9th International Workshop on Climate Informatics
October 02-04, 2019

Hosted by École Normale Supérieure, Paris, France

DATA-DRIVEN VS. PHYSICALLY-BASED STREAMFLOW PREDICTION MODELS

Martin Gauch,[1] Juliane Mai,[2] Shervan Gharari,[3] Jimmy Lin[1]

***Abstract*—Climate change leads to more frequent and severe floods and droughts. Precise water flow forecasts for rivers and streams help mitigate damage and are direly needed. We evaluate physically-based and data-driven models on the task of streamflow prediction in the Lake Erie region: Physically-based models capture simplified representations of the physical processes that underlie streamflow, while purely data-driven models encode no such knowledge explicitly. Experiments show that data-driven approaches can provide more accurate predictions than a physically-based model, suggesting potential in hybrid approaches that combine hydrological understanding with high prediction accuracy.**

I. INTRODUCTION

Accurate prediction of *streamflow*—the amount of water that flows through a river at a certain time—plays a vital role in managing extreme floods and droughts. Due to climate change, such disasters have become increasingly frequent and impact the lives of people around the world. Hydrology has a long history of developing streamflow prediction models: for different watersheds, based on different datasets, and based on different evaluation criteria. However, it often remains unclear which model is best under which conditions. The ongoing *Great Lakes Runoff Inter-comparison Project for Lake Erie* (*GRIP-E*) compares hydrologic models in the largest Canadian effort yet to overcome these issues [1].

GRIP-E mostly considers *physically-based* hydrologic and land-surface models; by this, we mean models that replicate simplified representations of the underlying physical processes to predict streamflow. We, however, believe that *data-driven*, machine-learning models can in fact provide meaningful contributions towards understanding streamflow, too. Although researchers have been hesitant to adopt machine-learning models that are often black-box predictors, our work shows that data-driven models can aid hydrologists' advancement in explaining the physical processes that underlie streamflow. Purely data-driven models reveal how much streamflow information is extractable from the datasets that are used in physically-based models.

Our study compares a physically-based model with data-driven linear and tree-based models in their ability to accurately predict the water flow of a stream at a particular gauging station, given meteorological data of the surrounding area. We use a five-year meteorological dataset of the Lake Erie watershed to predict the streamflow at gauging stations in sub-watersheds around the lake. The purely data-driven approaches predict streamflow more accurately than the physically-based model. To us, this is good news, as it shows that there is sufficient signal in existing data to make accurate predictions, and points to potential hybrid models that are both useful for advancing hydrological understanding and making high-quality forecasts.

II. DATA AND METHODS

As streamflow ground truth, we use daily measurements at 46 gauging stations in the Lake Erie region from 2010 to 2014. These stations divide the watershed into sub-watersheds, each comprised of the area in which all water flows towards the gauging station. Figure 1 shows a map of the gauging stations used for GRIP-E and their corresponding sub-watersheds.

Both physically-based and data-driven models use gridded meteorological *forcing data* as input. In hydrology, forcing data are time-series datasets that are required to run, or *force*, the model. Many physically-based models additionally use geophysical inputs such as soil and elevation maps that change little over time. The forcing data used in this study include hourly meteorological variables of temperature, precipitation, pressure, wind speed, specific humidity, and short- and longwave radiation with a spatial resolution of around 15 km spanning five years (2010 to 2014). Table I

Corresponding author: Martin Gauch, mgauch@uwaterloo.ca
[1]David R. Cheriton School of Computer Science, University of Waterloo, ON, Canada; [2]Civil and Environmental Engineering, University of Waterloo, ON, Canada; [3]Coldwater Lab, University of Saskatchewan, AB, Canada

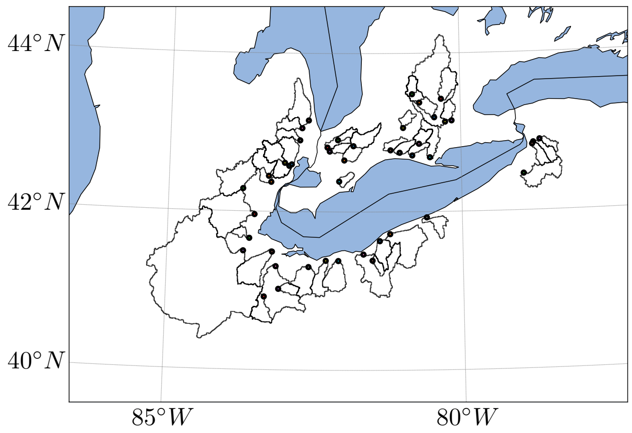

Fig. 1: Geographical outlines of the 46 sub-watersheds in our analysis, each draining towards a gauging station (black dots).

Variable	Explanation	Level	Unit
PR0	Quantity of precipitation	surface	m
TT	Air temperature	40 m	°C
FB	Downward solar flux	40 m	$\mathrm{W\,m^{-2}}$
FI	Surface incoming infrared flux	40 m	$\mathrm{W\,m^{-2}}$
P0	Surface pressure	surface	mbar
HU	Specific humidity	40 m	$\mathrm{kg\,kg^{-1}}$
UVC	Wind speed	40 m	kn

TABLE I: Meteorological forcing variables used in this study. Each variable covers the entire Lake Erie watershed at a resolution of around 15 km for the years 2010 to 2014 at an hourly resolution. The variables are available at the different vertical levels indicated.

summarizes more details about the variables. Figure 2 visualizes a snapshot of the temperature forcing data.

Formally, we aim to solve a regression problem. We predict the streamflow y^S_t at station S at time t, given the history of meteorological forcings $\mathbf{X}^S_{[1,t]}$:

$$\begin{aligned}\mathbf{X}^S_{[1,t]} &= \left[\mathbf{x}^1_1, \dots, \mathbf{x}^1_t, \dots, \mathbf{x}^{p_S}_1, \dots, \mathbf{x}^{p_S}_t\right] \\ &= \left[\mathbf{x}^1_{[1,t]}, \dots, \mathbf{x}^{p_S}_{[1,t]}\right] \in \mathbb{R}^{7\times(t\cdot p_S)} \qquad (1)\end{aligned}$$

The superscripts $1, \dots, p_S$ identify the p_S grid cells in the sub-watershed of station S, the subscripts $1, \dots, t$ represent time steps, and each $\mathbf{x}^c_i$ is a vector of seven forcing variables (Table I). Since we use machine-learning models that operate on vectors rather than matrices, we introduce the following vectorization:

$$\mathbf{x}^S_{[1,t]} = \mathrm{vec}(\mathbf{X}^S_{[1,t]}) = \begin{bmatrix}\mathbf{x}^1_1 \\ \vdots \\ \mathbf{x}^{p_S}_t\end{bmatrix} \in \mathbb{R}^{7\cdot(t\cdot p_S)} \qquad (2)$$

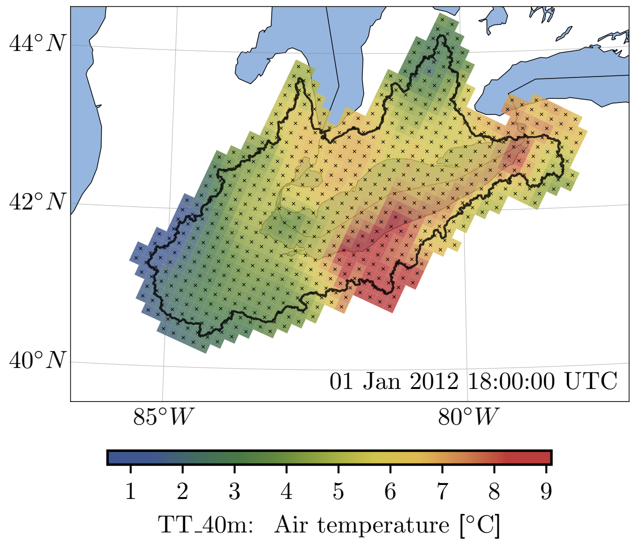

Fig. 2: Gridded forcing data for the Lake Erie watershed (black outline). As an example, the temperature of Jan 1, 2012 6pm (UTC) is depicted (colored tiles). Each tile is about $15 \times 15\,\mathrm{km}^2$ in size.

To obtain predictions, we train the parameters ω of a model f to output an estimate $\hat{y}^S_t = f(\mathbf{x}^S_{[1,t]}; \omega)$.

A. *Physically-Based Hydrologic Models*

Physically-based hydrologic models capture a simplified simulation of the underlying physical processes that result in streamflow and other hydrologic fluxes. These models often use various sources of data such as land cover maps, soil maps, or digital elevation models as their setup basis. Similar to data-driven models, the parameters ω of a hydrologic model are usually trained, or calibrated, against the observed streamflow time series at one or multiple gauges.

For this study, we make use of the *Variable Infiltration Capacity model based on Grouped Response Units* (*VIC-GRU*). This model originates from the *VIC* model [2], [3], which is a large-scale, semi-distributed hydrologic model that simulates each grid cell independently. VIC-GRU is a variant of VIC that processes spatial extents with similar characteristics as so-called *grouped response units* (GRUs). This grouping makes the simulation computationally more efficient, which allows us to use input data at a higher resolution.

We train one VIC-GRU model on all gauging stations, as the model already incorporates varying spatial characteristics through geospatial input information such as soil maps. As physically-based models approximate natural system states and fluxes, they need to attain realistic initial model conditions for the training period before generating accurate output. To evaluate the model's goodness-of-fit, we discard the first year of model simulations (2010) as the so-called *warm-up period*, and only use the NSE coefficient for the training period 2011 to 2012. We use the parameter set that generates the best NSE values in the training period to predict the test period 2013 to 2014.

B. Machine-Learning Models

We use a cross-validated random search to find suitable parameters for each model. To reduce dimensionality, we only use the meteorological forcing data of the p_S grid cells in sub-watershed S. As the models we use neither naturally incorporate temporally-distributed nor spatially-distributed input, we flatten the data to a fixed history window of eight days and train one model per gauging station. We further aggregate the hourly forcing data into daily values to match the target streamflow data resolution. This aggregation uses the minimum and maximum temperature per day and total precipitation on that day. Preliminary experiments show that the remaining forcing variables do not improve prediction accuracy (results not shown). We therefore exclude them from the inputs for the data-driven models.

As a baseline, we train a linear ridge regression model to predict streamflow. Linear regression finds a parameter vector $\omega \in \mathbb{R}^{7\cdot(8\cdot p_S)}$ that minimizes the sum of squared residual differences between target values y_t^S and predicted values $\hat{y}_t^S = \mathbf{x}_{[t-7,t]}^S \omega$. As our problem involves high-dimensional data, ridge regression is an appropriate choice because it includes a weighted regularization term to reduce overfitting.

We also employ XGBoost as a more sophisticated approach that trains gradient-boosted regression trees (GBRTs) [4]. GBRTs iteratively train K regression trees f_k and generate an overall prediction $\hat{y}_t^S$ as the sum of their outputs. Additionally, GBRTs provide regularization parameters such as a maximum tree depth to control overfitting. For more details on the objective function minimization, see Chen and Guestrin [4].

C. Evaluation

We split the available data into a training period from 2010 to 2012 and a test period from 2013 to 2014. Our data-driven models are trained using mean squared error (MSE). After fitting a model during the training phase, we apply it to the test period and evaluate its prediction accuracy. Following common practice in hydrology, we use the *Nash-Sutcliffe efficiency coefficient* (NSE) to evaluate the simulated streamflow $\hat{y}^S$ compared to the observed streamflow time series y^S for station S, defined as follows:

$$\begin{aligned} \text{NSE} &= 1 - \frac{\sum_{t=1}^{T} (\hat{y}_t^S - y_t^S)^2}{\sum_{t=1}^{T} (y_t^S - \bar{y}^S)^2} \\ &= 1 - \frac{\text{MSE}}{\frac{1}{T}\sum_{t=1}^{T} (y_t^S - \bar{y}^S)^2} \end{aligned} \tag{3}$$

where $\bar{y}^S$ is the mean observed streamflow at station S across all T time steps. Hence, the denominator is the variance of the streamflow observations y^S. Equation 3 shows that NSE and MSE are strongly correlated [5]. NSE values range between $-\infty$ and 1, with 1 representing perfect predictions. A score of 0 is obtained when predicting $\bar{y}^S$ at every time step.

Statistic	VIC-GRU	Ridge regression	XGBoost
p_0	-6.302	-1.677	-0.206
p_{25}	0.184	0.298	0.412
p_{50}	0.328	0.380	0.522
p_{75}	0.376	0.469	0.561
p_{100}	0.597	0.585	0.666
$p_{75} - p_{25}$	0.191	0.170	0.149

TABLE II: Minimum p_0, maximum p_{100}, quartiles p_{25} and p_{75}, median p_{50}, and interquartile range $p_{75} - p_{25}$ of the NSE distributions for the physically-based model VIC-GRU and the two data-driven models (ridge regression and XGBoost).

III. RESULTS

Table II outlines characteristics of the three models' NSE distributions across all 46 gauging stations. The third row (p_{50}) shows the median NSE values, and the other rows list the measure at different percentiles as well as the interquartile range (see table caption for details). We see that a simple ridge regression model outperforms the physically-based VIC-GRU model, and that XGBoost provides the most accurate predictions of the three examined models. The XGBoost model outperforms VIC-GRU in 40 of the 46 stations by an average NSE difference of 0.473. Not only are the XGBoost predictions more accurate overall, but they also show fewer outliers (i.e., make terribly bad predictions) and a smaller variation of NSE coefficients for different stations.

In Figure 3, we provide an overview of results for three sample gauging stations: one where all models perform relatively well, one where VIC-GRU performs

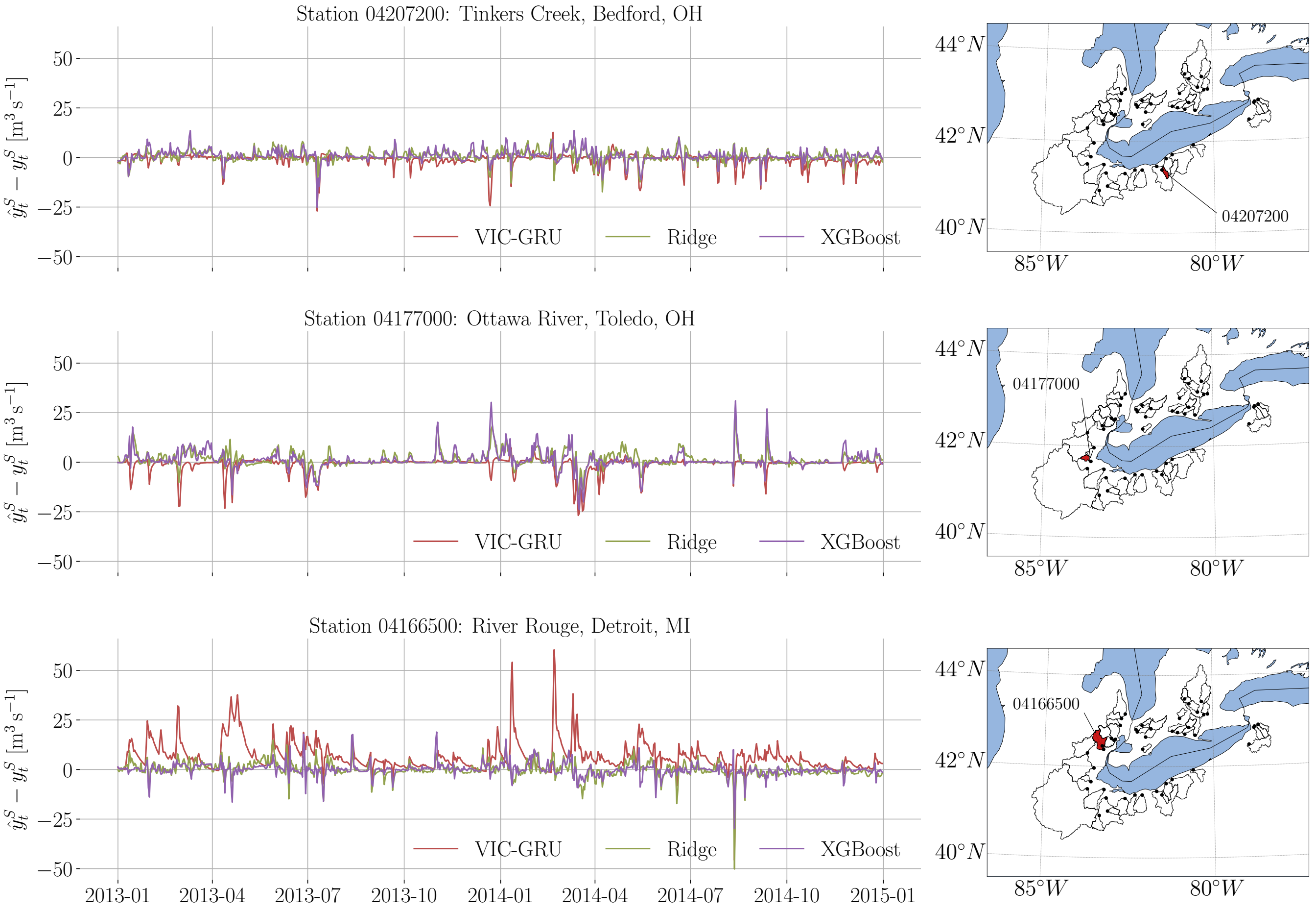

Fig. 3: Differences between actual streamflow y_t^S and predictions $\hat{y}_t^S$ for VIC-GRU (red), ridge regression (green), and XGBoost (purple) at gauging stations 04207200, 04177000, and 04166500 during the test period.

better than machine-learning models, and one where the machine-learning models outperform VIC-GRU. Each panel shows the differences between the actual and predicted streamflows at each gauging station based on the three models; the corresponding figures on the right highlight the stations' geographical locations on a map.

All three models provide rather good predictions for station 04207200 in Bedford, OH, USA (Figure 3, top), with NSE values between 0.44 (VIC-GRU) and 0.57 (XGBoost). For station 04177000 in Toledo, OH, USA (Figure 3, middle), VIC-GRU yields a better NSE than XGBoost. Largely, this seems to be due to a few overestimated streamflow spikes at the end of 2013 and around August 2014. In the third example, VIC-GRU makes very poor predictions for station 04166500 in Detroit, MI, USA (Figure 3, bottom), with an NSE score well below zero, while XGBoost and ridge regression provide far better results. VIC-GRU appears to struggle with the prediction of streamflow peaks, as it frequently incorrectly predicts high streamflows above $50\,\mathrm{m}^3\,\mathrm{s}^{-1}$ during the winter and spring months. XGBoost and ridge regression are more conservative and rarely predict streamflows above $30\,\mathrm{m}^3\,\mathrm{s}^{-1}$, likely because the station's training data contain few high-streamflow examples. This difference partly explains the more accurate predictions of the data-driven models, because the NSE calculation includes squared differences that emphasize outliers. Ridge regression often produces erratic predictions for periods of low streamflow, which explains the model's lower NSE coefficients compared to XGBoost.

It is further noteworthy that station 04166500 is located in the highly urbanized metropolitan area of Detroit. Such regions are more prone to human water regulation, which is often not included in the assumptions physically-based models make. In contrast, machine-learning techniques are able to implicitly capture water regulation policies.

IV. DISCUSSION

Our results show that, in relative terms, imprecise predictions by VIC-GRU cannot be solely explained by insufficient data, as the purely data-driven approaches outperform the physical model using only a subset of the data. In other words, it appears that the model does not yet fully exploit the available signals due to its design and the restrictive assumptions it makes. This is more pronounced in urban regions, where we envision great potential in augmenting physically-based models with machine-learning techniques.

We note, however, that the three-year time frame from 2010 to 2012 is a rather short training period for a physically-based model. As the GRIP-E project is still ongoing and awaiting an extended forcing dataset, we are unfortunately unable to train on a longer time period. Given the large differences in prediction accuracy, we however do not expect that additional data would fundamentally change our findings, especially since more data would benefit data-driven models also.

V. FUTURE WORK

Unfortunately, our data-driven models do not yet provide insights into *how* physically-based models could be improved. In future work, we therefore plan to study machine-learning models whose structures are more specifically targeted towards predictions on geospatial time series. Such models might allow for more interpretability as they more closely resemble the mechanics of physically-based models. With these models, we could further train one model to predict streamflow at arbitrary locations, which allows for more detailed comparisons to physically-based models.

Besides ridge regression and XGBoost, we have begun to explore predictions with neural networks, specifically simple long short-term memory cell (LSTM) architectures. In preliminary experiments, however, these neural models did not perform better than XGBoost. We assume that more sophisticated deep-learning approaches that are better suited for spatially-distributed input data would improve prediction accuracy.

VI. CONCLUSION

Our study shows that data-driven approaches predict streamflow more accurately than a physically-based model for the specific case study that we examined. The more rigid structure of a physically-based model prevents it from fully exploiting signals that are available in the input data. Especially for stations in highly urbanized regions, our data-driven models are better able to adapt to patterns of human water regulation.

From a high-level perspective, our project has the goal to both deliver more accurate streamflow predictions and advance hydrological understanding. We have taken only a small step in this direction, but are excited about future prospects.

ACKNOWLEDGMENTS

The authors would like to thank the Global Water Futures program and the Integrated Modeling Program for Canada (IMPC) for financial support of Juliane Mai, Shervan Gharari, and Martin Gauch. The authors wish to thank Nicolas Gasset and Vincent Fortin from Environment and Climate Change Canada for making the meteorological forcing dataset available for the GRIP-E project.

REFERENCES

[1] J. Mai, B. Tolson, H. Shen, E. Gaborit, V. Fortin, M. Dimitrijevic, N. Gasset, D. Durnford, Y. L. Shin, T. A. Stadnyk, L. M. Fry, T. Hunter, A. Gronewold, J. Smith, L. Mason, L. Read, K. FitzGerald, K. M. Sampson, A. F. Hamlet, F. Seglenieks, S. Gharari, S. Razavi, A. Haghnegahdar, D. G. Princz, and A. Pietroniro, “The Great Lakes Runoff Inter-comparison Project for Lake Erie (GRIP-E),” *AGU Fall Meeting Abstracts*, 2018.

[2] X. Liang, D. P. Lettenmaier, E. F. Wood, and S. J. Burges, “A simple hydrologically based model of land surface water and energy fluxes for general circulation models,” *Journal of Geophysical Research*, vol. 99, no. D7, pp. 14415–14428, 1994.

[3] J. J. Hamman, B. Nijssen, T. J. Bohn, D. R. Gergel, and Y. Mao, “The Variable Infiltration Capacity model version 5 (VIC-5): infrastructure improvements for new applications and reproducibility,” *Geoscientific Model Development*, vol. 11, no. 8, 2018.

[4] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, pp. 785–794, ACM, 2016.

[5] H. V. Gupta, H. Kling, K. K. Yilmaz, and G. F. Martinez, “Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling,” *Journal of Hydrology*, vol. 377, no. 1-2, pp. 80–91, 2009.