

Прогнозування рівня води засобами машинного навчання на основі дослідження річки Дунай

HydroEye

Зміст

I. Вступ	3
II. Методологія	8
III. Статистичний аналіз даних	11
IV. Результати.....	15
V. Обговорення.....	24
VI. Висновки.....	26
VII. Використані джерела.....	27

I. Вступ

Критичний імператив: Боротьба з глобальними ризиками повеней

Повені є однією з найбільш руйнівних природних катастроф, що загрожують мільйонам людей у всьому світі та спричиняють значні економічні збитки. Глобальні щорічні збитки від повеней оцінюються у 100 мільярдів доларів США [1], що становить 12,40 доларів на людину. За останнє десятиліття економічні втрати, пов'язані з природними небезпеками, в середньому становили близько 170 мільярдів доларів США на рік, з піками до 300 мільярдів доларів США в окремі роки [2].

Згідно з прогнозами, очікується, що водні ризики, включаючи повені, посухи та шторми поглинуть 5,6 трильйона доларів США світового ВВП до 2050 року, причому лише повені становитимуть 36% цих прямих втрат [3]. Економічні втрати від цих подій значно зросли з 2000-х років, що відповідає їх посиленій інтенсивності та частоті [2]. Дана тенденція, як очікується, продовжиться, причому показники збитків зростатимуть для більшості країн до 2100 року, особливо за більш серйозних сценаріїв глобального потепління.

За оцінками, 1,81 мільярда людей (23% населення світу) безпосередньо піддаються впливу затоплення глибиною понад 0,15 м під час повеней, які трапляються раз на 100 років, що становить значний ризик для життя, особливо для вразливих груп населення [4]. У Південній та Східній Азії значна частина населення вважається вразливою до потенційних повеней (1,24 мільярда), причому лише Китай (395 мільйонів) та Індія (390 мільйонів) становлять понад третину від загальної кількості [5]. У 2023 році повені та посухи змусили близько 8 мільйонів людей покинути свої домівки в десяти найбільш постраждалих країнах, що більш ніж удвічі перевищує показник десятирічної давнини [6].

Зростаючі витрати та людські втрати, незважаючи на поточні зусилля з управління ризиками повеней, вказують на те, що існуючі можливості прогнозування та управління повенями є недостатніми для того, щоб адаптуватись до прискорення зміни клімату та наслідків, спричинених даним

явищем. Це створює критичну та нагальну потребу в більш досконалих, точних та проактивних рішеннях.

У 2024 році в Європі спостерігались численні руйнівні повені, що підкреслює нагальну потребу в передових системах прогнозування.

Повені в Центральній Європі в 2024 році були спричинені рекордно сильними опадами, що виникли внаслідок шторму Борис, надзвичайно вологого генуезького циклону. Затоплення почалися в Австрії та Чехії, а потім поширилися на Польщу, Румунію, Словаччину, Німеччину та Угорщину [7]. Ця подія призвела до наймасштабніших повеней у Європі з 2013 року, причому рівні річок у 30% європейської річкової мережі перевищили поріг «високої» повені, а 12% перевищили поріг «сильної» повені. У вересні 2024 року шторм Борис спричинив екстремальні опади, причому деякі частини Польщі, Німеччини, Чехії та північно-східної Румунії отримали до трьох місяців дощу всього за 4-5 днів. Повінь призвела до піків, що принаймні вдвічі перевищували середньорічний максимум уздовж 8500 км річок, що значно вплинуло на місцеві громади, призвівши до жертв та значних збитків [8].

У жовтні 2024 року сильні та катастрофічні повені вразили регіон Валенсія в Іспанії. Ця руйнівна подія призвела до щонайменше 232 смертей лише у Валенсії, з додатковими жертвами в Альбасете, Куенці та Малазі [8]. Пошкодження інфраструктури та економічні втрати були значними і становили загалом близько 16,5 мільярдів євро, причому застраховані збитки оцінюються приблизно в 3,5 мільярда євро, а загальні збитки – близько 10,7 мільярда євро [9].

Крім того, 2024 рік відзначився численними іншими повенями по всій Європі, включаючи північну Францію, Англію, Уельс (шторм Генк у січні), північну Іспанію (лютий), Францію (березень, травень), південну Німеччину та Швейцарію (червень). Загалом, шторми та повені по всій Європі протягом року торкнулися близько 413 000 людей, призвели до загибелі щонайменше 335 людей та коштували щонайменше 18 мільярдів євро збитків [8]. Узагальнений аналіз впливу повеней наведено в таблиці 1.

Дані події слугують яскравим доказом руйнівних наслідків неготовності до повеней та нагальної потреби у покращеному прогнозуванні. Вони демонструють, що загроза не є абстрактною. Масштаби людських та економічних втрат від цих подій підкреслюють глибоку соціальну та комерційну цінність системи, здатної надавати точні та своєчасні попередження.

Таблиця 1. Статистика впливу повеней у світі та Європі

Категорія впливу	Деталі	Джерело
Глобальні економічні втрати	Щорічні збитки від повеней: ~100 млрд USD (2024)	Mitchell (2024)
	Середні щорічні збитки від природних небезпек (останнє десятиліття): ~170 млрд USD	(UNDRR 2022 Annual Report, 2023)
	Пікові щорічні збитки від природних небезпек: до 300 млрд USD	
Прогнозовані майбутні витрати	Водний ризик (повені, посухи, шторми) до 2050 року: ~5.6 трлн USD світового ВВП	Mitchell (2024)
	Частка повеней у прямих втратах (до 2050 року): 36%	
Людський вплив та переміщення	Населення, що безпосередньо піддається 100-річним повеням: 1.81 млрд людей (23% світового населення)	
	Переміщення через повені/посухи (2023): 8 млн людей (у 10 найбільш постраждалих країнах)	
Вплив повеней у Європі (2024)	Загальна кількість постраждалих: ~413 000 людей	Copernicus (2025)
	Загальна кількість загиблих: щонайменше 335 життів	
	Загальні оцінені збитки: щонайменше 18 млрд євро	

Повені в Іспанії (Валенсія, жовтень 2024)	Кількість загиблих: щонайменше 232 людини	
	Загальні збитки: ~16.5 млрд євро (за оцінками)	
	Застраховані збитки: ~3.5 млрд євро (~3.8 млрд USD)	Willard (2025)
	Загальні збитки: ~10.7 млрд євро (~11 млрд USD)	

Прогнозування рівня води в річках є критично важливим для адаптації до наслідків зміни клімату, таких як збільшення частоти екстремальних погодних явищ – повеней і посух. Розробка та впровадження подібного проєкту сприяє зміцненню потенціалу громад та місцевої влади щодо своєчасного реагування на кліматичні загрози. Це дозволяє зменшити збитки від стихійних лих, убезпечити населення та критичну інфраструктуру, а також ефективніше управляти водними ресурсами в умовах мінливості клімату.

Своєчасний прогноз змін рівня води в річках дозволяє оптимізувати управління водними ресурсами, зменшуючи ризики виснаження джерел водопостачання у випадку посухи та забруднення води під час повеней. Завдяки точним прогнозам можливо заздалегідь реалізовувати заходи із захисту водозаборів, резервування води або евакуації населення. Це особливо важливо для регіонів із нестабільним доступом до прісної води.

Міста та сільські громади, розташовані поблизу річок, найбільше страждають від раптових повеней або тривалих посух. Проєкт прогнозування водного рівня забезпечує кращу підготовленість місцевих органів влади до таких явищ, сприяє плануванню безпечної забудови, розвитку інфраструктури та зменшенню втрат житла, майна і людських життів. Це підвищує стійкість населених пунктів до природних катастроф і сприяє їх сталому розвитку.

Таким чином, проєкт безпосередньо відповідає Цілям сталого розвитку (ЦСР), що визначені ООН. Ключовою ціллю є ЦСР 13: Боротьба зі зміною клімату. Також проєкт тісно пов'язаний з ЦСР 6: Чиста вода та належні санітарні умови і ЦСР 11: Сталі міста та спільноти.

Слід зауважити, що проєкт не лише вирішує конкретну технічну задачу, але й комплексно підтримує досягнення одразу кількох глобальних цілей, що сприяють побудові безпечнішого, більш стійкого та екологічно відповідального світу.

Поточний стан досліджень

Традиційні гідрологічні моделі, хоча й ґрунтуються на фізичних принципах, часто стикаються зі значними проблемами. Вони є обчислювально інтенсивними, вимагають тривалого та точного калібрування (часто для кожного об'єкта окремо) і важко моделюють притаманну складність, нелінійність та нестаціонарні взаємозв'язки в гідрологічних системах [10]. Вони також можуть применшувати інтенсивність раніше небачених викидів [11] через відсутність широких фізичних знань, що виходять за межі навчальних даних.

Технології машинного навчання являють собою так звану «зміну парадигми» від традиційних фізично-орієнтованих моделей до підходів, керованих даними [10].

Основні переваги використання ШІ, і зокрема технологій машинного навчання та глибокого навчання включають:

- моделювання нелінійності та складних взаємозв'язків;
- адаптивність до різноманітних умов даних;
- автоматизація та прогнозування в режимі реального часу;
- мінімальні припущення, специфічні для предметної області;
- масштабованість та обчислювальна ефективність;
- краща обробка непередбачених викидів та екстремальних подій.

Алгоритми ШІ, наприклад, штучні нейронні мережі (ANN), мережі довгої короткострокової пам'яті (LSTM), випадкові ліси (RF) можуть ефективно вивчати та моделювати складні, нелінійні та динамічні взаємозв'язки між гідрологічними змінними без необхідності явного формулювання базових фізичних процесів. Це особливо корисно там, де на гідрологічну поведінку впливає безліч взаємозалежних факторів [10].

Моделі ШІ можуть інтегрувати різні типи даних, включаючи метеорологічні вимірювання, супутникові набори даних та навіть дані датчиків у реальному часі [12]. Ця адаптивність дозволяє розгортати їх як у середовищах з великою кількістю даних, так і в середовищах з обмеженими даними. Після навчання моделі ШІ можуть бути автоматизовані для прогнозування в реальному часі, що є критично важливим для систем раннього попередження [13].

На відміну від традиційних моделей, які вимагають детальної параметризації на основі фізичних знань про водозбір, моделі ШІ можуть працювати з мінімальними припущеннями [14]. Це особливо корисно в неконтрольованих або погано контрольованих басейнах, де фізичні характеристики можуть бути невідомі або важко кількісно оцінювані [15].

Хоча суто керовані даними моделі машинного навчання можуть мати обмеження щодо раніше небачених екстремальних подій, гібридні моделі, що поєднують машинне навчання з фізичними знаннями, пропонують покращену точність для таких сценаріїв. Ці гібридні моделі можуть зберігати фізичну інтерпретованість та описувати важливі особливості річкового басейну.

Найбільш поширеними методами є Random Forest, Gradient Boosting, XGBoost, LSTM, Prophet [16-18]. Детальний аналіз даних методів створить передумови до розробки інформаційної системи прогнозування рівня води, що особливо актуально в контексті досягнення ЦСР [19].

II. Методологія

Задля створення точного та надійного прогнозу пропонуємо провести три експерименти, для виявлення найбільш доцільних підходів в контексті поставленої задачі. Перший експеримент полягає в прогнозуванні майбутніх значень часового ряду на основі історичних даних. Другий експеримент полягає у прогнозуванні даних часового ряду на основі історичних значень та додаткових ознак, введених шляхом проведення статистичного аналізу даних. Третій експеримент ґрунтується на використанні підготовлених даних, отриманих у другому експерименті та трансформації даних часового ряду за допомогою вейвлет-перетворень та перетворень Фур'є. Порівняння прогнозу відбувається

на основі метрик зваженої середньоквадратичної похибки, середньої абсолютної похибки, середньої абсолютної відсоткової похибки та R^2 .

Датасет

В дослідженні використано дані рівня води в річці Дунай [20]. Значення вимірювань взято за період з 26 липня 2017 року до 9 жовтня 2024 року, які спостерігались в пункті спостереження Бернсдорф, Нижня Австрія. Обраний період зумовлено наявністю усіх значень вимірювань за вказаний період, задля кращого аналізу часового ряду, оскільки в даних за довші періоди часу спостерігається відсутність певних значень, що потенційно може вплинути на якість дослідження.

В контексті даного дослідження використано максимальне значення рівня води, що спостерігалось за добу. Проте, запропонований підхід може використовуватись для прогнозування середнього або мінімального значень, залежно від поставлених задач. Прогнозування максимальних значень є корисним при прогнозуванні повеней, а мінімальних – при прогнозуванні посух.

Метод ковзної середньої

Оскільки в часових рядах часто спостерігається шум, це суттєво ускладнює виявлення закономірностей. В контексті дослідження часового ряду рівня води шум може бути викликаним наступними факторами:

- короткочасні дощі;
- танення снігу;
- локальні гідротехнічні роботи.

Такі коливання можуть бути досить різкими, проте вони не відображають зміну загальною гідрологічної ситуації.

Для того, щоб уникнути хибного прогнозу, внаслідок подібних явищ, доцільно провести згладжування. Метод ковзної середньої дозволяє згладити часовий ряд шляхом заміни кожного значення середнім від кількох сусідніх значень.

Для часового ряду $\{x_t\}_{t=1}^N$ ковзна середня має вигляд:

$$\bar{x}_t^{(m)} = \frac{1}{m} \sum_{i=0}^{m-1} x_{t-i}, \quad (1)$$

де:

m – розмір вікна;

x_{t-i} – спостереження i кроків тому;

$\bar{x}_t^{(m)}$ – значення ковзної середньої у момент часу t з вікном m .

У машинному навчанні ковзна середня стає новою ознакою:

$$f_{SMA,m}(t) = \frac{1}{m} \sum_{i=0}^{m-1} x_{t-i} \quad (2)$$

Доцільною є побудова кількох ознак з різним розміром вікна, наприклад $m = 1$ для одного дня, $m = 7$ – для тижня, $m = 30$ – для місяця, $m = 365$ – для року тощо. Тоді масив ознак, сформованих за допомогою методу ковзної середньої буде визначатись як:

$$F_t = [f_{SMA,1}(t), f_{SMA,7}(t), f_{SMA,30}(t), f_{SMA,365}(t)] \quad (3)$$

Нові ознаки стають додатковими вхідними параметрами для моделі машинного навчання та здійснюють вплив на прогноз.

Перетворення Фур'є

Неперервне перетворення Фур'є проводить декомпозицію сигналу:

$$\mathcal{F}\{x(t)\} = X(f) = \int_{-\infty}^{\infty} x(t) \cdot e^{-2\pi i f t} dt, \quad (4)$$

де:

$x(t)$ – оригінальний сигнал у часовій області;

t – час (неперервна змінна);

$X(f)$ – спектральне представлення сигналу $x(t)$ на частоті f (комплексне число, що містить амплітуду й фазу);

f – частота;

$e^{-2\pi i f t}$ – комплексна експонента, яка задає синусоїдальні коливання;

i – уявна одиниця ($i^2 = -1$).

Оскільки в задачах машинного навчання операції проводяться з вимірюваннями у фіксовані моменти часу, необхідно використати дискретні перетворення. Дискретне перетворення Фур'є має вигляд:

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-2\pi i k n / N}, \quad (5)$$

де:

x_n – значення сигналу у момент n (дискретний час);

n – індекс часу;

k – індекс частоти;

X_k – спектральний коефіцієнт для k -ої частоти;

N – загальна кількість точок у вибірці;

$e^{-2\pi i k n / N}$ – комплексна експонента для побудови синусоїд з різними частотами.

Вейвлет перетворення

Неперервне Вейвлет перетворення має вигляд:

$$W_x(a, b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(t) \cdot \psi^*\left(\frac{t-b}{a}\right) dt, \quad (6)$$

де:

$W_x(a, b)$ – коефіцієнт вейвлет-перетворення для масштабу a і зсуву b ;

$x(t)$ – вхідний сигнал у часовій області;

t – час (неперервна змінна);

a – масштаб;

b – зсув у часі;

$\psi^*(\cdot)$ – материнський (базовий) вейвлет – функція з компактною підтримкою (короткий коливальний імпульс);

Дискретне Вейвлет перетворення має вигляд:

$$\psi_{j,k}(t) = 2^{j/2} \psi(2^j t - k), \quad (7)$$

$\psi(\cdot)$ – материнський вейвлет;

j – масштаб;

k – зсув.

III. Статистичний аналіз даних

Проведемо статистичний аналіз даних. В таблиці 2 наведено описові статистики.

Таблиця 2. Описові статистики

Name	Value for Danube
count	2628.0000
mean	411.6237
std	63.4742
min	338.0000
25%	370.0000
50%	395.0000
75%	430.2500
max	948.0000

Максимальний рівень води в Дунаї за період досліджень спостерігався 16 вересня 2024 року і становив 948 см. Мінімальний рівень максимального значення було зафіксовано 23 жовтня 2023 на рівні 338 см.

Візуалізацію часового ряду для досліджуваного періоду наведено на рисунку 1.

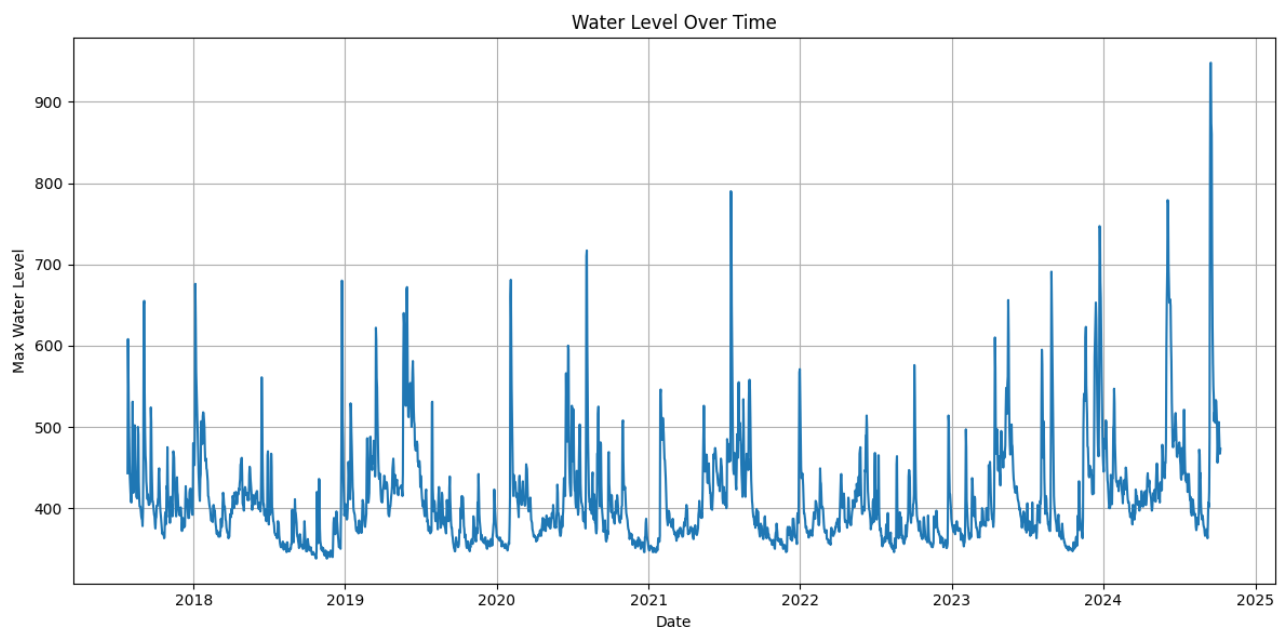


Рисунок 1. Значення максимального рівня води в річці Дунай з 2017 по 2025 рік

Середньомісячні максимуми рівня води представлено на рисунку 2.

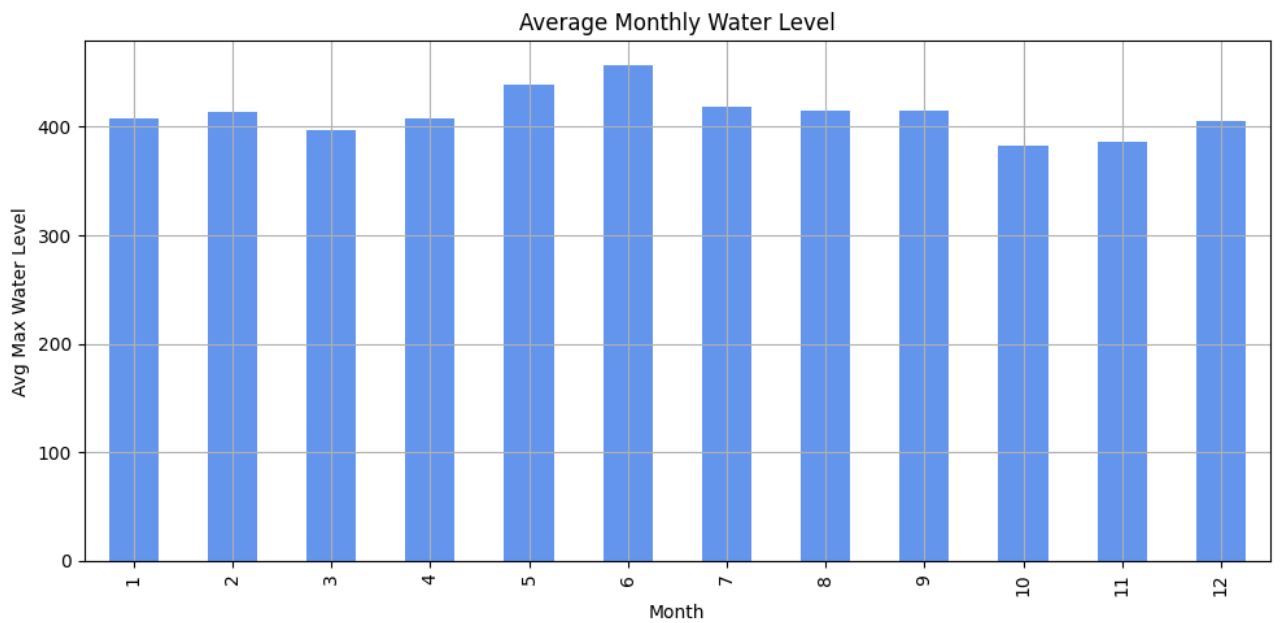


Рисунок 2. Середньомісячні максимуми
рівня води в річці Дунай з 2017 по 2025 рік
Розподіл рівня води за місяцями представлено на рисунку 3.

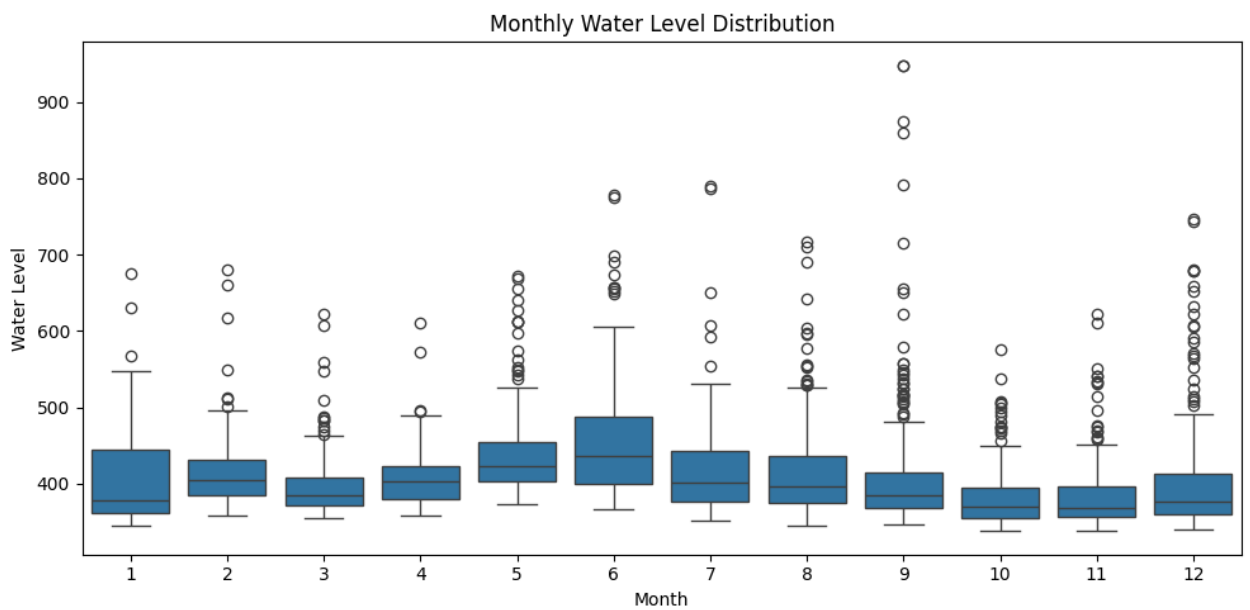


Рисунок 3. Середньомісячний розподіл максимумів
рівня води в річці Дунай з 2017 по 2025 рік
На рисунку 4 наведено декомпозицію часового ряду для даних річки Дунай
на тренд, сезонність та залишки.

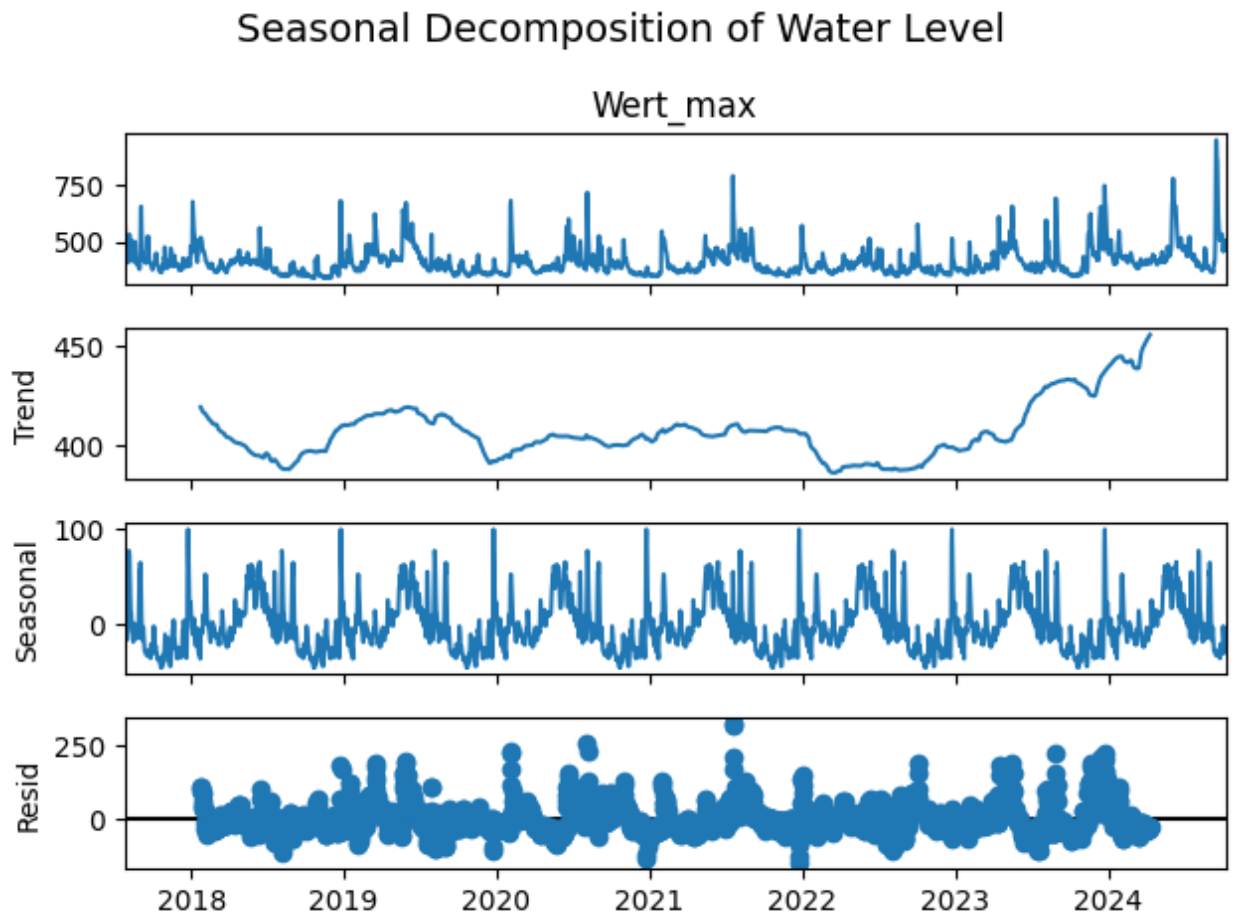


Рисунок 4. Декомпозиція часового ряду для річки Дунай

В таблиці 3 наведено результати проведення тестів ADF та KPSS.

Таблиця 3. Результати тестів ADF і KPSS

Value	ADF	KPSS
Statistic	-6.3632	0.4876
p-value	$2.4461 \cdot 10^{-8}$	0.0444
Stationarity	Stationary	Non-stationary

Відповідно до одержаних результатів наявний конфлікт між двома тестами. За результатами ADF тесту часовий ряд є стаціонарним, в той же час, згідно з результатом KPSS тесту, часовий ряд є нестаціонарним. Враховуючи природу обох тестів, це свідчить про те, що дані часового ряду знаходяться на межі стаціонарності, тобто часовий ряд є майже стаціонарним.

На рисунку 4 зображено функцію автокореляції рівня води.

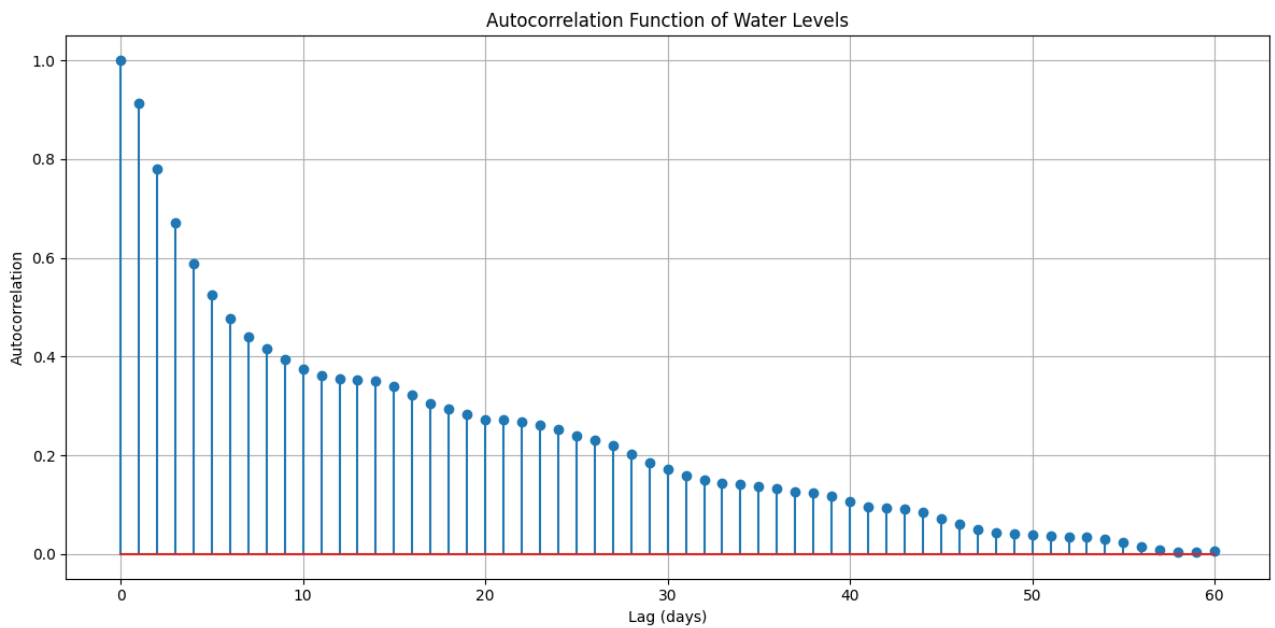


Рисунок 5. Функція автокореляції рівня води для річки Дунай

На рисунку 6 наведено візуалізацію 30-денної ковзної середньої.

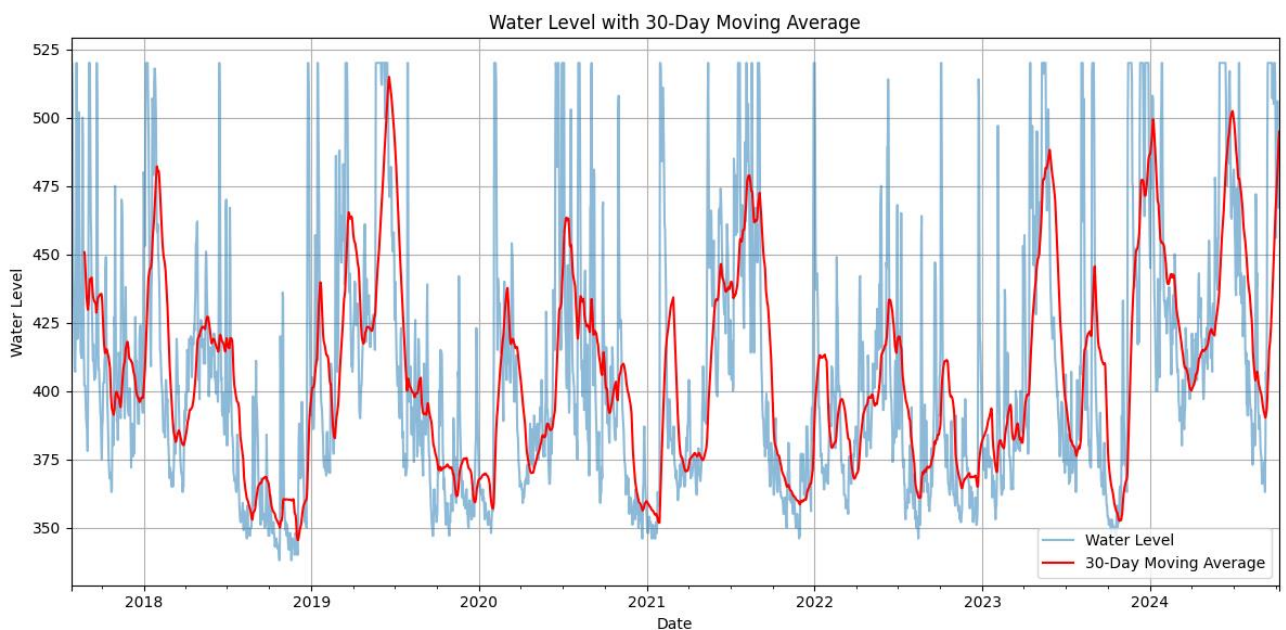


Рисунок 6. 30-денна ковзна середня часового ряду рівня води в річці Дунай

IV. Результати

В ході дослідження було проведено 3 експерименти для аналізу різних підходів до прогнозування часових рядів та визначення найдоцільнішого в контексті поставленої задачі.

Перший експеримент полягає в прогнозуванні часового ряду на основі історичних даних з використанням методів машинного навчання, зокрема випадкового лісу, посилення градієнту та XGBoost, глибокого навчання, а саме

рекурентної нейронної мережі з архітектурою LSTM та статистичного методу Prophet. Результати експерименту наведено в таблиці 4.

Таблиця 4. Метрики на тестовій вибірці для першого експерименту

Model	Metrics on test set			
	RMSE	MAE	MAPE	R ²
Random Forest	39.5011	19.3413	3.7589	0.7963
Gradient Boosting	38.0364	18.5053	3.5740	0.8111
XGBoost	44.5196	21.7971	4.2187	0.7413
LSTM	34.7914	15.9396	3.0582	0.8420
Prophet	91.7123	54.9689	10.7425	−0.1182

Візуалізація порівняння прогнозованих значень з реальними даними спостережень наведена на рис. 7-11.

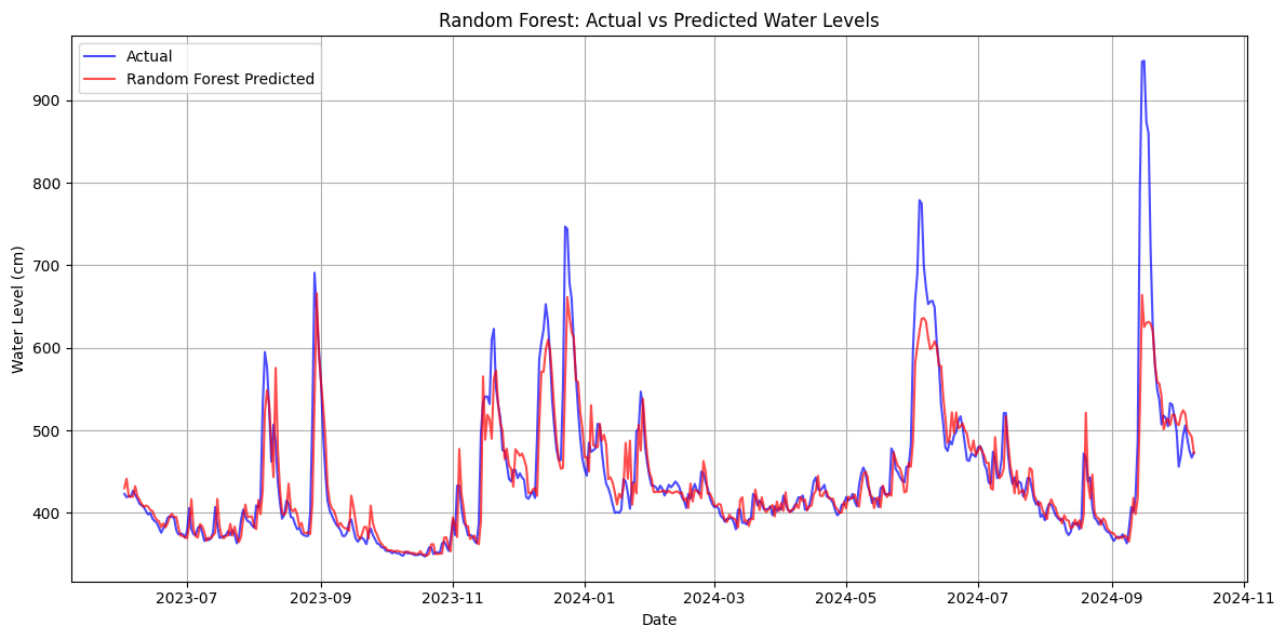


Рисунок 7. Результати прогнозування методу випадкового лісу

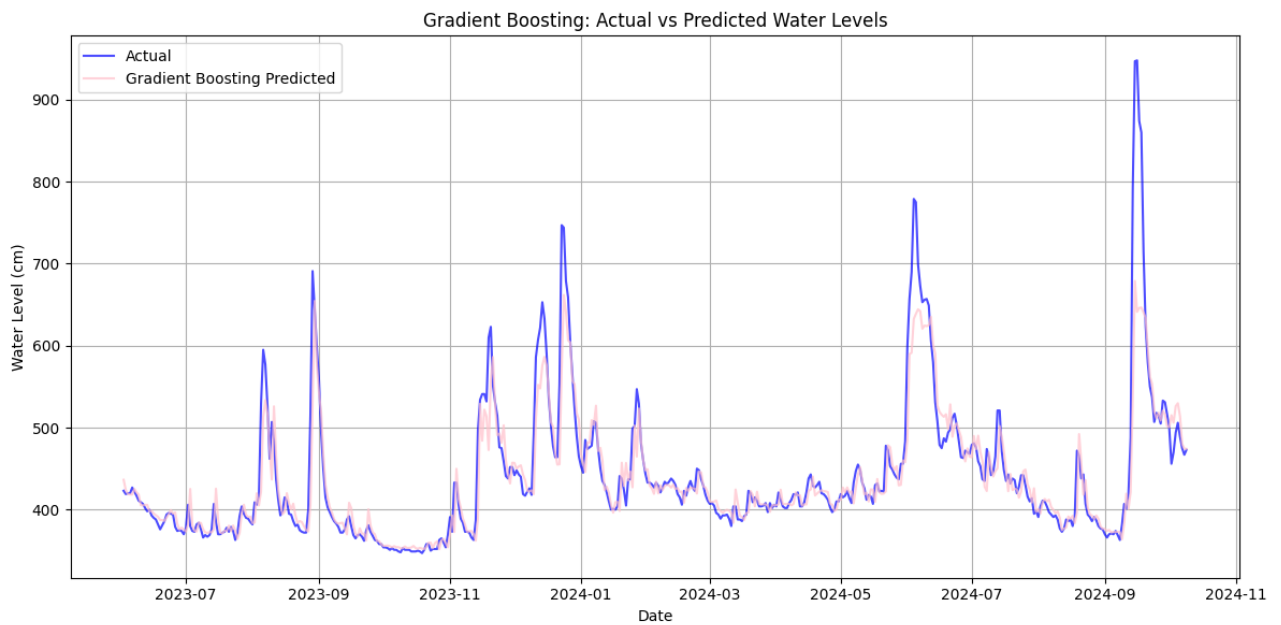


Рисунок 8. Результати прогнозування методу посилення градієнту

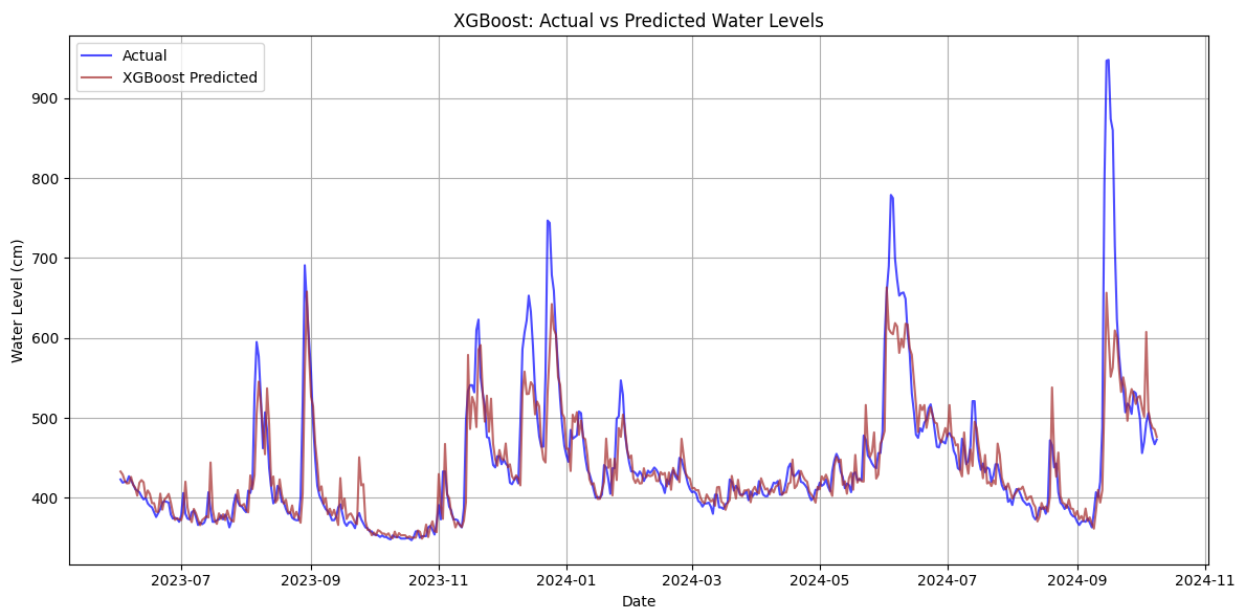


Рисунок 9. Результати прогнозування методу XGBoost

Відповідно до отриманих даних можна зробити висновок, що статистична модель Prophet не здатна вловити закономірності, що спостерігаються в коливаннях рівня води та не є придатною для прогнозування. На це вказують як високу значення похибок, так і від'ємне значення метрики R^2 . Таким чином, в подальших дослідженнях вважаємо за доцільне відмовитись від використання даного методу.

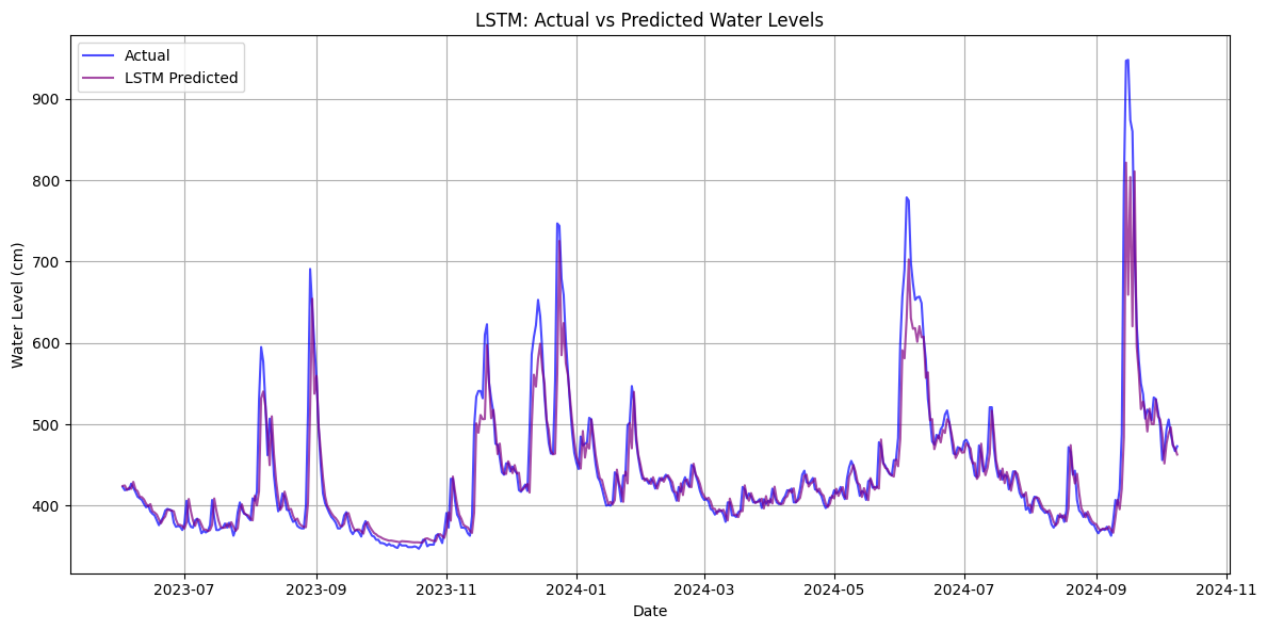


Рисунок 10. Результати прогнозування методу LSTM

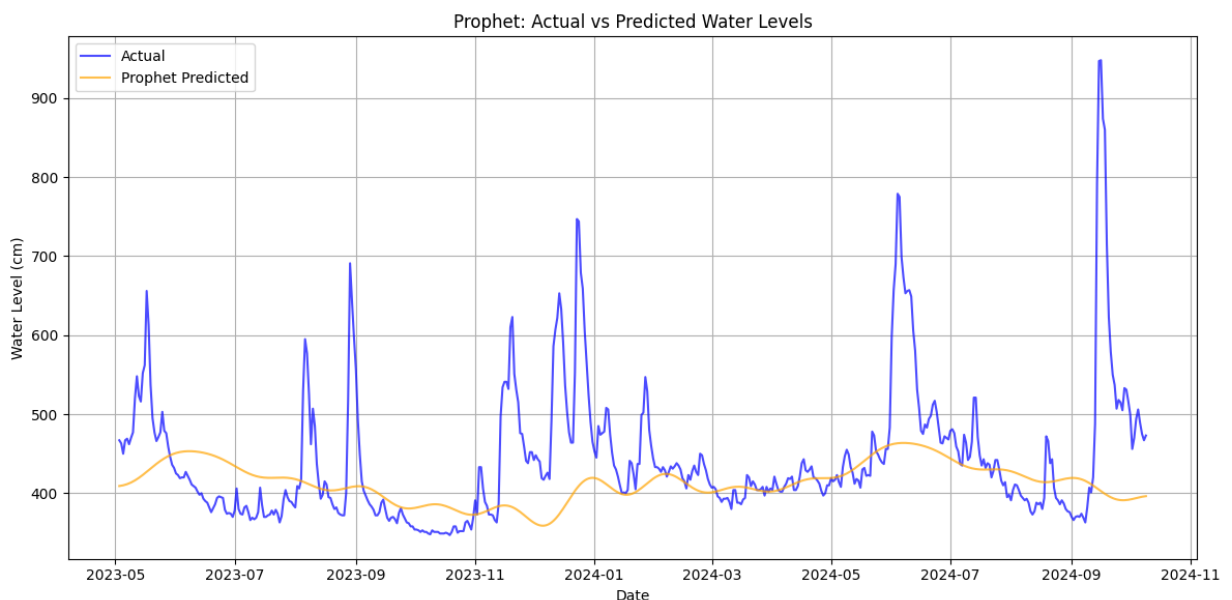


Рисунок 11. Результати прогнозування методу Prophet

Усі інші методи показали здатність виявлення закономірностей, проте метрики вказують на недостатню точність роботи методів, тому виникає необхідність у проведенні подальших досліджень.

Другий експеримент полягає у проведенні трансформації даних та введенні на їх основі додаткових ознак, задля створення більш точного прогнозу. Задля більш точного виявлення закономірностей створено ознаки затримки `Lag_1`, `Lag_7`, `Lag_30`, `Lag_90`. Дані ознаки утворені шляхом зсуву значення рівня води

на відповідну кількість днів. Це здійснюється задля фіксації минулих рівнів води для прогнозування майбутніх значень.

Для виявлення та фіксації короткострокових та довгострокових тенденцій та мінливості обчислено 7-денні та 30-денні ковзні середні значення, а також їх стандартні відхилення за ці періоди.

Результати експерименту наведено в таблиці 5.

Таблиця 5. Метрики на тестовій вибірці для другого експерименту

Model	Metrics on test set			
	RMSE	MAE	MAPE	R ²
Random Forest	17.9498	12.6382	2.8406	0.8777
Gradient Boosting	17.9371	11.6940	2.6048	0.8779
XGBoost	17.3757	11.1104	2.4712	0.8854
LSTM	16.9939	10.8641	2.4523	0.8910

Отримані дані вказують на суттєве покращення точності. За метрикою R² результат підвищився з 0.842 до 0.891 для моделі LSTM після оптимізації даних. Зважена середньоквадратична похибка знизилась більш ніж вдвічі: з 34.79 до 16.99 для найкращих моделей. MAPE не перевищує 3% для всіх оптимізованих моделей. Проведений експеримент вказує на стабільну роботу різних алгоритмів після оптимізації даних.

В таблиці 6 наведено метрики отримані на крос-валідації.

Таблиця 6. Метрики на крос-валідації для другого експерименту

Model	Metrics on cross-validation			
	RMSE	MAE	MAPE	R ²
Random Forest	18.4727	11.6601	2.7065	0.8376
Gradient Boosting	18.2924	11.6019	2.7059	0.8408
XGBoost	18.2369	11.4980	2.6679	0.8418
LSTM	17.7894	10.6537	2.4641	0.8512

В таблиці 7 наведено оцінку важливості ознак для моделі випадкового лісу. Порівняння ознак представлено на рисунку 12.

Таблиця 7. Оцінка важливості ознак для моделі випадкового лісу

Feature	Importance
Lag_1	0.8892
Rolling_Std_7	0.0340
Rolling_Mean_7	0.0181
Day	0.0088
DayOfYear	0.0085
Rolling_Mean_30	0.0076
Lag_90	0.0075
Lag_7	0.0075
Lag_30	0.0074
Rolling_Std_30	0.0074
Year	0.0025
Month	0.0013

Третій експеримент поділяється на дві частини. Дані часового ряду трансформовано з використанням вейвлет-перетворень та перетворень Фур'є, після чого проведено прогнозування. Враховуючи невелику різницю в результатах та обчислювальну складність методу довгої короткотривалої пам'яті, зосередимось на дослідженні методів випадкового лісу, посилення градієнту та XGBoost. Результати експерименту на тестовій вибірці наведено в таблиці 8.

Результати експерименту показують, що найуспішнішим підходом є введення додаткових ознак та трансформація часового ряду з використанням вейвлет-перетворень.

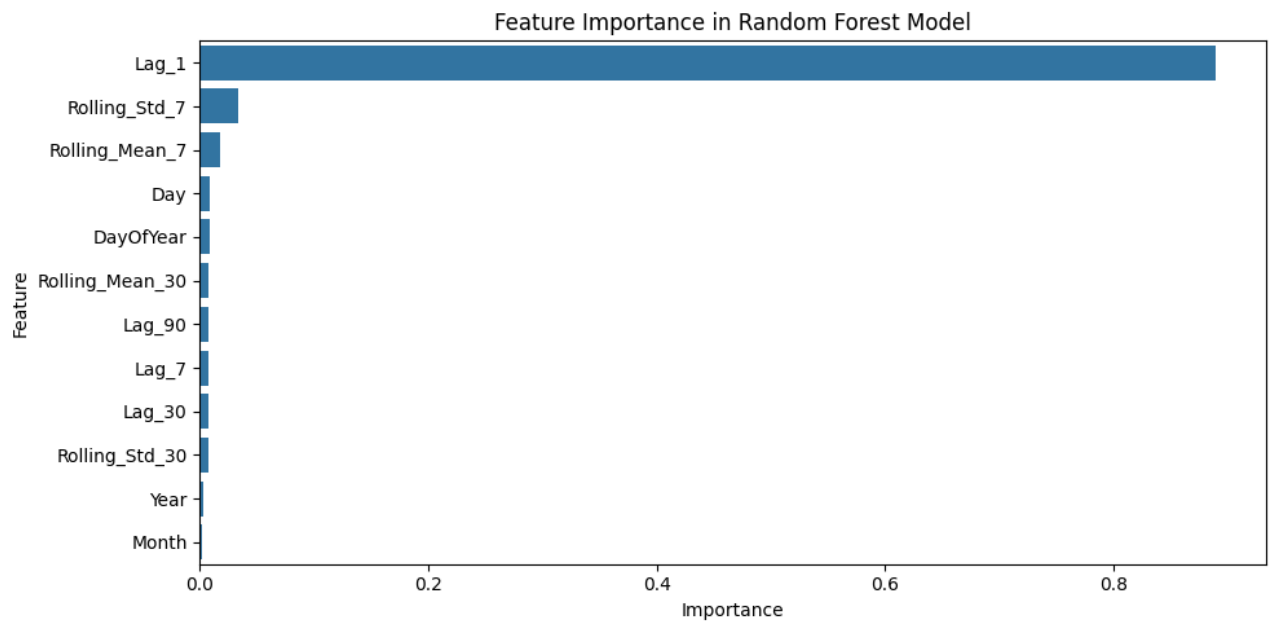


Рисунок 12. Оцінка важливості ознак для моделі випадкового лісу

Таблиця 8. Метрики на тестовій вибірці для третього експерименту

Model	Metrics on test set			
	RMSE	MAE	MAPE	R ²
Wavelet transformations				
Random Forest	16.6971	12.0994	2.7414	0.8942
Gradient Boosting	15.4447	10.4130	2.3661	0.9095
XGBoost	17.8202	12.6528	2.8396	0.8795
Fourier transformations				
Random Forest	18.0867	12.9376	2.9118	0.8759
Gradient Boosting	17.4450	11.6208	2.6078	0.8845
XGBoost	19.5773	13.5525	3.0096	0.8546

Результати експерименту на крос-валідації наведено в таблиці 9.

Таблиця 9. Метрики на крос-валідації для третього експерименту

Model	Metrics on cross-validation			
	RMSE	MAE	MAPE	R ²
Wavelet transformations				
Random Forest	17.3425	11.3582	2.6505	0.8567
Gradient Boosting	17.2847	11.4511	2.6984	0.8546
XGBoost	18.3742	12.2046	2.8636	0.8370
Fourier transformations				
Random Forest	18.4189	11.8118	2.7474	0.8388
Gradient Boosting	18.3493	11.6978	2.7367	0.8391
XGBoost	20.2478	13.3527	3.1224	0.8050

Порівняння прогнозованих значень моделі посилення градієнту з використанням вейвлет-перетворень з реальними даними спостережень наведено на рис. 13.

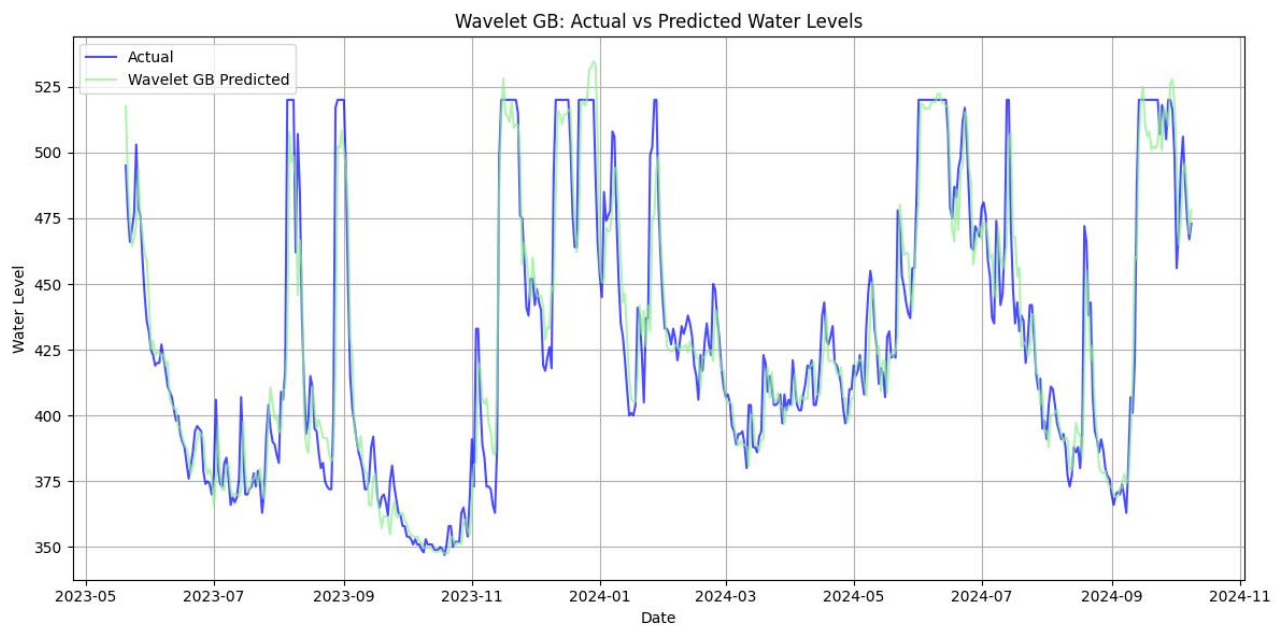


Рисунок 13. Результати прогнозування методу посилення градієнту з використанням вейвлет-перетворень

Результати порівняння метрик представлено на рис. 14, 15.

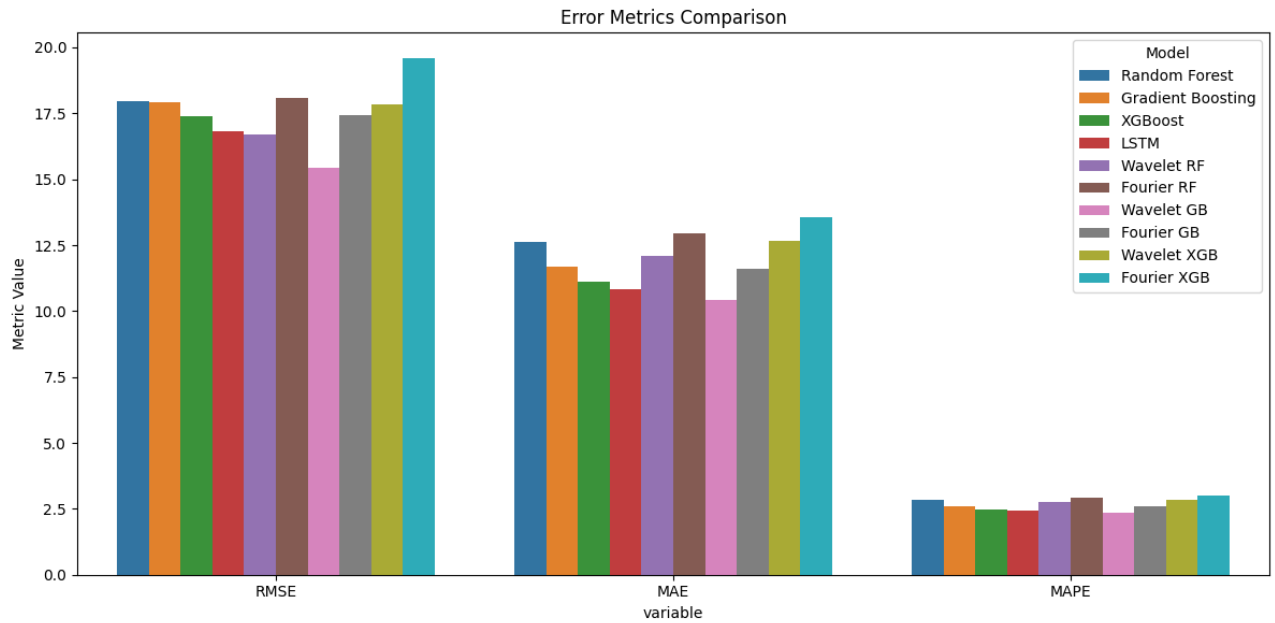


Рисунок 14. Порівняння похибок моделей

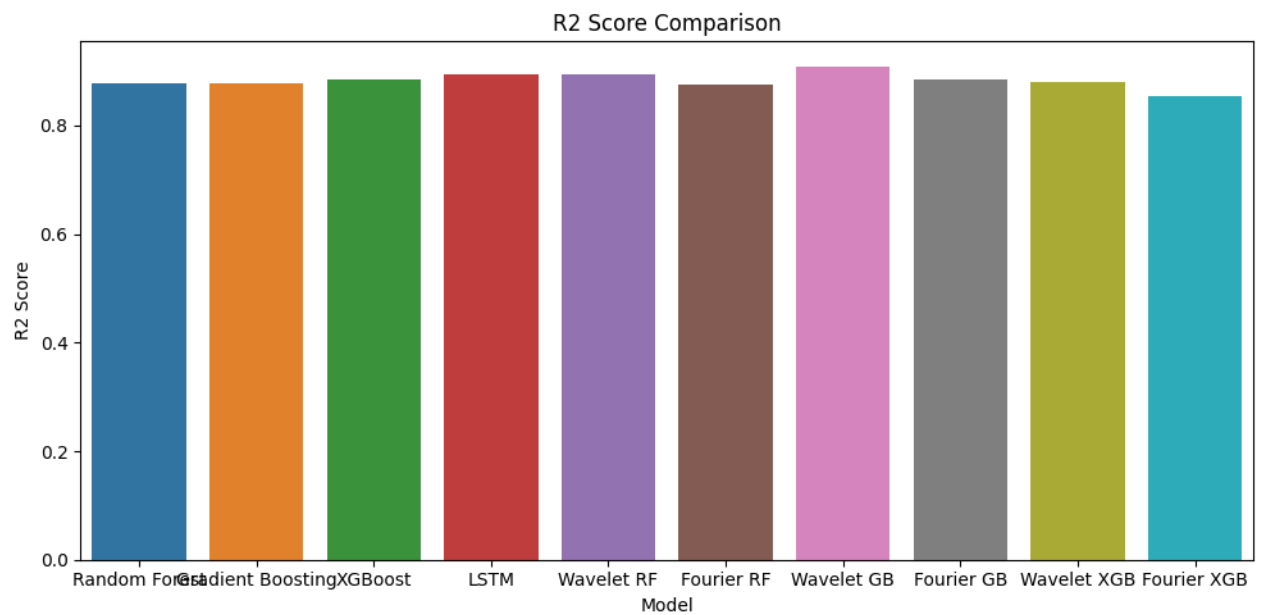


Рисунок 15. Порівняння метрики R^2

Залишки для усіх моделей представлено на рис. 16



Рисунок 16. Залишки для усіх моделей

V. Обговорення

Враховуючи отримані результати, використання інженерії ознак у розглянутій задачі дозволяє суттєво покращити результати моделей машинного навчання. В той же час, варто зауважити, що такі методи можуть мати як ряд переваг, так і деякі обмеження, залежно від характеру даних, які використовуються для прогнозування.

Ковзна середня – це інструмент попередньої обробки та інженерії ознак, який робить сигнал стабільнішим, допомагає виявити тренди, полегшує моделі пошук закономірностей.

Перевагами перетворень Фур'є є можливість виділення частотних ознак. Тобто з часового ряду можна отримати основні частоти коливань, такі як сезонність та цикли і додати їх як ознаки для моделі машинного навчання. Також, перетворення Фур'є дозволяє проводити фільтрацію шуму, тобто високочастотні складові можуть бути обнулені, а сигнал – відновлений для згладженого навчання. Іншим вагомим фактором є виявлення домінуючих періодів. Таким чином, модель може враховувати повторювані патерни.

Обмеження перетворень Фур'є полягають у тому, що метод не показує, коли у часі з'являється певна частота. Також перетворення Фур'є здійснюють припущення про стаціонарність сигналу. Класичні перетворення Фур'є добре працюють, коли статистика сигналу не змінюється з часом. При цьому, у реальних часових рядах частоти можуть змінюватися. Також метод має високу чутливість до шуму, тому навіть незначний шум у часовій області може дати значні зміни у спектрі. Перетворення Фур'є потребують рівномірно дискретизованих даних. Якщо дані нерівномірні по часу, їх треба інтерполювати, що може вносити похибки.

Вейвлет-перетворення дозволяють провести локальний аналіз трендів, тобто дозволяє виділяти короткотривалі події, зокрема піки рівня води та різкі перепади. Вейвлет-перетворення враховують мультимасштабні ознаки, оскільки розкладають ряд на низькочастотну (тренд) і високочастотну (деталі, шум) компоненти. Стиснення даних дозволяє залишити лише суттєві коефіцієнти та зменшити розмірність ознак. Суттєвою перевагою є виявлення аномалій. Нетипові патерни видно на певних масштабах вейвлет-перетворень. Перевагою вейвлет-перетворень над перетвореннями Фур'є є збереження інформації про час, коли відбулися події.

Проте, вейвлет-перетворення мають певні обмеження. Зокрема, різні материнські вейвлети дають різний результат. Не існує універсального критерію

вибору, отже материнський вейвлет потрібно підбирати експериментально. Іншим обмеженням є груба частотна роздільна здатність на високих масштабах. На великих масштабах (низьких частотах) часові інтервали великі, тому втрачаються деталі. На малих масштабах (високих частотах) погіршується частотна точність. Для сигналів з дуже складною структурою може спостерігатись погана інтерпретація. Якщо ряд має багато накладених процесів різної природи, розклад може бути важким для прямого використання у машинному навчанні.

VI. Висновки

Проведене дослідження продемонструвало високу ефективність застосування методів штучного інтелекту для прогнозування рівня води річок. Експериментальне порівняння п'яти різних алгоритмів (LSTM, Gradient Boosting, Random Forest, XGBoost та Prophet) на даних моніторингової станції Берндорф, річка Дунай, за період 2017-2024 рр. показало кардинальне покращення точності прогнозування після оптимізації підготовки даних.

Найкращі результати показала модель XGBoost з коефіцієнтом детермінації $R^2 = 0.8854$, RMSE = 17.38 см та MAPE = 2.47%, що свідчить про надзвичайно високу точність прогнозування. Додаткові експерименти з вейвлет-перетвореннями дозволили покращити результати до $R^2 = 0.9095$ для Gradient Boosting моделі.

Враховуючи розглянуті потенційні обмеження застосування запропонованих методів, в подальших дослідженнях необхідно провести верифікацію отриманих результатів на даних іншого характеру, наприклад прогнозування рівня води в річках іншого регіону.

Проект має чітко визначений план розвитку від локального case study до глобальної платформи. Короткострокові цілі включають розширення на інші річкові системи та створення веб-платформи, довгострокові – комерціалізацію та масштабування на міжнародний рівень.

Проект демонструє не лише високі наукові результати, але й значний потенціал для реального впливу на безпеку життя людей та економічну стабільність регіонів.

Отримані результати підтверджують гіпотезу про переваги методів машинного навчання над традиційними статистичними підходами у сфері гідрологічного прогнозування та відкривають нові можливості для розвитку систем раннього попередження про природні катастрофи.

VII. Використані джерела

- [1] Cred. (2023). EM-DAT – *The international disaster database*. <https://www.emdat.be/>
- [2] *UNDRR 2022 Annual Report*. (2023). United Nations Office for Disaster Risk Reduction. <https://www.undrr.org/annual-report/2022>
- [3] *2024 Sustainability Report | GHD – The Power of Commitment*. (2025). GHD. <https://www.ghd.com/en/2024-sustainability-report>
- [4] Rentschler, J., Salhab, M., & Jafino, B. A. (2022). Flood exposure and poverty in 188 countries. *Nature Communications*, 13(1). <https://doi.org/10.1038/s41467-022-30727-4>
- [5] Mitchell, C. (2024). Intersecting crises: A comparative analysis of global conflicts and the risk of flooding. *Journal of Flood Risk Management*, 17(4). <https://doi.org/10.1111/jfr3.13041>
- [6] *2024 Annual Report highlights*. (2025). Oxfam. <https://www.oxfamamerica.org/about-us/financial-information/2024-annual-report-highlights/>
- [7] Miħu-Pintilie, A., Urzicǎ, A., Stoleriu, C. C., & Pricop, C. I. (2025). Integrating LiDAR-derived DEM, rainfall radar data, and SAR imagery for 2D HEC-RAS modelling to assess the severity of pluvial flash floods induced by Storm Boris in SE Romania. *Geomatics Natural Hazards and Risk*, 16(1). <https://doi.org/10.1080/19475705.2025.2488190>
- [8] *Flooding*. (2025). Copernicus. <https://climate.copernicus.eu/esotc/2024/flooding>

[9] Willard, J. (2025, March 20). *PERILS ups storm Boris European flood industry loss estimate 5% to €2.19bn*. Artemis.bm – the Catastrophe Bond, Insurance Linked Securities & Investment, Reinsurance Capital, Alternative Risk Transfer and Weather Risk Management Site. <https://www.artemis.bm/news/perils-ups-storm-boris-european-flood-industry-loss-estimate-5-to-e2-19bn/>

[10] Dave Anny (2025). Machine Learning Applications in Hydrological Forecasting. *ResearchGate*.
https://www.researchgate.net/publication/392495308_Machine_Learning_Applications_in_Hydrological_Forecasting

[11] Shen, C., & Song, Y. (2025) Improving predictions of flood severity, place and time with AI. *Penn State University*.
<https://www.psu.edu/news/engineering/story/improving-predictions-flood-severity-place-and-time-ai>

[12] Xiao, Q., Zhou, L., Xiang, X., Liu, L., Liu, X., Li, X., & Ao, T. (2022). Integration of hydrological model and time series model for improving the runoff simulation: A case study on BTOP model in Zhou River Basin, China. *Applied Sciences*, 12(14), 6883. <https://doi.org/10.3390/app12146883>

[13] *Cohesion Policy Project deploys artificial intelligence to anticipate and mitigate flood risks*. (2025, February 17). European Commission.
https://ec.europa.eu/regional_policy/whats-new/newsroom/17-02-2025-cohesion-policy-project-deploys-artificial-intelligence-to-anticipate-and-mitigate-flood-risks_en

[14] Szczepanek, R. (2022). Daily streamflow forecasting in mountainous catchment using XGBoost, LightGBM and CatBoost. *Hydrology*, 9(12), 226. <https://doi.org/10.3390/hydrology9120226>

[15] Zakaria, M. N. A., Ahmed, A. N., Malek, M. A., Birima, A. H., Khan, M. M. H., Sherif, M., & Elshafie, A. (2023). Exploring machine learning algorithms for accurate water level forecasting in Muda river, Malaysia. *Heliyon*, 9(7), e17689. <https://doi.org/10.1016/j.heliyon.2023.e17689>

- [16] Hameed, M. M., Masood, A., Hamid, A., Elbeltagi, A., Razali, S. F. M., & Salem, A. (2025). Forecasting monthly runoff in a glacierized catchment: A comparison of extreme gradient boosting (XGBoost) and deep learning models. *PLoS ONE*, 20(5), e0321008. <https://doi.org/10.1371/journal.pone.0321008>
- [17] Manandhar, P., Rafiq, H., Rodriguez-Ubinas, E., & Palpanas, T. (2024). New forecasting metrics evaluated in Prophet, Random Forest, and Long Short-Term Memory models for load forecasting. *Energies*, 17(23), 6131. <https://doi.org/10.3390/en17236131>
- [18] Marr, P., Donato, Y. a. J., Carraro, E., Kanta, R., & Glade, T. (2023). The role of historical Data to investigate Slow-Moving Landslides by Long-Term Monitoring Systems in Lower Austria. *Land*, 12(3), 659. <https://doi.org/10.3390/land12030659>
- [19] Lambert, P. (2025, May 27). Early warning systems for floods based on AI: Protecting communities with cutting-edge technology. *globalsociety.earth*. <https://www.globalsociety.earth/post/early-warning-systems-for-floods-based-on-ai-protecting-communities-with-cutting-edge-technology>
- [20] *Wasserstandsnachrichten und Hochwasserprognosen*. (2025). <https://www.noel.gv.at/wasserstand/#/en/Messstellen/Map/Wasserstand>