Abstract. ....

1 Introduction

Watersheds, with their common points of drainage and collection into downstream environments, are important to the modeling of environmental change and runoff of pollutants. Software models have long been used to estimate changes to watersheds. These models typically involve defining physical characteristics of the watershed such as elevation, drainage patterns, and soil types, along with a calibration step against measured environmental data.

In this paper, we examine methods to encapsulate and package existing modeling approaches using a multivariate random forest regression approach. The aim is to predict yearly changes in runoff can be predicted from changes in precipitation or forestry. The end result captures the mathematical modeling of a traditional hydrology transport model, while allowing us to isolate specific predictors and their impact on a variety of measured downstream pollutants. Furthermore, we present a standard way of packaging these models so that end-users, such as policy makers, can make land-use decisions without the requisite technical knowledge to calibrate hydrology models.

2 Materials and Methods

2.1 Study Area

The South Nation Watershed spans 381,100 hectares between the Ottawa River to the watershed's North and the St. Lawrence River to the South. Mostly comprised of farm fields, the watershed drains both into the South Nation River, with eventual confluence at the Ottawa River, and into the Little Castor River. The watershed has been of increasing importance in recent years because of the wide-spread agricultural use and urbanization in towns along the South Nation River from such as Chesterville, Casselman, and Plantaganet along the South Nation River.

2.2 Hydrological Modeling

In order to encompass topographic and other physical characteristics of the watershed, we built a general purpose rainfall model using the Soil and Water Assessment Tool (SWAT) and used outputs from the SWAT model to train random forest models. The resulting steps were:

1. Setup of a rainfall model using the Soil and Water Assessment Tool (SWAT)
2. Calibration of the SWAT model using measured water quality data
3. Simulation of runoffs using the SWAT model
4. Fitting random forest models using the simulated values from SWAT
5. Evaluating the models' accuracies using a Time Series Split

6. Generating final random forest models and packaging into a application for end-users

After creating the SWAT model, it was calibrated using water quality data from the Ontario Provincial (Stream) Water Quality Monitoring Network, and outputs were simulated from the model. The final predictands from our SWAT model were: water outflow rate, evaporation rate, water transmission loss rate, sediment output, sediment concentration, organic nitrogen output, organic phosphorus output, nitrate output, ammonium output, nitrogen dioxide output, mineral phosphorus output, chlorophyll alpha output, carbonaceous biochemical oxygen demand of material output, dissolved oxygen output, and total nitrogen in surface runoff.

Using the simulated values from SWAT, we fitted a random forest model for each predictand. For predictors, we isolated the forest cover and precipitation for each subbasin in the watershed; with 31 subbasins, this creates 62 individual predictors for the random forest model, though not all predictors are included as nodes in the model after training. The random forest models were created using Scikit-Learn (SKLearn), a Python-based machine learning library. Flask, a server-side web framework, ReactJS, a client-side JavaScript library, and Docker, a virtualization program, was used to package the models for access through a web browser by end-users.

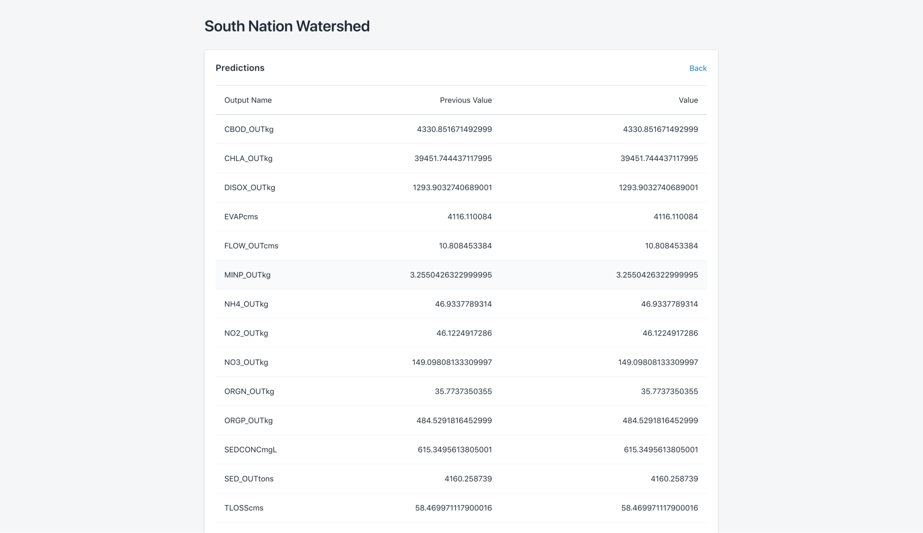
3 Results

3.1 Interface

The models are accessible to users via any standard web browser. Users can enter measured or expected precipitation metrics across the watershed, and individual forest cover metrics per subbasin. The software can then predict, within seconds, the expected downstream runoff measures for the aforementioned predictands and present the results to users as shown in Figure 1.

3.2 Projected Changes in Hydrological Variables

The random forest models were used to predict changes in downstream runoff values given a range of - 50% to + 50%change in forest cover or a range of -15% to + 15% change in precipitation across the watershed, which are within the range of observed values. Random forest regression models are inherently poor at extrapolating outside of data seen in its training set due to the branching and averaging of nodes, and the generated predictor values were chosen to be within this range. The individual predictions as a factor of percent change from baseline are provided in Figure 2 for changes in forest cover and Figure 3 for changes in precipitation.



South Nation Watershed

Predictions

Output Name	Previous Value	Value
CBOD_OUTkg	4330.851671492999	4330.851671492999
CHLA_OUTkg	39451.744437117995	39451.744437117995
DISOX_OUTkg	1293.9032740689001	1293.9032740689001
EVAPcms	4116.110084	4116.110084
FLOW_OUTcms	10.808453384	10.808453384
MINP_OUTkg	3.2550426322999995	3.2550426322999995
NH4_OUTkg	46.9337789314	46.9337789314
NO2_OUTkg	46.1224917286	46.1224917286
NO3_OUTkg	149.09808133309997	149.09808133309997
ORGN_OUTkg	35.7737350355	35.7737350355
ORGP_OUTkg	484.5291816452999	484.5291816452999
SEDCONCmgL	615.3495613805001	615.3495613805001
SED_OUTtons	4160.258739	4160.258739
TLOSScms	58.469971117900016	58.469971117900016

Fig. 1. The predicted changes in downstream runoff metrics based on precipitation and forest cover inputs, as presented to the user.

3.3 Model Accuracy

Table 1 provides a summary of the models' accuracies in predicting various downstream measures. The accuracies were determined using a Time Series Split where the temporal ordering is kept between training and test data sets. For example, with a dataset consisting of years 2000 to 2005 inclusive, a time series split might create a training set with year 2000, 2001, 2002 and test set with year 2003. As shown, the random forest models performed poorly for predicting water outflow rate (FLOW_OUTcms) and organic phosphorous outflow (ORGP_OUTkg). That is, if the models were trained with x sets of prior years, the models are poor at predicting a future year's water outflow or organic phosphorous outflow. The model performed significantly better for predicting the other measures: evaporation rate, water transmission loss rate, sediment output, sediment concentration, organic nitrogen output, nitrate output, ammonium output, nitrogen dioxide output, mineral phosphorus output, chlorophyll alpha output, carbonaceous biochemical oxygen demand of material output, dissolved oxygen output, and total nitrogen in surface runoff.

4 Conclusions

References

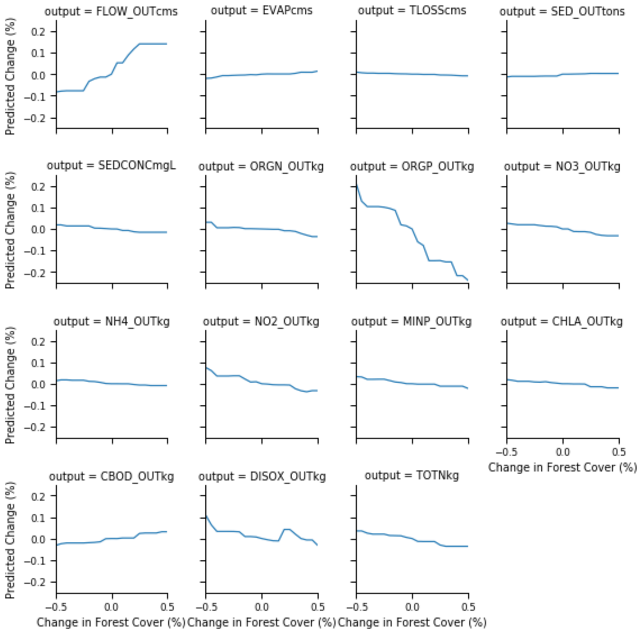


Fig. 2. Predicted percent changes given a -50% to +50% change in the watershed's overall forest cover.

Table 1. Accuracy of random forest models for the downstream predictands.

	accuracy	standard deviation
FLOW_OUTcms	0.77453974177041	0.2294149291032354
EVAPcms	0.9697462141575514	0.03342663731811072
TLOSScms	0.9849085443489066	0.015783200164635483
SED_OUTtons	0.9669386142355473	0.031369149259199765
SEDCONCmgL	0.9550856665204039	0.04719784299228629
ORGN_OUTkg	0.9485722009887052	0.0513537106801331
ORGP_OUTkg	0.8288877495883687	0.21551718529974206
NO3_OUTkg	0.960748053541662	0.03156239767529165
NH4_OUTkg	0.9505648610245114	0.04504648732478451
NO2_OUTkg	0.9166867771433014	0.08570031297945056
MINP_OUTkg	0.9692740147487756	0.027323554055449954
CHLA_OUTkg	0.9709013616816519	0.02554074062863536
CBOD_OUTkg	0.9707570596314152	0.02727914422698616
DISOX_OUTkg	0.9234165773438077	0.07587513874148259
TOTNkg	0.9172299325385997	0.09975026057106322

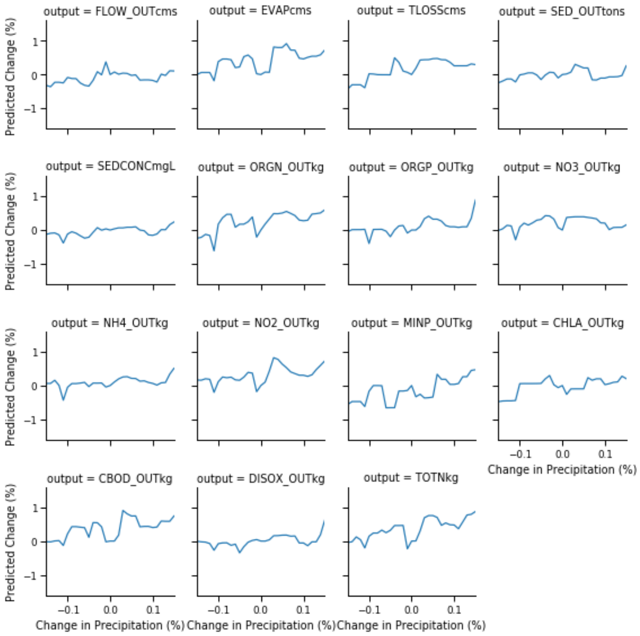


Fig. 3. Predicted percent changes given a -15% to +15% change in the watershed's overall forest cover.