Streamflow Prediction with Limited Spatially-Distributed Input Data

Martin Gauch,[1]* **Juliane Mai,**[2] **Shervan Gharari,**[3] **and Jimmy Lin**[1]
[1]David R. Cheriton School of Computer Science, University of Waterloo, ON, Canada
[2]Civil and Environmental Engineering, University of Waterloo, ON, Canada
[3]University of Saskatchewan Coldwater Lab, Canmore, AB, Canada

Abstract

Climate change causes more frequent and extreme weather phenomena across the globe. Accurate streamflow prediction allows for proactive and mitigative action in some of these events. As a first step towards models that predict streamflow in watersheds for which we lack ground truth measurements, we explore models that work on spatially-distributed input data. In such a scenario, input variables are more difficult to acquire, and thus models have access to limited training data. We present a case study focusing on Lake Erie, where we find that tree-based models can yield more accurate predictions than both neural and physically-based models.

1 Introduction

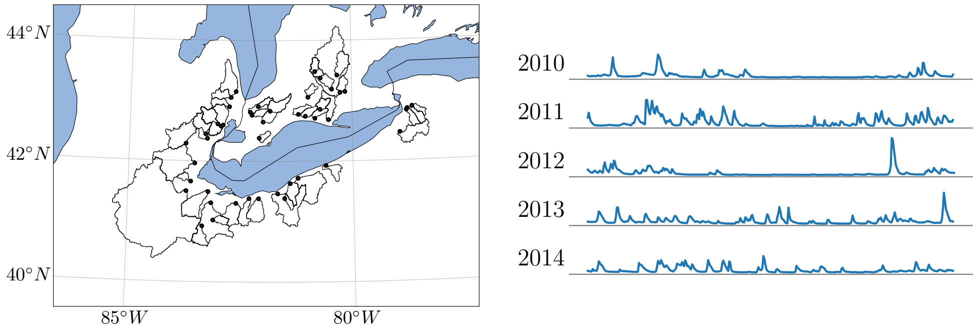
Water levels of lakes and rivers have a large impact on surrounding communities. Low water levels during droughts can jeopardize drinking water supply, industrial water use, and shipping routes. High water levels, on the other hand, pose the risk of flooding and endanger not just humans, but also flora and fauna. Researchers believe that climate change makes both extremes of high and low water levels more common [6]. As an example, a recent New York Times article reports not only on the danger to homes, businesses, and wildlife of the 2019 floods in the Great Lakes region, but also the huge societal costs that protective measures incur [14]. Precise forecasts of the flow of rivers or streams play a vital role in managing water levels and thus mitigating damage [3].

To this end, we focus on the problem of streamflow prediction: predicting the amount of water that flows past a gauging station (typically located in a river), based on meteorological variables and other information such as topology, soil, or land cover data. For example, Figure 1a shows a map of gauging stations around Lake Erie, and Figure 1b visualizes the hydrograph of a specific station. In this context, we are exploring the complementary roles of physically-based and data-driven models. Instances of the former type explicitly simulate simplified representations of the processes that underlie streamflow, while instances of the latter type do not. Numerous considerations include prediction accuracy, the ability to generalize to different watersheds and so-called ungauged basins, and uncovering scientific insights about underlying processes.

Applications of neural networks to streamflow prediction date back more than two decades [1, 2, 7]. The recent work by Kratzert et al. [9] takes advantage of abundant historical data—over a decade—to make accurate predictions across hundreds of locations using recurrent neural networks (RNNs), which do indeed benefit from abundant data [5]. We examine a slightly different setup, described by Gauch et al. [4]: instead of “pointwise” sources of input variables (corresponding to specific sensors in the environment; for example, temperature readings at a specific location), our models take as input spatially-distributed measurements (for example, temperature at points on a grid across a watershed). Such observations are more difficult to acquire and frequently not available across

*Corresponding author: Martin Gauch, `martin.gauch@uwaterloo.ca`

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.



(a) Outlines of the analyzed sub-watersheds, each draining towards a gauging station (black dots).

(b) Yearly streamflow time series for station 04212100 (Grand River near Painesville, OH).

long time spans. Nevertheless, we believe that this problem formulation yields greater potential for generalization to areas that lack any gauging stations. For a specific case study under this setup, we find that tree-based models can yield more accurate predictions than both neural and physically-based models. Hence—as with many applied machine learning problems—we believe that data collection and preparation form the bottleneck to progress.

2 Case Study: Lake Erie Watershed

In our study, we use daily streamflow measurements at 46 gauging stations in the Lake Erie region from 2010 to 2014 as ground truth. These stations divide the watershed into sub-watersheds, each comprised of the area in which all water flows towards the gauging station. Figure 1a shows a map of the 46 gauging stations and their corresponding sub-watersheds. As independent variables, we use gridded meteorological *forcing data*, which include hourly meteorological variables of temperature, precipitation, pressure, wind speed, specific humidity, and short- and longwave radiation with a spatial resolution of around $15\,\mathrm{km}$ spanning five years (2010 to 2014).

We split the available data into a training period from 2010 to 2012 and a test period from 2013 to 2014. After training a model, 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 compared to the observed streamflow time series for each station [11]. NSE values range between $-\infty$ and 1, where 1 represents perfect predictions. Scores below 0 indicate performance worse than predicting the station's mean streamflow at every time step.

Our baseline physically-based hydrological model is the *Variable Infiltration Capacity model based on Grouped Response Units* (*VIC-GRU*). VIC-GRU is a variant of the semi-distributed VIC model [8, 10] that processes spatial extents with similar characteristics combined and thus more efficiently. We train one VIC-GRU model on all gauging stations, as the model already incorporates varying spatial characteristics through its geophysical 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.

For the data-driven models, we compare a tree-based model with two LSTM-based neural networks:

XGBoost. Since XGBoost, a framework for gradient-boosted regression trees, does not naturally incorporate spatio-temporal input, we flatten the data to a fixed history window of eight days. We further aggregate (minimum, maximum, sum) the hourly forcing data into daily temperature and precipitation values to match the target streamflow data resolution and provide the model with one-hot vectors that encode the gauging station and the month. In preliminary experiments, the remaining forcing variables did not improve prediction quality (results not shown); we therefore excluded them from the input. To obtain suitable parameters, we perform a cross-validated random search.

Table 1: Minimum p_0, median p_{50}, and maximum p_{100} of the NSE distributions for the physically-based VIC-GRU model and the data-driven models XGBoost, LSTM, and ConvLSTM, either trained once for all stations or for each station individually. Best values are highlighted in bold.

Statistic	VIC-GRU	XGBoost		LSTM		ConvLSTM
	all stations	per-station	all stations	per-station	all stations	all stations
p_0	-6.302	-0.206	$\mathbf{0.207}$	-0.678	-0.322	-0.203
p_{50}	0.328	$\mathbf{0.522}$	0.493	0.124	0.180	0.427
p_{100}	0.597	$\mathbf{0.666}$	0.660	0.398	0.314	0.597

LSTM. Our first neural model is a standard LSTM network with two layers of 128 hidden states. To obtain a prediction, we feed the LSTM the previous five days of hourly temperature and precipitation, in addition to one-hot-encoded station and month identifiers. As loss function, we use $1 - \text{NSE}$.

ConvLSTM. A convolutional LSTM (ConvLSTM) network can better incorporate geospatial input data [13]. Our model consists of four convolutional LSTM layers followed by four convolutional but non-recurrent layers. To obtain predictions, we feed the history of the last eight days' precipitation, minimum and maximum temperature, combined with the one-hot-encoded month representation as a gridded matrix into the convolutional LSTM layers. We concatenate the last layer's last output with the station identifiers and pass the resulting tensor through the non-recurrent convolutions, using leaky ReLU activation functions. Finally, these layers output a prediction for each grid cell. We select the cells that correspond to gauging stations and calculate the loss for each station as $1 - \text{NSE}$.

For XGBoost and LSTM, we compare training one model per gauging station and one model for all stations together (the per-station LSTMs having only one layer and fewer hidden states).

Table 1 shows the prediction accuracy of each model on the test period. We find that XGBoost provides the best predictions. When we train one XGBoost model per gauging station, the median NSE score improves even further, but yields worse minimum NSE values. The convolutional LSTM model's predictions are worse than both XGBoost (all stations and per-station), likely due to the paucity of training data [5]. We did not train VIC-GRU and ConvLSTM for each station individually, as these models are already inherently spatially distributed. Compared to the physically-based model, the prediction quality of the data-driven models is more consistent across the stations. VIC-GRU struggles to predict certain stations, usually in highly urbanized areas, with NSE scores well below zero. The data-driven models' NSE distributions exhibit fewer and less pronounced outliers, even when we train one model for all stations. The code we use to run all our experiments is available at `https://github.com/gauchm/streamflow-ml`.

3 Discussion and Future Work

Our study represents a first step towards the goal of building streamflow prediction models for ungauged basins and locations for which we lack ground truth measurements. For this problem, we believe that spatially-distributed input data are key. Consider a simple example where we predict streamflow at a particular location in a river based on a few upstream gauges. Although accurate predictions would undoubtedly be useful, such a model cannot be applied at another location (even in the same river). Models that take spatially-distributed measurements as input, we believe, can better generalize to different locations—for example, by learning spatial and temporal structures. Also, with this approach we see a clearer path towards hybrid models that both capture an understanding of physical processes and leverage the advantages of data-driven techniques.

It appears that the machine learning mantra "there is no data like more data" also holds for streamflow prediction, because our approach is bottlenecked at data. To illustrate, consider a simple example: taking periodic air temperature readings at a location is relatively straightforward, and we have records in many locations dating back decades. Air temperature readings across large areas, in contrast, need to be acquired through computationally complex interpolation or remote sensing technology. Thus, such data are far more difficult to come by (particularly, historic data), which means that our models must make do with limited data. Amid the current excitement about the prospects of applying ML to tackle climate change [12], we argue that data collection, preparation, and curation efforts are equally important but often neglected contributions.

Acknowledgments

This research was undertaken thanks in part to funding from the Canada First Research Excellence Fund, and enabled by computational resources provided by Compute Ontario and Compute Canada. The authors further thank the Global Water Futures program and the Integrated Modeling Program for Canada for their financial support, as well as Nicolas Gasset and Vincent Fortin from Environment and Climate Change Canada for making the meteorological forcing dataset available upfront.

References

[1] E. C. Carcano, P. Bartolini, M. Muselli, and L. Piroddi. Jordan recurrent neural network versus IHACRES in modelling daily streamflows. *Journal of Hydrology*, 362(3):291–307, 2008.

[2] F.-J. Chang, L.-C. Chang, and H.-L. Huang. Real-time recurrent learning neural network for stream-flow forecasting. *Hydrological Processes*, 16(13):2577–2588, 2002.

[3] S. M. Eberts, M. D. Woodside, M. N. Landers, and C. R. Wagner. Monitoring the pulse of our nation's rivers and streams—the U.S. Geological Survey streamgaging network, 2019.

[4] M. Gauch, J. Mai, S. Gharari, and J. Lin. Data-driven vs. physically-based streamflow prediction models. *Proceedings of 9th International Workshop on Climate Informatics*, 2019.

[5] M. Gauch, J. Mai, and J. Lin. The proper care and feeding of CAMELS: How limited training data affects streamflow prediction. *arXiv:1911.07249*, 2019.

[6] A. D. Gronewold and R. B. Rood. Opinion: Climate change drives shifts between high, low Great Lakes water levels. Bridge Magazine, Aug 2019. `https://www.bridgemi.com/michigan-environment-watch/opinion-climate-change-drives-shifts-between-high-low-great-lakes-water`.

[7] A. H. Halff, H. M. Halff, and M. Azmoodeh. Predicting runoff from rainfall using neural networks. In *Proceedings of the Symposium Sponsored by the Hydraulics Division of ASCE*, pages 760–765, 1993.

[8] 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*, 11(8), 2018.

[9] F. Kratzert, D. Klotz, G. Shalev, G. Klambauer, S. Hochreiter, and G. Nearing. Benchmarking a catchment-aware Long Short-Term Memory network (LSTM) for large-scale hydrological modeling. *Hydrology and Earth System Sciences Discussions*, pages 1–32, 2019.

[10] 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*, 99(D7):14415–14428, 1994.

[11] J. E. Nash and J. V. Sutcliffe. River flow forecasting through conceptual models part I—a discussion of principles. *Journal of Hydrology*, 10(3):282–290, 1970.

[12] D. Rolnick, P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. S. Ross, N. Milojevic-Dupont, N. Jaques, A. Waldman-Brown, A. Luccioni, T. Maharaj, E. D. Sherwin, S. K. Mukkavilli, K. P. Kording, C. Gomes, A. Y. Ng, D. Hassabis, J. C. Platt, F. Creutzig, J. Chayes, and Y. Bengio. Tackling climate change with machine learning. *arXiv:1906.05433*, 2019.

[13] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems 28*, pages 802–810, 2015.

[14] M. Smith and L. French. Summer on the swollen Great Lakes. The New York Times, Aug 2019. `https://www.nytimes.com/2019/08/24/us/great-lakes-water-levels.html`.