC methods do not require likelihood function evaluations and, instead, they sample Bayesian posterior distributions at the expense of a higher number of model evaluations (Fenicia et al., 2018; Kavetski et al., 2018; Sadegh and Vrugt, 2014).

On the other side, deterministic methods are mainly comprised of optimization methods using single or multiple objective functions to generate parameter sets that provide the best fit between observations and simulations. Single-objective methods are rather unreliable for decision-making since they do not address equifinality and identifiability issues (Beven, 2006). In addition, they only provide point estimates of model parameters and predictions and ignore the distributional properties of the model residual errors (Farmer and Vogel, 2016). In contrast, multi-objective methods provide ranges of solutions (i.e., Pareto-optimal solutions). However, the resulting Pareto-optimal distribution of model parameters and predictions do not necessarily correspond to probabilistic solutions suitable for uncertainty analysis (Reichert and Schuwirth, 2012; Tang et al., 2018). In ecohydrological applications, replication of ERHIs using deterministic methods have been the rule rather than the exception when performing model calibration (Hallouin et al., 2020; Hernandez-Suarez et al., 2018; Parker et al., 2019; Pool et al., 2017; Sengupta et al., 2018; Shrestha et al., 2016, 2014; Vigiak et al., 2018; Vis et al., 2015; Zhang et al., 2016). However, it is worth noting that the reported prediction errors for some of these studies are originated based on the distribution of the relative differences between observed and point predictions of ERHIs across multiple locations (Vigiak et al., 2018), multiple calibration trials (Pool et al., 2018; Vis et al., 2015), or Pareto-optimal solutions (Hernandez-Suarez et al., 2018). These distributions are valuable since they can be used in other modeling

processes as prior knowledge, especially when using probabilistic methods (Almeida et al., 2013).

Improvements in ERHIs prediction have been mainly driven by the choice of objective functions on untransformed or transformed streamflows. These objective functions target either specific flow conditions (i.e., high, low flows), regime facets (i.e., magnitude, duration, frequency, rate of change, timing), or hydrologic indices (Hallouin et al., 2020). As a result, several calibration strategies have been devised, some of them resulting in ensembles of model solutions (Hernandez-Suarez et al., 2018; Kiesel et al., 2020; Sengupta et al., 2018). Among these strategies, those using multi-objective calibration gained popularity since they consider tradeoffs among different targets (Efstratiadis and Koutsoyiannis, 2010; Kollat et al., 2012). However, these methods do not provide formal uncertainty estimates for model parameters and outputs. These estimates are relevant when predicting streamflows and ERHIs within the spatial domain of distributed or semi-distributed models. In addition, it is important to estimate uncertainty when developing and evaluating regionalization schemes based on hydrological modeling results (Addor et al., 2018; Almeida et al., 2016; Guo et al., 2021; Mazor et al., 2018; Moges et al., 2021; Prieto et al., 2019).

Here, we evaluated the effect of prior knowledge obtained from multi-objective calibration on the resulting posterior parameter distributions and ERHIs predictions when using a time-domain Bayesian calibration method. This allows linking the advances in deterministic ERHIs prediction and uncertainty quantification. To the best of our knowledge, this is the first time that an evaluation of this kind is performed for predicting ERHIs. The objectives of this study were to 1) estimate the total uncertainty in predicting a set of ERHIs when targeting the overall streamflow time series, 2) compare the posterior model parameter distributions when

using non-informative versus Pareto-optimal priors, and 3) identify changes in parameter estimation performance when using non-informative versus Pareto-optimal priors. We performed this evaluation in an agriculture-dominated watershed in Michigan, US, using the Unified Non-dominated Sorting Algorithm III (U-NSGA-III) (Seada and Deb, 2016) for multi-objective calibration, the Soil and Water Assessment Tool (SWAT) as the hydrological model, and the multiple-try Differential Evolution Adaptive Metropolis (ZS) (MT-DREAM$_{(ZS)}$) algorithm (Laloy and Vrugt, 2012) for sampling the posterior distributions. For the likelihood function, we used a lumped residual errors model accounting for heteroscedasticity and autocorrelation (McInerney et al., 2017).

6.2 MATERIALS AND METHODS

We outlined two experiments using daily data to compare the performance of time-domain Bayesian calibration under different prior knowledge conditions. We employed the same likelihood function regardless of the experiment. For each experiment, we defined two time periods with the same number of consecutive streamflow observations. In *Experiment 1*, we employed non-informative priors for inferring model and error parameters for each time period. Meanwhile, *Experiment 2* was devised to evaluate the effect of prior knowledge obtained from multi-objective calibration in Bayesian parameter estimation. For this purpose, we obtained Pareto-optimal parameter distributions from one time period (*Period 1*) and used them as prior knowledge for calibrating the model using data for the other time period (*Period 2*). We compared the Pareto-optimal parameter distributions against the posterior distributions using Bayesian inference for *Period 1* with non-informative priors. Likewise, we compared the predictive distributions for model parameters and ERHIs using informative and non-informative

priors under *Period 2*. Finally, we assessed the reliability, precision, and bias of the streamflow and ERHIs predictions for each experiment.

6.2.1 Bayesian Parameter Estimation

In this study, we assumed that a streamflow observation $\tilde{Y}_t$ at time step t is linked to a deterministic hydrological model H with model parameters $\boldsymbol{\theta}_H$, and given forcing data $\widetilde{\boldsymbol{X}}$, as follows,

$$\tilde{Y}_t \leftarrow H_t\left(\boldsymbol{\theta}_H, \widetilde{\boldsymbol{X}}\right) + \varepsilon_t(\boldsymbol{\theta}_\varepsilon) \quad (1)$$

where, ε_t represents the raw residuals as an aggregated measure of predictive errors. We also assumed that the residuals follow a probability distribution with parameters $\boldsymbol{\theta}_\varepsilon$. Using the Bayes equation, the posterior probability distribution of hydrological and residual error model parameters can be obtained by conditioning the model to observations and the given forcing data (McInerney et al., 2017; Vrugt, 2016),

$$p\left(\boldsymbol{\theta}_H, \boldsymbol{\theta}_\varepsilon \middle| \widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{Y}}\right) \propto p\left(\widetilde{\boldsymbol{Y}} \middle| \boldsymbol{\theta}_H, \boldsymbol{\theta}_\varepsilon, \widetilde{\boldsymbol{X}}\right) p(\boldsymbol{\theta}_H, \boldsymbol{\theta}_\varepsilon) \quad (2)$$

where, $p(\boldsymbol{\theta}_H, \boldsymbol{\theta}_\varepsilon)$ is the joint prior distribution of hydrological and residual error model parameters, and $L(\boldsymbol{\theta}_H, \boldsymbol{\theta}_\varepsilon | \widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{Y}}) \equiv p\left(\widetilde{\boldsymbol{Y}} \middle| \boldsymbol{\theta}_H, \boldsymbol{\theta}_\varepsilon, \widetilde{\boldsymbol{X}}\right)$ is the likelihood function. In the following sections, we present the model of residual errors and corresponding likelihood function (section 6.2.1.1), the prior distributions we used (section 6.2.1.2), and the sampling procedure to approximate the posterior distributions (section 6.2.1.3).

6.2.1.1 Likelihood function

The likelihood function summarizes the distance between the model simulations and observations and is built on top of the model of residual errors (Vrugt, 2016). The error model used here corresponds to a typical formulation in hydrological sciences that describes the total effect of all sources of error (Ammann et al., 2019). We followed a transformational strategy to

account for heteroscedasticity and skewness in predictive errors (McInerney et al., 2017). For this purpose, we used the Box-Cox or power transformation with parameter λ (Box and Cox, 1964),

$$z[Y;\lambda] = \begin{cases} \left(Y^{\lambda}-1\right)/\lambda & \text{if } \lambda \neq 0 \\ \log Y & \text{otherwise} \end{cases} \quad (3)$$

Thus, the transformed residual η at time step t is obtained as follows,

$$\eta_t = z[\tilde{Y}_t,\lambda] - z\left[H_t\left(\boldsymbol{\theta}_H,\widetilde{\boldsymbol{X}}\right);\lambda\right] \quad (4)$$

Since errors in daily hydrological model outputs are usually highly autocorrelated, we used a first-order autoregressive (AR1) model to consider the temporal persistence of the (transformed) residual errors,

$$\eta_t = \phi\eta_{t-1} + W_t \quad (5)$$

where, ϕ is the autoregressive parameter and W_t is the disturbance or innovation. We assumed that innovations followed a truncated Gaussian distribution to avoid negative streamflow predictions with parameters $\mu = 0$, $\sigma = \sigma_W$, and lower bound $L_{W,t}$ (Fenicia et al., 2018),

$$W_t \sim \mathcal{TN}\left(0,\sigma_W,L_{W,t}\left(\boldsymbol{\theta}_H,\widetilde{\boldsymbol{X}},\eta_{t-1}\right)\right) \quad (6)$$

Note that $L_{W,t}$ is defined such that $z\left[H_t\left(\boldsymbol{\theta}_H,\widetilde{\boldsymbol{X}}\right);\lambda\right] + \eta_t(\boldsymbol{\theta}_\varepsilon) \geq z[0;\lambda]$, which makes it time-dependent. Assuming innovations are independent, the likelihood function is formulated as follows (Fenicia et al., 2018):

$$L(\boldsymbol{\theta}_H,\boldsymbol{\theta}_\varepsilon|\widetilde{\boldsymbol{X}},\widetilde{\boldsymbol{Y}}) = \prod_{t=1}^{N_t} \frac{f_N(W_t|0,\sigma_W)}{1-F_N\left(z[0;\lambda]-z\left[H_t\left(\boldsymbol{\theta}_H,\widetilde{\boldsymbol{X}}\right);\lambda\right]-\phi\eta_{t-1}\middle|0,\sigma_W\right)} \times \frac{\partial z[\tilde{Y}_t;\lambda]}{\partial Y} \quad (7)$$

where, $W_t = \eta_t - \phi\eta_{t-1}$, $f_N(v|\mu,\sigma)$ is the Gaussian probability distribution function with mean μ and standard deviation σ evaluated for v, $F_N(v|\mu,\sigma)$ is the corresponding cumulative

distribution function (CDF), and N_t is the total number of observations. It is worth noting that for large N_t, which is the case for hydrologic time series, the likelihood is a very small number that can result in arithmetic underflow. Thus, it is common to work with the log-likelihood, $\mathcal{L}(\boldsymbol{\theta}_H, \boldsymbol{\theta}_\varepsilon|\widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{Y}})$, instead,

$$\mathcal{L}\left(\boldsymbol{\theta}_H, \boldsymbol{\theta}_\varepsilon\middle|\widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{Y}}\right) \cong -\frac{N_t}{2}\log 2\pi - N_t \log \sigma_W - \frac{1}{2\sigma_W^2}\sum_{t=2}^{N_t}(\eta_t - \phi\eta_{t-1})^2 + \sum_{t=1}^{N_t}\log\frac{\partial z[\tilde{Y}_t;\lambda]}{\partial Y} - \sum_{t=2}^{N_t}\log\left\{1 - F_N\left(z[0;\lambda] - z\left[H_t\left(\boldsymbol{\theta}_H, \widetilde{\boldsymbol{X}}\right);\lambda\right] - \phi\eta_{t-1}\middle|0, \sigma_W\right)\right\} \quad (8)$$

The complete set of error model parameters is $\boldsymbol{\theta}_\varepsilon = \{\lambda, \phi, \sigma_W\}$. In this study, we fixed the values of λ and ϕ to 0.2 and 0.8, respectively, following Evin et al. (2014) and McInerney et al. (2017) recommendations for reducing parameter interactions during model calibration.

6.2.1.2 Prior distributions

6.2.1.2.1 Experiment 1 – Non-informative priors

In this experiment, we used uniform distributions that defined the feasible parameter space by providing the minimum and maximum values for the hydrological and error model parameters. Upper and lower limits for the hydrological model parameters were determined from the literature and previous modeling exercises in the study area (see section 6.2.4.3), whereas σ_W limits were defined from an initial screening of modeling results. The initial states for the chains used by the Markov chain Monte Carlo (MCMC) sampling algorithm for Bayesian inference (see section 6.2.1.3) were drawn using Latin-Hypercube Sampling (McKay et al., 1979) subject to the aforementioned parameter limits.

6.2.1.2.2 Experiment 2 – Multi-objective model calibration

A constrained, performance-based, multi-objective calibration targeting a set of ERHIs was executed to obtain near-optimal Pareto distributions of model parameters. We implemented the recently developed evolutionary multi-objective optimization algorithm U-NSGA-III (Seada and Deb, 2016). The calibration consisted in minimizing six objective functions $f(\boldsymbol{\theta}_H)$ derived from performance metrics $P_m(\boldsymbol{\theta}_H)$. Each $f(\boldsymbol{\theta}_H)$ is computed on transformed or untransformed streamflow values to accentuate different flow conditions or regime facets,

$$f_j(\boldsymbol{\theta}_H) = 1 - P_{m_j}(\boldsymbol{\theta}_H) \tag{9}$$

where, $j = 1, 2, \ldots, 6$. The performance metrics used in this study for calibration were the Kling-Gupta Efficiency (Gupta et al., 2009) computed on untransformed and inverse-transformed values (KGE and KGE_{inv}, respectively), the relative Index of Agreement d_{rel} (Krause et al., 2005), and the coefficient of determination computed on untransformed, inverse-, and square-root-transformed values (R^2, R^2_{inv}, and R^2_{sqrt}, respectively). An optimization constraint, which must not be greater than 0, was defined to limit all targeted ERHIs to not exceed a predefined acceptability threshold τ in terms of relative error $e_{rel}(\boldsymbol{\theta}_H)$,

$$e_{rel_i}(\boldsymbol{\theta}_H) = \frac{I_i\left(H\left(\boldsymbol{\theta}_H, \widetilde{\boldsymbol{X}}\right)\right) - I_i\left(\widetilde{\boldsymbol{Y}}\right)}{I_i\left(\widetilde{\boldsymbol{Y}}\right)} \tag{10}$$

where, I_i is the *i*-th ERHI evaluated for the simulation $H\left(\boldsymbol{\theta}_H, \widetilde{\boldsymbol{X}}\right)$ and observations $\widetilde{\boldsymbol{Y}}$. The constraint $CV(\boldsymbol{\theta}_H)$ was formulated to penalize high relative errors and for separating feasible from unfeasible solutions. An unfeasible solution results in ERHI values outside the predefined acceptability threshold,

$$CV(\boldsymbol{\theta}_H) = \sum_{i=1}^{m} k_i(\boldsymbol{\theta}_H)\left[1 + w_i\left(\frac{\left|e_{rel_i}(\boldsymbol{\theta}_H)\right|}{\tau} - 1\right)\right]$$

$$k_i(\boldsymbol{\theta}_H) = \begin{cases} 0 & \text{if } \dfrac{\left|e_{rel_i}(\boldsymbol{\theta}_H)\right|}{\tau} - 1 \leq 0 \\ 1 & \text{Otherwise} \end{cases}$$

$$w_i = \frac{1}{g \times h_i}$$

(11)

where, m is the total number of ERHIs, w_i is the weighting factor for the *i*-th ERHI, g is the number of ERHI categories, and h_i is the total number of ERHIs in the category that contains the *i*-th ERHI. The ERHIs used in this study and their categories are presented in section 6.2.4.4. The value of τ used in this study was 0.3 (Hernandez-Suarez et al., 2018), which means that all the targeted ERHIs must attain relative errors within ±30%. Once the near-optimal Pareto solutions were obtained, we computed σ_W for each individual solution using equations 3 – 5. A multivariate kernel distribution was generated to have a non-parametric representation of the joint distribution of parameters $\boldsymbol{\theta}_H$ and σ_W using the resulting near-optimal Pareto parameter sets. For this purpose, we employed the *mvksdensity* function in Matlab R2019b using a Gaussian kernel. An initial vector of bandwidths was defined using the Silverman's rule of thumb (Silverman, 1986),

$$b_k = \sigma_k \left[\frac{4}{(d+2)n}\right]^{1/(d+4)}$$

(12)

where, d is the number of dimensions (i.e., number of hydrological and error model parameters), n is the number of observations (i.e., Pareto-optimal solutions), $k = 1, 2, \ldots, d$, and σ_k is the standard deviation of the *k*-th variate (i.e., parameter). This vector of bandwidths was further

refined by maximizing the agreement between the marginal empirical CDF of each parameter and the corresponding marginal CDF obtained from the multivariate kernel density distribution using a 5-fold cross-validation and a genetic algorithm. The resulting optimized multivariate kernel distribution was then used as the prior distribution $p(\boldsymbol{\theta}_H, \boldsymbol{\theta}_\varepsilon)$ under this experiment. It is worth noting that the initial states for the chains used by the MCMC sampling algorithm were directly drawn from $p(\boldsymbol{\theta}_H, \boldsymbol{\theta}_\varepsilon)$.

6.2.1.3 Sampling algorithm

In this study, the MT-DREAM$_{(ZS)}$ algorithm (Laloy and Vrugt, 2012) was used to efficiently explore the posterior distribution of hydrological and error model parameters. MT-DREAM$_{(ZS)}$ is an adaptive MCMC algorithm using multiple-try sampling, snooker updating, and an archive of past states to improve the convergence of computationally intensive and high-dimensional models (Vrugt, 2016). This method belongs to the Differential Evolution Adaptive Metropolis (DREAM) multi-chain family of algorithms for Bayesian inference. The original DREAM algorithm automatically adjusts the scale and orientation of the proposal distribution used for posterior inference. In addition, it employs subspace sampling and outlier chain correction while maintaining balance and ergodicity (Vrugt et al., 2009). DREAM$_{(ZS)}$ introduced the use of past samples (from an external archive) into the jump distribution. As a result, the sampling procedure requires a smaller number of chains to explore the target distribution, outliers do not need a forceful treatment, and chains can run in a distributed manner (Vrugt, 2016). To ensure convergence (given the violation of Markovian principles introduced by adaptive Metropolis samplers), the adaptation rate decreases with the number of generations (Ter Braak and Vrugt, 2008). MT-DREAM$_{(ZS)}$ introduced multiple-try sampling in each of the chains,

creating mt different proposals in each chain that can be evaluated in parallel, which is more practical than running DREAM with large chain numbers (Laloy and Vrugt, 2012).

Convergence to a stationary posterior distribution was monitored using the multivariate $\hat{R}$-statistic proposed by Gelman and Rubin (1992). The multivariate $\hat{R}$-statistic, which is computed using the last 50% samples of each parallel chain, is used to evaluate whether the between- and within-covariance matrices of these chains are similar. When these matrices are very similar, the $\hat{R}$-statistic is close to unity. An $\hat{R}$-statistic below 1.2 is used in practice to declare convergence (Gelman et al., 2013).

6.2.2 Generation of Predictive Distributions of ERHIs

Predictive distributions were generated by propagating the posterior probability distributions for the model and error parameters through the hydrologic model and ERHIs computation methods. For an individual set of parameters $\boldsymbol{\theta}^i = (\boldsymbol{\theta}_H^i, \boldsymbol{\theta}_\varepsilon^i)$, a prediction of a given ERHI was obtained as follows (Fenicia et al., 2018; McInerney et al., 2017):

1) Sample W_t using equation 6.
2) Obtain η_t^i using equation 5. For $t = 1$, η_t^i is directly sampled and step 1 is ignored:

   $$\eta_1^i \leftarrow f_N\left(0, \sigma_\eta^i\right) \tag{13}$$

   where, $\sigma_\eta^2 = \sigma_W^2/(1-\phi^2)$
3) Compute the streamflow prediction at time step *t* for the given $\boldsymbol{\theta}^i$ as follows:

   $$Y_t^i\left(\boldsymbol{\theta}^i\right) = z^{-1}\left(z\left[H_t\left(\boldsymbol{\theta}_H^i, \widetilde{\boldsymbol{X}}\right); \lambda^i\right] + \eta_t^i\right) \tag{14}$$

   where, z^{-1} is the inverse Box-Cox transformation.
4) Once Y_t^i is obtained for the N_t time steps, the *j*-th ERHI is computed using equation 15:

   $$ERHI_j^i = I_j\left(Y^i(\boldsymbol{\theta}^i)\right) \tag{15}$$

The parameter sets were taken from the last 20% posterior samples obtained by the MT-DREAM$_{(ZS)}$ algorithm.

6.2.3 Performance evaluation

We computed three measures to quantify reliability, precision, and bias for evaluating the performance of the Bayesian parameter estimation under each experiment. A prediction is reliable when the observations can be considered samples of the predictive distribution (i.e., observations consistently fall within the prediction bounds). The reliability measure that was employed in this study represents the average absolute difference between the predictive quantile-quantile (PQQ) plot and a 1:1 line representing the CDF of a standard uniform distribution $U(0,1)$ (McInerney et al., 2017). Regarding precision, we determined the average coefficient of variation of the predicted streamflows using the observations as a proxy to the average streamflow at each time step (McInerney et al., 2017). The precision measure represents the width of the prediction bounds. The following equation was used to compute the precision:

$$\text{Precision}[Y,\tilde{Y}] = \frac{1}{N_t}\sum_{t=1}^{N_t}\frac{std(Y_t)}{\tilde{Y}_t} \tag{16}$$

where, $std(Y_t)$ is the standard deviation of streamflow predictions at time step t. Finally, we computed the absolute volumetric bias to evaluate the long-term water balance error of the predictions. For this purpose, we used the mean prediction value at each time step $\bar{Y}_t$:

$$\text{Bias}[Y,\tilde{Y}] = \left|\frac{\sum_{t=1}^{N_t}\tilde{Y}_t - \sum_{t=1}^{N_t}\bar{Y}_t}{\sum_{t=1}^{N_t}\tilde{Y}_t}\right| \tag{17}$$

The reliability, precision, and bias measures used herein targeted the overall streamflow predictions; for the ERHIs predictions, we directly compared the resulting predictive

distributions against the ERHIs obtained from the observations. For the latter, we visually inspected whether the ERHIs from observations fell within the corresponding predictive bounds. Also, we verified whether the median relative errors felt within the nominal ±30% relative error range. This error range has been reported in previous studies as a reference value for uncertainty in ERHIs due to data length effects when working with 15-year time series (Kennard et al., 2010a). In addition, we computed the coefficient of variation of each ERHI distribution using the observed ERHI as a reference and obtained the average relative error between each predicted and observed ERHI.

6.2.4 Case study

6.2.4.1 Study area and model

We executed the calibration experiments in the Honeyoey Creek-Pine Creek Watershed located in east-central Michigan, US (Figure 14). This watershed has a drainage area of 1010 km^2, and its land use is predominantly agriculture, covering about 50% of the total area, followed by forests (~24%) and wetlands (~16%). Developed areas account for less than 4% of the total watershed area (Hernandez-Suarez et al., 2018). SWAT 2012, Rev. 622 was used as the deterministic hydrological model in equation 1 for predicting streamflow at the watershed outlet. SWAT is a process-based model widely used to simulate daily water quantity and quality time series at the watershed scale (Arnold et al., 2012). In SWAT, a watershed is divided into subwatersheds, which are comprised of hydrologic response units (HRUs). An HRU is a homogeneous land unit concerning land use/cover, soil type, and slope. In this study, the Honeyoey Creek-Pine Creek Watershed was divided into 250 subwatersheds, each one comprised by a single HRU representing the dominant land use, soil type, and slope conditions (Einheuser et al., 2012). SWAT was used to simulate daily streamflows from 2003 to 2014 (*Period 1*) and from 1983 to 1994 (*Period 2*). Warm-up periods of 2 years (1981-1982, and

2001-2002) were used to reduce the effect of initial conditions in both periods. Potential evapotranspiration was estimated using the Penman-Monteith equation (Monteith, 1965). Surface runoff was obtained using the Soil Conservation Service (SCS) curve number method (USDA-SCS, 1972), and the selected routing method was the variable storage coefficient routine developed by Williams (1969).

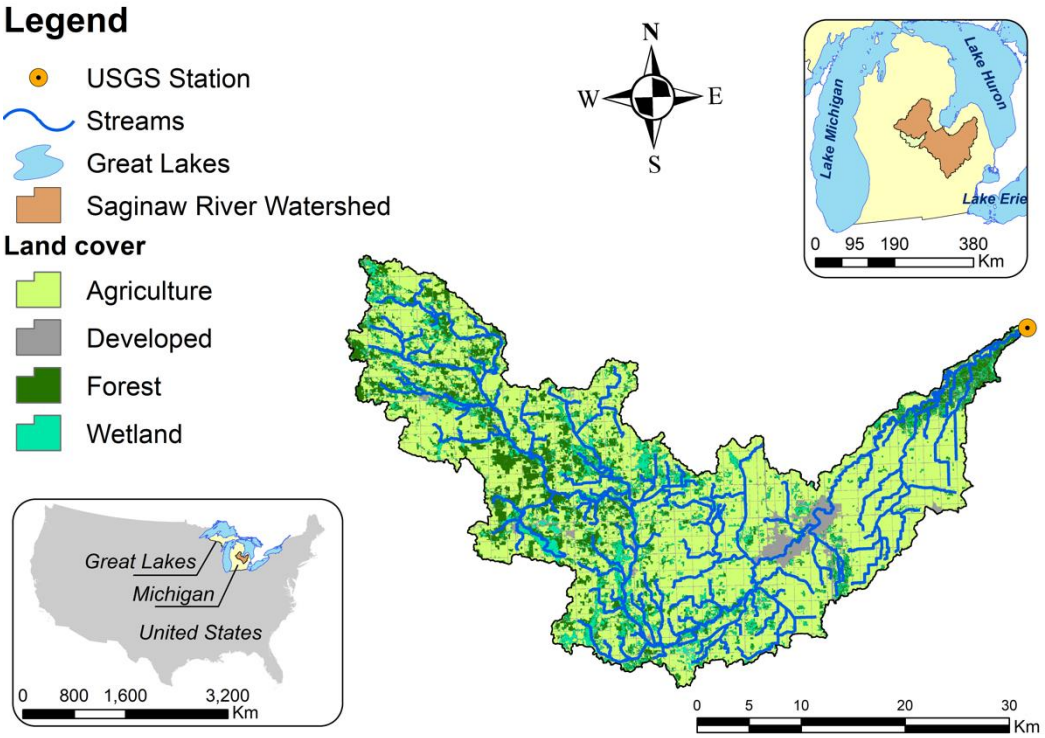


Figure 14 Study area location and major land uses

6.2.4.2 Data collection

Input data included the 30-m resolution National Elevation Dataset provided by the US Geological Survey (USGS, 2018), the 30-m resolution Cropland Data Layer provided by the National Agricultural Statistics Service of the US Department of Agriculture (USDA-NASS, 2012), soil properties extracted from the Soil Survey Geographic Database (SSURGO) from the USDA Natural Resources Conservation Service (USDA-NRCS, 2020), and daily precipitation and maximum and minimum air temperature time series from 1981 to 2014 collected at two land

stations from the National Centers for Environmental Information of the National Oceanic and Atmospheric Administration (NOAA-NCEI, 2020). Missing values and remaining input weather data such as solar radiation, wind speed, and relative humidity were estimated using the SWAT built-in WXGEN stochastic weather generator (Neitsch et al., 2011). Daily observed streamflow records were obtained at the watershed outlet for the period of study from the Pine River Near Midland USGS gauging station 04155500 (USGS, 2020).

6.2.4.3 Calibration parameters

The set of $\boldsymbol{\theta}_H$ to be estimated was comprised of 15 SWAT model parameters. Maximum and minimum limits for each parameter were defined following the model documentation (Neitsch et al., 2011) and previous studies (Herman et al., 2018; Hernandez-Suarez et al., 2018). Model parameters were adjusted during the calibration process either by replacing the original value with a new one or by perturbing the original value by a fraction. Most of the parameters were assumed to have the same value in every HRU. Only the Curve Number for moisture condition II (CN2) and the available water capacity of the soil layer (SOL_AWC) were assumed to spatially change and were adjusted by perturbing the initial default values by a global fraction. These global fractions were calibrated instead of estimating CN2 and SOL_AWC at each individual HRU. Calibration model parameters and their calibration ranges are presented in Table 11.

Table 11 Model calibration parameters and ranges

Parameter	Description	Calibration range
Biomix	Biological mixing efficiency	[0, 1]
CN2*	Initial Soil Conservation Service (SCS) runoff number for moisture condition II	[-0.25, 0.25]
Canmx	Maximum canopy storage (mm H_2O)	[0, 100]
Esco	Plant uptake compensation factor	[0, 1]
Epco	Soil evaporation compensation factor	[0, 1]
Alpha bf	Baseflow alpha factor ($days^{-1}$)	[0, 1]
Gw delay	Groundwater delay time (days)	[0, 500]
Gwqmn	Threshold depth of water in the shallow aquifer required for return flow to occur (mm H_2O)	[0, 5000]
Gw revap	Groundwater "revap" coefficient	[0.02, 0.2]
Revapmn	Threshold depth of water in the shallow aquifer for "revap" or percolation to the deep aquifer to occur (mm H_2O)	[0, 1000]
Rchrg dp	Deep aquifer percolation fraction	[0, 1]
Ch n2	Manning's *n* value for the main channel	[0, 0.3]
Ch k2	Effective hydraulic conductivity in main channel alluvium (mm h^{-1})	[0, 500]
Sol awc*	Available water capacity of the soil layer (mm H_2O mm^{-1} soil)	[-0.25, 0.25]
Surlag	Surface runoff lag coefficient	[1, 24]

Notes:
*These parameters were adjusted by perturbing the initial default spatially-varying values for each HRU by a global fraction.

6.2.4.4 Ecologically Relevant Hydrologic Indices

The selection of hydrologic indices depends on the ecohydrological application objectives. For instance, some studies target non-redundant hydrologic metrics for streamflow classification (Eng et al., 2017), others target specific hydrologic indices relevant to the condition of specific biological communities (George et al., 2021). Our goal in this study was to calibrate the hydrologic model targeting a balanced representation of the streamflow regime, which is of interest when defining environmental flows (Poff et al., 2010). Therefore, we selected 32 Indices of Hydrologic Alteration (IHA) (The Nature Conservancy, 2009), describing the central tendency of several streamflow regime characteristics. These indices were originally classified into five categories representing distinct regime facets such as magnitude, duration, frequency, timing, and rate of change of flows (Richter et al., 1997). In addition, we considered seven indices proposed by Archfield et al. (2014) (a.k.a., Magnificent seven) describing basic

properties of streamflow time series such as central tendency, variability, skewness, kurtosis, autocorrelation, and seasonality. The list of 39 ERHIs is presented in Table 12.

Table 12 List of ERHIs used in this study

Category	Index*	Description
Magnitude of monthly water conditions	MA12 – MA23	Mean monthly flows from January to December ($m^3 s^{-1}$)
Magnitude and duration of annual extreme water conditions	DL1 – DL5	Annual minimum with 1-, 3-, 7-, 30-, and 90-day moving average flow ($m^3 s^{-1}$)
	DH1 – DH5	Annual maximum with 1-, 3-, 7-, 30-, and 90-day moving average flow ($m^3 s^{-1}$)
	ML17	Baseflow index based on the 7-day minimum flow
Timing of annual extreme water conditions	TL1	Julian day of annual minimum
	TH1	Julian day of annual maximum
Frequency and duration of high and low pulses	FL1	Mean low flow pulse count per water year ($year^{-1}$)
	DL16	Mean low flow pulse duration (days)
	FH1	Mean high flow pulse count per water year with a threshold equal to the 75th percentile of the entire flow record ($year^{-1}$)
	DH15	Mean high flow pulse duration with a threshold equal to the 75th percentile of the entire flow record (days)
Rate and frequency of water condition changes	RA1	Rise rate ($m^3 s^{-1} d^{-1}$)
	RA3	Fall rate ($m^3 s^{-1} d^{-1}$)
	RA8	Reversals ($year^{-1}$)
Magnificent seven	MAG1 – MAG4	First four L–moments (mean, coefficient of variation, skewness, and kurtosis)
	MAG5	Autoregressive lag-one AR(1) correlation coefficient
	MAG6 – MAG7	Amplitude and phase of the seasonal signal

* Index abbreviations for Indicators of Hydrologic Alteration (IHA) as presented by Olden and Poff (2003).

6.2.4.5 Experiments set up

The U-NSGA-III algorithm was implemented using the *pymoo* library in Python 3.7 (Blank and Deb, 2020). We developed a Python interface to modify SWAT's input text files to link *pymoo* and SWAT. The multi-objective optimization algorithm was executed for 1000 generations, using 100 well-spaced reference directions obtained with the Riesz s-Energy method (Blank et al., 2021). This resulted in a total number of 100,000 model evaluations. Other U-NSGA-III parameters included the crossover probability, the distribution index for the Simulated Binary Crossover operator, the mutation probability, and the distribution index for the polynomial mutation operator, defined as 0.9, 10, 1/15 (i.e., the inverse of the number of the

hydrological model calibration parameters), and 20, respectively. The MT-DREAM$_{(ZS)}$ algorithm was executed in Matlab R2019b using the MT-DREAM$_{(ZS)}$ package developed by Vrugt (2016). The SWAT interface in Python was linked with Matlab to compute the log-likelihood function. MT-DREAM$_{(ZS)}$ was executed using three Markov chains and five multi-try proposals for 10,000 generations (for a total of 150,000 model evaluations). Additional MT-DREAM$_{(ZS)}$ parameters were assigned the default values reported by Vrugt (2016). The calibration experiments were executed using up to 20 threads in parallel on a machine equipped with two Intel® Xeon® CPU E5-2640 Processors at 2.5 GHz with 64 GB RAM running Ubuntu 16.04.7 LTS.

6.3 RESULTS AND DISCUSSION

6.3.1 Convergence of multi-objective and Bayesian calibration experiments

Convergence of the U-NSGA-III algorithm (*Experiment 2, Period 1*) was monitored using the Hypervolume Indicator (Auger et al., 2009), which started to show a steady behavior after 800 generations (i.e., 80,000 model evaluations). The first feasible solution (i.e., model simulation with all the selected ERHIs within ±30% relative error) was found after 4,800 model evaluations. The total computation time for the multi-objective calibration was 32.43 hours. Regarding the Bayesian calibration experiments, the MT-DREAM$_{(ZS)}$ algorithm converged after 5,400 generations (i.e., 81,000 model evaluations) for the *Experiment 1, Period 1*; 8,400 generations (i.e., 126,000 model evaluations) for the *Experiment 1, Period 2*; and 6,200 generations (i.e., 93,000 model evaluations) for the *Experiment 2, Period 2*. The total computation times for the Bayesian calibration experiments, which were simultaneously executed in the same machine, were 118.74 hours for *Experiment 1, Period 1*; 120.19 hours for *Experiment 1, Period 2*; and 118.97 hours for *Experiment 2, Period 2*. It is worth noting that Bayesian calibration using MCMC sampling had 50% more model evaluations than the multi-

objective calibration. According to the results, Bayesian calibration convergence using prior knowledge was 26% faster than the non-informative case. This faster convergence can be partially attributed to the further constrained search space for the informative case through the prior distribution.

6.3.2 Comparison between posterior parameter distributions using non-informative priors and Pareto-optimal results

The distribution of the hydrologic and error model parameters considered in this study are presented in Figure 15 for each experiment and calibration period. Under *Experiment 1*, some parameter distributions were similar for both calibration periods, whereas others showed a very distinct behavior. The latter group was comprised mostly of soil- and groundwater-related parameters: Epco, Gwqmn, Gw revap (which represents water movement from the shallow aquifer to the overlying unsaturated zone), Rchrg dp, and the Ch k2 (associated to transmission losses in the main channel). The standard deviation of the transformed autocorrelated residuals σ_W was greater than *period* 2, indicating a higher total uncertainty in streamflow predictions in *period 2* compared to *period 1*. These differences were perhaps originated by the time-varying nature of model parameters driven by non-stationarity effects from input weather data and changes in land use (Xiong et al., 2019).

Regarding *Experiment 2*, parameter distributions were more consistent across the two calibration periods compared with *Experiment 1*. This behavior revealed the strong influence of the prior distribution on the Bayesian calibration results for *period 2*. However, some parameters showed important differences, including Canmx (related to surface water interception) and Alpha bf (related to the shape of recession curves). These differences were originated from the contribution of new data under *period 2*. Excepting the global multiplier for Sol awc, the posterior parameter distributions from Bayesian calibration (i.e., *period 2*) were narrower

compared to the multi-objective calibration distributions (i.e., *period 1*) for *Experiment 2*. This reduction in parametric uncertainty resulted from the assimilation of a greater amount of information (i.e., *Experiment 2, period 2* ultimately used information from both *periods 1* and *2*).

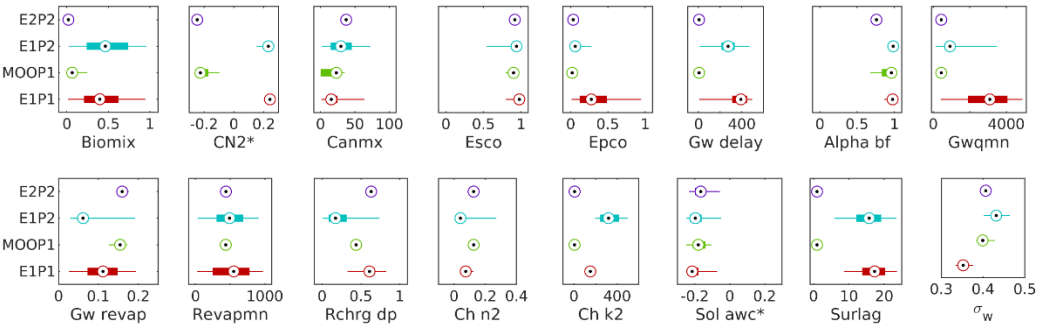


Figure 15 Distribution of model parameters obtained from multi-objective calibration (*Experiment 2, Period 1* – MOOP1) and Bayesian parameter estimation (*Experiment 1, Period 1* – E1P1; *Experiment 1, Period 2* – E1P2; *Experiment 2, Period 2* – E2P2). Box and whisker plots represent the 50% and 95% confidence limits, respectively; points represent median parameter values. Parameter descriptions are reported in Table 11. *These parameters were calibrated using global multipliers

In general, *Experiment 2* distributions were drastically narrower than *Experiment 1* distributions. Bayesian calibration with non-informative priors (especially *Experiment 1, period 1*) resulted in poorly-informative posteriors. These posteriors included distributions for Biomix, groundwater parameters such as Gwqmn, Gw revap, and Revapmn, and Surlag. Poorly-informative posteriors are related to the equifinality problem, indicating that different parameter combinations yield similar outputs (Beven, 2006). This may apply to the groundwater parameters interacting with each other and Surlag, which represents water storage. Likewise, non-informative posteriors are indicative of a low sensitivity of the modeling outputs to changes in those particular parameters, which can explain Biomix behavior. Nevertheless, *Experiment 1, period 1* attained the lowest σ_W, which can offset a lower total uncertainty in streamflow predictions.

The reduction in parametric uncertainty attained by *Experiment 2* was mainly driven by the constraints to ERHIs simulation accuracy, introduced by the proposed multi-objective calibration strategy. Another effect of the ERHIs constraints was the difference in the actual range of values for certain parameters. For instance, *Experiment 2* results indicated that the global multiplier perturbing initial CN2 values was around -25%, whereas *Experiment 1* results indicated the opposite (i.e., around +25%). Higher CN2 makes the watershed more impervious, resulting in higher runoff generation (and therefore higher streamflows). Similarly, while *Experiment 2* resulted in Surlag values close to unity, *Experiment 1* resulted in Surlag values mostly between 10 and 20. Surlag controls the fraction of total available water that enters the main channel on a daily basis; higher Surlag values result in higher fractions. Likewise, *Experiment 2* generated Gw delay close to zero, whereas *Experiment 1* yielded values greater than 200 days. Gw delay represents the time that water spends in the vadose zone before becoming shallow aquifer recharge. Gw delay ultimately affects groundwater contributions to the main channel, which also impacts low-flow dynamics.

6.3.3 Performance of uncertainty quantification of daily streamflows

Uncertainty quantification performance of streamflow predictions is presented in Figure 16 for each experiment and calibration period. As expected, the streamflow prediction bounds resulted in a lower parametric uncertainty for *Experiment 2*, which was consistent with the narrower parameter distributions presented in Figure 15 and the effects of the ERHIs' optimization constraints. In fact, parametric uncertainty for *Experiment 2*, *Period 2* (i.e., Bayesian parameter estimation using multi-objective calibration prior) was drastically lower compared to the other cases (~72% narrower compared to *Experiment* 1 for the same period). Regarding total uncertainty, the hydrographs in Figure 16 (left column) reveal wider limits for

low flows in *Experiment 1* (i.e., non-informative priors), and wider limits for high flows in *Experiment 2*.

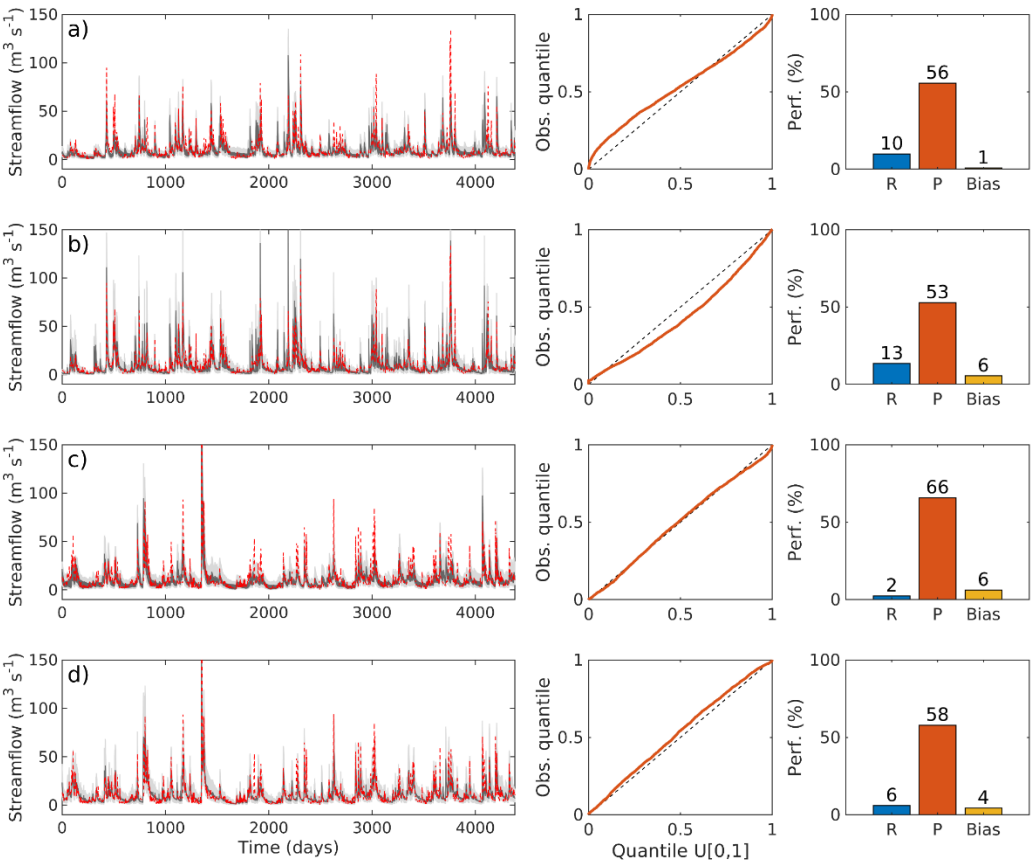


Figure 16 Uncertainty quantification performance using multi-objective calibration and Bayesian parameter estimation. The hydrographs (left column) represent the 95% prediction bounds for streamflow; light gray is for total uncertainty, dark gray is for parametric uncertainty, red line are observations. The middle column presents the corresponding quantile-quantile plots (PQQ) using a standard uniform distribution. The right column presents the overall performance indices for reliability (R), precision (P), and Bias. a) *Experiment 1, Period 1*; b) *Experiment 2, Period 1;* c) *Experiment 1, Period 2;* d) *Experiment 2, Period 2*

In terms of reliability, PQQ plots indicated that *Experiment 1* (Figure 16 row b, middle column) slightly over-estimated predictive uncertainty under *Period 1* (curve above the 1:1 diagonal at the lower-left corner and below the same line at the upper-right corner). Meanwhile, Pareto-optimal solutions (*Experiment 2*, Figure 16 row b, middle column) over-predicted streamflows under the same calibration period (curve was mostly below the 1:1 diagonal). The

latter occurred because total uncertainty was estimated using the σ_w obtained from the residuals of each Pareto-optimal solution. Since the multi-objective calibration strategy did not consider the additive error term included in the Bayesian calibration framework, the streamflow predictions resulted in higher values by explicitly adding the residual innovations to the Pareto-optimal simulations. Regarding *Period 2*, PQQ plots indicated excellent reliability for the streamflow predictions in both experiments (Figure 16 rows c and d, middle column).

When statistically comparing the observed quantiles from both experiments with a 95% confidence level, we did not find evidence of a significant difference in the reliability measure under *Period 2* (p = 0.070). For *Period 1*, the reliability was significantly different for both experiments (p = 1.7 x 10^{-5}), with *Experiment 1* presenting a better result overall (i.e., 10% vs. 13%, see Figure 16, right column). Regarding precision, no evidence of a significant difference was found between the two experiments under *Period 1* (p = 0.11). Meanwhile, *Experiment 2* resulted in a higher precision than *Experiment 1* under *Period 2* (58% vs 66%, p = 1.3 x 10^{-8}). Regarding bias in the long-term water balance, no significant difference was found for *Period 1* (p = 0.40), whereas for *Period 2*, *Experiment 2* attained a significantly lower bias compared with *Experiment 1* (4% vs. 6%, p = 0.0059). In other words, Bayesian calibration using multi-objective priors resulted in lower total uncertainty in streamflow predictions with lower bias in long-term water balance and no significant loss in reliability.

6.3.4 Performance of uncertainty quantification of ERHIs

Figure 17 shows the distribution of the relative errors for the 32 IHA and Magnificent seven indices selected in this study and reported in Table 12. This figure presents the parametric predictive distributions for the multi-objective calibration under *Period 1* because σ_w was not directly calibrated in this case. As expected, all Pareto-optimal predictions fell within the ±30%

relative error range due to the optimization constraints in ERHIs accuracy. However, note that only about 50% of the ERHIs computed on the observed streamflows fell within the Pareto-optimal predictive distributions. The situation was not better for the non-informative Bayesian calibration predictions under the same period. Meanwhile, the non-informative case under *Period 2* contained 64% of the observed ERHIs, against 56% for *Experiment 2* (see columns 3 and 4 in Table 13). Furthermore, all the predictive distributions for indices under the "frequency and duration of high and low pulses" category did not contain observed ERHIs under *Period 2*. The same situation was observed for DL1 (i.e., annual daily minimum), RA8 (i.e., reversals), and MAG3 (i.e., L-skewness). Distributions obtained with the Bayesian calibration using prior knowledge were particularly limited in the prediction of low-flow-related ERHIs across all the streamflow regime facets (i.e., magnitude, duration, frequency, rate of change, and timing). It is worth noting that Bayesian calibration using prior knowledge resulted in narrower predictive distributions compared with the non-informative case under *Period 2* (see columns 7 and 8 in Table 13), which explains the tradeoff between precision and reliability.

The aforementioned limitations in the accuracy of ERHIs predictive distributions were not surprising since we were trying to fit multiple streamflow facets simultaneously. Moreover, total uncertainty propagation through ERHIs computation affects precision and impacts the overall central tendency of the predictive distributions. Instead, our main interest was to limit the bias and the overall dispersion within the nominal ±30% relative error range. When comparing the median ERHIs relative errors against the acceptability threshold, we found that both experiments under *Period 2* yielded 90% of the median ERHIs within the acceptability threshold. For *Experiment 2*, ERHIs with the median outside the acceptability range include MA18-19 (i.e., monthly flows for July and August, summer flows), DL1, and FH1 (i.e., frequency of high flow

pulses). For *Experiment 1*. ERHIs with the median outside the acceptability range include MA12 and MA15 (i.e., monthly flows for January and April), TH1 (i.e., Julian day of annual maximum), and FH1.

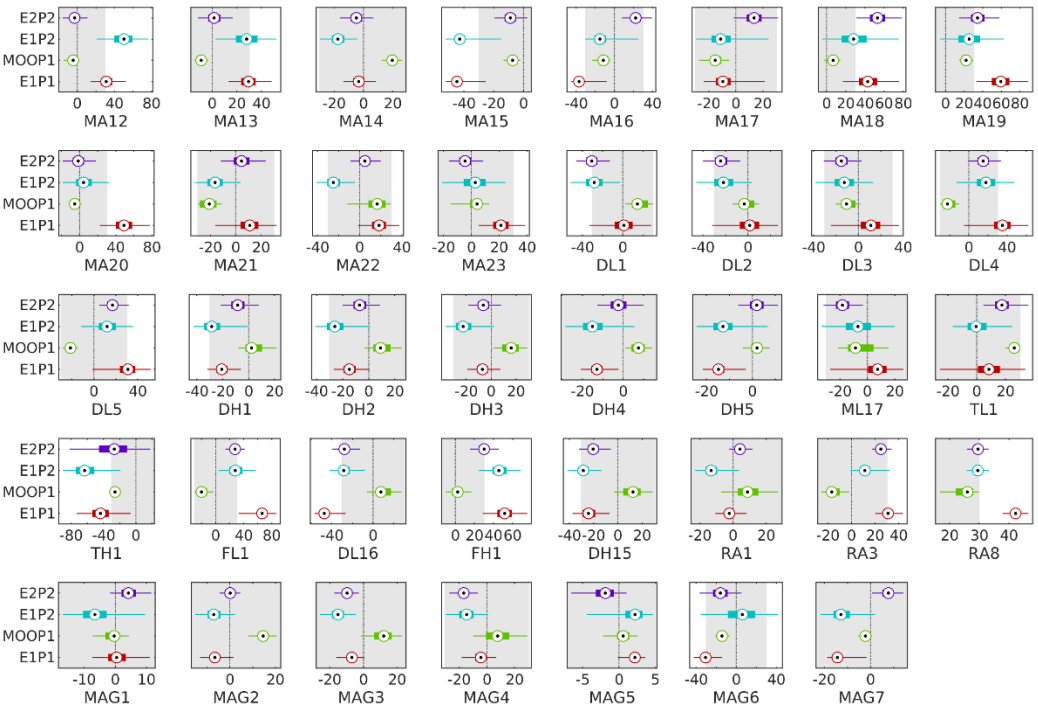


Figure 17 Distribution of relative errors of the selected ERHIs using multi-objective calibration (*Experiment 2, Period 1* – MOOP1) and Bayesian parameter estimation (*Experiment 1, Period 1* – E1P1; *Experiment 1, Period 2* – E1P2; *Experiment 2, Period 2* – E2P2). Box and whiskers represent the 50% and 95% confidence limits, respectively; points represent median relative error values; the vertical dotted line represents the zero axis, the gray area represent the nominal ±30% ERHI uncertainty. Index abbreviations are reported in Table 12

In general, *Experiment 2* exhibited a better performance in ERHIs prediction compared with *Experiment 1* because the overall precision had an average increase of ~32%. Regarding bias, ~59% ERHIs exhibited lower (better) values for *Experiment 2* compared with the non-informative case. It is worth noting that ERHIs predictions behaved differently depending on the calibration approach, as revealed by the differences in some posterior model parameter distributions. For instance, *Experiment 1* increased runoff generation and transmission losses and regulated groundwater return to the main channels. Meanwhile, *Experiment 2* decreased runoff

generation, increased water storage, and increased groundwater and lateral flow contributions. As a result, non-informative Bayesian calibration generated results with a lower bias for low flow ERHIs compared with *Experiment 2*. Likewise, *Experiment 2* resulted in a lower bias for high flow ERHIs compared with *Experiment 1*. The main factors contributing to these differences were the constraints inherited through the informative prior distribution and the likelihood function (particularly, the transformational approach selected to address heteroscedasticity). Deciding for a preferred calibration option ultimately requires a better understanding and assessment of the internal modeling processes using observations for variables of other water cycle components. Also, a better characterization of other uncertainty sources is required.

Table 13 Performance of predictive distributions of ERHIs obtained under *Period 2*. Reliability was evaluated by identifying whether the distributions contained the ERHIs from observations, and whether the median of the distributions was within the ±30% relative error range

ERHI Category	ERHI	Dist. contains observed ERHI		Median within ±30% range		Precision (%)		Bias (%)	
		Exp. 1	*Exp. 2*	*Exp. 1*	*Exp. 2*	*Exp. 1*	*Exp. 2*	*Exp. 1*	*Exp. 2*
Magnitude of monthly water conditions	MA12		✓		✓	14.4	6.8	49.6	-2.7
	MA13		✓	✓	✓	12.8	7.2	27.9	1.8
	MA14		✓	✓	✓	6.4	5.8	-17.4	-4.7
	MA15		✓		✓	7.7	5.7	-41.8	-8.8
	MA16	✓		✓	✓	12.6	8.1	-13.3	22.7
	MA17	✓	✓	✓	✓	13.1	8.4	-9.6	13.9
	MA18	✓		✓		20.2	12.1	29.9	53.5
	MA19	✓		✓		17.5	10.9	26.1	34.9
	MA20	✓	✓	✓	✓	12.9	9.4	5.2	-1.0
	MA21	✓	✓	✓	✓	9.3	9.0	-15.9	4.9
	MA22		✓	✓	✓	9.2	7.4	-24.5	5.0
	MA23	✓	✓	✓	✓	11.7	6.3	2.2	-4.3

Table 13 (cont'd).

Magnitude and duration of annual extreme water conditions (mean daily flow)	DL1			✓		12.3	8.3	-28.0	-30.5
	DL2	✓		✓	✓	12.6	8.3	-21.6	-24.0
	DL3	✓	✓	✓	✓	13.1	8.4	-12.5	-14.9
	DL4	✓	✓	✓	✓	15.0	8.8	17.9	15.2
	DL5	✓		✓	✓	11.9	7.0	12.2	17.3
	DH1		✓	✓	✓	10.6	7.6	-27.3	-8.0
	DH2	✓	✓	✓	✓	10.6	7.3	-24.8	-6.7
	DH3	✓	✓	✓	✓	9.9	6.7	-21.3	-5.9
	DH4	✓	✓	✓	✓	8.1	5.8	-14.4	-2.2
	DH5	✓	✓	✓	✓	7.1	4.5	-12.4	2.0
	ML17	✓		✓	✓	13.7	7.3	-6.6	-18.1
Timing of annual extreme water conditions	TL1	✓		✓	✓	11.2	8.1	0.8	18.0
	TH1		✓		✓	34.7	33.5	-59.2	-27.0
Frequency and duration of high and low pulses	FL1			✓	✓	13.8	7.2	29.0	27.6
	DL16			✓	✓	8.8	6.9	-27.6	-27.8
	FH1					11.2	8.0	46.0	30.5
	DH15			✓	✓	7.2	6.7	-28.5	-20.1
Rate and frequency of water condition changes	RA1	✓	✓	✓	✓	6.6	3.6	-12.6	4.3
	RA3	✓		✓	✓	8.1	4.2	12.1	25.1
	RA8			✓	✓	1.9	1.8	29.5	29.5
Magnificent seven	MAG1	✓	✓	✓	✓	6.4	3.4	-6.2	4.3
	MAG2	✓	✓	✓	✓	4.4	2.3	-6.9	0.1
	MAG3			✓	✓	5.4	3.7	-15.2	-9.8
	MAG4	✓		✓	✓	7.8	5.2	-14.7	-16.7
	MAG5	✓	✓	✓	✓	2.5	2.0	1.7	-2.1
	MAG6	✓	✓	✓	✓	20.0	10.7	4.9	-15.7
	MAG7	✓		✓	✓	5.9	3.4	-12.2	7.5

Note: *Exp. 1 = Experiment 1; Exp. 2 = Experiment 2*

6.4 CONCLUSIONS

In this study, we successfully linked multi-objective calibration with Bayesian parameter estimation for ERHIs uncertainty quantification. We achieved this by using a multivariate prior distribution of model parameters based on near-optimal Pareto solutions. The connection allowed the transfer of predefined ERHIs accuracy constraints into the overall Bayesian inference framework. The main advantage of the developed strategy is the use of multiple sources of information contained within a single continuously measured quantity for improving streamflow regimes prediction. Other benefits – compared with Bayesian calibration using non-informative

priors – included: 1) faster convergence to a stationary multivariate posterior distribution of model parameters using the MT-DREAM$_{(ZS)}$ algorithm., 2) drastic reduction of parametric uncertainty in streamflow predictions, 3) higher precision in streamflow predictive uncertainty with lower bias in the long-term water balance and no significant loss in reliability, and 4) higher precision in ERHIs' prediction.

It is worth noting that using prior knowledge had an important effect on how the hydrological model internally simulated surface runoff, interflow, and baseflow, as reflected by differences in related parameter posteriors. For example, when using non-informative priors, the chosen likelihood function favored the representation of low flows at the expense of high flows. Meanwhile, multi-objective calibration priors resulted in the opposite outcomes, yielding better results for high flows and behaving rather poorly for low flows. Therefore, further work is needed for understanding the tradeoffs between priors and likelihood functions for improving streamflow and ERHIs prediction. Ultimately, deciding on a particular modeling path depends on a better characterization of uncertainty sources (e.g., weather data, land use change, non-stationarity effects) and the incorporation of additional observations for other water cycle components (e.g., evapotranspiration, soil moisture, groundwater levels, leaf area index). Also, persistent limitations in ERHIs prediction for low flows, rate of change, and frequency and duration of high and low pulses require the revision of modeling workflows and structures to describe the ecohydrological behavior of freshwater ecosystems better.

7 CONCLUSIONS

This research developed different calibration strategies to improve the predictability of ERHIs using hydrological modeling. These strategies were tested in an agriculture-dominated watershed in Michigan, US. The first study evaluated the predictability of an exhaustive list of ERHIs by comparing the performance of two multi-objective and three single-objective formulations for model calibration. The second study improved the multi-objective formulations by explicitly targeting a subset of ERHIs, providing a balanced representation of the overall streamflow regime. Finally, the third study quantified the uncertainty in ERHIs prediction. For this purpose, multi-objective calibration and Bayesian parameter estimation were linked using near-optimal Pareto parameter distributions as prior knowledge. The overall process implemented watershed modeling, streamflow regime characterization, evolutionary multi-objective optimization, MCDM methods, and Bayesian inference using adaptive MCMC sampling. The following can be concluded from this research:

- The multi-objective calibration strategy based on NSE calculated on different streamflow transformations was superior to the RMSE-based strategy using flow separation. Specifically, the NSE-based strategy achieved a faster convergence, higher accuracy in ERHIs simulation, lower variability in ERHIs solutions, and narrower distributions of Pareto-optimal model parameters.
- NSE-based single-objective strategies calculated on untransformed and square-root-transformed streamflows outperformed the NSE- and RMSE-based multi-objective strategies in terms of accuracy in high-flow indices estimation.

- No calibration approach among the two multi-objective and three single-objective strategies tested in the first study was considered as the best for all ERHIs. Instead, they can be regarded as complementary to each other. However, the multi-objective strategies were preferred over the single-objective ones because they provided ranges of solutions while accounting for tradeoffs between different flow conditions.
- Outcomes from the first study help decision-makers in identifying which simulated ERHIs are more reliable when defining environmental standards and limits to human-driven hydrologic alteration. Having multiple optimal solutions provide natural resources managers with several options when defining these standards or limits under varying conditions (e.g., low, moderate, high flows).
- Obtaining a balanced representation of the overall streamflow regime is not directly related to improvements in a particular performance metric computed on streamflow time series. Instead, the balance responds to the interaction between different regime facets and flow conditions.
- It was possible to obtain consistent runoff simulations using multiple objective functions based on ERHIs that represent different streamflow regime facets. This was achieved without using any objective function computed on streamflow time series, at the expense of higher variability in the near-optimal Pareto solutions.
- Explicitly constraining or targeting ERHIs when calibrating a hydrological model, boosts its ability in representing those indices. Particularly, the performance-based strategy constraining the ERHIs accuracy was preferred for its lower variability in near-optimal Pareto solutions.
- Variabilities in the near-optimal Pareto solutions for both performance- and signature-based strategies were mainly driven by the model representation of low flows, as revealed by highly variable inverse-transformed KGE values and related Flow Duration Curve metrics.

- The performance-based strategy tested in the second study can incorporate acceptability thresholds for ERHIs prediction defined by decision-makers beforehand. These thresholds are generally based on additional information such as preferences of riverine species and socio-economic criteria. Therefore, the overall ecohydrological modeling process can be easily connected to broader decision-support tools.
- The overall bias and precision in streamflow predictions increased at the expense of a slight reduction in reliability. Still, the best precision in streamflow predictions was hardly below 60%.
- In general, ERHI predictive distributions were narrower and more accurate when using multi-objective calibration prior knowledge. However, ERHIs related to low flows presented a lower bias compared with the non-informative case.
- While most of the ERHIs predictive distributions fell within the nominal ±30% relative error range (i.e., expected uncertainty due to data length and non-stationarity), over 46% of ERHIs computed on streamflow observations did not fell within the predictive distributions, which can be related to model structure inadequacies.
- All the strategies developed in this research revealed limitations of the SWAT model structure when representing the vertical redistribution of soil moisture, rate of change, and timing of annual extremes. Particularly, the simulated interannual variability of low flow magnitude and timing, rise and fall rates, and duration and frequency of extreme flows was very inaccurate.
- Limitations in the representation of interannual variability have important repercussions in the definition of limits to hydrologic alteration, which is a major issue in freshwater systems protection and restoration. Thus, modelers and policymakers should account for these

limitations when implementing broader ecohydrological applications in ungauged and poorly gauged watersheds.

- Regarding the third study, simulated ERHIs with wide ranges of variability can be seen as less reliable for decision-makers. As a result, other ERHIs with narrower variability can be chosen for making decisions using modeling results, or additional efforts can be pursued to reduce the uncertainty in the ERHIs of interest.

8 FUTURE RESEARCH

This research presented novel calibration strategies based on multi-objective optimization and Bayesian inference to improve the predictability of ERHIs when using hydrological models. By linking multi-objective calibration results with Bayesian parameter estimation, multiple sources of information contained within a single continuously measured quantity (i.e., streamflow) can contribute to the overall uncertainty quantification process. In order to enhance the reliability of the proposed strategies, the following recommendations are provided here for further studies:

- *Extend the calibration strategies spatially and temporally*. The developed strategies were tested in a watershed with a single streamflow gauging station. Therefore, it is recommended that future studies implement these strategies considering multiple locations for calibration and validation purposes. For the former, the multi-objective calibration approach is flexible enough for incorporating multi-site calibration. Likewise, a third independent time period should be added to validate the predictive distributions that are built upon data from two other periods (one for multi-objective calibration and the other for Bayesian calibration using prior knowledge). In addition, the developed strategies should be extended to consider time-varying parameters and their effects on ERHIs predictability.
- *Explicitly consider other uncertainty sources.* In this research, only the parametric uncertainty was addressed explicitly. Additional uncertainty sources were aggregated into a lumped error model. Therefore, extending the Bayesian inference framework is recommended to account for other sources such as input (e.g., land use, weather, soil), model structure, measurement errors, data length, non-stationarity effects, and ERHIs computation methods. In addition, other error

models should be tested to evaluate whether a particular formulation is better suited for ERHIs prediction. A multi-level optimization approach can be suitable for trying different autocorrelation models/coefficients and transformational approaches, preventing parameter interaction issues.

- *Validate the modeling results with observations for other water cycle components.* Multi-variable calibration can be easily incorporated into the developed calibration strategies. With the advent of remotely sensed products and integrated modeling frameworks, it is recommended that future studies evaluate the predictability of ERHIs when considering additional variables representing other components of the hydrological cycle such as evapotranspiration, soil moisture, groundwater levels, and/or leaf area index.
- *Apply the calibration strategies with different model structures*. This research used SWAT as the hydrological model. Since several limitations were identified for this model, it is recommended to implement the developed strategies with other model structures and under different spatial scales (e.g., regional, national, continental) and time resolutions (e.g., sub-daily, monthly with disaggregation techniques) to evaluate any improvements in the prediction of ERHIs under different modeling paradigms.
- *Implement evolutionary multi-objective optimization with stochastic objective functions.* The connection between multi-objective optimization and Bayesian inference developed in this research was built upon the prior distribution. However, additional integration approaches should be examined, such as the formulation of stochastic objective functions that incorporate error models to quantify the total uncertainty.
- *Quantify uncertainty throughout broader ecohydrological applications.* In this research, we obtained prediction distributions for selected ERHIs. ERHIs are generally used as explanatory

variables for predicting other quantities of interest in ecohydrological applications, such as biological indicators or environmental flow settings. Therefore, future studies should consider these distributions for uncertainty quantification of ecohydrological variables.

- *Integrate surrogate modeling for computationally intensive models.* One of the main limitations of the integrated multi-objective calibration and Bayesian inference process is the requirement of large numbers of model executions. In general, MCMC methods are sequential approaches with limited parallelization capabilities. In order to address this issue, new strategies incorporating surrogate models should be considered.
- *Test uncertainty quantification results in practical decision-making scenarios.* Uncertainty analysis provides a higher transparency to the modeling process and to the discussion between modelers and policy and decision makers. Futures studies should consider the evaluation of modeling uncertainty effects on the definition of environmental standards which incorporate additional criteria regarding social and economic components. Other decision-making scenarios include prioritizing biological monitoring sites, definition of rules based on limits to hydrologic alteration, environmental impact assessment, and allocation of best management practices.

APPENDIX

Table A1 Description of ecologically-relevant hydrologic indices with all, high, medium, and low flow Pareto-optimal solutions having median relative errors outside the ±30% bound, for each multi-objective calibration strategy. Adapted from Olden and Poff (2003) and Henriksen et al. (2006)

Code	Hydrologic index	Units	Description	Calibration strategy
Magnitude of flow events				
Average flow conditions				
MA29	Variability in June flows	-	Coefficient of variation in monthly flows	Both NSE- and RMSE-based
MA30	Variability in July flows	-		Both NSE- and RMSE-based
MA31	Variability in August flows	-		Both NSE- and RMSE-based
MA32	Variability in September flows	-		Both NSE- and RMSE-based
MA33	Variability in October flows	-		Both NSE- and RMSE-based
MA34	Variability in November flows	-		Only RMSE-based
MA42	Variability across annual flows	-	Range of monthly flows divided by median monthly flows	Only NSE-based
MA44	Variability across annual flows	-	90^{th} - 10^{th} percentile of monthly flows divided by median monthly flows	Only NSE-based
MA45	Skewness in annual flows	-	(Mean annual flow - median annual flow)/median annual flow	Both NSE- and RMSE-based
Low flow conditions				
ML7	Mean minimum July monthly flow	$m^3 s^{-1}$	Mean minimum monthly flow	Both NSE- and RMSE-based
ML8	Mean minimum August monthly flow	$m^3 s^{-2}$		Both NSE- and RMSE-based
ML9	Mean minimum September monthly flow	$m^3 s^{-3}$		Only RMSE-based
ML14	Mean of annual minimum flows	-	Mean of the lowest annual daily flow divided by median annual daily flow averaged across all years	Both NSE- and RMSE-based

Table A1 (cont'd).

ML15	Low flow index	-	Mean of the lowest annual daily flow divided by mean annual daily flow averaged across all years	Both NSE- and RMSE-based
ML16	Median of annual minimum flows	-	Median of the lowest annual daily flow divided by median annual daily flow averaged across all years	Both NSE- and RMSE-based
ML17	Baseflow index 1	-	Seven-day minimum flow divided by mean annual daily flows averaged across all years	Only RMSE-based
ML19	Basefow index 2	-	Mean of the ratio of the lowest annual daily flow to the mean annual daily flow times 100 averaged across all years	Both NSE- and RMSE-based
ML21	Variability across annual minimum flows	-	Coefficient of variation in annual minimum flows averaged across all years	Both NSE- and RMSE-based
ML22	Specific mean annual minimum flows	$m^3 s^{-1} km^{-2}$	Mean annual minimum flows divided by catchment area	Both NSE- and RMSE-based
High flow conditions				
MH6	Mean maximum June monthly flow	$m^3 s^{-1}$	Mean of the maximum monthly flows	Both NSE- and RMSE-based
MH7	Mean maximum July monthly flow	$m^3 s^{-1}$		Both NSE- and RMSE-based
MH10	Mean maximum October monthly flow	$m^3 s^{-1}$		Both NSE- and RMSE-based
MH11	Mean maximum November monthly flow	$m^3 s^{-1}$		Only RMSE-based

Table A1 (cont'd)

MH21	High flow volume	days	Mean of the high flow volume (calculated as the area between the hydrograph and the upper threshold defined as the median annual flow) divided by median annual daily flow across all years	Both NSE- and RMSE-based
MH22	High flow volume	days	Mean of the high flow volume (calculated as the area between the hydrograph and the upper threshold defined as 3 times the median annual flow) divided by median annual daily flow across all years	Only NSE-based
MH23	High flow volume	days	Mean of the high flow volume (calculated as the area between the hydrograph and the upper threshold defined as 7 times the median annual flow) divided by median annual daily flow across all years	Only RMSE-based
Frequency of flow events				
Low flow conditions				
FL1	Low flood pulse count	$year^{-1}$	Average number of flow events below the 25^{th} percentile of the entire flow record	Both NSE- and RMSE-based
High flow conditions				
FH1	High flood pulse count 1	$year^{-1}$	Average number of flow events above the 75^{th} percentile of the entire flow record	Only RMSE-based

Table A1 (cont'd)

FH4	High flood pulse count 2	$year^{-1}$	Average number of days per year that the flow is above 7 times median daily flow of the entire record	Only NSE-based
FH5	Flood frequency 1	$year^{-1}$	Mean number of high flow events per year using a threshold equal to the median flow of the entire record	Both NSE- and RMSE-based
FH8	Flood frequency 2	$year^{-1}$	Mean number of high flow events per year using a threshold equal to the 25^{th} percentile of the entire flow record	Only RMSE-based
FH9	Flood frequency 2	$year^{-1}$	Mean number of high flow events per year using a threshold equal to the 75^{th} percentile of the entire flow record	Both NSE- and RMSE-based
Duration of flow events				
Low flow conditions				
DL1	Annual minimum daily flow	$m^3 s^{-1}$	Magnitude of minimum annual flow of 1-day duration	Both NSE- and RMSE-based
DL2	Annual minimum of 3-day moving average flow	$m^3 s^{-1}$	Magnitude of minimum annual flow of 3-day duration	Only RMSE-based
DL6	Variability of annual minimum daily average flow	-	Coefficient of variation in magnitude of minimum annual flow of 1-day duration	Both NSE- and RMSE-based
DL7	Variability of annual minimum of 3-day moving average flow	-	Coefficient of variation in magnitude of minimum annual flow of 3-day duration	Both NSE- and RMSE-based

Table A1 (cont'd)

DL8	Variability of annual minimum of 7-day moving average flow	-	Coefficient of variation in magnitude of minimum annual flow of 7-day duration	Both NSE- and RMSE-based
DL11	Mean of 1-day minima of daily discharge	-	Mean annual 1-day minimum, divided by median flow	Both NSE- and RMSE-based
DL12	Mean of 3-day minima of daily discharge	-	Mean annual 3-day minimum, divided by median flow	Only RMSE-based
DL16	Low flow pulse duration	days	Mean duration of flow events below the 25th percentile of the entire flow record	Both NSE- and RMSE-based
High flow conditions				
DH15	High flow pulse duration	days	Mean duration of flow events above the 75th percentile of the entire flow record	Both NSE- and RMSE-based
DH17	High flow duration 1	days	Mean duration of flow events above a threshold equal to the median flow of the entire record	Both NSE- and RMSE-based
DH19	High flow duration 1	days	Mean duration of flow events above a threshold equal to 7 times the median flow of the entire record	Only NSE-based
DH20	High flow duration 2	days	Mean duration of flow events above a threshold equal to the 75^{th} percentile value for the median annual flows	Only RMSE-based
DH21	High flow duration 2	days	Mean duration of flow events above a threshold equal to the 25^{th} percentile value for the median annual flows	Both NSE- and RMSE-based

Table A1 (cont'd)

DH23	Flood duration 2	days	Mean annual number of days that flows remain above the flood threshold (equal to the flow equivalent for a flood recurrence of 1.67 years) averaged across all years	Only RMSE-based
Timing of flow events				
Low flow conditions				
TL4	Seasonal predictability of non-low flow	-	Maximum proportion between the number of days that flow is above the 5-year flood threshold and 365 or 366 (leap year) among all years.	Only NSE-based
Rate of change in flow events				
Average flow conditions				
RA3	Fall rate	$m^3 s^{-1} d^{-1}$	Mean rate of negative changes in flow from one day to the next	Both NSE- and RMSE-based
RA6	Change of flow	$m^3 s^{-1}$	Median of difference between log10 of flows between two consecutive days with increasing flow	Both NSE- and RMSE-based
RA7	Change of flow	$m^3 s^{-1}$	Median of difference between log10 of flows between two consecutive days with decreasing flow	Both NSE- and RMSE-based

REFERENCES

REFERENCES

Abbasi, T., Abbasi, S.A., 2012. Multivariate Approaches for Bioassessment of Water Quality, in: Water Quality Indices. Elsevier, pp. 337–350. https://doi.org/10.1016/B978-0-444-54304-2.00015-4

Abouali, M., Daneshvar, F., Nejadhashemi, A.P., 2016a. MATLAB Hydrological Index Tool (MHIT): A high performance library to calculate 171 ecologically relevant hydrological indices. Ecol. Inform. 33, 17–23. https://doi.org/10.1016/j.ecoinf.2016.03.004

Abouali, M., Nejadhashemi, A.P., Daneshvar, F., Woznicki, S.A., 2016b. Two-phase approach to improve stream health modeling. Ecol. Inform. 34, 13–21. https://doi.org/10.1016/j.ecoinf.2016.04.009

Addor, N., Nearing, G., Prieto, C., Newman, A.J., Le Vine, N., Clark, M.P., 2018. A Ranking of Hydrological Signatures Based on Their Predictability in Space. Water Resour. Res. 54, 8792–8812. https://doi.org/10.1029/2018WR022606

Adriaenssens, V., De Baets, B., Goethals, P.L.M., De Pauw, N., 2004a. Fuzzy rule-based models for decision support in ecosystem management. Sci. Total Environ. 319, 1–12. https://doi.org/10.1016/S0048-9697(03)00433-9

Adriaenssens, V., Goethals, P.L.M., Charles, J., De Pauw, N., 2004b. Application of Bayesian Belief Networks for the prediction of macroinvertebrate taxa in rivers. Ann. Limnol. 40, 181–191. https://doi.org/10.1051/limn/2004016

Adriaenssens, V., Goethals, P.L.M., De Pauw, N., 2006. Fuzzy knowledge-based models for prediction of Asellus and Gammarus in watercourses in Flanders (Belgium). Ecol. Modell. 195, 3–10. https://doi.org/10.1016/j.ecolmodel.2005.11.043

Aguilera, P.A., Fernández, A., Fernández, R., Rumí, R., Salmerón, A., 2011. Bayesian networks in environmental modelling. Environ. Model. Softw. 26, 1376–1388. https://doi.org/10.1016/j.envsoft.2011.06.004

Ahmadi-Nedushan, B., St-Hilaire, A., Bérubé, M., Robichaud, É., Thiémonge, N., Bobée, B., 2006. A review of statistical methods for the evaluation of aquatic habitat suitability for instream flow assessment. River Res. Appl. 22, 503–523. https://doi.org/10.1002/rra.918

Allan, J.D., Yuan, L.L., Black, P., Stockton, T., Davies, P.E., Magierowski, R.H., Read, S.M., 2012. Investigating the relationships between environmental stressors and stream condition using Bayesian belief networks. Freshw. Biol. 57, 58–73. https://doi.org/10.1111/j.1365-2427.2011.02683.x

Almeida, D., Alcaraz-Hernández, J.D., Merciai, R., Benejam, L., García-Berthou, E., 2017. Relationship of fish indices with sampling effort and land use change in a large Mediterranean river. Sci. Total Environ. 605–606, 1055–1063.

https://doi.org/10.1016/j.scitotenv.2017.06.025

Almeida, S., Bulygina, N., McIntyre, N., Wagener, T., Buytaert, W., 2013. Improving parameter priors for data-scarce estimation problems. Water Resour. Res. 49, 6090–6095. https://doi.org/10.1002/wrcr.20437

Almeida, S., Le Vine, N., McIntyre, N., Wagener, T., Buytaert, W., 2016. Accounting for dependencies in regionalized signatures for predictions in ungauged catchments. Hydrol. Earth Syst. Sci. 20, 887–901. https://doi.org/10.5194/hess-20-887-2016

Álvarez-Cabria, M., González-Ferreras, A.M., Peñas, F.J., Barquín, J., 2017. Modelling macroinvertebrate and fish biotic indices: From reaches to entire river networks. Sci. Total Environ. 577, 308–318. https://doi.org/10.1016/j.scitotenv.2016.10.186

Ambelu, A., Lock, K., Goethals, P., 2010. Comparison of modelling techniques to predict macroinvertebrate community composition in rivers of Ethiopia. Ecol. Inform. 5, 147–152. https://doi.org/10.1016/j.ecoinf.2009.12.004

Ammann, L., Fenicia, F., Reichert, P., 2019. A likelihood framework for deterministic hydrological models and the importance of non-stationary autocorrelation. Hydrol. Earth Syst. Sci. 23, 2147–2172. https://doi.org/10.5194/hess-23-2147-2019

Andresen, J., Winkler, J., 2009. Weather and Climate, in: Schaetzl, R., Darden, J., Brandt, D. (Eds.), Michigan Geography and Geology. Pearson Custom Publishing, Boston, MA.

Araújo, M.B., New, M., 2007. Ensemble forecasting of species distributions. Trends Ecol. Evol. 22, 42–47. https://doi.org/10.1016/j.tree.2006.09.010

Archfield, S.A., Kennen, J.G., Carlisle, D.M., Wolock, D.M., 2014. An objective and parsimonious approach for classifying natural flow regimes at a continental scale. River Res. Appl. 30, 1166–1183. https://doi.org/10.1002/rra.2710

Arnold, J.G., Moriasi, D.N., Gassman, P.W., Abbaspour, K.C., White, M.J., Srinivasan, R., Santhi, C., Harmel, R.D., Griensven, a. Van, VanLiew, M.W., Kannan, N., Jha, M.K., 2012. Swat: Model Use, Calibration, and Validation. Trans. ASABE 55, 1491–1508. https://doi.org/ISSN 2151-0032

Arnold, J.G., Srinivasan, R., Muttiah, R.S., Williams, J.R., 1998. Large area hydrologic modeling and assesment Part I: Model development. JAWRA J. Am. Water Resour. Assoc. 34, 73–89. https://doi.org/10.1111/j.1752-1688.1998.tb05961.x

Arthur, D., Vassilvitskii, S., 2007. K-Means++: the Advantages of Careful Seeding. Proc. eighteenth Annu. ACM-SIAM Symp. Discret. algorithms 8, 1027–1025. https://doi.org/10.1145/1283383.1283494

Auger, A., Bader, J., Brockhoff, D., Zitzler, E., 2009. Theory of the hypervolume indicator. Proc. tenth ACM SIGEVO Work. Found. Genet. algorithms - FOGA '09 87. https://doi.org/10.1145/1527125.1527138

Austin, M.P., 1976. On non-linear species response models in ordination. Vegetatio 33, 33–41. https://doi.org/10.1007/BF00055297

Babbar-Sebens, M., Mukhopadhyay, S., Singh, V.B., Piemonti, A.D., 2015. A web-based software tool for participatory optimization of conservation practices in watersheds. Environ. Model. Softw. 69, 111–127. https://doi.org/10.1016/j.envsoft.2015.03.011

Barry, S., Elith, J., 2006. Error and uncertainty in habitat models. J. Appl. Ecol. 43, 413–423. https://doi.org/10.1111/j.1365-2664.2006.01136.x

Beechie, T.J., Sear, D.A., Olden, J.D., Pess, G.R., Buffington, J.M., Moir, H., Roni, P., Pollock, M.M., 2010. Process-based Principles for Restoring River Ecosystems. Bioscience 60, 209–222. https://doi.org/10.1525/bio.2010.60.3.7

Bejarano, M.D., Sordo-Ward, A., Gabriel-Martin, I., Garrote, L., 2019. Tradeoff between economic and environmental costs and benefits of hydropower production at run-of-river-diversion schemes under different environmental flows scenarios. J. Hydrol. 572, 790–804. https://doi.org/10.1016/j.jhydrol.2019.03.048

Bekele, E.G., Nicklow, J.W., 2007. Multi-objective automatic calibration of SWAT using NSGA-II. J. Hydrol. 341, 165–176. https://doi.org/10.1016/j.jhydrol.2007.05.014

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. Environ. Model. Softw. 40, 1–20. https://doi.org/10.1016/j.envsoft.2012.09.011

Beven, K., 2006. A manifesto for the equifinality thesis. J. Hydrol. 320, 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Beven, K., Binley, A., 2014. GLUE: 20 years on. Hydrol. Process. 28, 5897–5918. https://doi.org/10.1002/hyp.10082

Beven, K., Smith, P., 2015. Concepts of Information Content and Likelihood in Parameter Calibration for Hydrological Simulation Models. J. Hydrol. Eng. 20, 1–15. https://doi.org/10.1061/(asce)he.1943-5584.0000991

Blank, J., Deb, K., 2020. Pymoo: Multi-Objective Optimization in Python. IEEE Access 8, 89497–89509. https://doi.org/10.1109/ACCESS.2020.2990567

Blank, J., Deb, K., Dhebar, Y., Bandaru, S., Seada, H., 2021. Generating Well-Spaced Points on a Unit Simplex for Evolutionary Many-Objective Optimization. IEEE Trans. Evol. Comput. 25, 48–60. https://doi.org/10.1109/TEVC.2020.2992387

Blazkova, S., Beven, K., 2009. A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic. Water Resour. Res. 45, 1–12.

https://doi.org/10.1029/2007WR006726

Boavida, I., Dias, V., Ferreira, M.T., Santos, J.M., 2014. Univariate functions versus fuzzy logic: Implications for fish habitat modeling. Ecol. Eng. 71, 533–538. https://doi.org/10.1016/j.ecoleng.2014.07.073

Boets, P., Landuyt, D., Everaert, G., Broekx, S., Goethals, P.L.M., 2015. Evaluation and comparison of data-driven and knowledge-supported Bayesian Belief Networks to assess the habitat suitability for alien macroinvertebrates. Environ. Model. Softw. 74, 92–103. https://doi.org/10.1016/j.envsoft.2015.09.005

Booker, D.J., Snelder, T.H., Greenwood, M.J., Crow, S.K., 2015. Relationships between invertebrate communities and both hydrological regime and other environmental factors across New Zealand's rivers. Ecohydrology 8, 13–32. https://doi.org/10.1002/eco.1481

Borcard, D., Gillet, F., Legendre, P., 2011. Numerical Ecology with R, Applied Spatial Data Analysis with R. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4419-7976-6

Bowden, G.J., Dandy, G.C., Maier, H.R., 2005. Input determination for neural network models in water resources applications. Part 1 - Background and methodology. J. Hydrol. 301, 75–92. https://doi.org/10.1016/j.jhydrol.2004.06.021

Box, G.E.P., Cox, D.R., 1964. An Analysis of Transformations. J. R. Stat. Soc. Ser. B 26, 211–252. https://doi.org/10.2307/2287791

Branke, J., Deb, K., Miettinen, K., Słowiński, R., 2008. Multiobjective Optimization: Interactive and Evolutionary Approaches, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-88908-3

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and regression trees. CRC press.

Brown, L.R., May, J.T., Rehn, A.C., Ode, P.R., Waite, I.R., Kennen, J.G., 2012. Predicting biological condition in southern California streams. Landsc. Urban Plan. 108, 17–27. https://doi.org/10.1016/j.landurbplan.2012.07.009

Buchanan, B.P., Auerbach, D.A., McManamay, R.A., Taylor, J.M., Flecker, A.S., Archibald, J.A., Fuka, D.R., Walter, M.T., 2017. Environmental flows in the context of unconventional natural gas development in the Marcellus Shale: Ecol. Appl. 27, 37–55. https://doi.org/10.1002/eap.1425

Bunn, S.E., Arthington, A.H., 2002. Basic principles and ecological consequences of altered flow regimes for aquatic biodiversity. Environ. Manage. 30, 492–507. https://doi.org/10.1007/s00267-002-2737-0

Caldwell, P. V., Kennen, J.G., Sun, G., Kiang, J.E., Butcher, J.B., Eddy, M.C., Hay, L.E., Lafontaine, J.H., Hain, E.F., Nelson, S.A.C., Mcnulty, S.G., 2015. A comparison of hydrologic models for ecological flows and water availability. Ecohydrology 8, 1525–1546. https://doi.org/10.1002/eco.1602

Carlisle, D.M., Falcone, J., Meador, M.R., 2009a. Predicting the biological condition of streams: Use of geospatial indicators of natural and anthropogenic characteristics of watersheds. Environ. Monit. Assess. 151, 143–160. https://doi.org/10.1007/s10661-008-0256-z

Carlisle, D.M., Falcone, J., Wolock, D.M., Meador, M.R., Norris, R.H., 2009b. Predicting the natural flow regime: models for assessing hydrological alteration in streams. River Res. Appl. 30, n/a-n/a. https://doi.org/10.1002/rra.1247

Carlisle, D.M., Grantham, T.E., Eng, K., Wolock, D.M., 2017. Biological relevance of streamflow metrics: Regional and national perspectives. Freshw. Sci. 36, 927–940. https://doi.org/10.1086/694913

Carpenter, S.R., Stanley, E.H., Vander Zanden, M.J., 2011. State of the World's Freshwater Ecosystems: Physical, Chemical, and Biological Changes. Annu. Rev. Environ. Resour. 36, 75–99. https://doi.org/10.1146/annurev-environ-021810-094524

Casper, M.C., Grigoryan, G., Gronz, O., Gutjahr, O., Heinemann, G., Ley, R., Rock, A., 2012. Analysis of projected hydrological behavior of catchments based on signature indices. Hydrol. Earth Syst. Sci. 16, 409–421. https://doi.org/10.5194/hess-16-409-2012

Chee, Y.E., Elith, J., 2012. Spatial data for modelling and management of freshwater ecosystems. Int. J. Geogr. Inf. Sci. 26, 2123–2140. https://doi.org/10.1080/13658816.2012.717628

Chen, J., Zhong, P. an, Liu, W., Wan, X.Y., Yeh, W.W.G., 2020. A multi-objective risk management model for real-time flood control optimal operation of a parallel reservoir system. J. Hydrol. 590, 125264. https://doi.org/10.1016/j.jhydrol.2020.125264

Chen, S.H., Pollino, C.A., 2012. Good practice in Bayesian network modelling. Environ. Model. Softw. 37, 134–145. https://doi.org/10.1016/j.envsoft.2012.03.012

Chilkoti, V., Bolisetti, T., Balachandar, R., 2018. Multi-objective autocalibration of SWAT model for improved low flow performance for a small snowfed catchment. Hydrol. Sci. J. 63, 1482–1501. https://doi.org/10.1080/02626667.2018.1505047

Chinnayakanahalli, K.J., Hawkins, C.P., Tarboton, D.G., Hill, R.A., 2011. Natural flow regime, temperature and the composition and richness of invertebrate assemblages in streams of the western United States. Freshw. Biol. 56, 1248–1265. https://doi.org/10.1111/j.1365-2427.2010.02560.x

Chon, T.S., 2011. Self-Organizing Maps applied to ecological sciences. Ecol. Inform. 6, 50–61. https://doi.org/10.1016/j.ecoinf.2010.11.002

Cibin, R., Sudheer, K.P., Chaubey, I., 2010. Sensitivity and identifiability of stream flow generation parameters of the SWAT model. Hydrol. Process. 24, 1133–1148. https://doi.org/10.1002/hyp.7568

Clapcott, J.E., Collier, K.J., Death, R.G., Goodwin, E.O., Harding, J.S., Kelly, D., Leathwick, J.R., Young, R.G., 2012. Quantifying relationships between land-use gradients and structural and functional indicators of stream ecological integrity. Freshw. Biol. 57, 74–90. https://doi.org/10.1111/j.1365-2427.2011.02696.x

Clapcott, J.E., Goodwin, E.O., Snelder, T.H., Collier, K.J., Neale, M.W., Greenfield, S., 2017. Finding reference: a comparison of modelling approaches for predicting macroinvertebrate community index benchmarks. New Zeal. J. Mar. Freshw. Res. 51, 44–59. https://doi.org/10.1080/00288330.2016.1265994

Clapcott, J.E., Goodwin, E.O., Young, R.G., Kelly, D.J., 2014. A multimetric approach for predicting the ecological integrity of New Zealand streams. Knowl. Manag. Aquat. Ecosyst. 03. https://doi.org/10.1051/kmae/2014027

Clapcott, J.E., Young, R.G., Goodwin, E.O., Leathwick, J.R., 2010. Exploring the response of functional indicators of stream health to land-use gradients. Freshw. Biol. 55, 2181–2199. https://doi.org/10.1111/j.1365-2427.2010.02463.x

Coello Coello, C.A., Lamont, G.B., Veldhuizen, D. a Van, 2007. Evolutionary Algorithms for Solving Multi-Objective Problems, 2nd ed, Genetic and Evolutionary Computation Series. Springer US, Boston, MA.

Confesor, R.B., Whittaker, G.W., 2007. Automatic Calibration of Hydrologic Models With Multi-Objective Evolutionary Algorithm and Pareto Optimization1. JAWRA J. Am. Water Resour. Assoc. 43, 981–989. https://doi.org/10.1111/j.1752-1688.2007.00080.x

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random Forests for Classification in Ecology. Ecology 88, 2783–2792. https://doi.org/10.1890/07-0539.1

D'Ambrosio, J.L., Williams, L.R., Williams, M.G., Witter, J.D., Ward, A.D., 2014. Geomorphology, habitat, and spatial location influences on fish and macroinvertebrate communities in modified channels of an agriculturally-dominated watershed in Ohio, USA. Ecol. Eng. 68, 32–46. https://doi.org/10.1016/j.ecoleng.2014.03.037

D'Ambrosio, J.L., Williams, L.R., Witter, J.D., Ward, A., 2009. Effects of geomorphology, habitat, and spatial location on fish assemblages in a watershed in Ohio, USA. Environ. Monit. Assess. 148, 325–341. https://doi.org/10.1007/s10661-008-0163-3

Damanik-Ambarita, M.N., Everaert, G., Forio, M.A.E., Nguyen, T.H.T., Lock, K., Musonge, P.L.S., Suhareva, N., Dominguez-Granda, L., Bennetsen, E., Boets, P., Goethals, P.L.M., 2016. Generalized linear models to identify key hydromorphological and chemical variables determining the occurrence of macroinvertebrates in the Guayas River Basin (Ecuador). Water (Switzerland) 8. https://doi.org/10.3390/W8070297

Daneshvar, F., Nejadhashemi, A.P., Herman, M.R., Abouali, M., 2017a. Response of benthic macroinvertebrate communities to climate change. Ecohydrol. Hydrobiol. 17, 63–72. https://doi.org/10.1016/j.ecohyd.2016.12.002

Daneshvar, F., Nejadhashemi, A.P., Woznicki, S.A., Herman, M.R., 2017b. Applications of computational fluid dynamics in fish and habitat studies. Ecohydrol. Hydrobiol. 17, 53–62. https://doi.org/10.1016/j.ecohyd.2016.12.005

De'ath, G., 2007. Boosted regression trees for ecological modeling and prediction. Ecology 88, 243–251. https://doi.org/10.1890/0012-9658(2007)88[243:BTFEMA]2.0.CO;2

De'ath, G., 1999. Principal Curves : A New Technique for Indirect and Direct Gradient Analysis. Ecology 80, 2237–2253. https://doi.org/10.1890/0012-9658(1999)080[2237:PCANTF]2.0.CO;2

Death, R.G., Death, F., Stubbington, R., Joy, M.K., van den Belt, M., 2015. How good are Bayesian belief networks for environmental management? A test with data from an agricultural river catchment. Freshw. Biol. 60, 2297–2309. https://doi.org/10.1111/fwb.12655

Deb, K., 2001. Multi-objective optimization using evolutionary algorithms. John Wiley & Sons.

Deb, K., Agrawal, R.B., 1995. Simulated Binary Crossover for Continuous Search Space. Complex Syst. 9, 115–148.

Deb, K., Gupta, H., 2006. Introducing Robustness in Multi-Objective Optimization. Evol. Comput. 14, 463–494. https://doi.org/10.1162/evco.2006.14.4.463

Deb, K., Jain, H., 2014. An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems With Box Constraints. IEEE Trans. Evol. Comput. 18, 577–601. https://doi.org/10.1109/TEVC.2013.2281535

Deb, K., Pratap, A., Agarwal, S., Meyarivan, T., 2002a. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. 6, 182–197. https://doi.org/10.1109/4235.996017

Deb, K., Thiele, L., Laumanns, M., Zitzler, E., 2002b. Scalable multi-objective optimization test problems. Proc. 2002 Congr. Evol. Comput. CEC 2002 1, 825–830. https://doi.org/10.1109/CEC.2002.1007032

Dhungel, S., Tarboton, D.G., Jin, J., Hawkins, C.P., 2016. Potential Effects of Climate Change on Ecologically Relevant Streamflow Regimes. River Res. Appl. 32, 1827–1840. https://doi.org/10.1002/rra.3029

Donohue, I., McGarrigle, M.L., Mills, P., 2006. Linking catchment characteristics and water chemistry with the ecological status of Irish rivers. Water Res. 40, 91–98. https://doi.org/10.1016/j.watres.2005.10.027

Dudgeon, D., Arthington, A.H., Gessner, M.O., Kawabata, Z.-I., Knowler, D.J., Lévêque, C., Naiman, R.J., Prieur-Richard, A.-H., Soto, D., Stiassny, M.L.J., Sullivan, C.A., 2006. Freshwater biodiversity: importance, threats, status and conservation challenges. Biol. Rev. Camb. Philos. Soc. 81, 163–82. https://doi.org/10.1017/S1464793105006950

Dyer, F., ElSawah, S., Croke, B., Griffiths, R., Harrison, E., Lucena-Moya, P., Jakeman, A., 2014. The effects of climate change on ecologically-relevant flow regime and water quality attributes. Stoch. Environ. Res. Risk Assess. 28, 67–82. https://doi.org/10.1007/s00477-013-0744-8

Eckart, K., McPhee, Z., Bolisetti, T., 2017. Performance and implementation of low impact development – A review. Sci. Total Environ. 607–608, 413–432. https://doi.org/10.1016/j.scitotenv.2017.06.254

Efstratiadis, A., Koutsoyiannis, D., 2010. One decade of multi-objective calibration approaches in hydrological modelling: a review. Hydrol. Sci. J. 55, 58–78. https://doi.org/10.1080/02626660903526292

Einheuser, M.D., Nejadhashemi, A.P., Sowa, S.P., Wang, L., Hamaamin, Y.A., Woznicki, S.A., 2012. Modeling the effects of conservation practices on stream health. Sci. Total Environ. 435–436, 380–391. https://doi.org/10.1016/j.scitotenv.2012.07.033

Einheuser, M.D., Nejadhashemi, A.P., Wang, L., Sowa, S.P., Woznicki, S.A., 2013a. Linking Biological Integrity and Watershed Models to Assess the Impacts of Historical Land Use and Climate Changes on Stream Health. Environ. Manage. 51, 1147–1163. https://doi.org/10.1007/s00267-013-0043-7

Einheuser, M.D., Nejadhashemi, A.P., Woznicki, S.A., 2013b. Simulating stream health sensitivity to landscape changes due to bioenergy crops expansion. Biomass and Bioenergy 58, 198–209. https://doi.org/10.1016/j.biombioe.2013.08.025

Elias, C.L., Calapez, A.R., Almeida, S.F.P., Chessman, B., Simões, N., Feio, M.J., 2016. Predicting reference conditions for river bioassessment by incorporating boosted trees in the environmental filters method. Ecol. Indic. 69, 239–251. https://doi.org/10.1016/j.ecolind.2016.04.027

Elith, J., Graham, C.H., 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. Ecography (Cop.). 32, 66–77. https://doi.org/10.1111/j.1600-0587.2008.05505.x

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberon, J., Williams, S., Wisz, M.S., Zimmermann, N.E., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography (Cop.). 29, 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77, 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Eng, K., Grantham, T.E., Carlisle, D.M., Wolock, D.M., 2017. Predictability and selection of hydrologic metrics in riverine ecohydrology. Freshw. Sci. 36, 915–926. https://doi.org/10.1086/694912

Esselman, P.C., Infante, D.M., Wang, L., Cooper, A.R., Wieferich, D., Tsang, Y.P., Thornbrugh, D.J., Taylor, W.W., 2013. Regional fish community indicators of landscape disturbance to catchments of the conterminous United States. Ecol. Indic. 26, 163–173. https://doi.org/10.1016/j.ecolind.2012.10.028

Euser, T., Winsemius, H.C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., Savenije, H.H.G., 2013. A framework to assess the realism of model structures using hydrological signatures. Hydrol. Earth Syst. Sci. 17, 1893–1912. https://doi.org/10.5194/hess-17-1893-2013

Everaert, G., De Neve, J., Boets, P., Dominguez-Granda, L., Mereta, S.T., Ambelu, A., Hoang, T.H., Goethals, P.L.M., Thas, O., 2014. Comparison of the abiotic preferences of macroinvertebrates in tropical river basins. PLoS One 9. https://doi.org/10.1371/journal.pone.0108898

Evin, G., Thyer, M., Kavetski, D., McInerney, D., Kuczera, G., 2014. Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity. Water Resour. Res. 50, 2350–2375. https://doi.org/10.1002/2013WR014185

Falcone, J.A., Carlisle, D.M., Weber, L.C., 2010. Quantifying human disturbance in watersheds: Variable selection and performance of a GIS-based disturbance index for predicting the biological condition of perennial streams. Ecol. Indic. 10, 264–273. https://doi.org/10.1016/j.ecolind.2009.05.005

Fan, J., Wu, J., Kong, W., Zhang, Y.Y.Y., Li, M., Zhang, Y.Y.Y., Meng, W., Zhang, and M., 2017. Predicting Bio-indicators of Aquatic Ecosystems Using the Support Vector Machine Model in the Taizi River, China. Sustainability 9, 892. https://doi.org/10.3390/su9060892

Farmer, W.H., Vogel, R.M., 2016. On the deterministic and stochastic use of hydrologic models. Water Resour. Res. 52, 5619–5633. https://doi.org/10.1002/2016WR019129

Feio, M.J., Poquet, J.M., 2011. Predictive Models for Freshwater Biological Assessment: Statistical Approaches, Biological Elements and the Iberian Peninsula Experience: A Review. Int. Rev. Hydrobiol. 96, 321–346. https://doi.org/10.1002/iroh.201111376

Fenicia, F., Kavetski, D., Reichert, P., Albert, C., 2018. Signature-Domain Calibration of Hydrological Models Using Approximate Bayesian Computation: Empirical Analysis of Fundamental Properties. Water Resour. Res. 54, 3958–3987. https://doi.org/10.1002/2017WR021616

Fernandez-Palomino, C.A., Hattermann, F.F., Krysanova, V., Vega-Jácome, F., Bronstert, A.,

2020. Towards a more consistent eco-hydrological modelling through multi-objective calibration: a case study in the Andean Vilcanota River basin, Peru. Hydrol. Sci. J. 00, 1–16. https://doi.org/10.1080/02626667.2020.1846740

Fernando, T.M.K.G., Maier, H.R., Dandy, G.C., 2009. Selection of input variables for data driven models: An average shifted histogram partial mutual information estimator approach. J. Hydrol. 367, 165–176. https://doi.org/10.1016/j.jhydrol.2008.10.019

Forio, M.A.E., Landuyt, D., Bennetsen, E., Lock, K., Nguyen, T.H.T., Ambarita, M.N.D., Musonge, P.L.S., Boets, P., Everaert, G., Dominguez-Granda, L., Goethals, P.L.M., 2015. Bayesian belief network models to analyse and predict ecological water quality in rivers. Ecol. Modell. 312, 222–238. https://doi.org/10.1016/j.ecolmodel.2015.05.025

Fox, E.W., Hill, R.A., Leibowitz, S.G., Olsen, A.R., Thornbrugh, D.J., Weber, M.H., 2017. Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. Environ. Monit. Assess. 189. https://doi.org/10.1007/s10661-017-6025-0

Friedman, J.H., 2001. Greedy Function Approximation : A Gradient Boosting Machine. Ann. Stat. 29, 1189–1232.

Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Stat. 1–67.

Frimpong, E.A., Sutton, T.M., Engel, B.A., Simon, T.P., 2005. Spatial-scale effects on relative importance of physical habitat predictors of stream health. Environ. Manage. 36, 899–917. https://doi.org/10.1007/s00267-004-0357-6

Fukuda, S., De Baets, B., 2016. Data prevalence matters when assessing species' responses using data-driven species distribution models. Ecol. Inform. 32, 69–78. https://doi.org/10.1016/j.ecoinf.2016.01.005

Fukuda, S., De Baets, B., Waegeman, W., Verwaeren, J., Mouton, A.M., 2013. Habitat prediction and knowledge extraction for spawning European grayling (Thymallus thymallus L.) using a broad range of species distribution models. Environ. Model. Softw. 47, 1–6. https://doi.org/10.1016/j.envsoft.2013.04.005

Galelli, S., Humphrey, G.B., Maier, H.R., Castelletti, A., Dandy, G.C., Gibbs, M.S., 2014. An evaluation framework for input variable selection algorithms for environmental data-driven models. Environ. Model. Softw. 62, 33–51. https://doi.org/10.1016/j.envsoft.2014.08.015

Garcia, F., Folton, N., Oudin, L., 2017. Which objective function to calibrate rainfall–runoff models for low-flow index simulations? Hydrol. Sci. J. 62, 02626667.2017.1308511. https://doi.org/10.1080/02626667.2017.1308511

Gazendam, E., Gharabaghi, B., Ackerman, J.D., Whiteley, H., 2016. Integrative neural networks models for stream assessment in restoration projects. J. Hydrol. 536, 339–350. https://doi.org/10.1016/j.jhydrol.2016.02.057

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian data analysis. CRC press.

Gelman, A., Rubin, D.B., 1992. Inference from Iterative Simulation Using Multiple Sequences. Stat. Sci. 7, 15–51. https://doi.org/10.1214/ss/1177011136

George, R., McManamay, R., Perry, D., Sabo, J., Ruddell, B.L., 2021. Indicators of hydro-ecological alteration for the rivers of the United States. Ecol. Indic. 120, 106908. https://doi.org/10.1016/j.ecolind.2020.106908

Gevrey, M., Dimopoulos, I., Lek, S., 2003. Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecol. Modell. 160, 249–264. https://doi.org/10.1016/S0304-3800(02)00257-0

Gieswein, A., Hering, D., Feld, C.K., 2017. Additive effects prevail: The response of biota to multiple stressors in an intensively monitored watershed. Sci. Total Environ. 593–594, 27–35. https://doi.org/10.1016/j.scitotenv.2017.03.116

Goethals, P.L.M., Dedecker, A.P., Gabriels, W., Lek, S., De Pauw, N., 2007. Applications of artificial neural networks predicting macroinvertebrates in freshwaters. Aquat. Ecol. 41, 491–508. https://doi.org/10.1007/s10452-007-9093-3

Goldberg, D., 1991. Real-coded genetic algorithms, virtual alphabets, and blocking. Complex Syst. 5, 139–167.

Golden, H.E., Lane, C.R., Prues, A.G., D'Amico, E., 2016. Boosted Regression Tree Models to Explain Watershed Nutrient Concentrations and Biological Condition. J. Am. Water Resour. Assoc. 52, 1251–1274. https://doi.org/10.1111/1752-1688.12447

Grenouillet, G., Buisson, L., Casajus, N., Lek, S., 2011. Ensemble modelling of species distribution: The effects of geographical and environmental ranges. Ecography (Cop.). 34, 9–17. https://doi.org/10.1111/j.1600-0587.2010.06152.x

Grill, G., Lehner, B., Thieme, M., Geenen, B., Tickner, D., Antonelli, F., Babu, S., Borrelli, P., Cheng, L., Crochetiere, H., Ehalt Macedo, H., Filgueiras, R., Goichot, M., Higgins, J., Hogan, Z., Lip, B., McClain, M.E., Meng, J., Mulligan, M., Nilsson, C., Olden, J.D., Opperman, J.J., Petry, P., Reidy Liermann, C., Sáenz, L., Salinas-Rodríguez, S., Schelle, P., Schmitt, R.J.P., Snider, J., Tan, F., Tockner, K., Valdujo, P.H., van Soesbergen, A., Zarfl, C., 2019. Mapping the world's free-flowing rivers. Nature 569, 215–221. https://doi.org/10.1038/s41586-019-1111-9

Guo, C., Lek, S., Ye, S., Li, W., Liu, J., Li, Z., 2015a. Uncertainty in ensemble modelling of large-scale species distribution: Effects from species characteristics and model techniques. Ecol. Modell. 306, 67–75. https://doi.org/10.1016/j.ecolmodel.2014.08.002

Guo, C., Park, Y., Liu, Y., Lek, S., 2015b. Toward a new generation of ecological modelling techniques, in: Advanced Modelling Techniques for Studying Global Changes in Environmental Sciences. Elsevier B.V., pp. 11–44. https://doi.org/10.1016/B978-0-444-

63536-5.00002-8

Guo, J., Zhou, J., Lu, J., Zou, Q., Zhang, H., Bi, S., 2014. Multi-objective optimization of empirical hydrological model for streamflow prediction. J. Hydrol. 511, 242–253. https://doi.org/10.1016/j.jhydrol.2014.01.047

Guo, Y., Zhang, Y., Zhang, L., Wang, Z., 2021. Regionalization of hydrological modeling for predicting streamflow in ungauged catchments: A comprehensive review. Wiley Interdiscip. Rev. Water 8, 1–32. https://doi.org/10.1002/wat2.1487

Gupta, H. V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. 377, 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Gupta, H. V., Wagener, T., Liu, Y., 2008. Reconciling theory with observations: elements of a diagnostic approach to model evaluation. Hydrol. Process. 22, 3802–3813. https://doi.org/10.1002/hyp.6989

Guse, B., Kail, J., Radinger, J., Schröder, M., Kiesel, J., Hering, D., Wolter, C., Fohrer, N., 2015. Eco-hydrologic model cascades: Simulating land use and climate change impacts on hydrology, hydraulics and habitats for fish and macroinvertebrates. Sci. Total Environ. 533, 542–556. https://doi.org/10.1016/j.scitotenv.2015.05.078

Hall, K.R., Herbert, M.E., Sowa, S.P., Mysorekar, S., Woznicki, S.A., Nejadhashemi, P.A., Wang, L., 2017. Reducing current and future risks: Using climate change scenarios to test an agricultural conservation framework. J. Great Lakes Res. 43, 59–68. https://doi.org/10.1016/j.jglr.2016.11.005

Hallouin, T., Bruen, M., O'Loughlin, F.E., 2020. Calibration of hydrological models for ecologically relevant streamflow predictions: A trade-off between fitting well to data and estimating consistent parameter sets? Hydrol. Earth Syst. Sci. 24, 1031–1054. https://doi.org/10.5194/hess-24-1031-2020

Hassanzadeh, E., Elshorbagy, A., Nazemi, A., Jardine, T.D., Wheater, H., Lindenschmidt, K.E., 2017. The ecohydrological vulnerability of a large inland delta to changing regional streamflows and upstream irrigation expansion. Ecohydrology 10, 1–17. https://doi.org/10.1002/eco.1824

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, 2nd ed, Elements of Statistical Learning, Springer Series in Statistics. Springer-Verlag, New York, NY. https://doi.org/10.1198/jasa.2004.s339

Hawkins, C.P., Norris, R.H., Hogue, J.N., Feminella, J.W., 2000. Development and evaluation of predictive models for measuring the biological integrity of streams. Ecol. Appl. 10, 1456. https://doi.org/10.1890/1051-0761(2000)010{[}1456:DAEOPM]2.0.CO;2

He, Y., Wang, J., Lek-Ang, S., Lek, S., 2010. Predicting assemblages and species richness of endemic fish in the upper Yangtze River. Sci. Total Environ. 408, 4211–4220.

https://doi.org/10.1016/j.scitotenv.2010.04.052

Henriksen, J.A., Kennen, J, G., Nieswand, S., 2006. Users Manual for the Hydroecological Integrity Assessment Process Software (including the New Jersey Assessment Tools). Open-File Report 2006-1093.

Hering, D., Borja, A., Carstensen, J., Carvalho, L., Elliott, M., Feld, C.K., Heiskanen, A.S., Johnson, R.K., Moe, J., Pont, D., Solheim, A.L., de Bund, W. van, 2010. The European Water Framework Directive at the age of 10: A critical review of the achievements with recommendations for the future. Sci. Total Environ. 408, 4007–4019. https://doi.org/10.1016/j.scitotenv.2010.05.031

Herman, M.R., Hernandez-Suarez, J.S., Nejadhashemi, A.P., Kropp, I., Sadeghi, A.M., 2020. Evaluation of Multi- and Many-Objective Optimization Techniques to Improve the Performance of a Hydrologic Model Using Evapotranspiration Remote-Sensing Data. J. Hydrol. Eng. 25, 04020006. https://doi.org/10.1061/(asce)he.1943-5584.0001896

Herman, M.R., Nejadhashemi, A.P., 2015. A review of macroinvertebrate- and fish-based stream health indices. Ecohydrol. Hydrobiol. 15, 53–67. https://doi.org/10.1016/j.ecohyd.2015.04.001

Herman, M.R., Nejadhashemi, A.P., Abouali, M., Hernandez-Suarez, J.S., Daneshvar, F., Zhang, Z., Anderson, M.C., Sadeghi, A.M., Hain, C.R., Sharifi, A., 2018. Evaluating the role of evapotranspiration remote sensing data in improving hydrological modeling predictability. J. Hydrol. 556, 39–49. https://doi.org/10.1016/j.jhydrol.2017.11.009

Herman, M.R., Nejadhashemi, A.P., Daneshvar, F., Abouali, M., Ross, D.M., Woznicki, S.A., Zhang, Z., 2016. Optimization of bioenergy crop selection and placement based on a stream health indicator using an evolutionary algorithm. J. Environ. Manage. 181, 413–424. https://doi.org/10.1016/j.jenvman.2016.07.005

Herman, M.R., Nejadhashemi, A.P., Daneshvar, F., Ross, D.M., Woznicki, S.A., Zhang, Z., Esfahanian, A.-H.H., 2015. Optimization of conservation practice implementation strategies in the context of stream health. Ecol. Eng. 84, 1–12. https://doi.org/10.1016/j.ecoleng.2015.07.011

Hermoso, V., Linke, S., Prenda, J., Possingham, H.P., 2011. Addressing longitudinal connectivity in the systematic conservation planning of fresh waters. Freshw. Biol. 56, 57–70. https://doi.org/10.1111/j.1365-2427.2009.02390.x

Hernandez-Suarez, J.S., Nejadhashemi, A.P., 2018. A Review of Macroinvertebrate- and Fish-based Stream Health Modeling Techniques. Ecohydrology e2022. https://doi.org/10.1002/eco.2022

Hernandez-Suarez, J.S., Nejadhashemi, A.P., Kropp, I.M., Abouali, M., Zhang, Z., Deb, K., 2018. Evaluation of the impacts of hydrologic model calibration methods on predictability of ecologically-relevant hydrologic indices. J. Hydrol. 564, 758–772. https://doi.org/10.1016/j.jhydrol.2018.07.056

Hill, R.A., Fox, E.W., Leibowitz, S.G., Olsen, A.R., Thornbrugh, D.J., Weber, M.H., 2017. Predictive mapping of the biotic condition of conterminous U.S. rivers and streams. Ecol. Appl. 27, 2397–2415. https://doi.org/10.1002/eap.1617

Hipsey, M.R., Hamilton, D.P., Hanson, P.C., Carey, C.C., Coletti, J.Z., Read, J.S., Ibelings, B.W., Valesini, F.J., Brookes, J.D., 2015. Predicting the resilience and recovery of aquatic systems: A framework for model evolution within environmental observatories. Water Resour. Res. 51, 7023–7043. https://doi.org/10.1002/2015WR017175

Hoang, T.H., Lock, K., Mouton, A., Goethals, P.L.M., 2010. Application of classification trees and support vector machines to model the presence of macroinvertebrates in rivers in Vietnam. Ecol. Inform. 5, 140–146. https://doi.org/10.1016/j.ecoinf.2009.12.001

Holguin-Gonzalez, J.E., Boets, P., Alvarado, A., Cisneros, F., Carrasco, M.C., Wyseure, G., Nopens, I., Goethals, P.L.M., 2013a. Integrating hydraulic, physicochemical and ecological models to assess the effectiveness of water quality management strategies for the River Cuenca in Ecuador. Ecol. Modell. 254, 1–14. https://doi.org/10.1016/j.ecolmodel.2013.01.011

Holguin-Gonzalez, J.E., Boets, P., Everaert, G., Pauwels, I.S., Lock, K., Gobeyn, S., Benedetti, L., Amerlinck, Y., Nopens, I., Goethals, P.L.M., 2014. Development and assessment of an integrated ecological modelling framework to assess the effect of investments in wastewater treatment on water quality. Water Sci. Technol. 70, 1798–1807. https://doi.org/10.2166/wst.2014.316

Holguin-Gonzalez, J.E., Everaert, G., Boets, P., Galvis, A., Goethals, P.L.M., 2013b. Development and application of an integrated ecological modelling framework to analyze the impact of wastewater discharges on the ecological water quality of rivers. Environ. Model. Softw. 48, 27–36. https://doi.org/10.1016/j.envsoft.2013.06.004

Jähnig, S.C., Kuemmerlen, M., Kiesel, J., Domisch, S., Cai, Q., Schmalz, B., Fohrer, N., 2012. Modelling of riverine ecosystems by integrating models: conceptual approach, a case study and research agenda. J. Biogeogr. 39, 2253–2263. https://doi.org/10.1111/jbi.12009

Jain, H., Deb, K., 2014. An evolutionary many-objective optimization algorithm using reference-point based nondominated sorting approach, Part II: Handling constraints and extending to an adaptive approach. IEEE Trans. Evol. Comput. 18, 602–622. https://doi.org/10.1109/TEVC.2013.2281534

Jang, J.S.R., 1993. ANFIS: Adaptive-Network-Based Fuzzy Inference System. IEEE Trans. Syst. Man Cybern. 23, 665–685. https://doi.org/10.1109/21.256541

Jehn, F.U., Chamorro, A., Houska, T., Breuer, L., 2019. Trade-offs between parameter constraints and model realism: a case study. Sci. Rep. 9, 1–12. https://doi.org/10.1038/s41598-019-46963-6

Jelks, H.L., Walsh, S.J., Burkhead, N.M., Contreras-Balderas, S., Diaz-Pardo, E., Hendrickson, D.A., Lyons, J., Mandrak, N.E., McCormick, F., Nelson, J.S., Platania, S.P., Porter, B.A.,

Renaud, C.B., Schmitter-Soto, J.J., Taylor, E.B., Warren, M.L., 2008. Conservation Status of Imperiled North American Freshwater and Diadromous Fishes. Fisheries 33, 372–407. https://doi.org/10.1577/1548-8446-33.8.372

Jerves-Cobo, R., Everaert, G., Iñiguez-Vela, X., Córdova-Vela, G., Díaz-Granda, C., Cisneros, F., Nopens, I., Goethals, P.L.M., 2017. A methodology to model environmental preferences of EPT taxa in the Machangara River Basin (Ecuador), Water (Switzerland). https://doi.org/10.3390/w9030195

Johnson, L.B., Host, G.E., 2010. Recent developments in landscape approaches for the study of aquatic ecosystems. J. North Am. Benthol. Soc. 29, 41–66. https://doi.org/10.1899/09-030.1

Johnson, Z.C., Snyder, C.D., Hitt, N.P., 2017. Landform features and seasonal precipitation predict shallow groundwater influence on temperature in headwater streams. Water Resour. Res. 53, 5788–5812. https://doi.org/10.1002/2017WR020455

Kail, J., Guse, B., Radinger, J., Schröder, M., Kiesel, J., Kleinhans, M., Schuurman, F., Fohrer, N., Hering, D., Wolter, C., 2015. A Modelling Framework to Assess the Effect of Pressures on River Abiotic Habitat Conditions and Biota. PLoS One 10, e0130228. https://doi.org/10.1371/journal.pone.0130228

Kakouei, K., Kiesel, J., Kail, J., Pusch, M., Jähnig, S.C., 2017. Quantitative hydrological preferences of benthic stream invertebrates in Germany. Ecol. Indic. 79, 163–172. https://doi.org/10.1016/j.ecolind.2017.04.029

Kalteh, A.M., Hjorth, P., Berndtsson, R., 2008. Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application. Environ. Model. Softw. 23, 835–845. https://doi.org/10.1016/j.envsoft.2007.10.001

Karr, J., 1996. Ecological integrity and ecological health are not the same. Eng. within Ecol. constraints.

Karr, J.R., 1999. Defining and measuring river health. Freshw. Biol. 41, 221–234. https://doi.org/10.1046/j.1365-2427.1999.00427.x

Karr, J.R., 1981. Assessment of Biotic Integrity Using Fish Communities. Fisheries 6, 21–27. https://doi.org/10.1577/1548-8446(1981)006<0021:AOBIUF>2.0.CO;2

Karr, J.R., Dudley, D.R., 1981. Ecological perspective on water quality goals. Environ. Manage. 5, 55–68. https://doi.org/10.1007/BF01866609

Karr, J.R., Yoder, C.O., 2004. Biological assessment and criteria improve total maximum daily load decision making. J. Environ. Eng. June, 594–604.

Kavetski, D., Fenicia, F., Reichert, P., Albert, C., 2018. Signature-Domain Calibration of Hydrological Models Using Approximate Bayesian Computation: Theory and Comparison to Existing Applications. Water Resour. Res. 54, 4059–4083. https://doi.org/10.1002/2017WR020528

Kennard, M.J., Mackay, S.J., Pusey, B.J., Olden, J.D., Marsh, N., 2010a. Quantifying uncertainty in estimation of hydrologic metrics for ecohydrological studies. River Res. Appl. 26, 137–156. https://doi.org/10.1002/rra.1249

Kennard, M.J., Pusey, B.J., Olden, J.D., MacKay, S.J., Stein, J.L., Marsh, N., 2010b. Classification of natural flow regimes in Australia to support environmental flow management. Freshw. Biol. 55, 171–193. https://doi.org/10.1111/j.1365-2427.2009.02307.x

Kennen, J.G., Kauffman, L.J., Ayers, M.A., Wolock, D.M., Colarullo, S.J., 2008. Use of an integrated flow model to estimate ecologically relevant hydrologic characteristics at stream biomonitoring sites. Ecol. Modell. 211, 57–76. https://doi.org/10.1016/j.ecolmodel.2007.08.014

Kent, M., 2006. Numerical classification and ordination methods in biogeography. Prog. Phys. Geogr. 30, 399–408. https://doi.org/10.1191/0309133306pp489pr

Kerans, B.L.L., Karr, J.R., 1994. A Benthic Index of Biotic Integrity (B-IBI) for Rivers of the Tennessee Valley. Ecol. Appl. 4, 768–785. https://doi.org/10.2307/1942007

Kiesel, J., Guse, B., Pfannerstill, M., Kakouei, K., Jähnig, S.C., Fohrer, N., 2017. Improving hydrological model optimization for riverine species. Ecol. Indic. 80, 376–385. https://doi.org/10.1016/j.ecolind.2017.04.032

Kiesel, J., Kakouei, K., Guse, B., Fohrer, N., Jähnig, S.C., 2020. When is a hydrological model sufficiently calibrated to depict flow preferences of riverine species? Ecohydrology 13, 1–15. https://doi.org/10.1002/eco.2193

Kling, H., Fuchs, M., Paulin, M., 2012. Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. J. Hydrol. 424–425, 264–277. https://doi.org/10.1016/j.jhydrol.2012.01.011

Knight, R.R., Gain, W.S., Wolfe, W.J., 2012. Modelling ecological flow regime: An example from the Tennessee and Cumberland River basins. Ecohydrology 5, 613–627. https://doi.org/10.1002/eco.246

Kollat, J.B., Reed, P.M., Wagener, T., 2012. When are multiobjective calibration trade-offs in hydrologic models meaningful? Water Resour. Res. 48, 1–19. https://doi.org/10.1029/2011WR011534

Krause, P., Boyle, D.P., Base, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Adv. Geosci. 89–97.

Kropp, I., Nejadhashemi, A.P., Deb, K., Abouali, M., Roy, P.C., Adhikari, U., Hoogenboom, G., 2019. A multi-objective approach to water and nutrient efficiency for sustainable agricultural intensification. Agric. Syst. 173, 289–302. https://doi.org/10.1016/j.agsy.2019.03.014

Kuehne, L.M., Olden, J.D., Strecker, A.L., Lawler, J.J., Theobald, D.M., 2017. Past, present, and

future of ecological integrity assessment for fresh waters. Front. Ecol. Environ. 15, 197–205. https://doi.org/10.1002/fee.1483

Kuemmerlen, M., Schmalz, B., Guse, B., Cai, Q., Fohrer, N., Jähnig, S.C., 2014. Integrating catchment properties in small scale species distribution models of stream macroinvertebrates. Ecol. Modell. 277, 77–86. https://doi.org/10.1016/j.ecolmodel.2014.01.020

Kuemmerlen, M., Stoll, S., Sundermann, A., Haase, P., 2016. Long-term monitoring data meet freshwater species distribution models: Lessons from an LTER-site. Ecol. Indic. 65, 122–132. https://doi.org/10.1016/j.ecolind.2015.08.008

Kwon, Y.S., Bae, M.J., Hwang, S.J., Kim, S.H., Park, Y.S., 2015. Predicting potential impacts of climate change on freshwater fish in Korea. Ecol. Inform. 29, 156–165. https://doi.org/10.1016/j.ecoinf.2014.10.002

Laloy, E., Vrugt, J.A., 2012. High-dimensional posterior exploration of hydrologic models using multiple-try DREAM (ZS) and high-performance computing. Water Resour. Res. 48, 1–18. https://doi.org/10.1029/2011WR010608

Landuyt, D., Broekx, S., D'hondt, R., Engelen, G., Aertsens, J., Goethals, P.L.M., 2013. A review of Bayesian belief networks in ecosystem service modelling. Environ. Model. Softw. 46, 1–11. https://doi.org/10.1016/j.envsoft.2013.03.011

Leathwick, J.R., Elith, J., Francis, M.P., Hastie, T., Taylor, P., 2006a. Variation in demersal fish species richness in the oceans surrounding New Zealand: An analysis using boosted regression trees. Mar. Ecol. Prog. Ser. 321, 267–281. https://doi.org/10.3354/meps321267

Leathwick, J.R., Elith, J., Hastie, T., 2006b. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. Ecol. Modell. 199, 188–196. https://doi.org/10.1016/j.ecolmodel.2006.05.022

Leclere, J., Oberdorff, T., Belliard, J., Leprieur, F., 2011. A comparison of modeling techniques to predict juvenile 0+ fish species occurrences in a large river system. Ecol. Inform. 6, 276–285. https://doi.org/10.1016/j.ecoinf.2011.05.001

Lecun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539

Leigh, C., Stewart-Koster, B., Sheldon, F., Burford, M.A., 2012. Understanding multiple ecological responses to anthropogenic disturbance: Rivers and potential flow regime change. Ecol. Appl. 22, 250–263. https://doi.org/10.1890/11-0963.1

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagnier, S., 1996. Application of neural networks to modelling nonlinear relationships in ecology. Ecol. Modell. 90, 39–52. https://doi.org/10.1016/0304-3800(95)00142-5

Lek, S., Guégan, J.F., 1999. Artificial neural networks as a tool in ecological modelling, an

introduction. Ecol. Modell. 120, 65–73. https://doi.org/10.1016/S0304-3800(99)00092-7

Ley, R., Hellebrand, H., Casper, M.C., Fenicia, F., 2016. Comparing classical performance measures with signature indices derived from flow duration curves to assess model structures as tools for catchment classification. Hydrol. Res. 47, 1–14. https://doi.org/10.2166/nh.2015.221

Li, X., Maier, H.R., Zecchin, A.C., 2015. Improved PMI-based input variable selection approach for artificial neural network and other data driven environmental and water resource models. Environ. Model. Softw. 65, 15–29. https://doi.org/10.1016/j.envsoft.2014.11.028

Li, X., Wang, Y., 2013. Applying various algorithms for species distribution modelling. Integr. Zool. 8, 124–135. https://doi.org/10.1111/1749-4877.12000

Li, X., Zhang, Y., Guo, F., Gao, X., Wang, Y., 2018. Predicting the effect of land use and climate change on stream macroinvertebrates based on the linkage between structural equation modeling and bayesian network. Ecol. Indic. 85, 820–831. https://doi.org/10.1016/j.ecolind.2017.11.044

Lin, Y., Chen, Q., Chen, K., Yang, Q., 2016. Modelling the presence and identifying the determinant factors of dominant macroinvertebrate taxa in a karst river. Environ. Monit. Assess. 188. https://doi.org/10.1007/s10661-016-5322-3

Lu, S., Kayastha, N., Thodsen, H., Van Griensven, a, Andersen, H.E., 2014. Multiobjective calibration for comparing channel sediment routing models in the soil and water assessment tool. J. Environ. Qual. 43, 110–120. https://doi.org/10.2134/jeq2011.0364

Maddock, I., 1999. The importance of physical habitat assessment for evaluating river health. Freshw. Biol. 41, 373–391. https://doi.org/10.1046/j.1365-2427.1999.00437.x

Maier, H.R., Kapelan, Z., Kasprzyk, J., Kollat, J., Matott, L.S., Cunha, M.C., Dandy, G.C., Gibbs, M.S., Keedwell, E., Marchi, A., Ostfeld, A., Savic, D., Solomatine, D.P., Vrugt, J.A., Zecchin, A.C., Minsker, B.S., Barbour, E.J., Kuczera, G., Pasha, F., Castelletti, A., Giuliani, M., Reed, P.M., 2014. Evolutionary algorithms and other metaheuristics in water resources: Current status, research challenges and future directions. Environ. Model. Softw. 62, 271–299. https://doi.org/10.1016/j.envsoft.2014.09.013

Maloney, K.O., Schmid, M., Weller, D.E., 2012. Applying additive modelling and gradient boosting to assess the effects of watershed and reach characteristics on riverine assemblages. Methods Ecol. Evol. 3, 116–128. https://doi.org/10.1111/j.2041-210X.2011.00124.x

Maloney, K.O., Weller, D.E., Russell, M.J., Hothorn, T., 2009. Classifying the biological condition of small streams: an example using benthic macroinvertebrates. J. North Am. Benthol. Soc. 28, 869–884. https://doi.org/10.1899/08-142.1

Manel, S., Williams, H.C., Ormerod, S.J., 2001. Evaluating presence absence models in ecology; the need to count for prevalence. J. Appied Ecol. 38, 921–931.

https://doi.org/10.1046/j.1365-2664.2001.00647.x

Mantyka-Pringle, C.S., Jardine, T.D., Bradford, L., Bharadwaj, L., Kythreotis, A.P., Fresque-Baxter, J., Kelly, E., Somers, G., Doig, L.E., Jones, P.D., Lindenschmidt, K.-E., 2017. Bridging science and traditional knowledge to assess cumulative impacts of stressors on ecosystem health. Environ. Int. 102, 125–137. https://doi.org/10.1016/j.envint.2017.02.008

Mantyka-Pringle, C.S., Martin, T.G., Moffatt, D.B., Linke, S., Rhodes, J.R., 2014. Understanding and predicting the combined effects of climate change and land-use change on freshwater macroinvertebrates and fish. J. Appl. Ecol. 51, 572–581. https://doi.org/10.1111/1365-2664.12236

Marchini, A., Facchinetti, T., Mistri, M., 2009. F-IND: A framework to design fuzzy indices of environmental conditions. Ecol. Indic. 9, 485–496. https://doi.org/10.1016/j.ecolind.2008.07.004

Marcot, B.G., 2017. Common quandaries and their practical solutions in Bayesian network modeling. Ecol. Modell. 358, 1–9. https://doi.org/10.1016/j.ecolmodel.2017.05.011

Marcot, B.G., 2012. Metrics for evaluating performance and uncertainty of Bayesian network models. Ecol. Modell. 230, 50–62. https://doi.org/10.1016/j.ecolmodel.2012.01.013

Marcot, B.G., Holthausen, R.S., Raphael, M.G., Rowland, M.M., Wisdom, M.J., 2001. Using Bayesian belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement. For. Ecol. Manage. 153, 29–42. https://doi.org/10.1016/S0378-1127(01)00452-2

Marcot, B.G., Steventon, J.D., Sutherland, G.D., McCann, R.K., 2006. Guidelines for developing and updating Bayesian belief networks applied to ecological modeling and conservation. Can. J. For. Res. 36, 3063–3074. https://doi.org/10.1139/x06-135

Mateo-Sagasta, J., Zadeh, S.M., Turral, H., 2018. More people, more food, worse water? - a global review of water pollution from agriculture. Food and Agriculture Organization of the United Nations and International Water Management Institute, Rome.

Mathews, R., Richter, B.D., 2007. Application of the indicators of hydrologic alteration software in environmental flow setting. J. Am. Water Resour. Assoc. 43, 1400–1413. https://doi.org/10.1111/j.1752-1688.2007.00099.x

Mathon, B.R., Rizzo, D.M., Kline, M., Alexander, G., Fiske, S., Langdon, R., Stevens, L., 2013. Assessing Linkages in Stream Habitat, Geomorphic Condition, and Biological Integrity Using a Generalized Regression Neural Network. J. Am. Water Resour. Assoc. 49, 415–430. https://doi.org/10.1111/jawr.12030

May, J.T., Brown, L.R., Rehn, A.C., Waite, I.R., Ode, P.R., Mazor, R.D., Schiff, K.C., 2015. Correspondence of biological condition models of California streams at statewide and regional scales. Environ. Monit. Assess. 187, 1–21. https://doi.org/10.1007/s10661-014-4086-x

May, R.J., Maier, H.R., Dandy, G.C., Fernando, T.M.K.G., 2008. Non-linear variable selection for artificial neural networks using partial mutual information. Environ. Model. Softw. 23, 1312–1326. https://doi.org/10.1016/j.envsoft.2008.03.007

Mazor, R.D., May, J.T., Sengupta, A., McCune, K.S., Bledsoe, B.P., Stein, E.D., 2018. Tools for managing hydrologic alteration on a regional scale: Setting targets to protect stream health. Freshw. Biol. 63, 786–803. https://doi.org/10.1111/fwb.13062

McCann, R.K., Marcot, B.G., Ellis, R., 2006. Bayesian belief networks: applications in ecology and natural resource management. Can. J. For. Res. 36, 3053–3062. https://doi.org/10.1139/x06-238

McCullagh, P., Nelder, J.A., 1989. Generalized linear models, 2nd ed. Chapman & Hall/CRC Monographs on Statistics & Applied Probability.

McDonald, K.S., Ryder, D.S., Tighe, M., 2015. Developing best-practice Bayesian Belief Networks in ecological risk assessments for freshwater and estuarine ecosystems: A quantitative review. J. Environ. Manage. 154, 190–200. https://doi.org/10.1016/j.jenvman.2015.02.031

McInerney, D., Thyer, M., Kavetski, D., Lerat, J., Kuczera, G., 2017. Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors. Water Resour. Res. 53, 2199–2239. https://doi.org/10.1002/2016WR019168

McKay, M.D., Beckman, R.J., Conover, W.J., 1979. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. Technometrics 21, 239. https://doi.org/10.2307/1268522

McKay, S.K., Theiling, C.H., Dougherty, M.P., 2019. Comparing outcomes from competing models assessing environmental flows in the Minnesota River Basin. Ecol. Eng. X 4, 100014. https://doi.org/10.1016/j.ecoena.2019.100014

McLaughlin, D.B., Reckhow, K.H., 2017. A Bayesian network assessment of macroinvertebrate responses to nutrients and other factors in streams of the Eastern Corn Belt Plains, Ohio, USA. Ecol. Modell. 345, 21–29. https://doi.org/10.1016/j.ecolmodel.2016.12.004

Mcmanamay, R.A., Bevelhimer, M.S., Kao, S.C., 2014. Updating the US hydrologic classification: An approach to clustering and stratifying ecohydrologic data. Ecohydrology 7, 903–926. https://doi.org/10.1002/eco.1410

McMillan, H.K., 2020a. Linking hydrologic signatures to hydrologic processes: A review. Hydrol. Process. 34, 1393–1409. https://doi.org/10.1002/hyp.13632

McMillan, H.K., 2020b. A review of hydrologic signatures and their applications. Wiley Interdiscip. Rev. Water 1–23. https://doi.org/10.1002/wat2.1499

Merriam, E.R., Petty, J.T., Strager, M.P., Maxwell, A.E., Ziemkiewicz, P.F., 2015. Landscape-

based cumulative effects models for predicting stream response to mountaintop mining in multistressor Appalachian watersheds. Freshw. Sci. 34, 1006–1019. https://doi.org/10.1086/681970.

Merriam, E.R., Petty, J.T., Strager, M.P., Maxwell, A.E., Ziemkiewicz, P.F., 2013. Scenario analysis predicts context-dependent stream response to landuse change in a heavily mined central Appalachian watershed. Freshw. Sci. 32, 1246–1259. https://doi.org/10.1899/13-003.1

Mitchell, T.M., 1999. Machine learning and data mining. Commun. ACM 42, 30–36.

Mittal, N., Bhave, A.G., Mishra, A., Singh, R., 2016. Impact of human intervention and climate change on natural flow regime. Water Resour. Manag. 30, 685–699. https://doi.org/10.1007/s11269-015-1185-6

Mitteroecker, P., Bookstein, F., 2011. Linear Discrimination, Ordination, and the Visualization of Selection Gradients in Modern Morphometrics. Evol. Biol. 38, 100–114. https://doi.org/10.1007/s11692-011-9109-8

Mizukami, N., Rakovec, O., Newman, A.J., Clark, M.P., Wood, A.W., Gupta, H. V., Kumar, R., 2019. On the choice of calibration metrics for "high-flow" estimation using hydrologic models. Hydrol. Earth Syst. Sci. 23, 2601–2614. https://doi.org/10.5194/hess-23-2601-2019

Moges, E., Demissie, Y., Larsen, L., Yassin, F., 2021. Review: Sources of hydrological model uncertainties and advances in their analysis. Water (Switzerland) 13, 1–23. https://doi.org/10.3390/w13010028

Moisen, G.G., Freeman, E.A., Blackard, J.A., Frescino, T.S., Zimmermann, N.E., Edwards, T.C., 2006. Predicting tree species presence and basal area in Utah: A comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. Ecol. Modell. 199, 176–187. https://doi.org/10.1016/j.ecolmodel.2006.05.021

Monteith, J.L., 1965. Evaporation and environment, in: Symposia of the Society for Experimental Biology. pp. 205–234.

Moriasi, D.., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., Veith, T.L., D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, T. L. Veith, 2007. Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. Trans. ASABE 50, 885–900. https://doi.org/10.13031/2013.23153

Moriasi, D.N., Gitau, M.W., Pai, N., Daggupati, P., 2015. Hydrologic and Water Quality Models: Performance Measures and Evaluation Criteria. Trans. ASABE 58, 1763–1785. https://doi.org/10.13031/trans.58.10715

Mouton, A.M., De Baets, B., Goethals, P.L.M., 2009. Knowledge-based versus data-driven fuzzy habitat suitability models for river management. Environ. Model. Softw. 24, 1014–1018. https://doi.org/10.1016/j.envsoft.2009.02.005

Mouton, A.M., Dedecker, A.P., Lek, S., Goethals, P.L.M., 2010. Selecting variables for habitat suitability of Asellus (Crustacea, Isopoda) by applying input variable contribution methods to artificial neural network models. Environ. Model. Assess. 15, 65–79. https://doi.org/10.1007/s10666-009-9192-8

Moya, N., Hughes, R.M., Domínguez, E., Gibon, F.M., Goitia, E., Oberdorff, T., 2011. Macroinvertebrate-based multimetric predictive models for evaluating the human impact on biotic condition of Bolivian streams. Ecol. Indic. 11, 840–847. https://doi.org/10.1016/j.ecolind.2010.10.012

Muñoz-Mas, R., Fukuda, S., Pórtoles, J., Martínez-Capel, F., 2018. Revisiting probabilistic neural networks: a comparative study with support vector machines and the microhabitat suitability for the Eastern Iberian chub (Squalius valentinus). Ecol. Inform. 43, 24–37. https://doi.org/10.1016/j.ecoinf.2017.10.008

Muñoz-Mas, R., Martínez-Capel, F., Schneider, M., Mouton, A.M., 2012. Assessment of brown trout habitat suitability in the Jucar River Basin (SPAIN): Comparison of data-driven approaches with fuzzy-logic models and univariate suitability curves. Sci. Total Environ. 440, 123–131. https://doi.org/10.1016/j.scitotenv.2012.07.074

Muñoz-Mas, R., Vezza, P., Alcaraz-Hernández, J.D., Martínez-Capel, F., 2016. Risk of invasion predicted with support vector machines: A case study on northern pike (Esox Lucius, L.) and bleak (Alburnus alburnus, L.). Ecol. Modell. 342, 123–134. https://doi.org/10.1016/j.ecolmodel.2016.10.006

Murphy, J.C., Knight, R.R., Wolfe, W.J., S. Gain, W., 2013. Predicting ecological flow regime at ungaged sites: a comparison of methods. River Res. Appl. 29, 660–669. https://doi.org/10.1002/rra.2570

Mwiya, R.M., Zhang, Z., Zheng, C., Wang, C., 2020. Comparison of approaches for irrigation scheduling using AquaCrop and NSGA-III models under climate uncertainty. Sustain. 12. https://doi.org/10.3390/su12187694

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — A discussion of principles. J. Hydrol. 10, 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Neitsch, S.L., Arnold, J.G., Kiniry, J.R., Williams, J.R., 2011. Soil and Water Assessment Tool Theoretical Documentation Version 2009 Texas Water Resources Institute. Temple, Texas.

Nelder, J.A., Baker, R.J., 2006. Generalized Linear Models. Encycl. Stat. Sci. 4.

Niemi, G.J., McDonald, M.E., 2004. Application of Ecological Indicators. Annu. Rev. Ecol. Evol. Syst. 35, 89–111. https://doi.org/10.1146/annurev.ecolsys.35.112202.130132

NOAA-NCEI, 2020. Climate Data Online Data Tools [WWW Document]. URL https://www.ncdc.noaa.gov/cdo-web/datatools/ (accessed 7.12.20).

Nojavan A., F., Qian, S.S., Stow, C.A., 2017. Comparative analysis of discretization methods in Bayesian networks. Environ. Model. Softw. 87, 64–71. https://doi.org/10.1016/j.envsoft.2016.10.007

Ocampo-Duque, W., Ferré-Huguet, N., Domingo, J.L., Schuhmacher, M., 2006. Assessing water quality in rivers with fuzzy inference systems: A case study. Environ. Int. 32, 733–742. https://doi.org/10.1016/j.envint.2006.03.009

Ocampo-Duque, W., Schuhmacher, M., Domingo, J.L., 2007. A neural-fuzzy approach to classify the ecological status in surface waters. Environ. Pollut. 148, 634–641. https://doi.org/10.1016/j.envpol.2006.11.027

Olaya-Marín, E.J., Martínez-Capel, F., Soares Costa, R.M., Alcaraz-Hernández, J.D., 2012. Modelling native fish richness to evaluate the effects of hydromorphological changes and river restoration (Júcar River Basin, Spain). Sci. Total Environ. 440, 95–105. https://doi.org/10.1016/j.scitotenv.2012.07.093

Olaya-Marín, E.J., Martínez-Capel, F., Vezza, P., 2013. A comparison of artificial neural networks and random forests to predict native fish species richness in Mediterranean rivers. Knowl. Manag. Aquat. Ecosyst. 409, 07. https://doi.org/10.1051/kmae/2013052

Olden, J.D., Jackson, D.A., 2002. Illuminating the "black box": A randomization approach for understanding variable contributions in artificial neural networks. Ecol. Modell. 154, 135–150. https://doi.org/10.1016/S0304-3800(02)00064-9

Olden, J.D., Lawler, J.J., Poff, N.L., 2008. Machine Learning Methods Without Tears: A Primer for Ecologists. Q. Rev. Biol. 83, 171–193. https://doi.org/10.1086/587826

Olden, J.D., Poff, N.L., 2003. Redundancy and the choice of hydrologic indices for characterizing streamflow regimes. River Res. Appl. 19, 101–121. https://doi.org/10.1002/rra.700

Olsen, M., Troldborg, L., Henriksen, H.J., Conallin, J., Refsgaard, J.C., Boegh, E., 2013. Evaluation of a typical hydrological model in relation to environmental flows. J. Hydrol. 507, 52–62. https://doi.org/10.1016/j.jhydrol.2013.10.022

Oudin, L., Andréassian, V., Mathevet, T., Perrin, C., Michel, C., 2006. Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations. Water Resour. Res. 42, 1–10. https://doi.org/10.1029/2005WR004636

Palmer, M., Ruhi, A., 2019. Linkages between flow regime, biota, and ecosystem processes: Implications for river restoration. Science (80-.). 365. https://doi.org/10.1126/science.aaw2087

Parker, S.R., Adams, S.K., Lammers, R.W., Stein, E.D., Bledsoe, B.P., 2019. Targeted hydrologic model calibration to improve prediction of ecologically-relevant flow metrics. J. Hydrol. 573, 546–556. https://doi.org/10.1016/j.jhydrol.2019.03.081

Patrick, C.J., Yuan, L.L., 2017. Modeled hydrologic metrics show links between hydrology and the functional composition of stream assemblages. Ecol. Appl. 27, 1605–1617. https://doi.org/10.1002/eap.1554

Pearl, J., 1986. Fusion, propagation, and structuring in belief networks. Artif. Intell. 29, 241–288. https://doi.org/10.1016/0004-3702(86)90072-X

Pérez-Miñana, E., 2016. Improving ecosystem services modelling: Insights from a Bayesian network tools review. Environ. Model. Softw. 85, 184–201. https://doi.org/10.1016/j.envsoft.2016.07.007

Peters, D.L., Baird, D.J., Monk, W.A., Armanini, D.G., 2012. Establishing Standards and Assessment Criteria for Ecological Instream Flow Needs in Agricultural Regions of Canada. J. Environ. Qual. 41, 41–51. https://doi.org/10.2134/jeq2011.0094

Peterson, D.P., Wenger, S.J., Rieman, B.E., Isaak, D.J., 2013. Linking climate change and fish conservation efforts using spatially explicit decision support tools. Fisheries 38, 112–127. https://doi.org/10.1080/03632415.2013.769157

Pfannerstill, M., Bieger, K., Guse, B., Bosch, D.D., Fohrer, N., Arnold, J.G., 2017. How to Constrain Multi-Objective Calibrations of the SWAT Model Using Water Balance Components. J. Am. Water Resour. Assoc. 53, 532–546. https://doi.org/10.1111/1752-1688.12524

Pfannerstill, M., Guse, B., Fohrer, N., 2014. Smart low flow signature metrics for an improved overall performance evaluation of hydrological models. J. Hydrol. 510, 447–458. https://doi.org/10.1016/j.jhydrol.2013.12.044

Pham, H.V., Torresan, S., Critto, A., Marcomini, A., 2019. Alteration of freshwater ecosystem services under global change – A review focusing on the Po River basin (Italy) and the Red River basin (Vietnam). Sci. Total Environ. 652, 1347–1365. https://doi.org/10.1016/j.scitotenv.2018.10.303

Phan, T.D., Smart, J.C.R., Capon, S.J., Hadwen, W.L., Sahin, O., 2016. Applications of Bayesian belief networks in water resource management: A systematic review. Environ. Model. Softw. 85, 98–111. https://doi.org/10.1016/j.envsoft.2016.08.006

Phillips, S.B., Aneja, V.P., Kang, D., Arya, S.P., 2006. Modelling and analysis of the atmospheric nitrogen deposition in North Carolina. Int. J. Glob. Environ. Issues 6, 231–252. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Pilière, A., Schipper, A.M., Breure, T.M., Posthuma, L., de Zwart, D., Dyer, S.D., Huijbregts, M.A.J., 2014. Unraveling the relationships between freshwater invertebrate assemblages and interacting environmental factors. Freshw. Sci. 33, 1148–1158. https://doi.org/10.1086/677898

Poff, N.L., Allan, J.D., Bain, M.B., Karr, J.R., Prestegaard, K.L., Richter, B.D., Sparks, R.E., Stromberg, J.C., 1997. The Natural Flow Regime. Bioscience 47, 769–784.

https://doi.org/10.2307/1313099

Poff, N.L., Richter, B.D., Arthington, A.H., Bunn, S.E., Naiman, R.J., Kendy, E., Acreman, M., Apse, C., Bledsoe, B.P., Freeman, M.C., Henriksen, J., Jacobson, R.B., Kennen, J.G., Merritt, D.M., O'Keeffe, J.H., Olden, J.D., Rogers, K., Tharme, R.E., Warner, A., 2010. The ecological limits of hydrologic alteration (ELOHA): A new framework for developing regional environmental flow standards. Freshw. Biol. 55, 147–170. https://doi.org/10.1111/j.1365-2427.2009.02204.x

Poff, N.L., Zimmerman, J.K.H., 2010. Ecological responses to altered flow regimes: A literature review to inform the science and management of environmental flows. Freshw. Biol. 55, 194–205. https://doi.org/10.1111/j.1365-2427.2009.02272.x

Pond, G.J., Krock, K.J.G., Cruz, J. V., Ettema, L.F., 2017. Effort-based predictors of headwater stream conditions: comparing the proximity of land use pressures and instream stressors on macroinvertebrate assemblages. Aquat. Sci. 79, 765–781. https://doi.org/10.1007/s00027-017-0534-3

Pont, D., Hughes, R.M., Whittier, T.R., Schmutz, S., 2009. A Predictive Index of Biotic Integrity Model for Aquatic-Vertebrate Assemblages of Western U.S. Streams. Trans. Am. Fish. Soc. 138, 292–305. https://doi.org/10.1577/T07-277.1

Pool, S., Vis, M., Seibert, J., 2018. Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency. Hydrol. Sci. J. 63, 1941–1953. https://doi.org/10.1080/02626667.2018.1552002

Pool, S., Vis, M.J.P.P., Knight, R.R., Seibert, J., 2017. Streamflow characteristics from modeled runoff time series - Importance of calibration criteria selection. Hydrol. Earth Syst. Sci. 21, 5443–5457. https://doi.org/10.5194/hess-21-5443-2017

Poor, C.J., Ullman, J.L., 2010. Using regression tree analysis to improve predictions of low-flow nitrate and chloride in Willamette river basin watersheds. Environ. Manage. 46, 771–780. https://doi.org/10.1007/s00267-010-9550-y

Pourshahabi, S., Rakhshandehroo, G., Talebbeydokhti, N., Nikoo, M.R., Masoumi, F., 2020. Handling uncertainty in optimal design of reservoir water quality monitoring systems. Environ. Pollut. 266, 115211. https://doi.org/10.1016/j.envpol.2020.115211

Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. Ecosystems 9, 181–199. https://doi.org/10.1007/s10021-005-0054-1

Price, K., Purucker, S.T., Kraemer, S.R., Babendreier, J.E., 2012. Tradeoffs among watershed model calibration targets for parameter estimation. Water Resour. Res. 48, 1–16. https://doi.org/10.1029/2012WR012005

Prieto, C., Le Vine, N., Kavetski, D., García, E., Medina, R., 2019. Flow Prediction in Ungauged Catchments Using Probabilistic Random Forests Regionalization and New Statistical

Adequacy Tests. Water Resour. Res. 55, 4364–4392. https://doi.org/10.1029/2018WR023254

Pushpalatha, R., Perrin, C., Moine, N. Le, Andréassian, V., 2012. A review of efficiency criteria suitable for evaluating low-flow simulations. J. Hydrol. 420–421, 171–182. https://doi.org/10.1016/j.jhydrol.2011.11.055

Raschke, A., Hernandez-Suarez, J.S., Nejadhashemi, A.P., Deb, K., 2021. Multidimensional Aspects of Sustainable Biofuel Feedstock Production. Sustainability 13, 1424. https://doi.org/10.3390/su13031424

Reed, P.M., Hadka, D., Herman, J.D., Kasprzyk, J.R., Kollat, J.B., 2013. Evolutionary multiobjective optimization in water resources: The past, present, and future. Adv. Water Resour. 51, 438–456. https://doi.org/10.1016/j.advwatres.2012.01.005

Reichert, P., Schuwirth, N., 2012. Linking statistical bias description to multiobjective model calibration. Water Resour. Res. 48, 1–20. https://doi.org/10.1029/2011WR011391

Richter, B.D., Baumgartner, J., Wigington, R., Braun, D., 1997. How much water does a river need? Freshw. Biol. 37, 231–249. https://doi.org/10.1046/j.1365-2427.1997.00153.x

Richter, B.D., Baumgartner, J. V, Powell, J., Braun, D.P., 1996. A method for assessing hydrologic alteration within ecosystems. Conserv. Biol. 10, 1163–1174. https://doi.org/10.2307/2387152

Richter, B.D., Mathews, R., Harrison, D.L., Wigington, R., 2003. Ecologically sustainable water management: managing river flows for ecological integrity. Ecol. Appl. 13, 206–224. https://doi.org/10.1890/1051-0761(2003)013[0206:ESWMMR]2.0.CO;2

Riseng, C., Wiley, M., Black, R., Munn, M., 2011. Impacts of agricultural land use on biological integrity: a causal analysis. Ecol. Appl. 21, 3128–3146. https://doi.org/10.1890/11-0077.1

Sadegh, M., Vrugt, J.A., 2014. Approximate Bayesian Computation using Markov Chain Monte Carlo simulation. Water Resour. Res. 10, 6767–6787. https://doi.org/10.1002/2014WR015386.Received

Sadegh, M., Vrugt, J.A., Xu, C., Volpi, E., 2015. The stationarity paradigm revisited: Hypothesis testing using diagnostics, summary metrics, and DREAM (ABC). Water Resour. Res. 51, 9207–9231. https://doi.org/10.1002/2014WR016805

Sahraei, S., Asadzadeh, M., Unduche, F., 2020. Signature-based multi-modelling and multi-objective calibration of hydrologic models: Application in flood forecasting for Canadian Prairies. J. Hydrol. 588. https://doi.org/10.1016/j.jhydrol.2020.125095

Sanborn, S.C., Bledsoe, B.P., 2006. Predicting streamflow regime metrics for ungauged streamsin Colorado, Washington, and Oregon. J. Hydrol. 325, 241–261. https://doi.org/10.1016/j.jhydrol.2005.10.018

Sauer, J., Domisch, S., Nowak, C., Haase, P., 2011. Low mountain ranges: Summit traps for montane freshwater species under climate change. Biodivers. Conserv. 20, 3133–3146. https://doi.org/10.1007/s10531-011-0140-y

Schoups, G., Vrugt, J.A., 2010. A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. Water Resour. Res. 46, 1–17. https://doi.org/10.1029/2009WR008933

Seada, H., Deb, K., 2016. A Unified Evolutionary Optimization Procedure for Single, Multiple, and Many Objectives. IEEE Trans. Evol. Comput. 20, 358–369. https://doi.org/10.1109/TEVC.2015.2459718

Sengupta, A., Adams, S.K., Bledsoe, B.P., Stein, E.D., McCune, K.S., Mazor, R.D., 2018. Tools for managing hydrologic alteration on a regional scale: Estimating changes in flow characteristics at ungauged sites. Freshw. Biol. 63, 769–785. https://doi.org/10.1111/fwb.13074

Shafii, M., De Smedt, F., 2009. Multi-objective calibration of a distributed hydrological model (WetSpa) using a genetic algorithm. Hydrol. Earth Syst. Sci. Discuss. 6, 243–271. https://doi.org/10.5194/hessd-6-243-2009

Shafii, M., Tolson, B.A., 2015. Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives. Water Resour. Res. 51, 3796–3814. https://doi.org/10.1002/2014WR016520

Shenton, W., Hart, B.T., Chan, T.U., 2014. A Bayesian network approach to support environmental flow restoration decisions in the Yarra River, Australia. Stoch. Environ. Res. Risk Assess. 28, 57–65. https://doi.org/10.1007/s00477-013-0698-x

Shrestha, R.R., Peters, D.L., Schnorbus, M.A., 2014. Evaluating the ability of a hydrologic model to replicate hydro-ecologically relevant indicators. Hydrol. Process. 28, 4294–4310. https://doi.org/10.1002/hyp.9997

Shrestha, R.R., Schnorbus, M.A., Peters, D.L., 2016. Assessment of a hydrologic model's reliability in simulating flow regime alterations in a changing climate. Hydrol. Process. 30, 2628–2643. https://doi.org/10.1002/hyp.10812

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall/CRC. https://doi.org/10.1201/9781315140919

Sindhya, K., Miettinen, K., Deb, K., 2013. A hybrid framework for evolutionary multi-objective optimization. IEEE Trans. Evol. Comput. 17, 495–511. https://doi.org/10.1109/TEVC.2012.2204403

Smith, T., Marshall, L., Sharma, A., 2015. Modeling residual hydrologic errors with Bayesian inference. J. Hydrol. 528, 29–37. https://doi.org/10.1016/j.jhydrol.2015.05.051

Smucker, N.J., Becker, M., Detenbeck, N.E., Morrison, A.C., 2013. Using algal metrics and

biomass to evaluate multiple ways of defining concentration-based nutrient criteria in streams and their ecological relevance. Ecol. Indic. 32, 51–61. https://doi.org/10.1016/j.ecolind.2013.03.018

Snelder, T.H., J. Booker, D., 2013. Natural flow regime classifications are sensitive to definition procedures. River Res. Appl. 29, 822–838. https://doi.org/10.1002/rra.2581

Sofi, M.S., Bhat, S.U., Rashid, I., Kuniyal, J.C., 2020. The natural flow regime: A master variable for maintaining river ecosystem health. Ecohydrology 1–12. https://doi.org/10.1002/eco.2247

Sor, R., Park, Y., Boets, P., Goethals, P.L.M., Lek, S., 2017. Effects of species prevalence on the performance of predictive models. Ecol. Modell. 354, 11–19. https://doi.org/10.1016/j.ecolmodel.2017.03.006

Sowa, S.P., Herbert, M., Mysorekar, S., Annis, G.M., Hall, K., Nejadhashemi, A.P., Woznicki, S.A., Wang, L., Doran, P.J., 2016. How much conservation is enough? Defining implementation goals for healthy fish communities in agricultural rivers. J. Great Lakes Res. 42, 1302–1321. https://doi.org/10.1016/j.jglr.2016.09.011

Steel, A.E., Peek, R.A., Lusardi, R.A., Yarnell, S.M., 2017. Associating metrics of hydrologic variability with benthic macroinvertebrate communities in regulated and unregulated snowmelt-dominated rivers. Freshw. Biol. 1–15. https://doi.org/10.1111/fwb.12994

Steel, E.A., Hughes, R.M., Fullerton, A.H., Schmutz, S., Young, J.A., Fukushima, M., Muhar, S., Poppe, M., Feist, B.E., Trautwein, C., 2010. Are we meeting the challenges of landscape-scale riverine research? A review. Living Rev. Landsc. Res. 4, 1–60. https://doi.org/10.12942/lrlr-2010-1

Stockwell, D., Peters, D., 1999. The GARP modelling system: problems and solutions to automated spatial prediction. Int. J. Geogr. Inf. Sci. 13, 143–158. https://doi.org/10.1080/136588199241391

Strayer, D.L., Dudgeon, D., 2010. Freshwater biodiversity conservation: recent progress and future challenges. J. North Am. Benthol. Soc. 29, 344–358. https://doi.org/10.1899/08-171.1

Sui, P., Iwasaki, A., Saavedra V., O.C., Yoshimura, C., 2014. Modelling basin-scale distribution of fish occurrence probability for assessment of flow and habitat conditions in rivers. Hydrol. Sci. J. 59, 618–628. https://doi.org/10.1080/02626667.2013.827791

Surridge, B.W.J., Bizzi, S., Castelletti, A., 2014. A framework for coupling explanation and prediction in hydroecological modelling. Environ. Model. Softw. 61, 274–286. https://doi.org/10.1016/j.envsoft.2014.02.012

Sutela, T., Vehanen, T., Jounela, P., 2010. Response of fish assemblages to water quality in boreal rivers. Hydrobiologia 641, 1–10. https://doi.org/10.1007/s10750-009-0048-7

Tang, Y., Marshall, L., Sharma, A., Ajami, H., 2018. A Bayesian alternative for multi-objective ecohydrological model specification. J. Hydrol. 556, 25–38. https://doi.org/10.1016/j.jhydrol.2017.07.040

Tasdighi, A., Arabi, M., Harmel, D., 2018. A probabilistic appraisal of rainfall-runoff modeling approaches within SWAT in mixed land use watersheds. J. Hydrol. 564, 476–489. https://doi.org/10.1016/j.jhydrol.2018.07.035

Ter Braak, C.J.F., Vrugt, J.A., 2008. Differential Evolution Markov Chain with snooker updater and fewer chains. Stat. Comput. 18, 435–446. https://doi.org/10.1007/s11222-008-9104-9

The Nature Conservancy, 2009. Indicators of Hydrologic Alteration Version 7.1 User's Manual, The Nature Conservancy.

Tonkin, J.D., Stoll, S., Sundermann, A., Haase, P., 2014. Dispersal distance and the pool of taxa, but not barriers, determine the colonisation of restored river reaches by benthic invertebrates. Freshw. Biol. 59, 1843–1855. https://doi.org/10.1111/fwb.12387

Tsai, W.P., Chang, F.J., Herricks, E.E., 2016. Exploring the ecological response of fish to flow regime by soft computing techniques. Ecol. Eng. 87, 9–19. https://doi.org/10.1016/j.ecoleng.2015.11.015

Tucker, A., Duplisea, D., 2012. Bioinformatics tools in predictive ecology: applications to fisheries. Philos. Trans. R. Soc. B Biol. Sci. 367, 279–290. https://doi.org/10.1098/rstb.2011.0184

Turschwell, M.P., Stewart-Koster, B., Leigh, C., Peterson, E.E., Sheldon, F., Balcombe, S.R., 2017. Riparian restoration offsets predicted population consequences of climate warming in a threatened headwater fish. Aquat. Conserv. Mar. Freshw. Ecosyst. 1–12. https://doi.org/10.1002/aqc.2864

Tuulaikhuu, B.-A.. b, Guasch, H.., García-Berthou, E.., 2017. Examining predictors of chemical toxicity in freshwater fish using the random forest technique. Environ. Sci. Pollut. Res. 24, 10172–10181. https://doi.org/10.1007/s11356-017-8667-4

US EPA, 2011. A Primer on Using Biological Assessments to Support Water Quality Management. EPA 810-R-11-01. https://doi.org/10.1007/s13398-014-0173-7.2

USDA-NASS, 2012. CropScape - NASS CDL Program [WWW Document]. URL https://nassgeodata.gmu.edu/CropScape/ (accessed 8.5.18).

USDA-NRCS, 2020. Web Soil Survey [WWW Document]. URL https://websoilsurvey.sc.egov.usda.gov/App/WebSoilSurvey.aspx (accessed 7.12.20).

USDA-SCS, 1972. National engineering handbook, section 4: Hydrology. Washington, DC.

USEPA, 2015. Saginaw River and Bay Area of Concern [WWW Document]. URL https://www.epa.gov/saginaw-river-bay-aoc (accessed 7.12.17).

USGS, 2020. National Water Information System: Web Interface [WWW Document]. URL https://waterdata.usgs.gov/nwis (accessed 7.12.20).

USGS, 2018. The National Map: Elevation [WWW Document]. URL https://nationalmap.gov/elevation.html (accessed 10.5.18).

Van Broekhoven, E., Adriaenssens, V., De Baets, B., Verdonschot, P.F.M., 2006. Fuzzy rule-based macroinvertebrate habitat suitability models for running waters. Ecol. Modell. 198, 71–84. https://doi.org/10.1016/j.ecolmodel.2006.04.006

Van Echelpoel, W., Boets, P., Landuyt, D., Gobeyn, S., Everaert, G., Bennetsen, E., Mouton, A., Goethals, P.L.M.M., Echelpoel, W. Van, Boets, P., Landuyt, D., Gobeyn, S., Everaert, G., Bennetsen, E., Mouton, A., Goethals, P.L.M.M., Van Echelpoel, W., Boets, P., Landuyt, D., Gobeyn, S., Everaert, G., Bennetsen, E., Mouton, A., Goethals, P.L.M.M., 2015. Species distribution models for sustainable ecosystem management, 1st ed, Developments in Environmental Modelling. Elsevier B.V. https://doi.org/10.1016/B978-0-444-63536-5.00008-9

Van Sickle, J., Baker, J., Herlihy, A., Bayley, P., Gregory, S., Haggerty, P., Ashkenas, L., Li, J., 2004. Projecting the biological condition of streams under alternative scenarios of human land use. Ecol. Appl. 14, 368–380. https://doi.org/10.1890/02-5009

Van Sickle, J., Burch Johnson, C., 2008. Parametric distance weighting of landscape influence on streams. Landsc. Ecol. 23, 427–438. https://doi.org/10.1007/s10980-008-9200-4

van Werkhoven, K., Wagener, T., Reed, P., Tang, Y., 2009. Sensitivity-guided reduction of parametric dimensionality for multi-objective calibration of watershed models. Adv. Water Resour. 32, 1154–1169. https://doi.org/10.1016/j.advwatres.2009.03.002

Vander Laan, J.J., Hawkins, C.P., Olson, J.R., Hill, R.A., 2013. Linking land use, in-stream stressors, and biological condition to infer causes of regional ecological impairment in streams. Freshw. Sci. 32, 801–820. https://doi.org/10.1899/12-186.1

Vapnik, V.N., 2000. The Nature of Statistical Learning Theory, Second. ed, Farming for Health: Green-care farming across Europe and the United States of America, Graduate Texts in Contemporary Physics. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4757-3264-1

Vezza, P., Muñoz-Mas, R., Martinez-Capel, F., Mouton, A., 2015. Random forests to evaluate biotic interactions in fish distribution models. Environ. Model. Softw. 67, 173–183. https://doi.org/10.1016/j.envsoft.2015.01.005

Vigiak, O., Lutz, S., Mentzafou, A., Chiogna, G., Tuo, Y., Majone, B., Beck, H., de Roo, A., Malagó, A., Bouraoui, F., Kumar, R., Samaniego, L., Merz, R., Gamvroudis, C., Skoulikidis, N., Nikolaidis, N.P., Bellin, A., Acuňa, V., Mori, N., Ludwig, R., Pistocchi, A., 2018. Uncertainty of modelled flow regime for flow-ecological assessment in Southern Europe. Sci. Total Environ. 615, 1028–1047. https://doi.org/10.1016/j.scitotenv.2017.09.295

Vilizzi, L., Price, A., Beesley, L., Gawne, B., King, A.J., Koehn, J.D., Meredith, S.N., Nielsen, D.L., 2013. Model development of a Bayesian Belief Network for managing inundation events for wetland fish. Environ. Model. Softw. 41, 1–14. https://doi.org/10.1016/j.envsoft.2012.11.004

Villeneuve, B., Piffady, J., Valette, L., Souchon, Y., Usseglio-Polatera, P., 2018. Direct and indirect effects of multiple stressors on stream invertebrates across watershed, reach and site scales: A structural equation modelling better informing on hydromorphological impacts. Sci. Total Environ. 612, 660–671. https://doi.org/10.1016/j.scitotenv.2017.08.197

Villeneuve, B., Souchon, Y., Usseglio-Polatera, P., Ferréol, M., Valette, L., 2015. Can we predict biological condition of stream ecosystems? A multi-stressors approach linking three biological indices to physico-chemistry, hydromorphology and land use. Ecol. Indic. 48, 88–98. https://doi.org/10.1016/j.ecolind.2014.07.016

Vis, M., Knight, R., Pool, S., Wolfe, W., Seibert, J., 2015. Model calibration criteria for estimating ecological flow characteristics. Water (Switzerland) 7, 2358–2381. https://doi.org/10.3390/w7052358

Vogel, R.M., Fennessey, N.M., 1994. Flow-Duration Curves. I: New Interpretation and Confidence Intervals. J. Water Resour. Plan. Manag. 120, 485–504. https://doi.org/10.1061/(ASCE)0733-9496(1994)120:4(485)

Vogel, R.M., Sieber, J., Archfield, S.A., Smith, M.P., Apse, C.D., Huber-Lee, A., 2007. Relations among storage, yield, and instream flow. Water Resour. Res. 43, 1–12. https://doi.org/10.1029/2006WR005226

Vörösmarty, C.J., McIntyre, P.B., Gessner, M.O., Dudgeon, D., Prusevich, A., Green, P., Glidden, S., Bunn, S.E., Sullivan, C.A., Liermann, C.R., Davies, P.M., 2010. Global threats to human water security and river biodiversity. Nature 467, 555–561. https://doi.org/10.1038/nature09440

Vrugt, J.A., 2016. Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. Environ. Model. Softw. 75, 273–316. https://doi.org/10.1016/j.envsoft.2015.08.013

Vrugt, J.A., Beven, K.J., 2018. Embracing equifinality with efficiency: Limits of Acceptability sampling using the DREAM(LOA) algorithm. J. Hydrol. 559, 954–971. https://doi.org/10.1016/j.jhydrol.2018.02.026

Vrugt, J.A., Ter Braak, C.J.F., Diks, C.G.H., Robinson, B.A., Hyman, J.M., Higdon, D., 2009. Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. Int. J. Nonlinear Sci. Numer. Simul. 10, 273–290. https://doi.org/10.1515/IJNSNS.2009.10.3.273

Wagenhoff, A., Liess, A., Pastor, A., Clapcott, J.E., Goodwin, E.O., Young, R.G., 2016. Thresholds in ecosystem structural and functional responses to agricultural stressors can inform limit setting in streams. Freshw. Sci. 36, 000–000. https://doi.org/10.1086/690233

Waite, I.R., Brown, L.R., Kennen, J.G., May, J.T., Cuffney, T.F., Orlando, J.L., Jones, K.A., 2010. Comparison of watershed disturbance predictive models for stream benthic macroinvertebrates for three distinct ecoregions in western US. Ecol. Indic. 10, 1125–1136. https://doi.org/10.1016/j.ecolind.2010.03.011

Waite, I.R., Kennen, J.G., May, J.T., Brown, L.R., Cuffney, T.F., Jones, K.A., Orlando, J.L., 2014. Stream macroinvertebrate response models for bioassessment metrics: Addressing the issue of spatial scale. PLoS One 9. https://doi.org/10.1371/journal.pone.0090944

Waite, I.R., Kennen, J.G., May, J.T., Brown, L.R., Cuffney, T.F., Jones, K.A., Orlando, J.L., 2012. Comparison of Stream Invertebrate Response Models for Bioassessment Metrics. J. Am. Water Resour. Assoc. 48, 570–583. https://doi.org/10.1111/j.1752-1688.2011.00632.x

Waite, I.R., Van Metre, P.C., 2017. Multistressor predictive models of invertebrate condition in the Corn Belt, USA. Freshw. Sci. 36, 000–000. https://doi.org/10.1086/694894

Waldron, A., Miller, D.C., Redding, D., Mooers, A., Kuhn, T.S., Nibbelink, N., Roberts, J.T., Tobias, J.A., Gittleman, J.L., 2017. Reductions in global biodiversity loss predicted from conservation spending. Nature 551, 364–367. https://doi.org/10.1038/nature24295

Walker, K.F., Sheldon, F., Puckridge, J.T., 1995. A perspective on dryland river ecosystems. Regul. Rivers Res. Manag. 11, 85–104. https://doi.org/10.1002/rrr.3450110108

Walters, D.M., Roy, A.H., Leigh, D.S., 2009. Environmental indicators of macroinvertebrate and fish assemblage integrity in urbanizing watersheds. Ecol. Indic. 9, 1222–1233. https://doi.org/10.1016/j.ecolind.2009.02.011

Wang, L., Robertson, D.M., Garrison, P.J., 2007. Linkages between nutrients and assemblages of macroinvertebrates and fish in wadeable streams: Implication to nutrient criteria development. Environ. Manage. 39, 194–212. https://doi.org/10.1007/s00267-006-0135-8

Wenger, S.J., Luce, C.H., Hamlet, A.F., Isaak, D.J., Neville, H.M., 2010. Macroscale hydrologic modeling of ecologically relevant flow metrics. Water Resour. Res. 46, 1–10. https://doi.org/10.1029/2009WR008839

Westerberg, I.K., McMillan, H.K., 2015. Uncertainty in hydrological signatures. Hydrol. Earth Syst. Sci. 19, 3951–3968. https://doi.org/10.5194/hess-19-3951-2015

Westerberg, I.K., Wagener, T., Coxon, G., McMillan, H.K., Castellarin, A., Montanari, A., Freer, J., 2016. Uncertainty in hydrological signatures for gauged and ungauged catchments. Water Resour. Res. 52, 1847–1865. https://doi.org/10.1002/2015WR017635

While, L., Bradstreet, L., Barone, L., 2016. Walking Fish Group: Hypervolume Project [WWW Document]. URL http://www.wfg.csse.uwa.edu.au/hypervolume/ (accessed 10.10.17).

While, L., Bradstreet, L., Barone, L., 2012. A fast way of calculating exact hypervolumes. IEEE Trans. Evol. Comput. 16, 86–95. https://doi.org/10.1109/TEVC.2010.2077298

Williams, J.R., 1969. Flood Routing With Variable Travel Time or Variable Storage Coefficients. Trans. ASAE 12, 100–103. https://doi.org/10.13031/2013.38772

Willmott, C.J., 1981. On the validation of models. Phys. Geogr. 2, 184–194. https://doi.org/10.1080/02723646.1981.10642213

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: A basic tool of chemometrics. Chemom. Intell. Lab. Syst. 58, 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

Woznicki, S.A., Nejadhashemi, A.P., Abouali, M., Herman, M.R., Esfahanian, E., Hamaamin, Y.A., Zhang, Z., 2016a. Ecohydrological modeling for large-scale environmental impact assessment. Sci. Total Environ. 543, 274–286. https://doi.org/10.1016/j.scitotenv.2015.11.044

Woznicki, S.A., Nejadhashemi, A.P., Ross, D.M., Zhang, Z., Wang, L., Esfahanian, A.-H.H., 2015. Ecohydrological model parameter selection for stream health evaluation. Sci. Total Environ. 511, 341–353. https://doi.org/10.1016/j.scitotenv.2014.12.066

Woznicki, S.A., Nejadhashemi, A.P., Tang, Y., Wang, L., 2016b. Large-scale climate change vulnerability assessment of stream health. Ecol. Indic. 69, 578–594. https://doi.org/10.1016/j.ecolind.2016.04.002

WWAP-UN, 2017. The United Nations World Water Development Report 2017, Wastewater: The Untapped Resource. UNESCO, Paris.

Xiong, M., Liu, P., Cheng, L., Deng, C., Gui, Z., Zhang, X., Liu, Y., 2019. Identifying time-varying hydrological model parameters to improve simulation efficiency by the ensemble Kalman filter: A joint assimilation of streamflow and actual evapotranspiration. J. Hydrol. 568, 758–768. https://doi.org/10.1016/j.jhydrol.2018.11.038

Yang, H.C., Suen, J.P., Chou, S.K., 2016. Estimating the Ungauged Natural Flow Regimes for Environmental Flow Management. Water Resour. Manag. 30, 4571–4584. https://doi.org/10.1007/s11269-016-1437-0

Yi, Y., Cheng, X., Yang, Z., Wieprecht, S., Zhang, S., Wu, Y., 2017. Evaluating the ecological influence of hydraulic projects: A review of aquatic habitat suitability models. Renew. Sustain. Energy Rev. 68, 748–762. https://doi.org/10.1016/j.rser.2016.09.138

Yilmaz, K.K., Gupta, H. V., Wagener, T., 2008. A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. Water Resour. Res. 44, 1–18. https://doi.org/10.1029/2007WR006716

You, G.J.Y., Thum, B.H., Lin, F.H., 2014. The examination of reproducibility in hydro-ecological characteristics by daily synthetic flow models. J. Hydrol. 511, 904–919. https://doi.org/10.1016/j.jhydrol.2014.02.047

Zadeh, L. a., 1994. Soft computing and fuzzy logic. Software, IEEE 48–56. https://doi.org/10.1109/52.329401

Zadeh, L.A., 1998. Roles of Soft Computing and Fuzzy Logic in the Conception, Design and Deployment of Information/Intelligent Systems, in: Kaynak, O., Zadeh, L.A., Türk\csen, B., Rudas, I.J. (Eds.), Computational Intelligence: Soft Computing and Fuzzy-Neuro Integration with Applications. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–9.

Zeleny, M., 2011. Multiple Criteria Decision Making (MCDM): From Paradigm Lost to Paradigm Regained? J. Multi-Criteria Decis. Anal. 18, 77–89. https://doi.org/10.1002/mcda.473

Zhang, Y., Shao, Q., Zhang, S., Zhai, X., She, D., 2016. Multi-metric calibration of hydrological model to capture overall flow regimes. J. Hydrol. 539, 525–538. https://doi.org/10.1016/j.jhydrol.2016.05.053

Zhao, J., Cao, J., Tian, S., Chen, Y., Zhang, S., Wang, Z., Zhou, X., 2014. A comparison between two GAM models in quantifying relationships of environmental variables with fish richness and diversity indices. Aquat. Ecol. 48, 297–312. https://doi.org/10.1007/s10452-014-9484-1

Zuur, A.F., Ieno, E.N., Smith, G.M., 2007. Analysing Ecological Data, Statistics for Biology and Health. Springer New York, New York, NY, NY. https://doi.org/10.1007/978-0-387-45972-1

Zuur, A.F., Leno, E.N., Walker, N.J., Saveliev, A.A., Smith, G.M., Ieno, E.N., Walker, N.J., Saveliev, A.A., Smith, G.M., 2009. Mixed effects models and extensions in ecology with R, Public Health, Statistics for Biology and Health. Springer New York, New York, NY. https://doi.org/10.1007/978-0-387-87458-6