
COMPARATIVE ANALYSIS OF DATA-DRIVEN APPROACHES FOR HYDROLOGICAL FORECASTING

A PREPRINT

Eldar Khuzin
Chair of Data Analysis
Moscow Institute of Physics and Technology
Moscow, Russia
khuzin.er@phystech.su

Novikov Ivan
Affiliation
Address
email

Abramov Dmitrii
Affiliation
Address
email

ABSTRACT

В данной работе проводится сравнительный анализ методов машинного обучения для решения задачи прогнозирования речного стока и уровня воды. Рассматриваются классические алгоритмы (Random Forest и CatBoost), отличающиеся сравнительно низкой вычислительной сложностью и быстрым обучением, рекуррентные нейронные сети с долговременной памятью (LSTM), способные учитывать долгосрочные зависимости в данных, и современные архитектуры на основе трансформеров, которые демонстрируют перспективные результаты. Эксперименты проводятся на гидрологических данных рек России и Франции, содержащие динамические временные ряды гидрологических измерений и статические характеристики водосборных бассейнов. Цель работы — выявить наиболее эффективные подходы к прогнозированию.

Keywords Flood discharge · Rainfall-runoff modeling · More

1 Introduction

Прогнозирование речного стока и уровня воды является важной задачей для предупреждения наводнений и эффективного использования водных ресурсов в таких областях как сельское хозяйство и гидроэнергетика.[1]

Современные подходы к моделированию гидрологических процессов можно разделить на два основных типа: физически обоснованные (physical-based) и основанные на данных (data-driven). Физические модели, несмотря на свою интерпретируемость и потенциальную точность, имеют ряд существенных недостатков: для их использования требуются обширные данные о речном бассейне и значительные вычислительные ресурсы, что ограничивает их практическое применение в задачах краткосрочного прогнозирования [2]. Также построение таких моделей требует глубокого понимания гидрологических процессов и сталкивается с проблемами обобщаемости.

Data-driven подходы привлекают отсутствием явного моделирования физических процессов. Классические алгоритмы машинного обучения, такие как Random Forest, показывают себя конкурентно способными в определённых условиях, выделяясь при этом быстрой обучаемостью и небольшими требованиями к вычислительным ресурсам[3]. Рекуррентные нейронные сети (Recurrent Neural Networks, RNN), в частности, Long Short-Term Memory (LSTM), также показывают потенциал за счет способности учитывать долгосрочные зависимости, такие как таяние снежного покрова или наличие накопленной влаги в

почве[4, 5]. Недавние исследования демонстрируют, что и трансформеры могут успешно применяться для гидрологических прогнозов[6, 7]. Однако сравнительный анализ на разных типах гидрологических данных и в различных географических условиях представлен недостаточно.

В данной работе авторы ставят перед собой следующие цели: провести сравнительный анализ data-driven подходов, в частности с желанием побить метрики LSTM, а также исследовать влияние физико-географических атрибутов на результат. Оценка производится через такие метрики, как среднеквадратичная ошибка (RMSE) и коэффициент эффективности Нэша-Сатклиффа (Nash-Sutcliffe Efficiency, NSE).

2 Problem statement

Используется метрика среднеквадратичной ошибки (MSE). А именно:

$$\text{MSE}(f) = \frac{1}{N} \sum_{i=1}^N (y_i - f(x_i))^2,$$

где y_i — истинное значение наблюдения, а $f(x_i)$ — предсказанное значение.

И ставится множество оптимизационных задач по нахождению оптимальной модели f^* в своём классе моделей \mathcal{F} , т.е.:

$$f^* = \arg \min_{f \in \mathcal{F}} \text{MSE}(f).$$

Основное множество данных состоит из динамических временных рядов гидрологических измерений (уровня воды, речного стока, а также температуры) и статических характеристик водосборных бассейнов.

References

- [1] David A. Lavers, Maria-Helena Ramos, Linus Magnusson, Ilias Pechlivanidis, Bastian Klein, Christel Prudhomme, Louise Arnal, Louise Crochemore, Bart Van Den Hurk, Albrecht H. Weerts, Shaun Harrigan, Hannah L. Cloke, David S. Richardson, and Florian Pappenberger. A Vision for Hydrological Prediction. 11(3):237. ISSN 2073-4433. doi:10.3390/atmos11030237. URL <https://www.mdpi.com/2073-4433/11/3/237>.
- [2] Vasilis Bellos and George Tsakiris. A hybrid method for flood simulation in small catchments combining hydrodynamic and hydrological techniques. 540:331–339. ISSN 00221694. doi:10.1016/j.jhydrol.2016.06.040. URL <https://linkinghub.elsevier.com/retrieve/pii/S0022169416304024>.
- [3] Lukas Schoppa, Markus Disse, and Sophie Bachmair. Evaluating the performance of random forest for large-scale flood discharge simulation. 590:125531. ISSN 00221694. doi:10.1016/j.jhydrol.2020.125531. URL <https://linkinghub.elsevier.com/retrieve/pii/S0022169420309914>.
- [4] Frederik Kratzert, Daniel Klotz, Claire Brenner, Karsten Schulz, and Mathew Herrnegger. Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks. 22(11):6005–6022, . ISSN 1027-5606. doi:10.5194/hess-22-6005-2018. URL <https://hess.copernicus.org/articles/22/6005/2018/>.
- [5] Frederik Kratzert, Martin Gauch, Daniel Klotz, and Grey Nearing. HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin. 28(17):4187–4201, . ISSN 1027-5606. doi:10.5194/hess-28-4187-2024. URL <https://hess.copernicus.org/articles/28/4187/2024/>.
- [6] Wenzhong Li, Chengshuai Liu, Yingying Xu, Chaojie Niu, Runxi Li, Ming Li, Caihong Hu, and Lu Tian. An interpretable hybrid deep learning model for flood forecasting based on Transformer and LSTM. 54:101873. ISSN 22145818. doi:10.1016/j.ejrh.2024.101873. URL <https://linkinghub.elsevier.com/retrieve/pii/S2214581824002210>.
- [7] Jiangtao Liu, Yuchen Bian, and Chaopeng Shen. Probing the limit of hydrologic predictability with the Transformer network. URL <http://arxiv.org/abs/2306.12384>.

Ask:
Требуется ли описать каждую из введённых моделей

Comment:
Поговорить насчёт того, как каждую из моделей адаптировать под данную задачу.