April 15, 2016

Dr. Dominic Mazvimavi
Department of Earth Sciences
Institute for Water Studies
University of the Western Cape
Bellville 7535
Republic of South Africa

Dear Dr. Mazvimavi,

We are submitting a revised manuscript for HESS-2015-413, originally entitled "Empirical streamflow simulation for water resource management in data-scarce seasonal watersheds" but now with the revised title (in response to reviewer comments) of "Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability and uncertainty in seasonal watersheds." This revision addresses the comments provided by the two referees in the most recent iteration of review. Additionally, we've summarized how we responded to the specific comments in the text below, and attached a marked up version of the manuscript in track changes. We thank you for your time and consideration of our manuscript, and look forward to your decision.

Sincerely,

Julie Shortridge
Johns Hopkins University

Reviewer 1:

1. I do not see a change to Figure 3 that demonstrates greater inter-annual variability.
We respectfully disagree with the reviewer that Figure 3 does not show inter-annual variability; for instance, the observed peak wet season flow in 1986 was approximately 30 CMS, while the peak wet season flow in 1990 was almost twice that at approximately 60 CMS. While this variability may be relatively minor compared to seasonal variability in the flow regime, we think that the figure in its current form provides an accurate representation of both seasonal and inter-annual variability, as well as the ability of different model formulations in capturing this variability.

2. The rainfall intensity calculation is simplistic and that could affect its value as an input. It would be better just to say that intensity data are not available.
The text referring rainfall intensity has been removed from Section 2.2.

3. I still do not think that models that respond to higher rainfall with higher runoff (or similar statements) can be considered physical realism and I would like to see such comments changed. If a model does not respond in a sensible way that is just related to really bad

model structure. The statements in the paper represent a very broad interpretation of physical realism and perhaps this should be made a bit clearer.
The references to "physical realism" have been removed from the second and third paragraphs of discussion section, as well as the conclusion section. Instead, we just refer to models as performing in a "reasonable" fashion.

Reviewer 2:

The authors have addressed most of the comments, so I believe that the manuscript should be considered for publication as long as the following minor aspects are first taken into account.

1. Title. The manuscript describes a hydrological modelling problem (i.e., fitting models, studying interpretability, bias and uncertainty), not a water resources management one. Water management entails the use of decision-making techniques to influence the way water resources are operated or planned in a given basin. None of these techniques is used in the study. I understand that the authors have a different opinion, but I believe that the title is misleading. I would leave this to the editor.
We propose the revised title "Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability and uncertainty in seasonal watersheds"

2. Line 24-27, page 5. The study is not only about comparing six models in terms of their predictive ability, as clearly explained in the last paragraph of the introduction. The authors may want to anticipate this important point.
This text has been revised, removing the reference to predictive ability so it focuses only on the case study location.

3. Line 12, page 9. Please correct the typo 'of of'.
This has been corrected.

4. Line 15-16, page 9. Please provide the technical details of the log transformation (see reviewer #1, comment no. 8. How are the streamflow data distributed?
The distributions of streamflow data, log-transformed streamflow, and streamflow anomalies in each basin are shown in supplementary Figure S1.

5. Line 25-28, page 9. More technical details are needed here. How are the 'anomaly themselves calculated based on climatic and land cover conditions that are non-stationary through time?
Text has been added elaborating on the flow calculations in the anomaly model formulation. The anomaly value for a given month in the time series (for instance, June 1990) are calculated based on temperature, precipitation, and land cover values for that specific month. The long-term average and standard deviation values (estimated from all observed June flows from 1961 to 2004) are only used to convert this anomaly value back into a flow value.

6. Figure 3 illustrates the predicted streamflow. Which model (of the six proposed) issues these predictions? It does not emerge from the caption. The same comment applies to Table 4.
   The model types have been added to Table 4 and to the caption of Figure 3.

Machine learning methods for e~~E~~mpirical streamflow simulation: a comparison of model accuracy, interpretability and uncertainty in ~~for water resource management in data-scarce~~ seasonal watersheds

J. E. Shortridge,[1] S. D. Guikema,[2] and B. F. Zaitchik[3]

[1]{Department of Geography and Environmental Engineering, Johns Hopkins University, Baltimore, USA}

[2]{Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, USA}

[3]{Department of Earth and Planetary Sciences, Johns Hopkins University, Baltimore, USA}

Correspondence to: J. E. Shortridge (jshortridge@jhu.edu)

Abstract

In the past decade, machine-learning methods for empirical rainfall-runoff modeling have seen extensive development and been proposed as a useful complement to physical hydrologic models, particularly in basins where data to support process-based models are limited. However, the majority of research has focused on a small number of methods, such as artificial neural networks, despite the development of multiple other approaches for non-parametric regression in recent years. Furthermore, this work has often evaluated model performance based on predictive accuracy alone, while not considering broader objectives such as model interpretability and uncertainty that are important if such methods are to be used for planning and management decisions. In this paper, we use multiple regression and machine-learning approaches (including generalized additive models, multivariate adaptive regression splines, artificial neural networks, random forests, and M5 cubist models) to simulate monthly streamflow in five highly-seasonal rivers in the highlands of Ethiopia and compare their performance in terms of predictive accuracy, error structure and bias, model interpretability, and uncertainty when faced with extreme climate conditions. While the relative predictive performance of models differed across basins, data-driven approaches were able to achieve reduced errors when compared to physical models developed for the region. Methods such as random forests and generalized additive models may have advantages in terms of visualization and interpretation of model structure, which can be useful in providing insights into physical watershed function. However, the uncertainty associated with model predictions under extreme climate conditions should be carefully evaluated, since certain models (especially generalized additive models and multivariate adaptive regression splines) become highly variable when faced with high temperatures.

1 Introduction

Hydrologists and water managers have made use of observed relationships between rainfall and runoff to predict streamflow ever since the creation of the rational method in the 19th century (Beven, 2011). However, the development of increasingly sophisticated machine learning techniques, combined with rapid increases in computational ability, has prompted extensive research into advanced methods for data-driven streamflow prediction in the past decade. Artificial neural networks (ANNs), regression trees, and support vector machines have been shown to be powerful tools for predictive modeling and exploratory data analysis, particularly in systems that exhibit complex, non-linear behavior (Solomatine and Ostfield, 2008; Abrahard and See, 2007).

While distributed physical models that accurately represent hydrologic processes can still be considered the gold standard for rainfall runoff modeling, empirical models can be a useful tool in contexts where there is limited data on physical watershed processes but long time-series of precipitation and streamflow (Iorgulescu and Beven, 2004). The development of historical data centers and more recent efforts to merge satellite data with *in situ* observations to monitor climate and hydrology has made acceptable climate and streamflow data more widely available in data poor regions. Because obtaining measurement-based estimates of soil hydraulic parameters or details on hydrologically-relevant land management activities can be more difficult, empirical models may be particularly useful in these locations.While many criticize these approaches as "black boxes" with no relationship to underlying physical processes (See et al., 2007), a number of studies have demonstrated how empirical approaches can be used to gain insights about physical system function (e.g., Han et al., 2007; Galelli and Castelletti, 2013a). Additionally, improvements in interpretation and visualization methods can make complex models more easily interpretable (Sudheer and Jain, 2004; Jain et al., 2004). Finally, data-driven models can be useful in identifying situations where observed data disagree with what would be predicted based on conceptual models, and thus identify assumptions regarding runoff generation processes that may be incorrect (Beven 2011).

While there have been some applications of alternative machine learning methods, such as support vector machines (Asefa et al., 2006; Lin et al., 2006) and regression-tree based approaches (Iorgulescu and Beven, 2004; Galelli and Castelletti, 2013a) for streamflow simulation, the vast majority of research has focused on artificial neural networks (Solomatine and Ostfield, 2008). While they have demonstrated impressive predictive accuracy in a number of different contexts, excessive parameterization of ANNs can result in overfit models that are not generalizable to unseen data (Iorgulescu and Beven, 2004; Gaume and Gosset, 2003). While methods exist to avoid overfitting, such as cross validation and bootstrapping, these methods are not always employed (Solomatine and Ostfield, 2008). A review by Maier et al. (2010) found that relatively few studies evaluated model performance based on parameters such as Akaike information criterion that would lead to parsimonious models that are likely to be more generalizable and interpretable. This can lead to complex models that only result in modest improvements (or no improvements at all) over much simpler approaches (Gaume and Gosset, 2003; Han et al., 2007).

Even outside of a hydrology context, it has been argued that ANNs are better suited for problems aimed at prediction without any need for model interpretation, rather than those where understanding the process generating predictions and the role of input variables is important (Hastie et al., 2009). Given the importance that this interpretation plays in understanding the contexts in which a hydrologic model is appropriate and reliable, the strong opinions surrounding the use of ANNs for water resources management are perhaps not surprising. To address this issue, a number of studies have focused on highlighting the structure and mechanism by which machine learning models make predictions to confirm their physical realism and gain insight into physical watershed function. For example, some studies have demonstrated how internal ANN structure corresponds to physical hydrologic processes (Wilby et al. 2003; Jain et al., 2004; Sudheer and Jain, 2004), while others have shown how variable selection and importance can be used to gain insights about model structure and runoff generating processes (Galelli and Castelletti, 2013a and 2013b). While these studies demonstrate that a number of methods exist for characterizing model structure, they generally focus on a single model type and thus provide little insight into the comparative ease with which different model types can be interpreted.

While a number of comparison studies exist that apply multiple empirical models to a given problem, finding generalizable insights from these studies is hindered because of the limited number of models and datasets evaluated. Perhaps the most comprehensive comparison to date is that of Elshorbagy et al. (2010a and 2010b), who compared six methods for data-driven modeling of daily discharge in the Ourthe River in Belgium. This work found that linear models were able to perform comparably to much more complex methods when the data content of the models were limited, or when system input-output behavior was close to linear. However, other studies have demonstrated the value of using more complex approaches when modeling more complex rainfall-runoff behavior (e.g., Abrahart and See, 2007; Asefa et al., 2006). The differing results obtained across these studies indicate that no single method is likely to be suitable for all basins, timescales, or applications.

However, it is important to recognize that predictive accuracy alone is not necessarily sufficient justification for applying a model to a given problem. Models should not only be accurate, but also be fit-for-purpose (Beven, 2011; Van Griensven et al., 2012). For instance, accurate representation of low return period flows is more important in a flood forecasting model than one aimed at predicting average amounts of water available for withdrawal and human consumption. Similarly, the ability to provide insights into physical watershed function may be more important in basins where land-use change could alter the hydrologic regime, compared to a basin that is heavily urbanized and expected to remain so. The use of multiple objective functions in training data-driven models can address this to some degree by identifying models that provide sufficient balance between different performance objectives, such as accurate representation of different portions of the flow hydrograph (De Vos and Rientjes, 2008). However, more refined model training procedures will not necessarily address other aspects of model performance that make it suitable for planning purposes, such as interpretability (Solomatine and Ostfield, 2008). More comprehensive consideration of model strengths and limitations should be standard practice in model development and selection, rather than simply evaluating global error metrics.

In this work, we compare six methods for empirical streamflow ~~prediction~~ simulation (linear models, generalized additive models, multivariate adaptive regression splines, random forests, M5 model trees and ANNs) ~~in their ability to predict monthly streamflow~~ in five rivers in the Lake Tana basin in Ethiopia. This study region was selected as it provides insights into the use of data-driven models for streamflow simulation in tropical regions of the world that are underrepresented in existing studies; for instance, a review of 210 articles on water resource applications of ANNs found that over three quarters of the studies evaluated were conducted in North America, Europe, Australia, or temperate East Asia (Maier et al., 2010). Existing studies conducted in tropical regions generally apply a single methodology to the basin of interest and evaluate predictive accuracy alone (see for instance, Machado et al., 2011; Chibanga et al., 2003; Antar et al., 2006; Aqil et al., 2007), making it difficult to find generalizable insights into the relative advantages of different modeling approaches in these regions. Better development of data-driven models for these regions has the potential to be particularly valuable because data limitations and complex hydrodynamic processes often hinder the use of physical watershed models, but relatively long time series of streamflow, precipitation and temperature may be available at a monthly timescale. These data, combined with information on relevant landscape change (in particular, the expansion of agricultural land cover), can be leveraged to create reasonably accurate empirical models.

Models are compared not only in terms of their predictive accuracy, but also in terms of model error structure and the implications that this structure may have for water resource applications. Additionally, we evaluate the methods by which model structure and predictor variable influence can be evaluated to gain insights into physical system function for each model type. Finally, we assess the suitability of using different model types for climate change impact assessment by comparing model uncertainty in projections made for increasingly extreme climate conditions. The overall objective of this research is not to identify a single "best" model, but rather to highlight some of the strengths and limitations of different approaches, as well as demonstrate important issues that should be kept in mind for model comparisons in the future

2 Data and Methods

2.1 Study Area

Lake Tana is located at an elevation of approximately 1800 meters in the highlands of northwest Ethiopia (Fig. 1). The catchment draining to the lake encompasses approximately 12,000 square kilometers, and the four main tributaries providing water to the lake are the Gilgel Abbay (including its tributary, the Koga River), Ribb, Gumara, and Megech Rivers. Collectively, these rivers account for 93% of the inflow to the lake (Alemayehu et al., 2010). Ninety percent of rainfall in the basin occurs during the wet season from May until October, and there is significant interannual variability in precipitation with annual rainfall levels ranging from below 1000 mm to over 1800 mm (Achenef et al., 2013). Population growth and expansion of agricultural and pastoral land use in the region has resulted in substantial deforestation and land degradation, with agricultural, pastoral and settled land cover comprising over 70% of the basin's surface area (Rientjes et al., 2011; Garede and Minale, 2014; Gebrehiwot et al., 2010). There is some evidence that this has impacted the hydrology of the rivers draining into the lake (Gebrehiwot et al., 2010). A summary of basin characteristics for the evaluation period of 1960-2004 is presented in Table 1.

Approximately 2.6 million people live in the basin, and are largely settled in rural areas and reliant on rainfed subsistence agriculture. This makes the region quite vulnerable to climate variability and change, and a number of water resources infrastructure projects are planned to better manage this vulnerability and support economic development (Alemayehu et al., 2010). This includes the recent construction of the Tana-Beles hydropower transfer tunnel and the Koga River irrigation reservoir, as well as five other reservoirs planned for construction in the next 10 to 20 years (Alemayehu et al., 2010). To better understand the potential implications of this development, extensive effort has been put towards developing rainfall-runoff models for the Lake Tana basin, as well as other areas of the Ethiopian highlands with similar characteristics (van Griensven et al., 2012). Many of these studies rely on Soil and Water Assessment Tool (SWAT) models, although there are some that use water balance approaches (Van Griensven et al., 2012). While these models have in some cases demonstrated reasonably high accuracy, previous evaluations were largely based on Nash-Sutcliffe Efficiency (NSE; Nash and Sutcliffe, 1970) which can be a flawed performance metric in highly seasonal watersheds (Schaefli and Gupta, 2007; Legates and McCabe, 1999). More importantly, the limited data available for physical parameterization of these models required a heavy reliance on model calibration, which sometimes resulted in parameterization schemes that are inconsistent with physical understanding of the region's hydrology (Steenhuis et al., 2009; van Griensven et al., 2012). Furthermore, a number of studies relied on empirical relationships such as curve numbers and the Hargreave's equation that were developed for temperate regions (e.g., Mekonnen et al., 2009; Setegne et al., 2009). While these limitations are likely to introduce considerable uncertainty into model projections, particularly in situations where climatic or environmental conditions differ from those experienced in the calibration period, few studies from this region of Ethiopia include any sort of uncertainty analysis in model predictions. Empirical models could provide a useful complement to physical models developed for the region by providing insights into physical system function and allowing for more comprehensive uncertainty analysis.

2.2 Data and Model Development

Models were developed using monthly streamflow, climate, and land cover data for the period from 1961 to 2004, resulting in 528 monthly observations. In each of the five major rivers in the basin, we developed empirical models that estimated monthly streamflow as a function of climate conditions and agricultural land cover in each basin. Monthly streamflow data were taken from historic stream gauge records for each basin, as reported in feasibility studies developed for proposed irrigation projects (Alemayehu, 2010). Historic data for monthly average temperature and~~,~~ monthly total precipitation~~, and monthly wet days~~ in each river basin were derived from the University of East Anglia Climate Research Unit (CRU) TS3.10 gridded meterological fields (Harris et al., 2014), which are based on meteorological station observations. ~~Historic estimates of rainfall intensity were also calculated by dividing monthly total precipitation by CRU TS3.10 records of the number of wet days in that month, but was found to be highly correlated with monthly precipitation and did not result in significant improvements to the predictive accuracy of tested models. Thus, it was not included in the final model formulations.~~ Finally, to account for historic increases in agricultural and pastoral land cover that have occurred in the basin, the percentage of land cover used for any crop or grazing was estimated from historic land cover analyses described by Rientjes et al. (2011), Gebrehiwot et al. (2010), and Garede and Minale (2014). These studies used historic aerial photos and satellite images to estimate land cover changes in the Ribb, Gilgel Abbay, and Koga basins from the periods of 1957 to 2011. The percentage of agricultural land cover was interpolated for years when data weren‘t available, and the value of agricultural land cover in the two basins without data was assumed to be equal to average agricultural land cover in the basins with data. Land cover was assumed to change on an annual, rather than monthly basis. While this approach is prone to errors that could stem from differing rates of land use change through time and between basins, it does provide a mechanism for capturing the long-term trend of expanding agricultural land cover that has been observed throughout the Ethiopian highlands when detailed land-cover data are unavailable. Including this data improved out-of-sample predictive accuracy of the models, further suggesting that it was a valuable addition.

Two general formulations for the empirical models were evaluated. The first (referred to below as the standard model formulation) was

$$\log(Q_{b,t}) = f\left(P_{b,t}, P_{b,t-1}, P_{b,t-2}, T_{b,t}, T_{b,t-1}, T_{b,t-2}, AgLC_{b,t}\right) + \varepsilon_{b,t} \qquad (1)$$

where $Q_{b,t}$ is the monthly streamflow in river b at time period t, $P_{b,t}$ and $T_{b,t}$ are the monthly total precipitation and average temperature in river basin b at time period t, $AgLC_{b,t}$ is the total percentage of agricultural land cover in basin b at time t, and $\varepsilon_{b,t}$ is the model error. The subscripts $t-1$ and $t-2$ indicate lagged measurements from one and two months prior, and were included to roughly account for storage times longer than one month that could impact streamflow in each river. While the exact time of concentration is not known in each basin, the minor influence of ~~of~~ climate conditions at two months prior suggest that climate conditions from beyond this time period do not contribute significantly to flow variability. The function f represents a general function that differed between the specific models assessed and is discussed in more detail below. The logarithm of monthly streamflow was used as a response variable to keep model predictions positive. The distribution of streamflow data and log-transformed streamflow values in each basin are shown in supplementary Fig. S1.

In the second formulation, streamflow and climate anomalies were used as the response and predictor variables to better account for the highly seasonal nature of streamflow and precipitation in the region. Streamflow anomalies were calculated for each observation by subtracting the long-term average streamflow for that month (m) from the observed value and dividing this number by the long-term standard deviation of that month's streamflow as in Eq. (2). Anomaly values thus represent how streamflow in a given month compares to the long-term average flow for that month; for instance, an anomaly value of 1.0 for June of 1990 would indicate that streamflow in that month was one standard deviation higher than the average June flow from 1961 to 2004. This procedure was repeated for precipitation and temperature, and these values were then used to fit models of the form described in Eq. (3). In each month of the time series, the model estimates the flow relative to the long-term average flow for that month, based on whether temperature and precipitation values were greater or less than their long-term averages, as well as the percentage of agricultural land cover in that month of the time series. In this sense, the anomaly values are calculated based on climatic and land cover conditions that vary through time. These anomaly values are then converted back to raw flow values based on the long-term average and standard deviation of flow for that month. The distribution of streamflow anomaly values in each basin are shown in supplementary Fig. S1. ~~It should be noted that although this formulation uses long-term averages and standard deviations to convert anomaly values to flow volumes, the anomaly values themselves are calculated based on climatic and land cover conditions that are nonstationary through time.~~

$$Q_{b,t}^{AN} = \frac{Q_{b,t} - \bar{Q}_{b,m}}{sd(Q_{b,m})} \tag{2}$$

$$Q_{b,t}^{AN} = f\left(P_{b,t}^{AN}, P_{b,t-1}^{AN}, P_{b,t-2}^{AN}, T_{b,t}^{AN}, T_{b,t-1}^{AN}, T_{b,t-2}^{AN}\ AgLC_{b,t}\right) + \varepsilon_{b,t} \qquad (3)$$

Six different types of models were compared using each formulation in each basin:

1. A Gaussian linear regression model (GLM) using the basic stats package in the R statistical computing software (R Development Core Team, 2014)
2. Gaussian generalized additive model (GAM): GAMs are a semi-parametric regression approach where the response variable is estimated as the sum of smoothing functions applied over predictor variables. These functions allow the model to capture non-linear relationships between the predictor and response variables without *apriori* assumptions about the form (eg., quadratic, logarithmic) of these functions, and are fit using penalized likelihood maximization to prevent model overfitting (Hastie and Tibshirani, 1990). GAMs were fit using the mgcv package in R (Wood, 2011).
3. Multivariate adaptive regression splines (MARS): MARS are a non-parametric regression approach where the response variable is estimated as the sum of basis functions fit to recursively partitioned segments of the data (Friedman, 1991). MARS models were fit using the earth package in R (Milborrow, 2015).
4. Artificial neural network (ANN): ANNs are a non-parametric regression approach represented by a network of nodes and links that connects predictor variables to the response variable. Each link in the network represents a function that maps the input nodes into the output node (Ripley, 1996). ANN models were fit using the nnet package in R (Venables and Ripley, 2013).
5. Random forest (RF): Random forests are a rule-based, non-parametric regression approach where the model prediction is created by averaging the predicted value from multiple regression trees which are trained on separate bootstrapped resamples of the data. Each tree is fit using a small, randomly selected subset of predictor variables, resulting in reduced correlation between trees (Breiman, 2001). Random forest models were fit using the randomForest package in R (Liaw and Wiener, 2002).
6. M5 model: M5 models are a rule-based, non-parametric regression approach that fits a linear regression model to each terminal node of a regression tree (Quinlan, 1992). M5 models were fit using the Cubist package in R (Kuhn et al., 2014).

7. Climatology model: A climatology model that simply predicted each month's streamflow as equivalent to the long-term average streamflow for that month was included for comparison purposes.

2.3 Model Evaluation

When using non-parametric regression approaches, it is important to avoid overfitting a model to a given dataset because this can result in large errors in out-of-sample predictions (Hastie et al., 2009). To avoid model overfit, the caret package in R (Kuhn, 2015) was used to determine model parameters for the MARS, ANN, RF and M5 models. This package uses resampling to evaluate the effect that model parameters have on the model's predictive performance and chooses the set of parameters that minimizes out-of-sample error (Kuhn 2015). In this evaluation, 25 bootstrap resamples of the training dataset were generated for each parameter value to be assessed. A model was fit using each bootstrap sample and used to predict the remaining observations, and the parameter values that minimized average RMSE across all resamples. Details on the specific parameters evaluated for each model are presented in Table 2. While the development of more complex structures are possible for some models, this process can result in over-parameterization and poor model performance (Gaume and Gosset, 2003; Han et al., 2007). Additionally, the use of a standardized parameterization procedure allows for a more even comparison between different model types.

The predictive ability of each model was assessed using 50 random holdout cross-validation samples. In each sample, a random selection of years were chosen, and observations from these years were removed ("held-out") from the dataset. The size of the held-out sample ranged from 1 to 9 years. Each model was then fit to the remaining portion of the data, using the caret package described above to determine model parameters for the MARS, ANN, RF and M5 models. These models were then used to predict streamflow for the held-out portion of the data, and both the mean absolute error (MAE) and NSE were calculated after transforming model predictions after back to the original streamflow units. Mean MAE and NSE were calculated for each model across the 50 cross-validation samples and used to choose the model with the highest predictive accuracy in each basin. This cross-validation procedure provides a mechanism for evaluating how well a model will generalize to an unseen set of data while avoiding some of the problems that can arise from the use of a single calibration and validation dataset (Elshorbagy et al., 2010a; Han et al., 2007).

MAE was included as an error metric because it provides a simple and easily interpretable measure of error on the same scale as observed flow volumes. While NSE values are acknowledged to be a flawed performance metric in highly seasonal watersheds where seasonal fluctuations contribute to a substantial portion of flow variability (Schaefli and Gupta, 2007; Legates and McCabe, 1999), this metric was included to provide a rough comparison of how empirical model performance compared to the performance of physical models developed for the region. The use of alternative error metrics has been discussed extensively in the literature (for instance Pushpalatha et al., 2012; Mathevet et al., 2006; Criss and Winston, 2008), and could provide additional insights into what contributes to predicitive capabilities of different model formulations. However, this work examined predicitve accuracy based on MAE and NSE alone to allow for greater focus on how models differ in terms of error structure and uncertainty.

As a rough point of comparison for the statistical models developed in this research, we also evaluated discharge estimates derived from a process-based hydrological model. The model used in this application is the Noah Land Surface Model version 3.2 (Noah LSM; Ek et. al, 2003; Chen et al., 1996). Noah LSM was implemented for offline simulations of the Lake Tana basin at a gridded spatial resolution of 5km for the period 1979-2010 using a time step of 30 minutes. Meteorological forcing was drawn from the Princeton 50-year reanalysis dataset (Sheffield et al. 2006), downscaled to account for Ethiopia's steep terrain using MicroMet elevation correction equations (Liston & Elder 2006). The Princeton reanalysis was selected because it provides relatively high resolution meteorological fields, including all variables required to run a water and energy balance LSM like Noah, for the period 1948-present. While higher resolution and possibly higher quality datasets are available for recent years, this longer dataset was utilized to compare the process-based model to statistical models developed for a long historical period. Soil parameters for the Noah simulation were drawn from the FAO global soil database, land use was defined according to the United States Geological Survey (USGS) global 1km land cover product, and vegetation fraction was derived from MODerate Imaging Spectroradiometer (MODIS) imagery. Land cover was treated as a static parameter over the full length of the simulation, as spatially complete estimates of historical land use were not available at the required resolution and specificity.

The highest performing model in each basin based on MAE was retained for more detailed evaluation of model error structure, covariate influence, and uncertainty in climate change sensitivity analysis. To generate a complete time-series of out-of-sample model predictions for error analysis, the holdout cross validation procedure was repeated for the highest performing standard-formulation and anomaly-formulation models for each basin, but this time holding out a single year of observations in each iteration. The predictions from this cross validation were used to evaluate the how model error structure might impact model predictions used for water resource applications. The influence of different predictor variables on model predictions was also assessed for the highest performing model in each basin after being fit to the complete dataset. Each predictor variable was assessed using metrics for covariate importance and influence that are unique to that model type, demonstrating how models could be used to gain physical insights about data-scarce regions and the mechanisms for generating these insights for each type of model. Partial dependence plots (Hastie et al., 2009) were also generated for each covariate for the highest performing model in each basin to provide insights about how covariate influence compared across different basins and model types.

Finally, two evaluations were conducted to assess uncertainty in model projections of streamflow under increasingly extreme climate conditions to better understand the implications of using different model formulations for climate change impact studies. Model projections of streamflow in different climate conditions are likely to be accompanied by considerable uncertainty, particularly when climate conditions exceed those experienced historically. To assess this uncertainty, the best performing model in each basin was used to generate streamflow predictions for 1) changes in temperature from 0 to 5° C, 2) changes in precipitation from -30 to +30%, 3) an increase in temperature to 5° C combined with a decrease in precipitation to -30%, and 4) an increase in temperature to 5° C combined with an increase in precipitation to +30%. For each of the four assessments, the models generated predictions for the 45-year historic climate record adjusted for a given degree of climate change using the delta-change method (Gleick, 1986), while holding agricultural land cover constant at 60%. In this method, monthly temperature values are simply added to the temperature change value, and monthly precipitation values are multiplied by the precipitation change percentage. Model predictions for the altered climate record were then used to calculate the average annual streamflow in each river. This process was repeated 100 times for models fit on random bootstrap resamples of the historic dataset to generate uncertainty bounds surrounding model predictions and evaluated how the uncertainty in these predictions increased as climate conditions became more extreme. It is important to recognize that these should not be interpreted as a prediction or assessment of actual climate change impacts, but rather a measurement of the sensitivity of modeled streamflow in the basin to different climate conditions. Since one of the key motivations for using rainfall-runoff models is to understand how climate change may impact water resources, it is important to understand how model formulation contributes to this sensitivity and uncertainty.

3 Results

3.1 Model Accuracy and Error Structure

Table 3 shows the out-of-sample cross validation errors for each model assessed in each basin. The random forest model had the lowest mean absolute error for the standard-formulation model in four of the five basins, with the M5 model performing best for the Koga basin. These models outperformed the Noah LSM simulations in all basins assessed. The Noah LSM errors are for a single period of analysis and thus don't present an exact corollary to the cross validation performed for the empirical models. Nevertheless, the significant increases in errors associated with the Noah LSM model demonstrates the difficulty associated with the use of process-based models in the region, particularly when relying on global datasets that may be unreliable at the spatial and temporal resolutions required for physical modeling. Physical models developed for monthly streamflow prediction in other basins within the Ethiopian highlands have reported NSE values ranging from 0.53 to 0.92 (van Griensven et al., 2012), compared to values ranging from 0.71 to 0.87 for the random forest models developed here. If this measure alone was used for model evaluation, these empirical models would generally be classified as having good performance based on the guidelines suggested by Moraisi et al. (2007). However, the climatology model outperforms the best standard formulation models in all basins except Megech, indicating that in the majority of basins the errors from the fitted empirical models are higher than those that result from simply using the long-term monthly average for each month's prediction. This is due to the fact that seasonality accounts for such a large portion of the variability in monthly flow values, and demonstrates how high NSE values can be quite easy to obtain in seasonal basins.

Evaluation of anomaly model errors indicates that the models using this formulation achieve better predictive accuracy than those using the standard formulation, and are able to outperform the climatology model based on both NSE and MAE in all basins. However, the highest performing models in each basin varies more when the anomaly formulation is used, with the GLM, GAM, random forest, and M5 models all minimizing MAE in different basins. In all basins except Koga, the highest performing model significantly outperformed the climatology model based on paired Wilcoxon rank-sum tests (Bonferroni-corrected p-value < 0.01).

Further exploration of model residuals indicates another important advantage of using the anomaly model formulation. In the standard model formulation, model residuals appear to be non-random. Example autocorrelation plots are shown for the Gilgel Abbay and Ribb Rivers in Fig. 2, and demonstrate that a positive autocorrelation exists at the 12 month time lag. For brevity, only plots for two rivers are shown, although this autocorrelation existed in the standard-formulation models for all basins except Megech (Table 4). This autocorrelation occurs because the standard-formulation models consistently underestimate wet-season streamflow while overestimating dry-season flows, as is apparent in hydrographs of observed and predicted streamflow (Fig. 3). Because wet-season flows contribute such a large portion of the total annual flow volume, this results in regular underestimation of aggregate values such as mean annual flow (Table 4). This autocorrelation is reduced in the anomaly-formulation models, meaning that they are better able to capture the peak flow volumes experienced in the wet season and do not underestimate mean annual flow to the same degree that the standard formulation models do.

3.2 Model Structure and Covariate Influence

Evaluating the relationship between predictor covariates and streamflow response can lend insight into the physical processes underlying runoff generation in each basin. There are two components of this relationship that can be evaluated: how much each covariate contributes to model accuracy (covariate importance), and the direction and nature of the relationship between covariate values and model response (covariate influence). In many machine-learning models, complete description of the all of the mathematical relationships within the model (for instance, through description of each tree comprising a random forest model) is infeasible, requiring the use of other mechanisms for understanding covariate importance and influence. However, because each model type is structured in a different way, these mechanisms differ. This section first describes the mechanisms available for obtaining insights about covariate influence in each of the highest performing models. To provide a mechanism for comparing results across different basins, each basin model is then assessed using the general approach of partial dependence plots.

In the Gilgel Abbay and Koga basins, the highest performing model was a simple linear regression model. These models can be evaluated by reviewing model coefficients and associated p-values, as shown in Table 5. In a standard linear regression, model coefficients can be interpreted as the mean change in the response variable that results from a unit change in that covariate when all others are held constant. These coefficients are for streamflow anomalies rather than raw values, making their immediate interpretation less intuitive. For instance, in the Gilgel Abbay model an increase of one standard deviation in precipitation results in an increase of 0.22 standard deviations in flow. The associated p-value for each coefficient evaluates a null hypothesis that the true coefficient value is equal to zero given the other covariates in the model, and thus has no influence on the response variable.

Evaluating model structure based on regression coefficients is appealing due to their simplicity and familiarity. However, it is important to keep in mind that the above interpretations rely on specific assumptions regarding model error distributions. Examination of fitted model residuals from both basins indicate that errors are autocorrelated in the Koga basin and not normally distributed due to the presence of outliers in both basins. Non-normality and autocorrelation both impact the t statistics and f statistics used to test for the significance of model coefficients, and thus the p-values for these models are likely biased (Montgomery et al., 2012).

Interpretation of variable influence in GAMs is based on the estimated degrees of freedom (EDF) a covariate's smoothing function $s(X_i)$ uses within a model (Hastie and Tibushini, 1986). An EDF value of one or below indicates a linear function relating the response variable to that covariate, while values greater than one represent a non-linear smoothing function. An EDF value of zero indicates that the covariate smoothing function is penalized to zero (meaning it has no influence on model predictions). In the model for the Megech River, the terms for lagged temperature at one and two months, as well as precipitation lagged at two months were all smoothed to zero. Of the remaining covariates, lagged precipitation has a linear impact on model response, while precipitation, temperature and land cover have non-linear impacts. Smoothing functions can be plotted to gain more insight about these relationships (Fig. 4). The functions for precipitation anomaly, lagged (one month) precipitation anomaly, and agricultural land cover show a positive relationships with streamflow, while the function for temperature anomaly predicts low streamflow at both high and low anomalies.

P-values test the null hypothesis that a covariate's smoothing function is equal to zero, but rest on the assumption that model residuals are homoscedastic and independent (Wood, 2012). Similar to the linear models, residuals in the Megech GAM model appear to be both autocorrelated and heteroscedastic, meaning that a formal statistical interpretation of this value may be inappropriate and that confidence bounds around smoothing functions might be misleading.

The M5 cubist model fit for the Gumara basin is an ensemble of 100 small M5 regression trees. In each tree, the model splits observations based on logical rules related to one or more covariates and fits a linear regression model to each set of observations. The final model prediction is the average across all of the individual trees. Using this sort of ensemble approach can reduce model variance and improve accuracy if the individual trees are unbiased, uncorrelated predictors (Breiman 1996). This can be useful in avoiding models that are overfit to the data, but can reduce model interpretability since direct visualization of model structure becomes impractical as the number of trees increases. However, the frequency with which individual covariates are used as splitting points within trees and as regression coefficients can provide some insights about covariate importance (Table 5; note that because multiple covariates can be used for rules and linear models, these don't necessarily add to 100%). Model rules were largely based on land cover, with some rules based on precipitation. These two covariates were also used most frequently in linear regressions at model nodes, followed by temperature (current and 1-month lag) and 1-month lagged precipitation. Notably, climate data from 2 months lagged were not used at all. While this can be useful in identifying which covariates have the largest impact on model predictions, it doesn't provide any information regarding the nature or direction of that influence.

Similarly, the random forest model developed for the Ribb basin is an ensemble of regression trees in which the final model prediction is the average of the predictions from each individual tree. However, random forests use standard regression trees that do not incorporate linear regression models at terminal nodes. Variable importance within the final model is measured by recording the increase in out-of-sample MSE that results when a covariate is randomly permuted for each tree in the ensemble. This increase in error is then averaged across all trees in the ensemble. In our model, the largest increases in error resulted from permutation of land cover and temperature, followed by 2-month lagged temperature and precipitation. Covariate influence can be evaluated through the use of partial dependence plots, which measure the change in model predictions that result from changing the value of one parameter while leaving all other covariates constant (Hastie et al., 2009). Partial dependence plots indicate that model predictions of streamflow are higher when the percent of agricultural land cover is greater than approximately 75%, when temperatures anomalies are low, and when precipitation anomalies are high. However, it appears that the plot for lagged temperature might be sensitive to outliers at high temperature anomalies as evidenced by the large increase that occurs above an anomaly of +2, in a region where very few data points are present.

Many of the measures used to evaluate covariate importance and influence are model specific, making inter-basin and inter-model comparisons difficult. However, the partial dependence plots used in the randomForest R package can be developed for any model and provide a mechanism for comparing the influence that covariates have in the different models and basins (Shortridge et al., 2015). Partial dependence plots were generated for each basin's best performing model and results are shown for climatic variables in Fig. 6. As expected, models generally respond positively to increases in precipitation and negatively to increases in temperature, with the greatest influence in the current month and decreasing influence at one and two months prior. The influence of the current month's precipitation is linear in three of the five basins; while this is constrained to the be the case in the Gilgel Abbay and Koga basins due to the use of a linear model, the linear response in Gumara is not required from the M5 model structure. Interestingly, both Megech and Ribb demonstrate a linear response to negative precipitation anomalies, but little response to positive anomalies. Streamflow response to temperature is strongest in the Gumara basin; interestingly, this is the basin with the smallest response to precipitation.

The partial dependence plots for the percentage of the basin classified as agricultural land cover indicates a positive relationship between agricultural land cover and streamflow in all basins except for the Gilgel Abbay (Fig. 7). This would be expected if deforestation had contributed to a decrease in evapotranspiration in the contributing watersheds. The exact nature of this response differs across the different rivers, with the relatively minor responses in Koga and Ribb, and much stronger responses in the Gumara and Megech basins. However, this plot also demonstrates some of the limitations associated with different model structures. The plot for Gumara is highly erratic, indicating that the M5 model might be overfit to the training dataset, despite the use of model averaging to reduce model variance. Additionally, the GAM used in the Megech basin was only trained on agricultural land cover values up to 77%; while this model may be accurately representing the impact of land cover changes within this range, extrapolating this relationship to higher values leads to predictions that may not be physically realistic.

3.3 Climate Change Sensitivity and Uncertainty Assessment

Fig. 8 shows the results of the climate change sensitivity analysis for total flow from all five tributaries, with dashed lines representing 95% confidence intervals obtained through 100 bootstrapped resamples of the data set. As would be expected, increasing temperature independently of precipitation results in decreasing total flows while increasing precipitation results in higher flows. However, the uncertainty surrounding temperature sensitivity increases at higher changes in temperature, while the uncertainty surrounding precipitation sensitivity remains relatively constant, even at extreme changes in annual precipitation. The bottom panels of the figure show the sensitivity of total inflows to concurrent changes in temperature and precipitation. Unsurprisingly, decreasing precipitation combined with higher temperatures results in greater decreases in total flow than when temperature and precipitation are varied independently. However, even if temperature increases are combined with higher precipitation, total flows decline in the majority of bootstrap resamples.

The uncertainty surrounding temperature sensitivity is a key limitation to using data-driven approaches for climate impact assessment. To better understand which models and basins are contributing to this uncertainty, Fig. 9 shows how the coefficient of variation (the standard deviation of predictions from all bootstrap samples divided by the mean of these predictions) varies as a function of temperature change in each basin. From this figure, it is apparent that the Megech model is by far the largest contributor to model uncertainty; however, it is not clear whether this contribution is due to model structure (the GAM model used for the Megech River) or characteristics associated with the basin itself. To investigate how different model structures contributed to this uncertainty, the bootstrap resampling procedure was used to assess uncertainty in streamflow predictions in the Gumara River from all model types. This basin was chosen because all six models were able to outperform the climatology model, and thus could be considered good choices for model selection based on predictive accuracy alone. The results indicate that the increase in uncertainty is highest, and increases non-linearly, in the GLM, GAM, and MARS models. Uncertainty increases more slowly in the ANN and M5 models, and no noticeable increase in uncertainty is apparent in the random forest model.

4 Discussion

The objective of this study was not to identify the "best" approach for empirical rainfall-runoff modeling, as this is likely to be highly specific to the basin and problem to which a model is applied. However, we hope that the comparison conducted here can highlight some of the strengths and limitations of different approaches, as well as demonstrate some important issues that should be kept in mind for model comparisons in the future. One important finding was the limitation with using NSE as an error metric. Our results confirm previous studies that found that even uninformative models able to capture basic seasonality are able to achieve high NSE values (Legates and McCabe, 1999; Schaefli and Gupta, 2007), and provide further evidence indicating that high NSE values should be considered a necessary but not sufficient requirement for model usage in planning situations. For instance, the simple climatology model used for comparison purposes here is able to achieve high NSE values, but would be unsuitable for planning since it does not account for any interannual variability nor the possibility for non-stationary conditions caused by changing climate and land cover. In particular, understanding error structure can be valuable in evaluating whether model biases might undermine the model's suitability for management activities. In our example, the autocorrelation present in the standard-formulation models meant that these models were consistently underestimating wet-season flows, resulting in low estimates of the total annual flow in the rivers. Since multiple reservoirs are planned for construction on these rivers to support irrigation activities, this bias could lead to poor estimates of how much water is available for agricultural use in the short term (ie., seasonal forecasting) and long-term (due to climate change). Interestingly, difficulties in accurately capturing high flows has been observed in physical hydrologic models for Ethiopia (e.g., Setegne et al., 2011; Mekonnen et al., 2009) and more generally (e.g., Wilby, 2005). The implications of this limitation should be carefully evaluated before using models for water resource planning or (more importantly) flood risk evaluation.

Depending on the model type used, different mechanisms are available to evaluate covariate importance and influence within the model. This evaluation can be useful in confirming that the model is replicating ~~physically realistic~~ relationships between input and output variables in a reasonable manner. While the relationships identified in this evaluation are fairly straightforward (for example, increasing runoff with higher precipitation and lower temperatures), these simple relationships are still important in highlighting the mechanisms by which the models make predictions so that they are not "black boxes." For instance, Han et al. (2007) explore how ANN flood forecasting models responds to a double-unit input of rain, finding that some formulations respond in a hydrologically meaningful way to increased rainfall intensity, while others do not. Similarly, Galelli and Castelletti (2013a) describe how input variable importance can be used to highlight differences in hydrologic processes between an urbanized and forested watershed. The easy manner in which covariate relationships within the GAM and random forest models can be visualized using a single command within their respective R packages is a strong advantage to these approaches compared to methods such as M5 model trees and artificial neural networks. Of course, partial dependence plots can be developed for any model type (as was done in this research), but code must be written by the user and thus requires a higher degree of effort than is necessary for in-package functions. A downside to most machine-learning models is that they do not support the statistical formalism in assessing variable importance that is possible when linear models and GAMs are used. However, this formalism often rests on assumptions regarding model residuals that are unlikely to be met in many hydrologic models (Sorooshian and Dracup, 1980).

Within the Lake Tana basin, evaluation of covariate influence indicates that each basin's model is performing in a ~~physically realistic~~reasonable manner, with ~~a~~ runoff increasing with higher precipitation levels and decreasing with higher temperatures. The influence of precipitation and temperature is greatest in the current month, and progressively declines to a very small influence after two months. This suggests that long-term (multi-month) storage does not significantly contribute to variability in flow volumes. One interesting finding is the non-linear relationship between concurrent month precipitation and runoff that exists in the Megech and Ribb basins, which suggests that above a certain point increasing rainfall does not result in a commiserate increase in streamflow. Other studies have noted the dampening effect that wetlands and floodplains have had on river flows in the region (Dessie et al., 2014; Gebrehiwot et al., 2010); this phenomenon could explain the non-linear relationship identified in this work. The clearly negative relationship between temperature and runoff demonstrates the degree to which upstream evapotranspiration impacts streamflow and suggests that evapotranspiration is largely energy-limited, rather than water-limited. Increasing agricultural land-use appears to be associated with higher runoff in all rivers except for Gilgel Abbay (where no clear relationship between land cover and runoff was observed), and suggests that agricultural expansion at the expense of forest cover has reduced the evaporative component of the water balance in these basins. Finally, the relative performance of different model formulations themselves can also be informative. For instance, the improved performance of the anomaly-formulation models indicates that the relationship between precipitation and runoff varies throughout the year and could point towards differences in runoff-generating mechanisms in the wet and dry seasons that have been observed in other case studies (Wilby, 2005).

One limitation with data-driven approaches for streamflow prediction is that the relationships they model can only generate reliable predictions for conditions that are comparable to those experienced historically. Using these models to generate predictions for conditions that exceed historic variability is likely to introduce considerable uncertainty into their projections. Our results indicate that uncertainty in projections of streamflow under changing precipitation is relatively constant, whereas uncertainty increases markedly in projections of streamflow under increasing temperature. This result is not surprising when one considers the basin's climate, which is characterized by highly variable rainfall but fairly consistent temperatures (Table 6). A temperature increase of 3° C equates to almost two standard deviations beyond the historic mean, whereas a change in precipitation of 30% is well within the range of conditions experienced historically. One would expect that in other climates (for example, temperate watersheds with only minor changes in rainfall throughout the year), this relationship could be reversed. Despite the uncertainty that exists in projections of streamflow under changing temperature, total annual flow appears to be quite sensitive to increasing temperatures. In fact, the decreases in streamflow due to increasing temperature appears likely to be more than enough to counteract any increases in streamflow resulting from higher precipitation that is projected for the region in some global circulation models (GCMs). This is consistent with the work of Setegne et al. (2011), who used projections from multiple GCMs as input for a SWAT model developed for the region and found that streamflow decreased in the majority of emissions scenarios and models, even when precipitation increased. Unfortunately, this suggests that any hopes for a "windfall" of additional water to support agriculture and hydropower in the region under climate change may be unfounded.

Repeating the climate change sensitivity experiment with multiple models fit to the Gumara watershed indicated that the MARS, GAM, and linear models all result in the largest increase in uncertainty at high temperatures. This indicates that when models are fit to slightly different bootstrap resamples of the historic dataset, the projected changes in streamflow at high temperature changes can be highly erratic. This is likely due to the fact that extrapolating the relationships that are observed between historic temperature and streamflow to higher temperatures can lead to very large changes in streamflow. Fitting the models to bootstrap resamples of the data results in minor changes to these relationships that can result in widely varying projections when the models are used to predict streamflow at higher temperatures, particularly when these relationships are nonlinear (as in the GAM). At the other end of the spectrum, the random forest model exhibits almost no increase in uncertainty at high temperatures, meaning that projections of streamflow at high temperatures are consistent across the bootstrap resamples. This is likely the result of the random forest model structure. The predicted value for each of a regression tree's terminal nodes is the average of all observations that meet the conditions described for that node. Thus, the model will not predict values beyond those experienced historically, even if covariate values exceed those contained within the historic dataset. Thus, this model is likely to underestimate the change in streamflow that results from increasing temperatures.

5 Conclusions

In this work, we compared multiple methods for data-driven rainfall-runoff modeling in their ability to simulate streamflow in five highly-seasonal watersheds in the Ethiopian highlands. Despite the popularity of ANNs in research on streamflow prediction to date, ANNs were not found to be the most accurate model in any of the five basins evaluated. Other methods, in particular GAMs and random forests, are able to capture non-linear relationships effectively and lend themselves to simpler visualization of model structure and covariate influence, making it easier to gain insights on physical watershed functions and confirm that the model is operating in a ~~physically realistic~~reasonable manner. However, it is important to carefully evaluate model structure and residuals, as these can contribute to biased estimates of water availability and uncertainty in estimating sensitivity to potential future changes in climate. In particular, autocorrelation in model residuals can result in underestimation of aggregate metrics such as annual flow volumes, even in models with high NSE performance. Uncertainty in GAM projections was found to rapidly increase at high temperatures, whereas random forest projections may be underestimating the impact of high temperatures on river flows. Thorough consideration of this uncertainty and bias is important any time that models are used for water planning and management, but especially crucial when using such models to generate insights about future streamflow levels. By considering multiple model formulations and carefully assessing their predictive accuracy, error structure and uncertainties, these methods can provide an empirical assessment of watershed behavior and generate useful insights for water management and planning. This makes them a valuable complement to physical models, particularly in data-scarce regions with little data available for model parameterization, and warrants additional research into their development and application.

Acknowledgements

We would like to gratefully acknowledge the Ethiopian Ministry of Water and Energy, the Tana Sub Basin Organization, and the International Water Management Institute for making available the data used to perform this analysis. All data for this paper are properly cited and referred to in the reference list. The source code for the models developed in this study is available from the authors upon request. Empirical modeling work was supported by a National Defense Science and Engineering Graduate Fellowship and by National Science Foundation Grant 1069213 (IGERT). Noah LSM simulations presented here were performed under NASA Applied Sciences Program grant NNX09AT61G. This research was conducted while Dr. Guikema was affiliated with the Department of Geography and Environmental Engineering at Johns Hopkins University. This support is gratefully acknowledged. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding sources.

References

Abrahart, R. J. and See, L. M.: Neural network modelling of non-linear hydrological relationships, Hydrol. Earth Syst. Sci., 11(5), 1563–1579, doi:10.5194/hess-11-1563-2007, 2007.

Achenef, H., Tilahun, A. and Molla, B.: Tana Sub Basin Initial Scenarios and Indicators Development Report, Tana Sub Basin Organization, Bahir Dar, Ethiopia., 2013.

Alemayehu, T., McCartney, M. and Kebede, S.: The water resource implications of planned development in the Lake Tana catchment, Ethiopia, Ecohydrology & Hydrobiology, 10(2-4), 211–221, doi:10.2478/v10104-011-0023-6, 2010.

Antar, M. A., Elassiouti, I. and Allam, M. N.: Rainfall-runoff modelling using artificial neural networks technique: a Blue Nile catchment case study, Hydrol. Process., 20(5), 1201–1216, doi:10.1002/hyp.5932, 2006.

Aqil, M., Kita, I., Yano, A. and Nishiyama, S.: Neural Networks for Real Time Catchment Flow Modeling and Prediction, Water Resour. Manag., 21(10), 1781–1796, doi: 10.1007/s11269-006-9127-y, 2007.

Asefa, T., Kemblowski, M., McKee, M. and Khalil, A.: Multi-time scale stream flow predictions: The support vector machines approach, J. Hydrol., 318(1-4), 7–16, doi:10.1016/j.jhydrol.2005.06.001, 2006.

Beven, K. J.: Rainfall-Runoff Modelling: The Primer, John Wiley & Sons., 2011.

Breiman, L.: Bagging predictors, Mach Learn, 24(2), 123–140, doi:10.1007/BF00058655, 1996.

Breiman, L.: Random forests, Mach Learn, 45(1), 5–32, 2001.

Chen, F., Mitchell, K., Schaake, J., Xue, Y., Pan, H.-L., Koren, V., Duan, Q. Y., Ek, M. and Betts, A.: Modeling of land surface evaporation by four schemes and comparison with FIFE observations, J. Geophys. Res., 101(D3), 7251–7268, doi:10.1029/95JD02165, 1996.

Chibanga, R., Berlamont, J. and Vandewalle, J.: Modelling and forecasting of hydrological variables using artificial neural networks: the Kafue River sub-basin, Hydrolog. Sci. J., 48(3), 363–379, doi:10.1623/hysj.48.3.363.45282, 2003.

Criss, R. E. and Winston, W. E.: Do Nash values have value? Discussion and alternate proposals, Hydrol. Process., 22(14), 2723–2725, doi:10.1002/hyp.7072, 2008.

De Vos, N. J. and Rientjes, T. H. M.: Multiobjective training of artificial neural networks for rainfall-runoff modeling, Water Resour. Res., 44(8), W08434, doi:10.1029/2007WR006734, 2008.

Dessie, M., Verhoest, N. E. C., Admasu, T., Pauwels, V. R. N., Poesen, J., Adgo, E., Deckers, J. and Nyssen, J.: Effects of the floodplain on river discharge into Lake Tana (Ethiopia), J. Hydrol., 519, 699–710, doi:10.1016/j.jhydrol.2014.08.007, 2014.

Ek, M. B., Mitchell, K. E., Lin, Y., Rogers, E., Grunmann, P., Koren, V., Gayno, G. and Tarpley, J. D.: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model, J. Geophys. Res., 108(D22), 8851, doi:10.1029/2002JD003296, 2003.

Elshorbagy, A., Corzo, G., Srinivasulu, S. and Solomatine, D. P.: Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology - Part 1: Concepts and methodology, Hydrol. Earth Syst. Sci., 14(10), 1931–1941, doi:10.5194/hess-14-1931-2010, 2010a.

Elshorbagy, A., Corzo, G., Srinivasulu, S. and Solomatine, D. P.: Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology - Part 2: Application, Hydrol. Earth Syst. Sci., 14(10), 1943–1961, doi:10.5194/hess-14-1943-2010, 2010b.

Friedman, J. H.: Multivariate adaptive regression splines, The annals of statistics, 1–67, 1991.

Galelli, S. and Castelletti, A.: Assessing the predictive capability of randomized tree-based ensembles in streamflow modelling, Hydrol. Earth Syst. Sci., 17(7), 2669–2684, doi:10.5194/hess-17-2669-2013, 2013.

Galelli, S. and Castelletti, A.: Tree-based iterative input variable selection for hydrological modeling, Water Resour. Res., 49(7), 4295–4310, doi:10.1002/wrcr.20339, 2013.

Garede, N. M. and Minale, A. S.: Land Use/Cover Dynamics in Ribb Watershed, North Western Ethiopia, Journal of Natural Sciences Research, 4(16), 9–16, 2014.

Gaume, E. and Gosset, R.: Over-parameterisation, a major obstacle to the use of artificial neural networks in hydrology?, Hydrol. Earth Syst. Sci. Discussions, 7(5), 693–706, 2003.

Gebrehiwot, S. G., Taye, A. and Bishop, K.: Forest Cover and Stream Flow in a Headwater of the Blue Nile: Complementing Observational Data Analysis with Community Perception, Ambio, 39(4), 284–294, doi:10.1007/s13280-010-0047-y, 2010.

Han, D., Kwong, T. and Li, S.: Uncertainties in real-time flood forecasting with neural networks, Hydrol. Process., 21(2), 223–228, doi:10.1002/hyp.6184, 2007.

Harris, I., Jones, P. d., Osborn, T. j. and Lister, D. h.: Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 Dataset, Int. J. Climatol., 34(3), 623–642, doi:10.1002/joc.3711, 2014.

Hastie, T. and Tibshirani, R.: Generalized Additive Models, Statistical Science, 1(3), 297–310, 1986.

Hastie, T. and Tibshirani, R.: Generalized additive models. Chapman, Hall, London, 1990.

Hastie, T., Tibshirani, R. and Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference and Prediction, Second Ed., Springer, New York., 2009.

Iorgulescu, I. and Beven, K. J.: Nonparametric direct mapping of rainfall-runoff relationships: An alternative approach to data analysis and modeling?, Water Resour. Res., 40(8), W08403, doi:10.1029/2004WR003094, 2004.

Jain, A., Sudheer, K. P. and Srinivasulu, S.: Identification of physical processes inherent in artificial neural network rainfall runoff models, Hydrol. Process., 18(3), 571–581, doi:10.1002/hyp.5502, 2004.

Kuhn, M.: caret: Classification and regression training, Available from: http://CRAN.R-project.org/package=caret, 2015.

Kuhn, M., Weston, S., Keefer, C. and Coulter, N.: Cubist: Rule- and instance-based regression modeling, Available from: http://CRAN.R-project.org/package=Cubist, 2014.

Legates, D. R. and McCabe Jr, G. J.: Evaluating the use of" goodness-of-fit" measures in hydrologic and hydroclimatic model validation, Water Resour. Res., 35(1), 233–241, 1999.

Liaw, A. and Wiener, M.: Classification and regression by randomForest, R News, 2(3), 18–22, 2002.

Lin, J.-Y., Cheng, C.-T. and Chau, K.-W.: Using support vector machines for long-term discharge prediction, Hydrological Sciences Journal, 51(4), 599–612, doi:10.1623/hysj.51.4.599, 2006.

Liston, G. E. and Elder, K.: A Meteorological Distribution System for High-Resolution Terrestrial Modeling (MicroMet), J. Hydrometeor, 7(2), 217–234, doi:10.1175/JHM486.1, 2006.

Machado, F., Mine, M., Kaviski, E. and Fill, H.: Monthly rainfall–runoff modelling using artificial neural networks, Hydrolog. Sci. J., 56(3), 349–361, doi:10.1080/02626667.2011.559949, 2011.

Maier, H. R., Jain, A., Dandy, G. C. and Sudheer, K. P.: Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions, Environ. Modell. Softw., 25(8), 891–909, doi:10.1016/j.envsoft.2010.02.003, 2010.

Mathevet, T., Michel, C., Andreassian, V. and Perrin, C.: A bounded version of the Nash-sutcliffe criterion for better model assessment on large sets of basins, in IAHS-AISH publication, pp. 211–219, International Association of Hydrological Sciences. [online] Available from: http://cat.inist.fr/?aModele=afficheN&cpsidt=18790113 (Accessed 10 February 2016), 2006.

Mekonnen, M. A., Wörman, A., Dargahi, B. and Gebeyehu, A.: Hydrological modelling of Ethiopian catchments using limited data, Hydrol. Process., 23(23), 3401–3408, doi:10.1002/hyp.7470, 2009.

Milborrow, S.: earth: Multivariate Adaptive Regression Splines, Available from: http://CRAN.R-project.org/package=earth, 2015.

Montgomery, D. C., Peck, E. A. and Vining, G. G.: Introduction to Linear Regression Analysis, John Wiley & Sons., 2012.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D. and Veith, T. L.: Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, Trans. Asabe, 50(3), 885–900, 2007.

Pushpalatha, R., Perrin, C., Moine, N. L. and Andréassian, V.: A review of efficiency criteria suitable for evaluating low-flow simulations, J. Hydrol., 420–421, 171–182, doi:10.1016/j.jhydrol.2011.11.055, 2012.

Quinlan, J. R.: Learning with Continuous Classes, in Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, World Scientific, Singapore., 1992.

R Development Core Team: R: A language and environment for statistical computing., R Foundation for Statistical Computing, Vienna, Austria. Available from: http://www.R-project.org, 2014.

Rientjes, T. H. M., Haile, A. T., Kebede, E., Mannaerts, C. M. M., Habib, E. and Steenhuis, T. S.: Changes in land cover, rainfall and stream flow in Upper Gilgel Abbay catchment, Blue Nile basin – Ethiopia, Hydrol. Earth Syst. Sci., 15(6), 1979–1989, doi:10.5194/hess-15-1979-2011, 2011.

Ripley, B. D.: Pattern Recognition and Neural Networks, Cambridge University Press., 1996.

Schaefli, B. and Gupta, H. V.: Do Nash values have value?, Hydrol. Process., 21(15), 2075–2080, doi:10.1002/hyp.6825, 2007.

See, L., Solomatine, D., Abrahart, R., and Toth, E.: Hydroinformatics: computational intelligence and technological developments in water science applications—Editorial, Hydrological Sciences Journal, 52(3), 391–396, doi:10.1623/hysj.52.3.391, 2007.

Setegn, S. G., Srinivasan, R., Melesse, A. M. and Dargahi, B.: SWAT model application and prediction uncertainty analysis in the Lake Tana Basin, Ethiopia, Hydrol. Process., doi:10.1002/hyp.7457, 2009.

Setegn, S. G., Rayner, D., Melesse, A. M., Dargahi, B. and Srinivasan, R.: Impact of climate change on the hydroclimatology of Lake Tana Basin, Ethiopia, Water Resour. Res, 47(4), doi:10.1029/2010WR009248, 2011.

Sheffield, J., Goteti, G. and Wood, E. F.: Development of a 50-Year High-Resolution Global Dataset of Meteorological Forcings for Land Surface Modeling, J. Climate, 19(13), 3088–3111, doi:10.1175/JCLI3790.1, 2006.

Shortridge, J. E., Falconi, S. M., Zaitchik, B. F. and Guikema, S. D.: Climate, agriculture, and hunger: statistical prediction of undernourishment using nonlinear regression and data-mining techniques, Journal of Applied Statistics (ahead of press), doi:10.1080/02664763.2015.1032216, 2015.

Solomatine, D. P. and Ostfeld, A.: Data-driven modelling: some past experiences and new approaches, Journal of Hydroinformatics, 10(1), 3, doi:10.2166/hydro.2008.015, 2008.

Sorooshian, S. and Dracup, J. A.: Stochastic parameter estimation procedures for hydrologie rainfall-runoff models: Correlated and heteroscedastic error cases, Water Resour. Res., 16(2), 430–442, doi:10.1029/WR016i002p00430, 1980.

Steenhuis, T. S., Collick, A. S., Easton, Z. M., Leggesse, E. S., Bayabil, H. K., White, E. D., Awulachew, S. B., Adgo, E. and Ahmed, A. A.: Predicting discharge and sediment for the Abay (Blue Nile) with a simple model, Hydrol. Process., doi:10.1002/hyp.7513, 2009.

Sudheer, K. P. and Jain, A.: Explaining the internal behaviour of artificial neural network river flow models, Hydrol. Process., 18(4), 833–844, doi:10.1002/hyp.5517, 2004.

Van Griensven, A., Ndomba, P., Yalew, S. and Kilonzo, F.: Critical review of SWAT applications in the upper Nile basin countries, Hydrol. Earth Syst. Sci., 16(9), 3371–3381, doi:10.5194/hess-16-3371-2012, 2012.

Venables, W. N. and Ripley, B. D.: Modern Applied Statistics with S-PLUS, Springer Science & Business Media., 2013.

Wilby, R. L.: Uncertainty in water resource model parameters used for climate change impact assessment, Hydrol. Process., 19(16), 3201–3219, doi:10.1002/hyp.5819, 2005.

Wood, S.: Generalized Additive Models: An Introduction with R, CRC Press., 2006.

Wood, S. N.: Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(1), 3–36, doi:10.1111/j.1467-9868.2010.00749.x, 2011.

Wood, S. N.: On p-values for smooth components of an extended generalized additive model, Biometrika, 100(1), 221–228 doi:10.1093/biomet/ass048, 2012.

Table 1. Study basin characteristics over the evaluation period of 1961 to 2004.

Basin	Drainage area above gauge (km^2)	Average annual streamflow at gauge (MCM)	Standard deviation of annual streamflow (MCM)	Coefficient of variation of annual streamflow	Average temp (°C)	Average monthly rainfall [mm]	
						May-Oct	Nov-Apr
Gilgel Abbay	2664	1883	217	0.12	15.7	206	39.3
Gumara	385	236	71	0.30	17.7	186	29
Koga	200	114	31	0.27	15.7	206	39.3
Megech	424	172	66	0.31	20.6	234	41.4
Ribb	677	210	83	0.36	18.2	263	45.8

Table 2. Model parameters evaluated through cross validation.

Model type	R package	Parameters defined in model formulation	Parameters selected through cross validation
GLM	stats	family = Gaussian	NA
GAM	mgcv	family = Gaussian	
		method = generalized cross validation	
		variable selection = true	
		basis dimension k = 3	
		epsilon = 10^{-7}	
		maxit = 200	
MARS	earth	nk = 21	degree = {1, 2, 3}
		thresh = 0.001	nprune = {5, 10, 15, 20, 25}
		fast.k = 20	
		pmethod = backward	
ANN	nnet	weights = 1	size = {1, 2, 4, 8, 20}
		rang = 0.7	decay = {0.0, 0.1, 0.5, 1.0, 2.0}
		maxit = 100	
		maxNWts = 1000	
		abstol = 10^{-4}	
		reltol = 10^{-8}	
RF	randomForest	ntree = 500	mtry = {2, 3, 4, 5, 6, 7}
		sampsize = 528	
		nodesize = 5	
		nPerm = 1	
M5	Cubist	rules = 100	committees = {10, 50, 100}
		extrapolation = 100	neighbors = {0, 5, 9}
		sample = 0	

Table 3. Cross validation errors for each assessed model.

Standard Formulation		GLM	GAM	MARS	RF	M5	ANN	Climatology	Noah LSM
	Gilgel Abbay	30.78	18.54	16.75	14.89	15.11	17.22	10.42	28.11
	Gumara	4.29	3.41	3.28	2.67	2.96	3.15	2.57	3.95
MAE	Koga	1.50	1.30	1.38	1.20	1.17	1.23	1.06	1.97
	Megech	4.45	2.64	2.83	2.37	2.53	3.04	2.54	4.09
	Ribb	4.69	2.98	3.50	2.97	3.27	3.17	2.81	7.01
	Gilgel Abbay	-0.02	0.81	0.83	0.87	0.86	0.84	0.95	0.59
	Gumara	0.04	0.51	0.61	0.80	0.66	0.70	0.81	0.48
NSE	Koga	0.45	0.71	0.65	0.76	0.77	0.76	0.83	0.25
	Megech	-1.85	0.63	0.46	0.73	0.65	0.52	0.71	0.41
	Ribb	-1.14	0.71	0.39	0.71	0.31	0.67	0.73	-0.75
Anomaly Formulation		GLM	GAM	MARS	RF	M5	ANN	Climatology	Noah LSM
	Gilgel Abbay	9.73	9.82	10.10	10.12	9.94	9.79	10.42	28.11
	Gumara	2.22	2.25	2.43	2.23	2.16	2.22	2.57	3.95
MAE	Koga	1.03	1.06	1.08	1.09	1.05	1.05	1.06	1.97
	Megech	2.49	2.48	2.63	2.66	2.69	2.50	2.54	4.09
	Ribb	2.79	2.76	2.84	2.70	2.78	2.77	2.81	7.01
	Gilgel Abbay	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.59
	Gumara	0.85	0.85	0.82	0.85	0.86	0.86	0.81	0.48
NSE	Koga	0.83	0.82	0.81	0.81	0.82	0.82	0.83	0.25
	Megech	0.73	0.72	0.65	0.66	0.61	0.72	0.71	0.41
	Ribb	0.73	0.75	0.72	0.75	0.73	0.74	0.73	-0.75

Table 4. Residual autocorrelation factors at a 12-month lag for the highest performing standard formulation and anomaly formulation models in each basin (with model type in parenthesis), and resulting mean annual observed and predicted flow.

	Autocorrelation Factors		Mean Annual Flow (MCM)		
	Standard	Anomaly	Observed	Standard	Anomaly
Gilgel	0.33 (RF)	0.11 (GLM)	22,925	20,703	22,958
Gumara	0.29 (RF)	0.07 (M5)	2,870	2,392	2,734
Koga	0.04 (M5)	0.10 (GLM)	1,383	1,333	1,386
Megech	0.05 (RF)	0.04 (GAM)	2,035	1,637	2,028
Ribb	0.21 (RF)	-0.01 (RF)	2,575	1,969	2,615

Table 5. Covariate importance measurements from each basin's model

Model type	Linear model				Generalized additive model		M5 model tree		Random forest
Measure of influence	Linear regression coefficients and associated p-values				Estimated degrees of freedom (EDF) and associated p-values		Covariate usage in tree rules and model coefficients		Increase in MSE when covariate is randomly permuted
Basin	Gilgel Abbay		Koga		Megech		Gumara		Ribb
Covariate	Coefficient estimate	P-value	Coefficient estimate	P-value	EDF	P-value	Tree rules	Model coefficients	Percent increase in MSE
Prec	0.22	< 0.01	0.24	< 0.01	1.346	< 0.01	5%	58%	7.71%
Prec (lag 1)	0.10	0.03	0.16	< 0.01	0.624	0.08	0%	19%	2.79%
Prec (lag 2)	0.01	0.74	0.05	0.26	0	0.29	0%	0%	1.10%
Temp	-0.09	0.08	-0.07	0.17	1.023	0.07	0%	47%	12.74%
Temp (lag 1)	-0.04	0.49	-0.06	0.22	0	0.32	0%	46%	4.97%
Temp (lag 2)	-0.01	0.81	-0.09	0.08	0	0.56	0%	0%	8.16%
Agr. LC	0.00	0.33	0.02	0.01	1.986	< 0.01	86%	73%	15.21%

Table 6. Mean and standard deviation values for temperature, wet-season rainfall, and dry-season rainfall in each basin.

	Temperature (°C)		Wet season rainfall (mm/month)		Dry season rainfall (mm/month)	
	Mean	SD	Mean	SD	Mean	SD
Gilgel Abbay	15.7	1.54	206	145	39.3	56.5
Gumara	17.7	1.55	186	137	29.0	43.6
Koga	15.7	1.54	206	145	39.3	56.5
Megech	20.6	1.75	234	118	41.4	60.9
Ribb	18.2	1.61	263	115	45.8	57.0

Figure 1. Map of Lake Tana and surrounding rivers

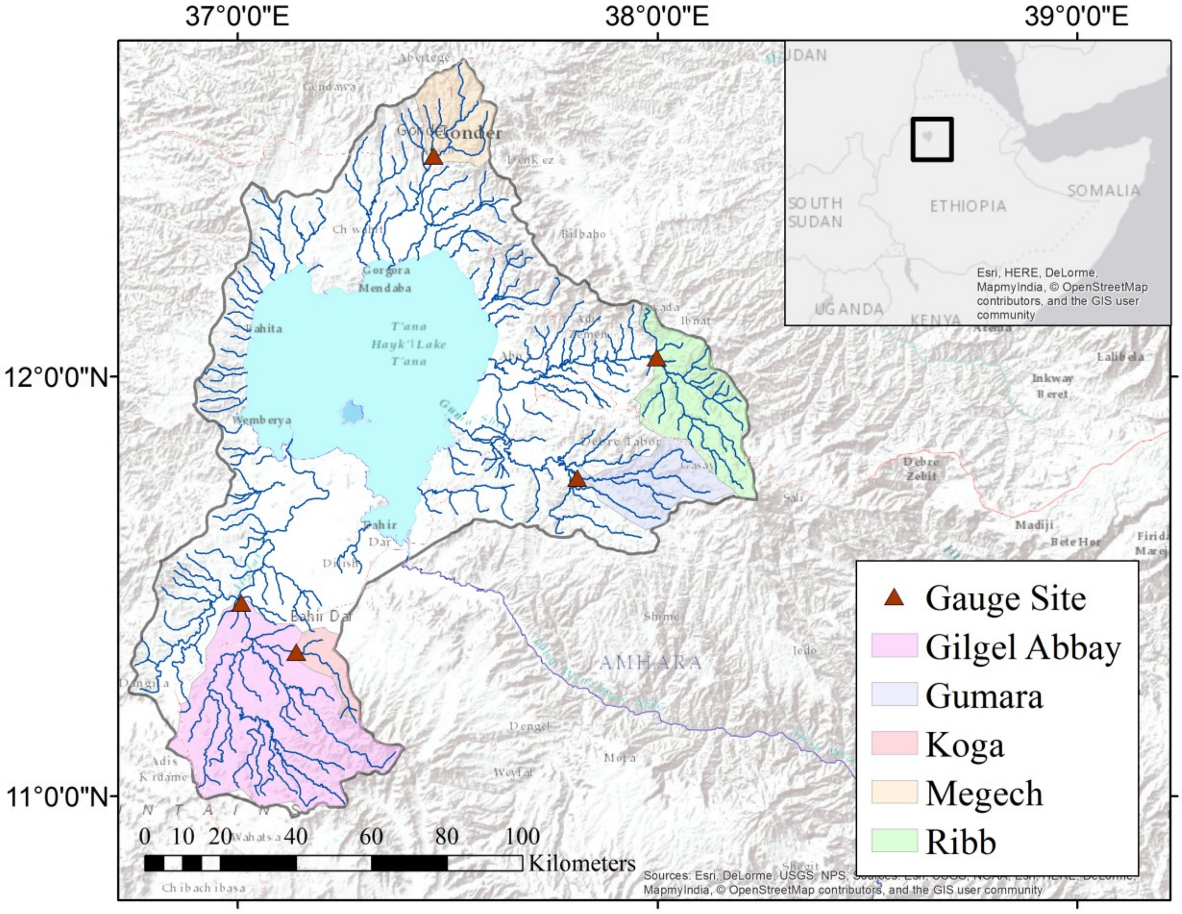

Figure 2. Autocorrelation in model residuals for the Gilgel Abbay and Ribb Rivers

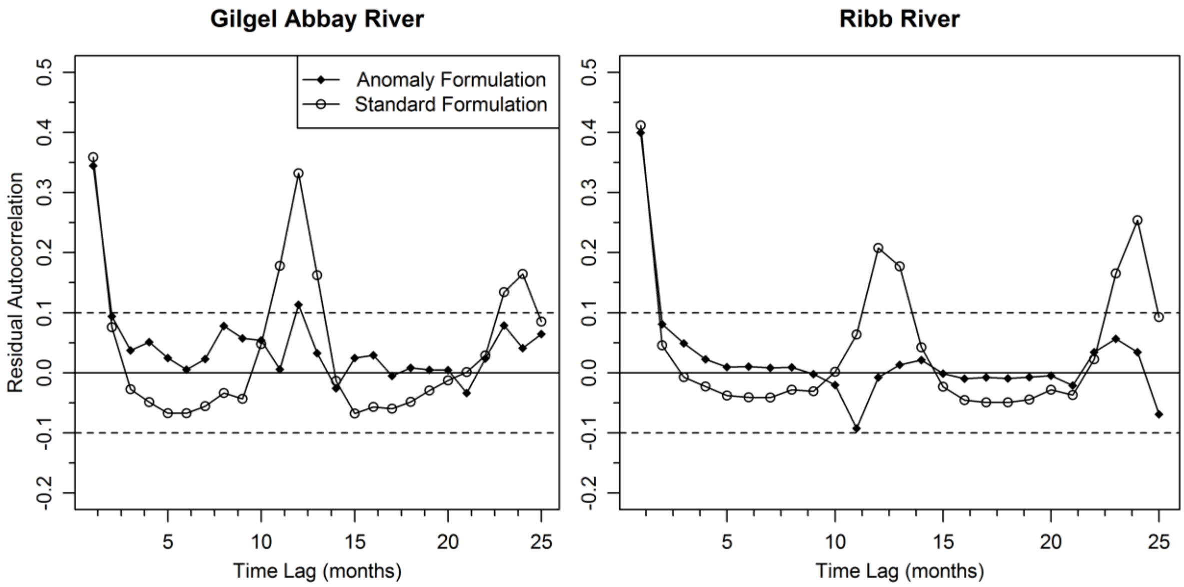

Figure 3. Example observed and predicted flows from the standard formulation RF model and anomaly formulation M5 model for the Gumara River from 198~~9~~5 to 1991~~2000~~.

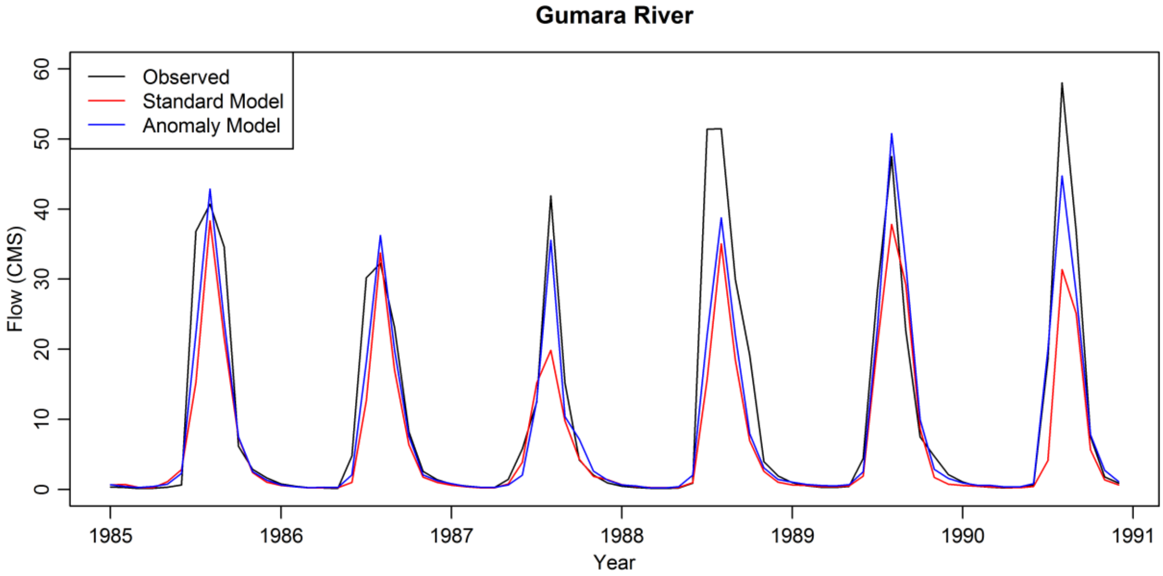

Figure 4. Plots of the smoothing functions used in the Megech River GAM. Hash marks along the x-axis indicate observation values of each covariate.

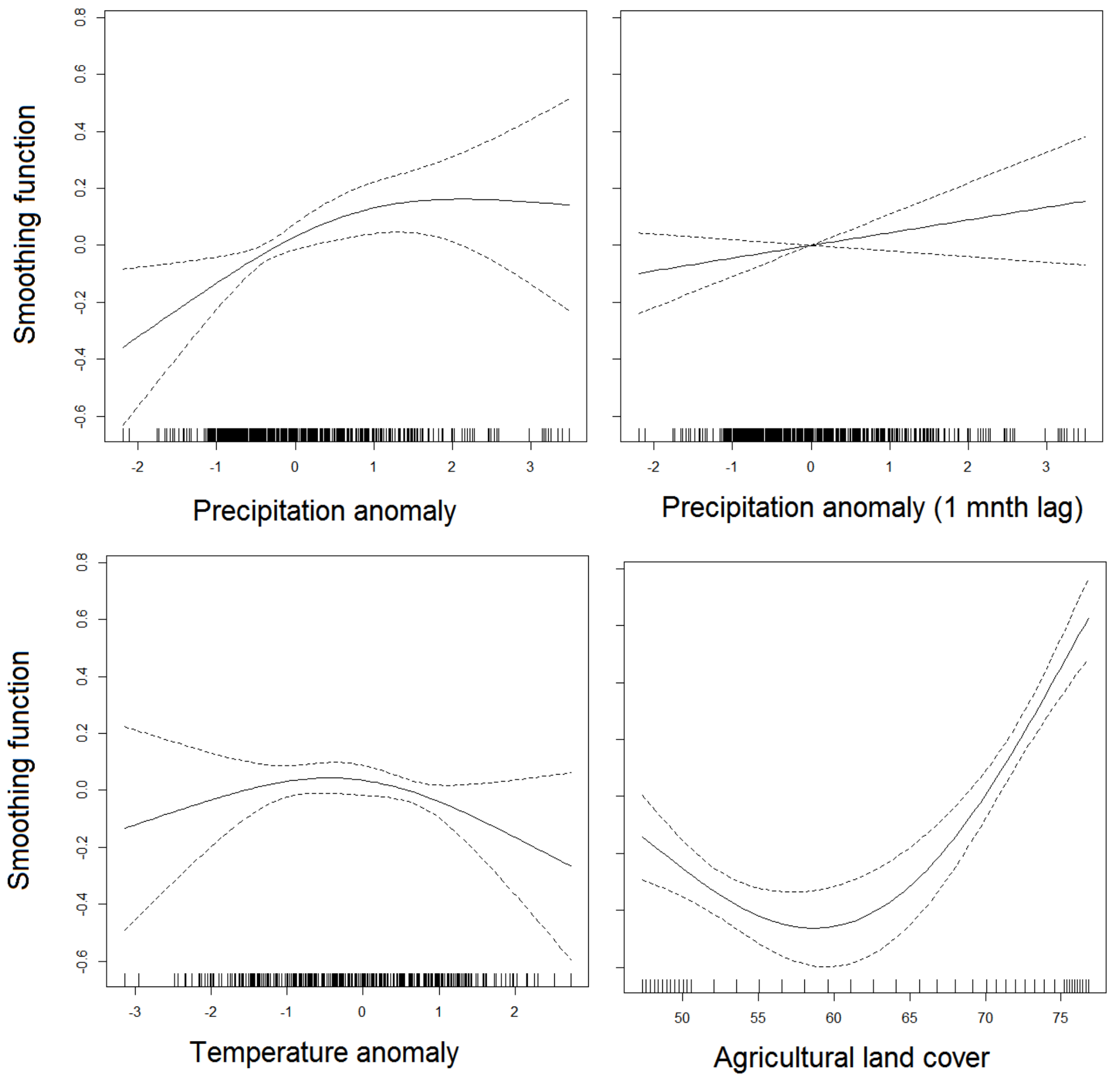

Figure 5. Partial dependence plots for the Ribb River random forest model. Hash marks along the x-axis show covariate sample decile values.

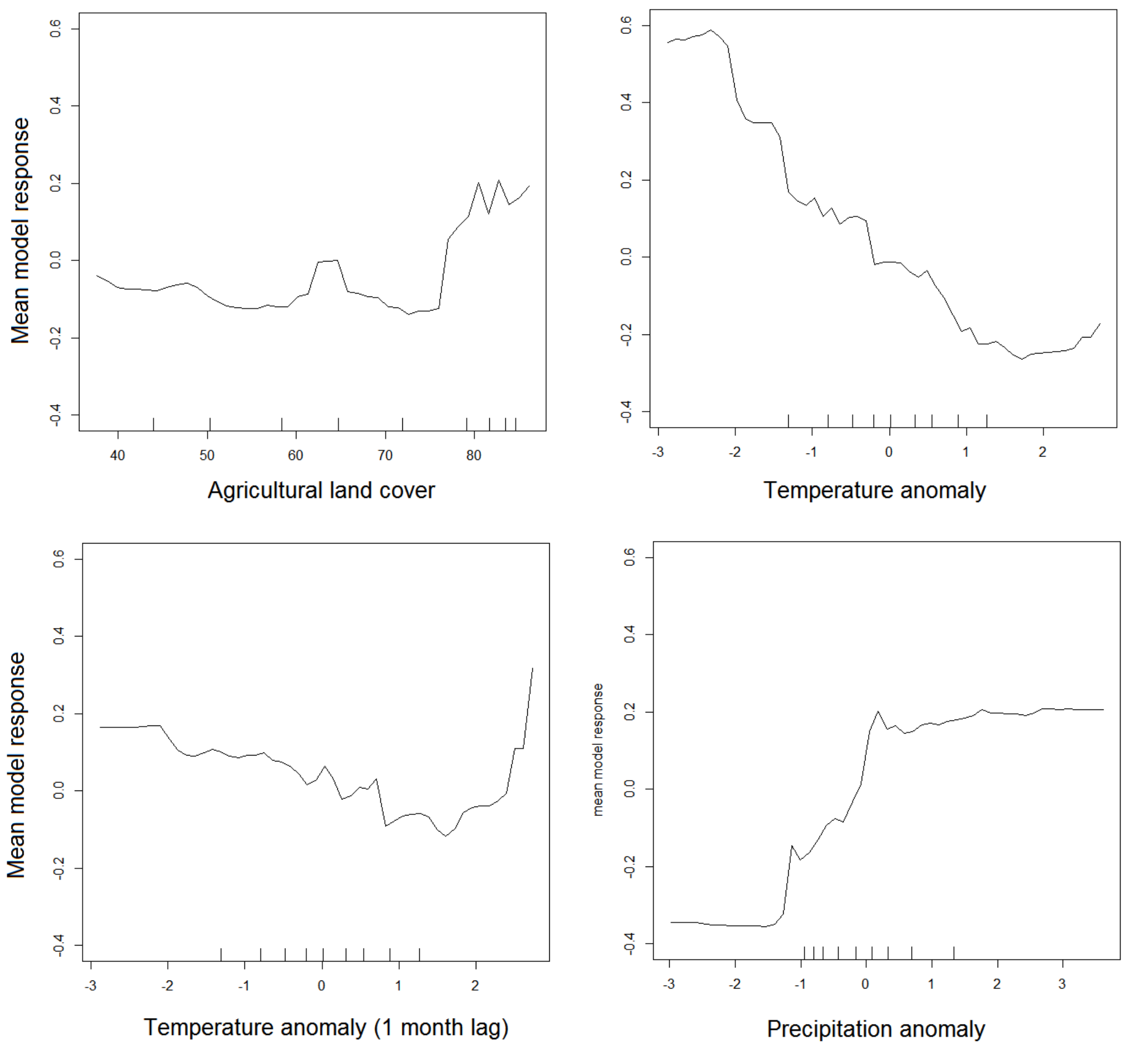

Figure 6. Partial dependence plots for climate covariates in the highest performing model in each basin. Model type is indicated in parentheses.

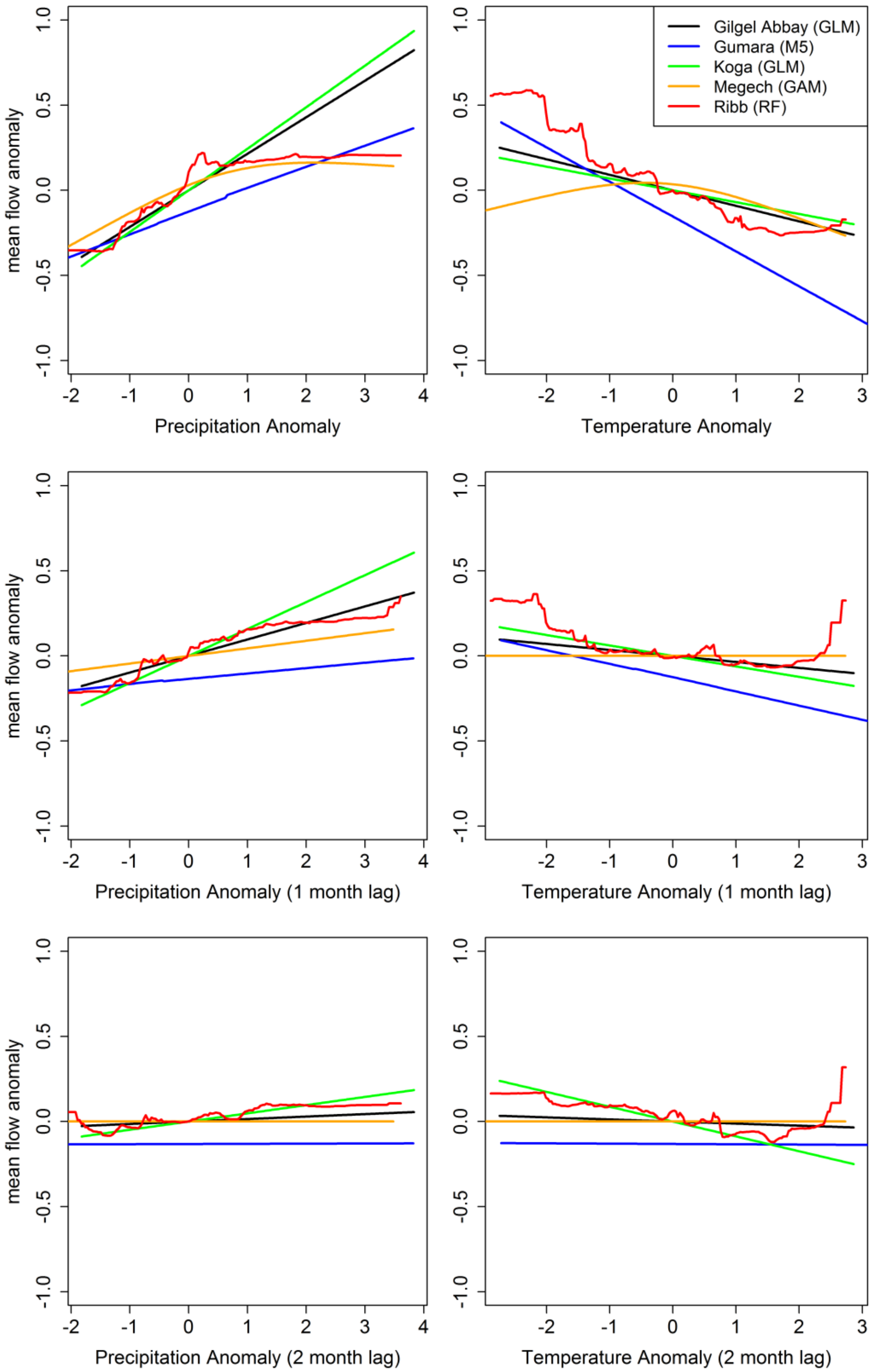

Figure 7. Partial dependence plot for agricultural land cover in the highest performing model in each basin. Model type is listed in parentheses for each basin. Dashed lines indicate values that exceed historic levels of agricultural land cover experienced in that basin.

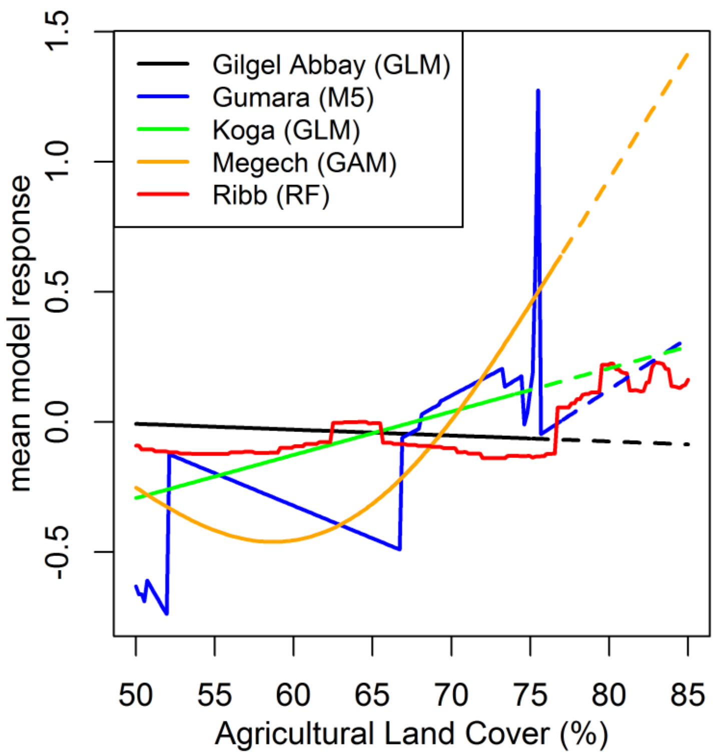

Figure 8. Projected changes in total streamflow (relative to current long-term average) under changing climate conditions. The top two panels show the sensitivity to changes in temperature and precipitation when they are varied independently. The bottom panel shows sensitivity to changing temperature in conjunction with decreasing (left panel) and increasing (right panel) precipitation. Dashed lines represent 95% confidence bounds from bootstrap resampling.

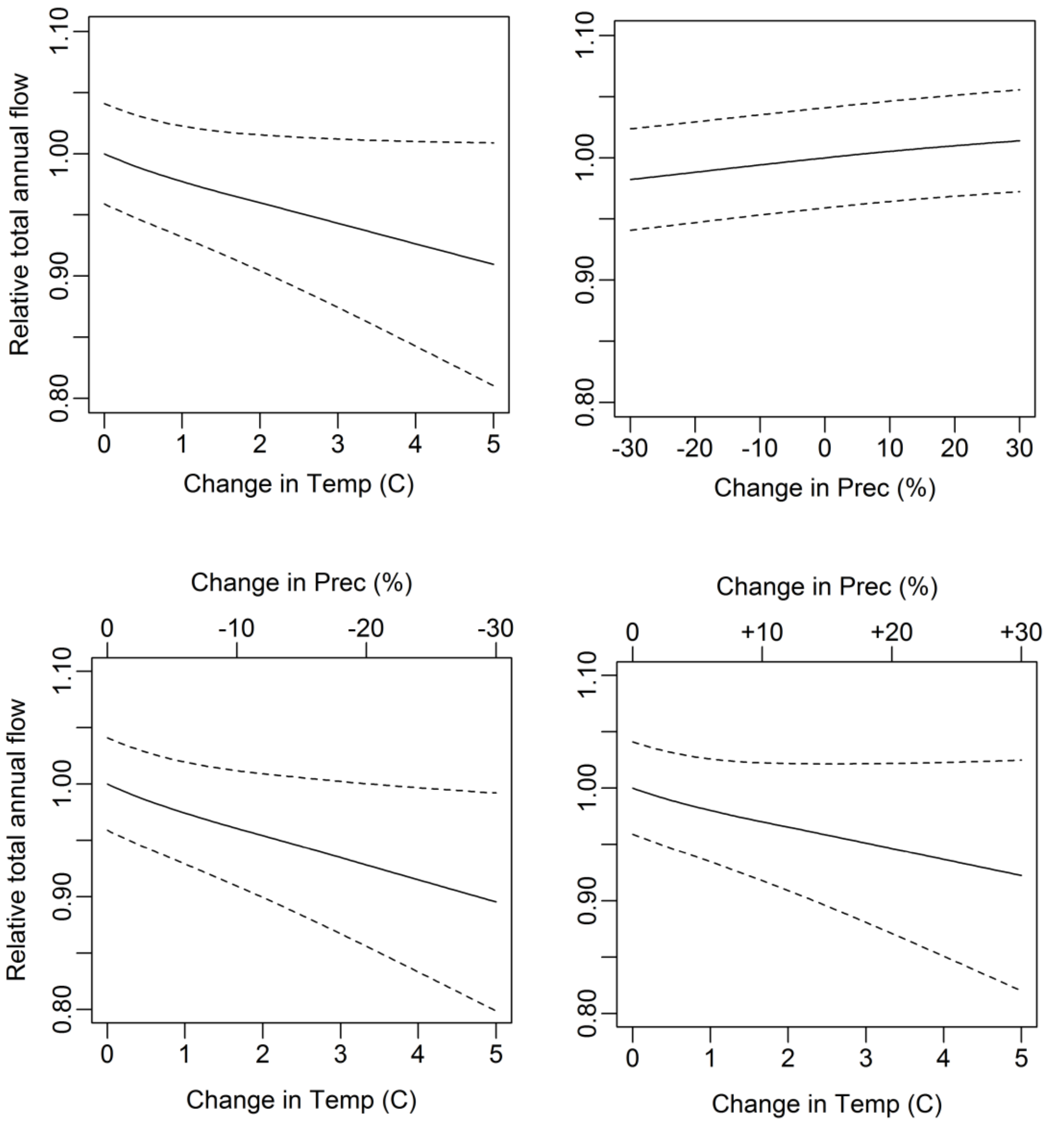

Figure 9. Changes in the coefficient of variation across bootstrap resamples from the highest performing model in each basin (left panel) and multiple models all applied to the Gumara basin (right panel).