**ДЕРЖАВНИЙ УНІВЕРСИТЕТ**
**ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ**
**НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ**
**ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ**
**КАФЕДРА ІНФОРМАЦІЙНИХ СИСТЕМ ТА ТЕХНОЛОГІЙ**

# КВАЛІФІКАЦІЙНА РОБОТА

на тему: «Методика для аналізу і передбачення погодних умов за допомогою нейронної мережі»

на здобуття освітнього ступеня магістра
зі спеціальності 124 Системний аналіз
освітньо-професійної програми «Інтелектуальні системи управління»

*Кваліфікаційна робота містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело*

_____ Микола ЛЯШУК
*(підпис)*

Виконав: здобувач вищої освіти групи САДМ-61

Микола ЛЯШУК

Керівник: Вікторія ШКАПА
*к.ф-м.н., доцент*

Рецензент: _____
*науковий ступінь, вчене звання* *Ім'я, ПРІЗВИЩЕ*

**Київ 2024**

**ДЕРЖАВНИЙ УНІВЕРСИТЕТ**
**ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ**

**Навчально-науковий інститут інформаційних технологій**

Кафедра інформаційних систем та технологій

Ступінь вищої освіти магістр

Спеціальність 124 Системний аналіз

Освітньо-професійна програма Інтелектуальні системи управління

**ЗАТВЕРДЖУЮ**
Завідувач кафедрою ІСТ
_____Каміла СТОРЧАК
«_____» ___20___р.

# ЗАВДАННЯ
# НА КВАЛІФІКАЦІЙНУ РОБОТУ

Ляшук Микола Анатолійович
*(прізвище, ім'я, по батькові здобувача)*

1. Тема кваліфікаційної роботи: Методика для аналізу і передбачення погодних умов за допомогою нейронної мережі

керівник кваліфікаційної роботи Вікторія ШКАПА к.ф-м.н, доцент,
*(Ім'я, ПРІЗВИЩЕ, науковий ступінь, вчене звання)*

затверджені наказом Державного університету інформаційно-комунікаційних технологій

від «____» _____20___р. № ____

2. Строк подання кваліфікаційної роботи «___» _____20___р.

3. Вихідні дані до кваліфікаційної роботи: методи машинного навчання, штучний інтелект, прогнозування погодних умов, статистичний аналіз, бінарні дерева даних.

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити)
   1. Дослідження предметної області
   2. Методи прогнозування погоди
   3. Методи прогнозування швидкості вітру в аеропортах

5. Перелік ілюстративного матеріалу: *презентація*

6. Дата видачі завдання «_____» _____20____р.

# КАЛЕНДАРНИЙ ПЛАН

| № з/п | Назва етапів кваліфікаційної роботи | Строк виконання етапів роботи | Примітка |
|---|---|---|---|
| 1 | Підбір науково-технічної літератури | | Виконано |
| 2 | Дослідження предметної області | | Виконано |
| 3 | Методи прогнозування погоди | | Виконано |
| 4 | Методи прогнозування швидкості вітру в аеропортах | | Виконано |
| 5 | Висновки, вступ, реферат | | Виконано |
| 6 | Розробка презентації | | Виконано |
| | | | |
| | | | |

Здобувач(ка) вищої освіти _____ Микола ЛЯШУК
*(підпис)* *(Ім'я, ПРІЗВИЩЕ)*

Керівник
кваліфікаційної роботи _____ Вікторія ШКАПА
*(підпис)* *(Ім'я, ПРІЗВИЩЕ)*

# РЕФЕРАТ

Текстова частина кваліфікаційної роботи на здобуття освітнього ступеня магістра: 109 стор., 38 рис., 21 табл., 30 джерел.

*Мета роботи* – дослідити методики аналізу і прогнозування погодних умов з використанням сучасних методів машинного навчання, зокрема нейронних мереж, для підвищення точності прогнозів і покращення ефективності роботи систем моніторингу погоди.

*Об'єкт дослідження* – процеси аналізу і прогнозування погодних умов на основі метеорологічних даних.

*Предмет дослідження* – методи і алгоритми машинного навчання, зокрема нейронні мережі, які використовуються для обробки та аналізу метеорологічних даних з метою прогнозування погодних умов.

*Короткий зміст роботи:* Комп'ютерне моделювання відіграє важливу роль у вивченні стану та розвитку атмосфери, проте бракує методів автоматизації інтерпретації даних, отриманих із цих моделей. У цій роботі розглядається застосування методів машинного навчання для вирішення специфічних завдань метеорології, з особливим акцентом на вдосконаленні точності чисельних моделей прогнозування погоди. Для виконання метеорологічної регресії використовуються класичні підходи, такі як багатовимірна непараметрична регресія та бінарні дерева рішень. Аналіз даних про погоду здійснюється як загальна задача структурованого прогнозування із застосуванням глибоких нейронних мереж. Згорткові та рекурентні нейронні мережі дозволяють ефективно враховувати просторово-часову структуру, характерну для моделей прогнозування погоди. У дослідженні вивчається потенціал глибоких згорткових нейронних мереж у вирішенні складних завдань метеорології.

КЛЮЧОВІ СЛОВА: МЕТОДИ МАШИННОГО НАВЧАННЯ, ШТУЧНИЙ ІНТЕЛЕКТ, ПРОГНОЗУВАННЯ ПОГОДНИХ УМОВ, СТАТИСТИЧНИЙ АНАЛІЗ, БІНАРНІ ДЕРЕВА ДАНИХ

## ABSTRACT

Text part of the master's qualification work:
109 pages, 38 pictures, 21 table, 30 sources.

*The purpose of the work* is to explore methods for analyzing and forecasting weather conditions using modern machine learning techniques, particularly neural networks, to enhance forecast accuracy and improve the efficiency of weather monitoring systems.

*Object of research* – processes of analyzing and forecasting weather conditions based on meteorological data.

*Subject of research* – methods and algorithms of machine learning, particularly neural networks, used for processing and analyzing meteorological data to forecast weather conditions.

*Summary of the work:* Computer modeling plays a crucial role in studying the state and development of the atmosphere; however, there is a lack of methods for automating the interpretation of data generated by these models. This work examines the application of machine learning methods to address specific challenges in meteorology, with a particular focus on improving the accuracy of numerical weather prediction models. Classical approaches, such as multivariate nonparametric regression and binary decision trees, are used for meteorological regression. Weather data analysis is approached as a structured prediction task using deep neural networks. Convolutional and recurrent neural networks effectively capture the spatial and temporal structures inherent in weather prediction models. This study explores the potential of deep convolutional neural networks for solving complex problems in meteorology.



KEYWORDS: MACHINE LEARNING METHODS, ARTIFICIAL INTELLIGENCE, WEATHER CONDITION FORECASTING, STATISTICAL ANALYSIS, BINARY DATS TREES

**ДЕРЖАВНИЙ УНІВЕРСИТЕТ**
**ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ**

**Навчально-науковий інститут інформаційних технологій**

**ПОДАННЯ**
**ГОЛОВІ ЕКЗАМЕНАЦІЙНОЇ КОМІСІЇ**
**ЩОДО ЗАХИСТУ КВАЛІФІКАЦІЙНОЇ РОБОТИ**
**на здобуття освітнього ступеня магістра**

Направляється здобувач(ка) *Ляшук М. А.* до захисту кваліфікаційної роботи
*(прізвище та ініціали)*
за спеціальністю *124 Системний аналіз*
*(код, найменування спеціальності)*
освітньо-професійної програми Інтелектуальні системи управління
*(назва)*
на тему: «Методика для аналізу і передбачення погодних умов за допомогою нейронної мережі».

Кваліфікаційна робота і рецензія додаються.

Директор ННІІТ _____ Катерина НЕСТЕРЕНКО
*(підпис)* *(Ім'я, ПРІЗВИЩЕ)*

**Висновок керівника кваліфікаційної роботи**

Здобувач Ляшук Микола Анатолійович під час написання кваліфікаційної роботи показав гарну теоретичну підготовку, володіння необхідними знаннями з системного аналізу, користуватися навчальною, довідковою і науково-технічною літературою. Працюючи над завданнями, які були дані керівником, проявляв ініціативність, сумлінність та хист до роботи.

Все це дозволяє оцінити виконану кваліфікаційну роботу здобувача(ки) Миколи ЛЯШУКА на оцінку «_____» та присвоїти йому(їй) кваліфікацію магістр з системного аналізу.

Керівник кваліфікаційної роботи _____ Вікторія ШКАПА
*(підпис)* *(Ім'я, ПРІЗВИЩЕ)*

«_____» ____20_____року

**Висновок кафедри про кваліфікаційну роботу**

Кваліфікаційна робота розглянута. Здобувач(ка) *Ляшук М. А.* допускається до захисту даної роботи в Екзаменаційній комісії.

Завідувач кафедрою ІСТ _____ Каміла СТОРЧАК
*(назва)* *(підпис)* *(Ім'я, ПРІЗВИЩЕ)*

**ВІДГУК РЕЦЕНЗЕНТА**

**на кваліфікаційну роботу**

**на здобуття освітнього ступеня магістра**

здобувача(ки) вищої освіти Ляшука Миколи Анатолійовича
(прізвище, ім'я, по батькові)

на тему «Методика для аналізу і передбачення погодних умов за допомогою нейронної мережі»

**Актуальність.**

Сучасні погодні умови мають значний вплив на різні сфери життя, включаючи сільське господарство, транспорт, енергетику та урбаністику, військові цілі. Використання нейронних мереж дозволяє обробляти великі обсяги даних і виявляти складні закономірності, що забезпечує більш точне та адаптивне прогнозування.

**Позитивні сторони.**

1. В роботі досліджена регресія метеорологічних даних, що використовує класичні методології, такі як багатовимірна непараметрична регресія та бінарні дерева.
2. Проведено аналіз даних про погоду як загальну задачу структурованого прогнозування з використанням глибоких нейронних мереж. Нейронні мережі, такі як згорткові та рекурентні мережі, забезпечують метод для захоплення просторової та часової структури, властивої моделям прогнозування погоди.
3. Дослідження, що лежить в основі цієї роботи, є прикладом того, як співпраця між дослідницькими спільнотами машинного навчання та метеорології є взаємовигідною та веде до прогресу в обох дисциплінах.

**Недоліки.**

1. Недостатньо повно описаний порівняльний аналіз досліджуваних методів
2. Не в повній мірі описаний вибір всіх параметрів при розрахунках.

Відзначені зауваження не впливають на загальну позитивну оцінку кваліфікаційної роботи магістерської

**Висновок:** *кваліфікаційна робота на здобуття ступеня магістра заслуговує оцінку "_____", а здобувач(ка) Микола ЛЯШУК заслуговує присвоєння кваліфікації магістр з системного аналізу*

Рецензент:

*науковий ступінь, вчене звання* _____ *підпис* _____ *Ім'я, ПРІЗВИЩЕ*

**ЗМІСТ**

# ВСТУП

Прогнозування погоди навіть сьогодні – це діяльність, в основному виконана людьми. Незважаючи на те, що комп'ютерне моделювання відіграє важливу роль у моделюванні стану та еволюції атмосфери, не вистачає методологій для автоматизації інтерпретації інформації, отриманої за допомогою цих моделей. У цій кваліфікаційній роботі досліджується використання методологій машинного навчання для вирішення конкретних проблем у метеорології та особливо зосереджується на вивченні методологій для підвищення точності числових моделей прогнозування погоди. Робота, представлена в цьому рукописі, містить два різні підходи до застосування машинного навчання до проблем прогнозування погоди. У першій частині для виконання регресії метеорологічних даних використовуються класичні методології, такі як багатовимірна непараметрична регресія та бінарні дерева. Ця перша частина особливо зосереджена на прогнозуванні вітру, кругова природа якого створює цікаві проблеми для класичних алгоритмів машинного навчання. Друга частина цієї дисертації досліджує аналіз даних про погоду як загальну задачу структурованого прогнозування з використанням глибоких нейронних мереж. Нейронні мережі, такі як згорткові та рекурентні мережі, забезпечують метод для захоплення просторової та часової структури, властивої моделям прогнозування погоди. У цій частині досліджується потенціал глибоких згорткових нейронних мереж для вирішення складних проблем у метеорології, таких як моделювання опадів на основі полів базової числової моделі. Дослідження, що лежить в основі цієї роботи, є прикладом того, як співпраця між дослідницькими спільнотами машинного навчання та метеорології є взаємовигідною та веде до прогресу в обох дисциплінах. Моделі прогнозування погоди та дані спостережень являють собою унікальні приклади великих (петабайт), структурованих і високоякісних наборів даних, які потрібні спільноті машинного навчання для розробки наступного покоління масштабованих алгоритмів.

# РОЗДІЛ 1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

## 1.1 Витоки та еволюція прогнозування погоди

Прогнозування погоди визначається як застосування науки і техніки для визначення умов атмосфери для даного місця і часу в майбутньому. З історичної точки зору, люди намагалися зрозуміти поведінку атмосфери, вивчаючи закономірності та взаємозв'язки між явищами та пов'язуючи їх з майбутні події. Наприклад, протягом століть було добре відомо, що раптове зниження барометричного тиску часто супроводжується випаданням опадів.

У 1922 р. Л. Ж. Річардсон запропонував використовувати основні рівняння механіки рідини для моделювання рухів атмосфери. У той час не було можливості автоматизувати обчислення, тому цьому автору прийшла в голову ідея розбити поверхню Землі на осередки і використовувати людей для розв'язування диференціальних рівнянь, що описують рухи атмосфери. За його оцінками, для створення оновленого прогнозу для всієї планети знадобиться армія з 64 000 людських комп'ютерів. На рисунку 1.1 представлено задумане будівля Річардсона, щоб розмістити свою ідею людської «фабрики електронного лиття». На жаль, через масштаби цього проекту ідея Річардсона так і не була реалізована, і глобальних прогнозів погоди довелося чекати ще два десятиліття.

Наприкінці Другої світової війни, коли стали доступними електронні комп'ютери та дані атмосферного радіозондування, Сполучені Штати очолили амбітний проект із створення першого автоматизованого система прогнозування. У 1950 році метеорологи Жюль Шарні, Агнар Фьортофф і математик Джон фон Нейман опублікували статтю під назвою «Чисельне інтегрування баротропіс V рівняння оптичності», яка встановлює основу для комп'ютерних моделей погоди, задаючи основи числової дикції (NWP), як ми його знаємо. Результатом цієї роботи стала реалізація першого прогнозу погоди електронними комп'ютерами. Модель була реалізована і запущена в операційному режимі на комп'ютері ENIAC в

Університеті Пенсільванії. Через обмеження ємності цього комп'ютера для створення 24-годинного прогнозу знадобилося 24 години обробки, що обмежило його практичне застосування, оскільки він не міг ефективно передбачати майбутнє.

Рис. 1.1 - Представлення ідеї Річардсона про "прогнозну - торію", в якій було б зайнято близько шістдесяти чотирьох тисяч людей-комп'ютерів, що сидять ярусами по колу гігантський глобус, що розв'язує рівняння для прогнозування погоди.

З 1950-х років комп'ютери використовуються для моделювання стану та еволюції атмосфери за допомогою моделей NWP. Ці моделі використовують набір нелінійних диференціальних рівнянь для апроксимації стану та еволюції атмосферної сфери, які відомі як примітивні рівняння. Ці примітивні рівняння визначають збереження маси, імпульсу і теплової енергії. Примітивні рівняння розв'язуються за допомогою моделей NWP з використанням скінченних різниць, або спектральних методів, для трьох вимірів простору і часу. Моделі NWP ініціалізують ці рівняння, використовуючи спостережувані дані, щоб створити знімок стану атмосфери. Цей процес відомий як «аналіз». Щоб інтегрувати параметри, змодельовані за допомогою цих рівнянь, Земля розділена на дискретну сітку, яка використовується для представлення еволюції регіонів атмосфера крізь час. На рисунку 1.2 представлено розподіл точок сітки в регулярній гауссовій сітці.

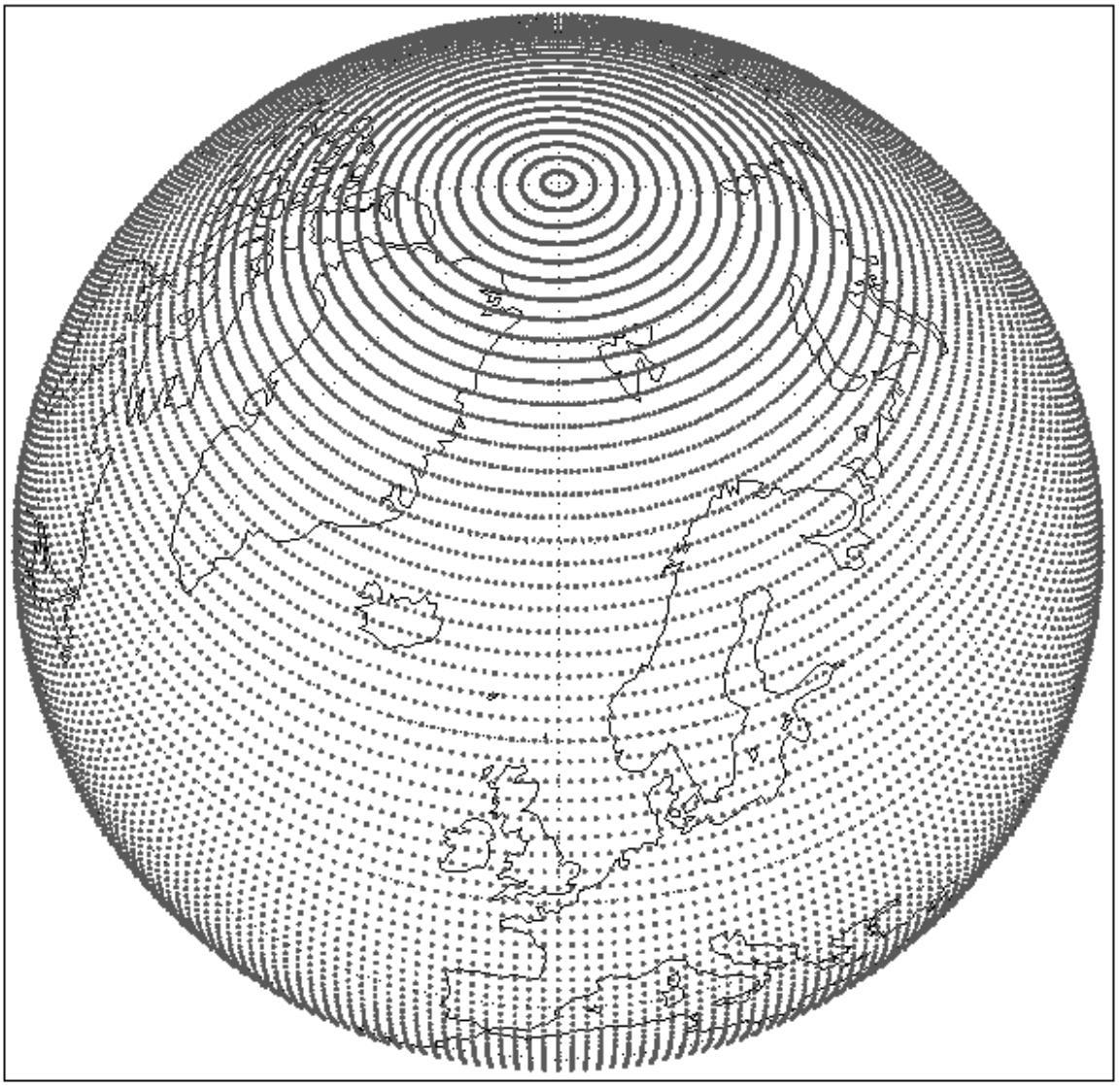


Рис. 1.2 - Приклад регулярної гауссової сітки F80, що використовується деякими NWP для представлення атмосфери Землі

Тому моделі NWP будуються з використанням математичних рівнянь, які описують динамічні фізичні процеси в атмосфері. Щоб змоделювати майбутній стан атмосфери, той же набір рівнянь вирішується використовуючи вихідні дані попереднього кроку. Цей процес повторюється до тих пір, поки розчин не досягне потрібного прогнозованого часу. Однак помилки в модельованих змінних накопичуються з часом, і точність обчисленого прогнозу погіршується на кожному кроці. Використання занадто грубої сітки клітини або тривалі кроки часу інтеграції є основними причинами помилок NWP. Значним стає залучення дрібномасштабних або підгріткових процесів, які явно не розв'язуються фізичними

рівняннями в моделі. Збільшення просторової і темпоральної роздільної здатності NWP частково вирішує цю проблему, але ціною значного підвищення обчислювальних вимог моделі.

Одним з основних обмежень, що обмежують NWP, є роздільна здатність сітки, що використовується для розв'язання рівнянь, яка зазвичай становить кілометри. Ця роздільна здатність, як правило, менша, ніж природний масштаб деяких важливих атмосферних процесів. Такі процеси, як конвенція, радіаційне перенесення або утворення хмар, відбуваються в менших масштабах, ніж моделі можуть явно вирішити. Для цих дрібномасштабних або складних процесів моделі NWP використовують спрощені або наближені процеси, які називаються параметризацією.

Параметризація є одним з найбільш комплікованих компонентів NWP. Правильна параметризація дрібномасштабних процесів набуває вирішального значення при прогнозуванні подій більш ніж за 48 годин, коли ефект від цих процесів зазвичай стає значним. Параметризацій, як правило, будуються з використанням спрощених моделей для апроксимації атмосферних процесів. Ці моделі часто визначаються шляхом поділу процесу на різні категорії, де в кожному випадку застосовуються різні лінійні моделі.

Параметризації для різних процесів іноді тісно пов'язані між собою і повинні бути вказані в сукупності, що призводить до нетривіальних взаємодій і залежностей. На рисунку 1.3 представлено результати Системи прогнозування клімату (CFS) Національного управління океанів і атмосферних ресурсів (NOAA), що представляє глобальні атмосферні опади вода, яка є параметризованою змінною.

Ще одним наслідком роздільної здатності сітки NWP є обмеження в обліку ефекту топографу. У моделях Global NWP використовується груба оцінка форми та особливостей поверхні Землі, усереднення берегових ліній та гір висоти над протяжністю відповідної комірки сітки, яка зазвичай становить від 10 до 100 кілометрів. Як наслідок, вони не в змозі відобразити важливі локальні ефекти, які відбуваються в масштабі підсітки.

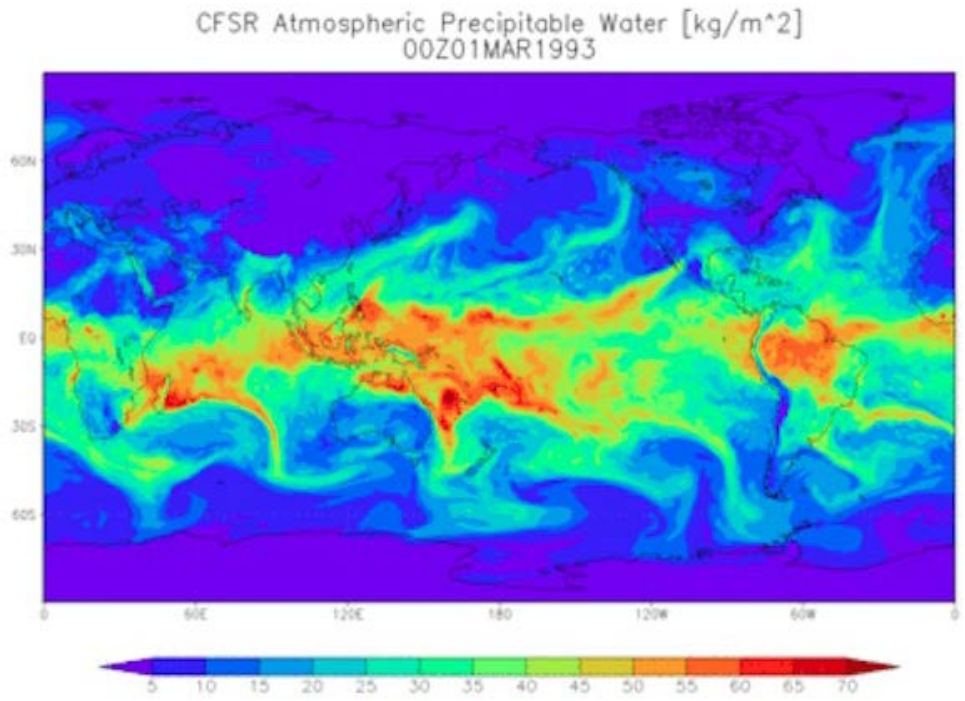



Рис 1.3 - Вихід NOAA CFS для глобальних атмосферних опадів води 15 березня 1993 року за 12-годинний період накопичення

Ще одним серйозним обмеженням, з яким стикаються фізичні моделі NWP, є складність представлення хаотичної природи атмосфери. Математичні рівняння, що регулюють динаміку атмосферних потоків, нелінійні і являють собою нестійкі гідродинамічні і термодинамічні процеси. Невеликі відмінності в початковому стані можуть посилюватися в міру розвитку системи, що призводить до значних відмінностей сценаріїв. Моделі NWP ініціалізуються на етапі аналізу з використанням даних спостережень, які мають внутрішню невизначеність. Ця невизначеність може бути опанована в часі, що призводить до мінливості результатів. Нестабільність атмосферної сфери визначає верхню і нижню межу передбачуваності миттєвих погодних умов.

На початку 1990-х років метеорологічне співтовариство запропонувало ансамблеве прогнозування як методологію вимірювання передбачуваності атмосфери за допомогою NWP. Замість того, щоб робити одну симуляцію з високою роздільною здатністю, ансамблеві прогнози запускають набір прогнозів з дещо іншими початковими умовами. Цей набір прогнозів дає уявлення про діапазон можливих майбутніх станів атмосфери та можливі сценарії розвитку подій. На рисунку 1.4 представлена еволюція температурної змінної, представлена у вигляді

перетворення у вигляді її ймовірного розподілу в часі. Члени зазвичай агрегуються навколо значень, які відповідають найбільш ймовірним сценаріям для даного фізичного параметра.

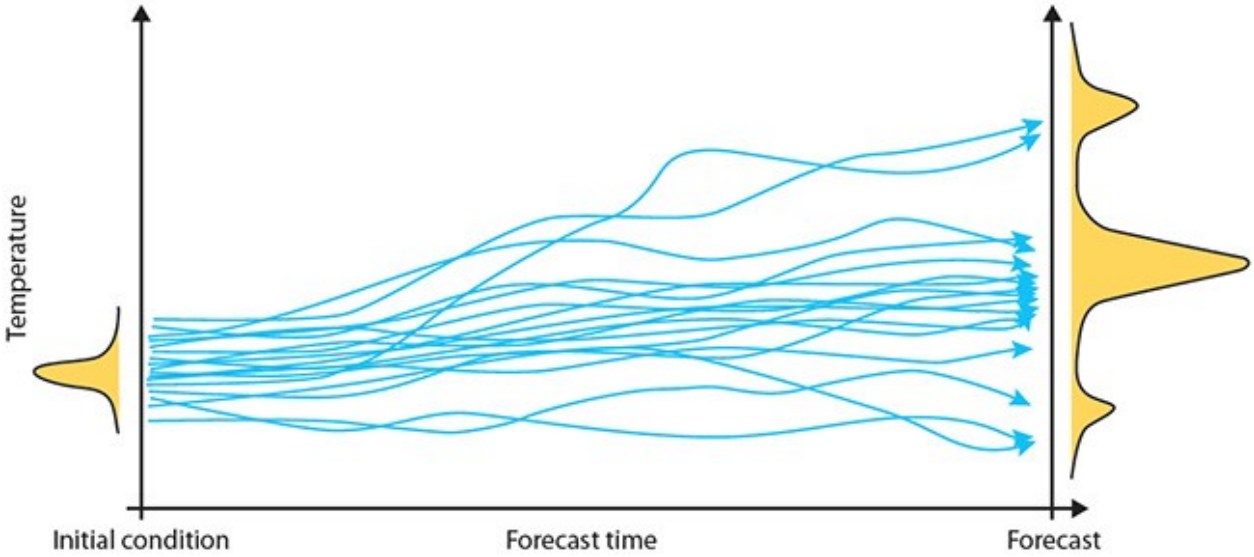



Рис. 1.4 - Ансамбль прогнозів виробляє ряд можливих сценаріїв з урахуванням початкового розподілу ймовірностей прогнозованого параметра. Різні учасники ансамблю дають уявлення про можливі результуючі сценарії на основі розподілу ймовірностей.

Як ми бачили, представлення нелінійних підгрід-процесів, а також кількісна оцінка і поширення невизначеності є центральними в процесі прогнозування погоди. Перше тісно пов'язане з масштабом рівнянь, що використовуються для відтворення атмосферних фізичних процесів, а друге - з точністю датчиків і впевненість, яку ми маємо в їхніх вимірах.

## 1.2 Чисельний прогноз погоди

Як було представлено в попередньому розділі, комп'ютери використовуються з 1950-х років для моделювання стану та еволюції атмосфери. З тих пір потужність комп'ютерів подвоювалася кожні 18 місяців відповідно до закону Мура і, відповідно, роздільної здатності погодних моделей. Хоча фізичні рівняння і методології їх розв'язання залишилися принципово такими ж, як і в перших моделях погоди, просторова і часова роздільна здатність. Так як частота пробігів моделей, постійно збільшувалася з плином часу. З точки зору точності – навички

роботи з літературою з прогнозування погоди – з прогнозів показники NWP покращувалися на два дні за декаду, як показано на рисунку 1.5. Це означає, що точність 4-денного прогнозу сьогодні така ж хороша, як і 10 років тому для 2-денного прогнозу.

Сфера верифікації NWP широко вивчена в прогнозуванні погоди. Потреба в методах, здатних виміряти точність і якість отриманої інформації, є принциповою для виявлення слабких місць в моделюванні атмосферних процесів. Двома базовими лініями, які часто використовуються для вимірювання якості прогнозів, є стійкість у короткострокових (0-2 дні) прогнозах та кліматологія в середньострокових (2-5 днів) до довгострокових (5+ днів) прогнозів. Наполегливість – це припущення, що метеорологічні умови залишаться незмінними. У цьому контексті модель NWP повинна краще прогнозувати погоду, ніж проста модель, яка зберігає змінні постійними в часі. З іншого боку, кліматологічний базовий рівень припускає, що погода буде поводитися як усереднені історичні записи для цієї конкретної місцевості.

Одним з основних обмежень моделей NWP була відсутність спостережень за великими регіонами світу, що обмежувало ініціалізацію початкового стану моделей, називається «аналіз». Довгий час умови над Тихим океаном або південною півкулею були в основному невідомі через відсутність наземних станцій і даних про зондування атмосфери. Це обмеження було значно покращено з появою супутників і дистанційних датчиків. З 1990-х років супутникові дані стали доступними, і асиміляція цих даних моделями NWP призвела до значного поліпшення якості їх прогнозу. Очікується, що ця тенденція збережеться, оскільки з'являться нові супутники, оснащені більш потужними датчиками. Наприклад, у серпні 2018 року за спільною ініціативою Європейського Союзу та Європейського космічного агентства (ESA) було запущено Aeolus – перший супутник, здатний стежити за земними вітрами. Очікується, що лише цей супутник значно підвищить точність моделей NWP, надаючи цінну інформацію про вітри з глобальним покриттям.

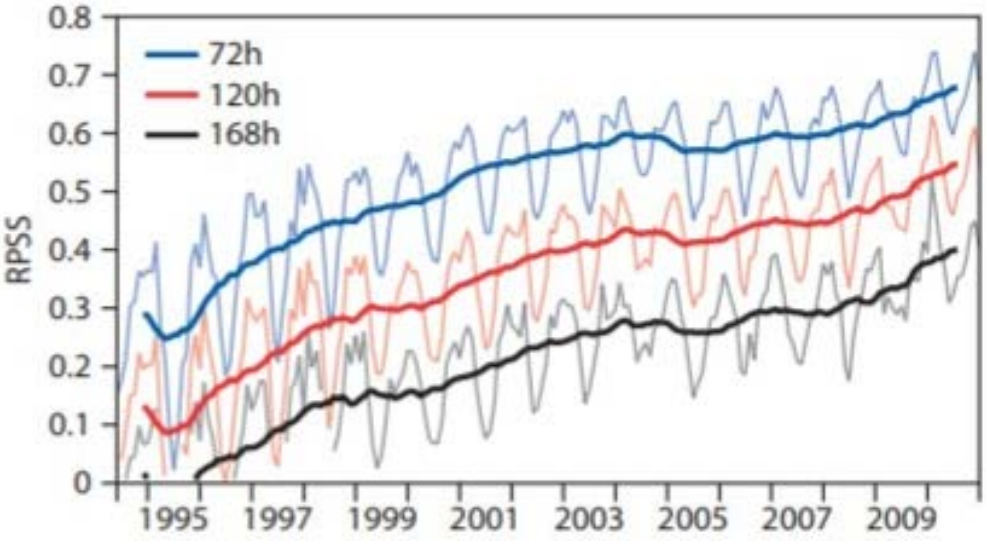



Рис.1.5 - Продуктивність Системи прогнозування ансамблів (EPS)

Синоптична шкала в метеорології (також відома як великомасштабна або циклонічна шкала) - це горизонтальна шкала довжин порядку 1000 кілометрів і більше. Моделі NWP змогли змоделювати присутність та еволюцію синоптичних атмосферних систем з самого початку. Це включає такі явища, як западини середніх широт, фронти та більшість областей високого та низького тиску, які можна побачити на метеорологічних картах. Оскільки розмір комірок сітки, які використовуються для імітації погоди, був зменшений, моделі отримали можливість моделювати атмосферні процеси, що відбуваються в менших масштабах. Сучасні моделі здатні явно розв'язувати процеси, що відбуваються на мезомасштабному рівні, що включає явища, що проявляються в масштабах від 5 до 100 км, такі як морський бриз, шквалові лінії або конвертині осередки середнього та великого розміру. Поточні глобальні моделі погоди працюють з роздільною здатністю, яка коливається від 25 до 100 км, що дозволяє частково врахувати деякі з цих мезомасштабів системи. Такі важливі явища, як конвекція, турбулентність, радіаційні процеси або злиття крапель дощу, відбуваються на мікрорівні, який включає масштаби розміром менше 5 км. Моделі NWP не можуть явно розв'язувати фізичні рівняння, що керують цими процесами.

Фізичні змінні, що моделюються NWP, можна розділити на два основних типи, залежно від характеру рівнянь, що використовуються для їх моделювання.

Базові поля - це поля, які явно обчислюються за допомогою НВП для розв'язання фізичних рівнянь. Приклади таких атмосферних явищ як вітер, температура або вологість. Похідні поля - це поля, які явно не розв'язуються NWP, але є похідними від базових полів за допомогою параметризації. Опади, конвекція, теплові радіаційні процеси або турбулентність є прикладами похідних полів.

У зв'язку з відносно грубою роздільною здатністю моделей NWP необхідно враховувати репрезентативність її змінних та їх інтерпретацію на конкретних комірках сітки. На рисунку 1.6 порівнюється топографія Європи з використанням двох різних роздільних здатностей. Як видно, рівень деталізації значно відрізняється в різних представленнях, і важливі деталі в гірських і прибережних районах не можуть бути представлені за допомогою грубих представлень. Це означає, що моделі NWP не можуть враховувати вплив топографії на погодні змінні в масштабах, менших за розмір комірки сітки.

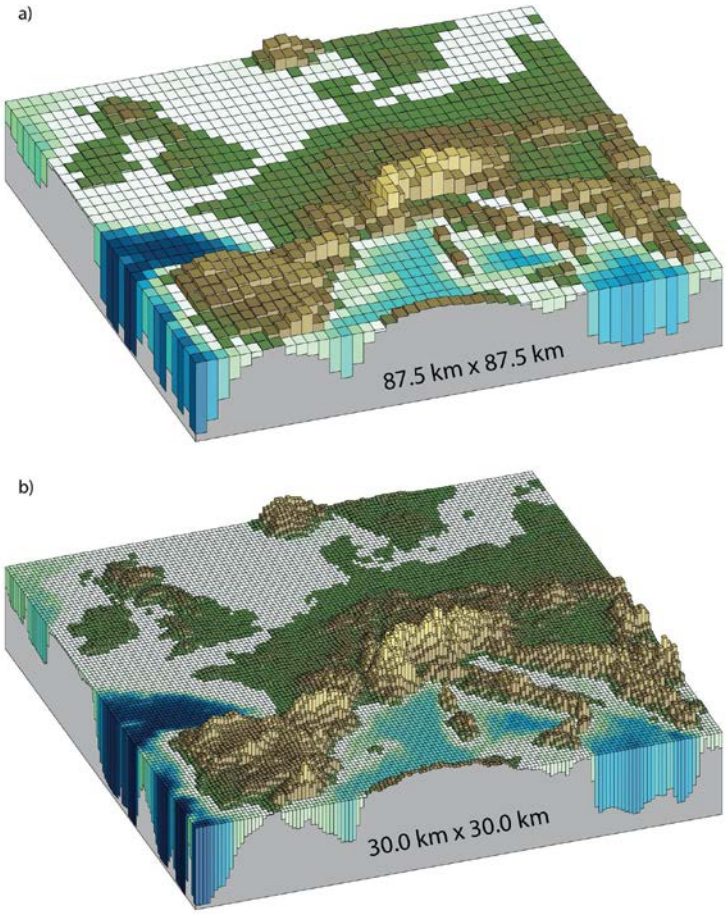



Рис.1.6 - Порівняння горизонтальної роздільної здатності, що використовується в глобальних моделях сьогодні [30 км] (b) та еквівалентних моделях 10 років тому [87,5 км] (a).

Поширеною проблемою в інтерпретації даних NWP є висновок про формування високої роздільної здатності зі змінних з низькою роздільною здатністю. Цей процес називається «даунскейлінгом» у таких дисциплінах, як метеорологія, кліматологія або ремот-зондування. Цей метод може ґрунтуватися на динамічних або статистичних підходах, таких як сплайни, крігинг або вкладеність у моделі з вищою роздільною здатністю. Методи, засновані на машинному навчанні, відносяться до категорії статистичних методів, і вони можуть бути використані для поліпшення результатів навчання НВП і вилучення закономірностей зв'язок між історичними результатами моделі та відповідними спостережуваними значеннями.

Моделі НВП - це складні системи, які, як правило, розробляються великими організаціями протягом десятиліть. Їх розробка вимагає високого ступеня спеціалізації, а окремі компоненти присвячені цілим командам висококваліфікованих вчених моделі. Моделі часто проектуються з використанням модулів, які можуть запускатися з певною незалежністю від решти системи. Однак атмосфера є складною системою, в якій більшість змінних взаємопов'язані, визначаючи залежності і процеси зворотного зв'язку між ними. NWP оптимізовані для роботи на об'єктах високопродуктивних обчислень з використанням Fortran і C і таких бібліотек, як OpenMP і MPI для розподілу обчислень між великими обчислювальними кластерами і прискорення генерації їх результатів.

## 1.3 Джерела даних про погоду

Отже, ми зосередили нашу увагу на NWP, оскільки це єдиний доступний інструмент, який здатний «передбачити» еволюцію та майбутній стан фізичних параметрів, що описують атмосферу. Однак вихід ПНП - не єдине джерело інформації при вивченні атмосфери. Датчики, у багатьох формах, забезпечують точні значення спостережуваних умов у певній частині атмосфери. Прикладами датчиків, що використовуються в прогнозуванні погоди, є: наземні станції, повітряні кулі атмосферного зондування, метеорологічні радари або супутники.

Кожен із цих типів датчиків має різні характеристики з точки зору просторової та часової роздільної здатності, а також протяжності, яку вони можуть охопити.

Набори даних спостережень мають вирішальне значення в процесі прогнозування погоди, оскільки вони забезпечують живий потік інформації, що описує поточний стан атмосфери. Прогнозисти використовують ці дані в операційному плані для перевірки результатів NWP і корекції можливих помилок або локальних ефектів.

Особливе значення для цієї дипломної роботи мають METAR, які є звітами про спостереження за погодою, що створюються в більшості аеропортів світу. Ці звіти використовуються для планування контролю за повітряним рухом в аеропортах і є одним із найякісніших джерел даних, що спостерігаються. Наведений нижче код є звітом METAR для аеропорту Доностія, Іспанія. Код починається з коду аеропорту Міжнародної організації цивільної авіації, дати аеропорту, вітрових умов, видимості, хмарності, температури та закінчується умовами тиску.

Моделі NWP також об'єднують всі джерела спостережуваних даних для встановлення початкових умов моделі. Аналіз NWP є складним процесом, в якому всі джерела інформації разом з попереднім результатом моделі повинні бути об'єднані в узгоджену спосіб. На рисунку 1.7 представлені деякі з найважливіших даних датчиків, які використовуються для ініціалізації моделей NWP.

Для аналізу моделей NWP використовуються дві основні методики: ансамблевий фільтр Калмана (EKF) і 4-вимірна варіаційна асиміляція (4D-Var). Ці методи в основному інтегрують змінні в просторі та часі, мінімізуючи функцію витрат, яка вимірює різницю між спостережуваними значеннями та значеннями в історичних наборах даних.

Історичні набори даних спостережень зберігаються в записах метеорологічних організацій для проведення кліматологічних та дослідницьких досліджень. Особливий вид моделі NWP, званий кліматичним або повторним аналізом, моделює стан атмосфери в минулому, а не в майбутньому. Ці моделі створюють фізичні параметри, які відображають погоду минулого. Моделі

повторного аналізу в основному використовуються в дослідницьких цілях і вивчають еволюцію атмосфери протягом багатьох років або століть, залежно від протяжності та роздільної здатності моделі. Прикладами таких моделей є ERA-Interim і CMIP5.

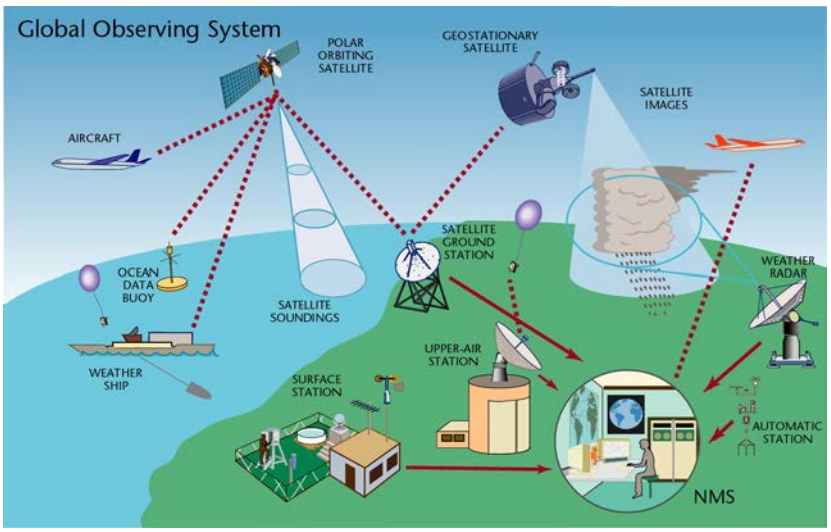



Рис 1.7 - Глобальна система спостережень (GOS)

Метеорологічні агентства по всьому світу надають суспільству доступ до прогнозів погоди, попереджень про суворі погодні умови, спостережень, інформації про повені та кліматичну інформацію. Обсяг інформації, що генерується цими центрами, дуже великий, а її зберігання та управління вимагає високої потужності допоміжної інфраструктури. Обсяг деяких з цих колекцій знаходиться в порядку петабайт інформації. Наприклад, ERA5, один з останніх наборів даних повторного аналізу, згенерував майже 6 петабайт даних. Незважаючи на те, що ці набори даних історично підтримувалися в стрічкових системах зберігання, зростає попит на споживання цих даних у режимі реального часу, що вимагає високої продуктивності файлові системи зберігання.

Дані про погоду в основному представляються за допомогою багатовимірних числових масивів. Існують спеціальні формати файлів, такі як NetCDF-4 або HDF5, які оптимізовані для забезпечення швидкого доступу до всіх даних або Його підмножини реалізують передові методи стиснення для зменшення розміру файлів. Ці формати також впроваджують стандарти метаданих, щоб забезпечити сумісність і сумісність файлів між організаціями.

Великі обсяги даних, отриманих при моделюванні атмосфери, і зростаючий інтерес до цих даних з боку таких секторів, як сільське господарство, повітряний і морський транспорт або відновлювані джерела енергії, створює проблему щодо того, як найкраще зробити ці дані доступними. Численні національні та міжнародні зусилля, такі як Європейська програма «Коперник» (*«Очі Європи Коперника* на *Землі*»), в даний час зосереджені на розробці та впровадженні нових системи, здатні обробляти та витягувати цінність з даних про погоду та клімат.

У зв'язку з тим, що доступність даних про Землю пробудила інтерес до інтеграції нових джерел даних у моделей та аналізу, ввівши термін *злиття даних*. Однією з найбільших проблем при об'єднанні даних з різних джерел є неоднорідність їх представлення. Дані, що надходять з моделей NWP або супутників, представляються за допомогою різних сіток, одиниць, проектів, просторових і часових роздільних здатностей і форматів файлів. Процесу злиття даних, як правило, передує трудомісткий і схильний до помилок процес гомогенізації даних в загальне представлення. Цей процес широко відомий як *суперечка про дані* і, на жаль, є послідовним впливом у науках про погоду. Наприклад, зменшення обсягу випуску НВП є активною областю досліджень у галузі медицини, яка зазвичай є комбінацією низького рішення сітчасті дані NWP з іншими джерелами з вищою роздільною здатністю, такими як місцеві метеостанції або дистанційне зондування.

Протягом останнього десятиліття зростає інтерес з боку приватних компаній до наборів даних про погоду та дистанційне зондування. Набори даних про погоду стали доступними для великих компаній, що займаються хмарними обчисленнями, таких як Amazon або Google. Ці компанії інвестують у системи доступу та обробки цих великих наборів даних за допомогою розподіленої інфраструктури. Це відкриває новий шлях для використання цих даних, який історично залежав від підтримки великих національних агентств і високопродуктивних обчислень центрів.

Ці ініціативи популяризують та демократизують доступ широкої громадськості до наборів даних про погоду. Однак хмара представляє зовсім іншу

архітектуру в порівнянні з традиційними центрами HPC. Системи, моделі та формати файлів були розроблені протягом останніх півстоліття років на основі технічних та стабільних технологій, таких як файл POSIX системи та мікропроцесора x86 CPU. В даний час існує безпрецедентна можливість перепроектувати системи, які перенесуть дані про погоду в нові додатки і повсюдно використовуються в суспільстві.

Цей перехід, що робить дані про погоду більш доступними, відбувається одночасно з розвитком нового машинного навчання та великомасштабних аналітичних алгоритмів. Завдяки розміру цих наборів даних комп'ютери замінили людей як основних споживачів даних. Очікується, що ця тенденція збережеться з випуском нових алгоритмів, здатних аналізувати дані про погоду та виробляти продукти з доданою вартістю на вимогу.

## 1.4 Машинне навчання

Термін «машинне навчання» з'явився ще в середині минулого століття. У 1959 році вчений-комп'ютерник Артур Семюел визначив машинне навчання як «здатність навчатися без явного програмування». Машинне навчання також можна розглядати як особливий спосіб вирішення більш загального завдання: «штучний інтелект», який передбачає здатність машин думати та міркувати в подібно до людини. Концепція «штучного інтелекту» з'явилася на кілька років раніше поняття «машинне навчання». У 1950 році Алан Тюрінг опублікував новаторську статтю під назвою «Обчислювальна техніка та інтелект», в якій було поставлено питання про те, чи можуть машини мислити було формально піднято. У 1956 році Джон Маккарт, науковий співробітник Університету Стенфорду, використав цей термін, щоб висловити ідею про те, що Машини можуть імітувати будь-яку форму людського навчання.

Існують різні способи визначення та представлення перетину між «штучним інтелектом» та «машинним навчанням». Залежно від автора та наукового дослідження, ці терміни часто зустрічаються в поєднанні з іншими, такими як

«обчислювальна статистика» або «наука про дані». У контексті цієї тези ми використовуємо термін «машинне навчання» для позначення галузі комп'ютерних наук, яка використовує статистичні методи для надання комп'ютерних програм здатність «вчитися» на даних. Цей розділ знайомить з деякими основними областями та алгоритмами машинного навчання, з особливим акцентом на тих, які були застосовані в цій роботі для розв'язувати конкретні задачі прогнозування погоди.

Однією з головних проблем машинного навчання є розробка загальних алгоритмів, які можуть знаходити значущі представлення в даних. На практиці вхідні дані зазвичай адаптуються за допомогою ряду перетворень відповідно до вимог конкретного алгоритму. Наприклад, розробка алгоритму, здатного витягти час із зображень настінних годинників, є нетривіальною задачею. Якщо обробляються одні й ті ж зображення настінного годинника, в результаті чого виходить новий набір даних, який містить кути стрілок годинника для кожного зображення, складність проблем – вочевидь значна. Ця проблема називається репрезентацією або навчанням функцій.

З точки зору природи проблем, які намагаються вирішити алгоритми машинного навчання, ми можемо виділити дві основні категорії проблем: контрольовані та неконтрольовані. У задачах з учителем алгоритм навчання забезпечується навчальними даними, що містять вихідні значення, які використовуються для навчання моделі. Після того, як модель навчена, вона може бути використана для попереднього визначення виведення нових зразків вхідних даних, мітка або вихідне значення яких невідомі. З іншого боку, навчання без учителя вирішує проблеми, коли результат залишається невідомим на етапі навчання моделі. Ці алгоритми виконують дослідницький аналіз даних, намагаючись знайти властиву їм структуру або взаємозв'язки між вхідними зразками. Він використовується для висновків із наборів даних, що складаються з вхідних даних без позначених відповідей.

Центральне застосування навчання без учителя знаходиться в області оцінки щільності в статистиці, хоча навчання без учителя охоплює багато інших областей,

пов'язаних з узагальненням і пояснення особливостей даних. Прикладами неконтрольованих завдань є кластеризація та зменшення розмірності. Кластеризація полягає у визначенні груп елементів у наборі даних, які мають певні спільні характеристики. Однією з найпопулярніших методик виконання кластеризації є k-середні. Мікстурні моделі надають ряд методологій для виконання мереж зменшення розмірності та глибоких переконань та автокодерів дозволяють виконувати задачі на розмірність r на основі нейронних мереж.

Контрольовані задачі можна розділити на дві категорії: регресійні та класифікаційні, залежно від природи вихідної змінної, яку потрібно передбачити. У регресії вихідні дані представляються за допомогою неперервної змінної, тоді як у класифікаційних проблемах вихідна змінна містить дискретне число можливих значень або етикетки. Як класифікаційні, так і регресійні задачі можуть бути формально виражені за допомогою векторів випадкових величин для представлення вхідних і вихідних даних. Для цілей цієї роботи, беручи до уваги характер проблем, з якими ми працювали, ми розробляємо теорію задач регресії. Задачі класифікації можуть бути виражені за допомогою подібних рівнянь, змінюючи вихідну змінну на дискретну випадкову величину, а не на безперервну.

Задача контрольованої регресії може бути описана узагальнено відповідністю між набором $n$ прогностичних змінних $X = (X_1, \ldots, X_n)$ і множина $m$ безперервних пояснювальних або вихідних змінних $Y = (Y_1, \ldots, Y_m)$. Коли n і $m$ більше одиниці, задача називається «багатовимірною множинною регресією», або commonly, «багатовимірною regression». Кожна вибірка випадкового вектора (**X**, **Y**) дорівнює repre- за формулою (x, **y**) = (**x** $_1, \ldots, x_n, y_1, \ldots, y_m$). Область кожної вхідної змінної $X$ і позначається $X_i$, а область кожної вихідної змінної Yj позначається як $Y_j$. Отже, область випадкового вектора (**x**, **y**) дорівнює (X, Y) = $X_1 \rightarrow \ldots \rightarrow X_n \rightarrow Y_1 \rightarrow \ldots \rightarrow Y_m$, який містить множину всіх можливих екземплярів (**x**, **y**).

Регресійна модель визначає функцію **F**, яка відображає кожен екземпляр вхідного векторного простору в простір вихідних змінних:

$$\Phi : \quad \mathcal{X} \to \mathcal{Y}$$
$$(x_1, \ldots, x_n) \mapsto (y_1, \ldots, y_m)$$

Аналогічно, класифікація problem відображає вхідний простір на множину дискових варіабельних C.

Однак не завжди під час навчання можна мати повністю контрольований набір даних, що вводить концепцію слабоконтрольованих проблем. Навчання зі слабким наглядом має справу зі сценаріями, в яких набори даних представляють відсутні або невідповідні мітки або цільові значення. Слабко контрольовані алгоритми досліджують методи навчання для точного прогнозування вихідних значень нових «немічених» зразків.

Найпростішим алгоритмом для виконання регресії є лінійна регресія, яка використовує лінійну функцію для зв'язку вхідних або незалежних змінні з вихідною або залежною змінною. У математиці лінійну систему можна розв'язати, коли ми знаємо кількість рівнянь, що дорівнює кількості змінних у системі. Лінійна регресія представляє проблему наявності набору даних, який зазвичай містить набагато більшу кількість точок, ніж змінні в даних, що можна розглядати як розв'язання надмірно визначена система. Регресія знаходить рішення для таких систем, розглядаючи лінійну функцію, яка пов'язує різні змінні та мінімізує загальну похибку при прогнозуванні вихідного значення.

Наступне рівняння представляє лінійну залежність між одним вихідним елементом набору даних і набором *n* вхідних змінних:

$$y = b\,1\,x\,1 + \; b\,2\; x2 \; + \; ... + \; бнх\,н + e$$

Існує багато підходів до розв'язання лінійних регресійних моделей, які ґрунтуються на мінімізації похибки, представленої значенням *e* у рівнянні, для всього набору даних. Методи найменших квадратів, як правило, використовуються для виконання лінійної регресії. Більшість реалізацій вимагають інверсії матриці, що змушує компутаційну вартість цих методів кубічно зростати зі збільшенням розмірності даних. Альтернативним підходом є використання ітеративних методів, таких як градієнтний спуск, щоб змусити модель зближуватися, мінімізуючи значення похибки.

Лінійні регресійні моделі, хоча і прості і ефективні в багатьох ситуаціях, не забезпечують хорошої точності, коли дані представляють нелінійні зв'язки. Методи

нелінійної регресії дозволяють навчати моделі, які здатні представляти нелінійні зв'язки між вхідними змінними. Існує багато методів для визначення нелінійних методів, таких як Тейлора або синусоїдальних рядів, або шляхом сегментації простору та використання серії лінійних регресійних моделей, також званих кусковою регресією.

Нелінійні регресійні моделі краще здатні представляти зв'язки в даних, ніж лінійні моделі. Однак більша репрезентативна здатність нелінійних регресійних моделей може призвести до представлення даних з надмірною деталізацією. Якщо модель навчиться відтворювати з високою точністю взаємозв'язки в навчальному наборі даних, іноді це є небажаним ефектом, оскільки може не піддаватися узагальненню нові точки даних, невидимі під час етапу навчання. Ця проблема відома як перенавчання, а її наслідки представлені на рисунку 1.8. На цьому малюнку ми можемо побачити нелінійну регресійну модель, яка точно представляє всі точки в наборі даних, на відміну від лінійної моделі, яка апроксимує його цінності. У деяких випадках лінійна модель може забезпечити краще представлення проблеми, ніж нелінійна версія, яка переповнює дані. Іноді перенавчання можна легко вирішити, додавши більше даних до моделі. В інших ситуаціях уникнення перенавчання вимагає ретельного розгляду конструкції моделі та її параметрів.

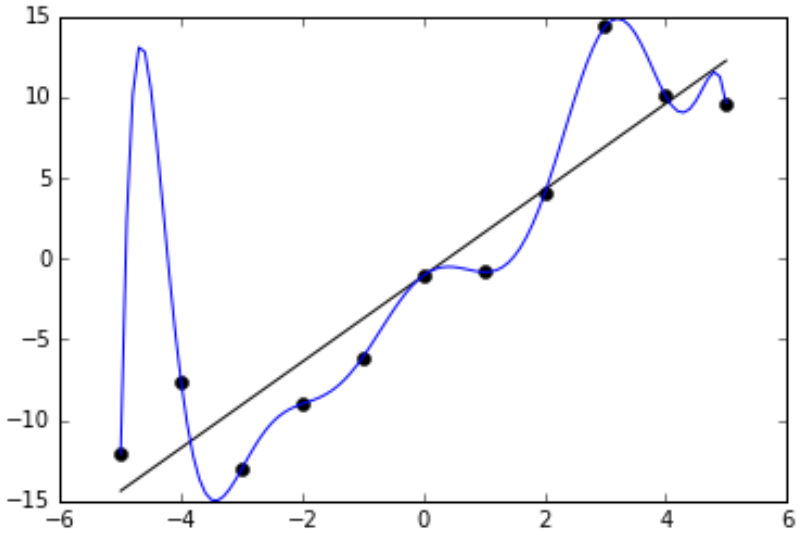



Рис 1.8 - Порівняння нелінійної регресійної моделі (синя) та лінійної моделі (чорна), що представляє 2-вимірний набір даних.

Іншою альтернативою виконання нелінійної регресії є так звані непараметричні методи регресії. У непараметричної регресії немає жодної модель, яка представляє весь набір даних. Модель будується ситуативно для кожного окремого випадку, відповідно до інформації, отриманої з отриманих даних. Непараметричні регресійні моделі здійснюють пошук у наборі даних, шукаючи елементи, які знаходяться «ближче» до вхідного значення. Тому вони часто вимагають більших наборів даних, ніж еквівалентні параметричні методи.

Регресія ядра є особливим методом непараметричної регресії. Регресія ядра використовує математичні віконні функції, які називаються ядрами, для зважування внеску вхідних точок даних. Цей метод оцінює неперервну залежність, що варіюється з обмеженого набору точок даних, які ядро вибирає, зважуючи внесок кожної точки даних, надання більшої ваги прилеглим місцям. На рисунку 1.9 показано сім різних функцій ядра, які застосовують різні шаблони зважування даних.

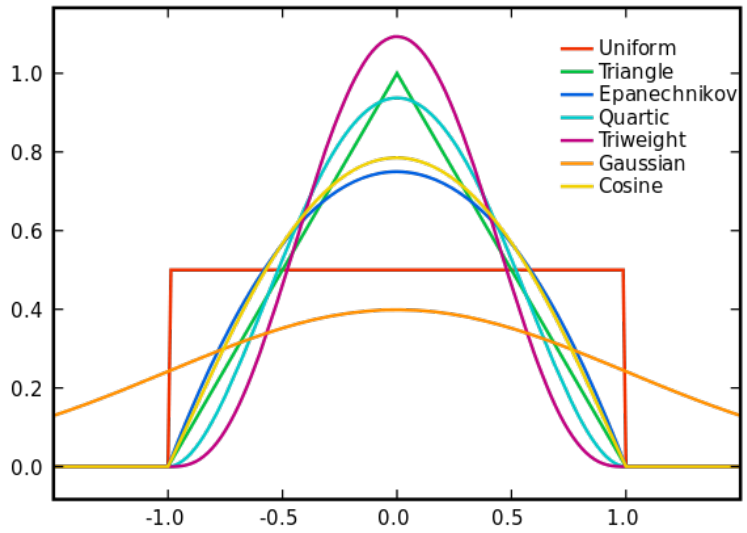



Рис. 1.9 - Порівняння форми семи загальних віконних функцій, що використовуються в регресії ядра.

Ми розглянули деякі методології, які використовуються для проведення статистичної регресії на основі даних на основі методів найменших квадратів і непараметричних. Однак існують і інші добре відомі способи побудови моделей на основі даних, що представляють неперервні змінні. Дерева рішень, наприклад, надають інструмент для моделювання зв'язків у даних, розділяючи простір набору

даних на незалежні області. Алгоритм дерев класифікації та регресії (CART), запропонований Лео Брейманом у 1984 році, заснований на використанні бінарного рішення дерев для генерації моделей класифікації та регресії.

Регресійні дерева виконують рекурсивне розбиття набору даних на основі метрики, яка максимізує однорідність у результуючих дочірніх вузлах, наприклад, дисперсію. У кожному вузлі двійкового дерева регресії вибирається одна змінна для виконання розділу; Як правило, точки даних зі значеннями, більшими за значення розбиття, потрапляють до одного дочірнього вузла, а менші присвоюються іншому дочірньому вузлу. Цей процес розбиття виконується до тих пір, поки не будуть виконані критерії зупинки, зазвичай засновані на кількості точок даних або мінімальному значенні дисперсії на вузол. Коли вузол більше не розділений, він називається «листом».

Дерева регресії є простим та інтуїтивно зрозумілим способом моделювання даних. Вони швидко навчаються, добре масштабуються з великими наборами даних і легко інтерпретуються та розуміються, дивлячись на рішення, прийняті в кожному вузлі дерева. Ці фактори сприяли їх популярності, і в наш час їх застосування можна знайти практично в будь-якій галузі науки. Існує багато версій і реалізацій регресійних дерев і, з моменту їх винаходу, багато авторів представили альтернативні методи побудови таких дерев, представили нові метрики або способи обрізки дерев для кращого представлення даних.

Алгоритми дерева регресії працюють значно краще при навчанні в групах або ансамблях. Мішки, також відомі як початкова агрегація, — це метод ансамблевого дерева, який використовується для зменшення дисперсії окремих дерев рішень. Елементи дерева ансамблю створюються за допомогою піднаборів даних шляхом випадкового вибору точок даних, із заміною, з даних навчання. Середнє арифметичне з усіх індивідуальних прогнозів з ансамблевих дерев дає більш надійний прогноз, ніж будь-яке з окремих дерев регресії. Ансамблеві методології також стали дуже популярними, і в літературі доступний широкий спектр алгоритмів, таких як бустинг або випадковий ліс.

В якості альтернативи розглянутим методам існують інші види алгоритмів

для перформуючої регресії, які часто більш здатні представляти високорозмірні набори даних і нелінійності в дані. Двома дуже популярними прикладами є Support Vector Machines (SVM) і багатошарові Штучні нейронні мережі (ШНМ). Обидві ці методології використовують серію «базисних функцій» для визначення гіперплощин і представлення надісланих даних. Нелінійність даних може бути представлена за допомогою «трюку ядра» в SVM і функцій активації в ШНМ. Хоча між SVM і багатошаровим ШНМ прямого зв'язку є схожість, перший є непараметричним і має змінний розмір, тоді як другий є параметричним і має фіксовані розміри, що визначаються кількістю і розмірами його шарів. Далі в цьому розділі ми повернемося до ШНМ у контексті аналізу зображень та глибокого навчання.

У галузі машинного навчання, яка досліджує застосування методологій SVM та ШНМ, проблеми, пов'язані зі структурованими та високорозмірними даними множини широко відомі як «структуроване передбачення». Структуроване прогнозування зосереджує свою увагу на методологіях, які мають справу з регулярними та великими вихідними просторами, такими як набори даних, що представляють часові або просторові виміри.

В області аналізу просторових даних комп'ютерний зір традиційно зосереджується на розробці методів вилучення закономірностей з даних зображень. З огляду на складність розробки загальних алгоритмів, здатних інтерпретувати дані зображення, комп'ютерний зір традиційно приділяє значну увагу розробці методів інженерії, таких як SIFT і HOG, які описують особливості з базових будівельних блоків, визначених людиною. За останнє десятиліття нові методології, засновані на використанні згорткових нейронних мереж, виявилися більш загальними і пропонують значні переваги в порівнянні з попередніми методологіями.

Методи глибокого навчання нещодавно досягли безпрецедентних результатів у різних контрольованих завданнях класифікації та регресії, використовуючи дані високої розмірності набори, як-от зображення або аудіо. У цих методах використовуються великі ШНМ, що складаються з десятків, а в деяких випадках і сотень шарів. Нещодавні дослідження представили мережі, які здатні перевершити

продуктивність людського рівня в складних завданнях, таких як класифікація зображень або семантичного опису. Мережі глибокого навчання для аналізу зображень зазвичай базуються на згорткових нейронних мережах (CNN), які були виявився дуже ефективним у фіксації внутрішніх особливостей, представлених у різних масштабах зображення. CNN вивчають кілька шарів згорткових ядер, які здатні встановлювати локальні зв'язки між сусідніми пікселями зображення. Кожен шар в мережі виконує операцію дискретизації відразу після згортки, зменшуючи розміри зображення. Ядра в один шар покривають більші площі, ніж в попередньому шарі. Мережа здатна поступово агрегувати просторові особливості зображень, створюючи представлення високого рівня.

Згорткові мережі кодерів-декодерів - це тип CNN, які забезпечують найсучасніші результати при вирішенні таких завдань, як сегментація зображень, знешумлення зображень або регресія від зображення до зображення. Ці мережі засновані на автокодерах, які використовують CNN для вивчення обґрунтованих, але точних оцінок зображень, узагальнення його використання для виконання регресії між зображеннями. Згортки в енкодерній половині мережі виконують процес вибору ознак, зменшуючи розмірність даних. Частина декодера збільшує простір об'єктів, відображаючи його на вихідний простір. Мережі кодер-декодер пропонують ефективний метод вивчення взаємозв'язку між вхідними та вихідними просторами високої розмірності, такими як ті, що визначаються зображеннями або відео.

В рамках регресії існує область досліджень, спеціально орієнтована на аналіз даних часових рядів, яка тісно пов'язана з темою прогнозування погоди. Метою цих методів є прогнозування майбутніх значень на основі короткострокових і довгострокових тенденцій і закономірностей у наборах даних часових рядів. До цієї проблеми статистичне співтовариство традиційно підходить, застосовуючи авторегресійні методи, такі як авторегресивна ковзна середня (ARMA) або авторегресійний інтегрований рух середньої (ARIMA), для прогнозування майбутнього стану або переходів часових і послідовних змін.

## 1.5 Прогнозування погоди: підхід до машинного навчання

До появи моделей NWP люди використовували просту техніку для прогнозування погоди на основі спостережуваних даних, зібраних протягом багатьох років для певного регіону, який є відомий як кліматологу. Приблизні оцінки далекосяжних t r кінців і кумулятивних значень змінних, таких як температура або опади, можуть бути виведені, застосовуючи базову статистику до спостережуваних погодних явищ протягом досить тривалих періодів часу.

На початку 20-го століття з появою технології для точного вимірювання атмосфери і передачі цих значень по всій географії, у віддалених місцях стали доступні карти погоди. Ці карти спочатку почали відображати положення і форму систем низького і високого тиску і дозволили розробити методології прогнозування погоди, які передбачали рух і вплив цих систем тиску в атмосфері. Застосування статистичних моделей до прогнозування погоди було запропоновано ще в 1950-х роках. У той же час, коли були розроблені перші моделі NWP, Мелоун висунув аргумент, що «статистика повинна в кінцевому підсумку відігравати певну роль» в симуляторі. Автор продемонстрував методологію, що використовувала множинну лінійну регресію, для прогнозування поля тиску на рівні моря. На рисунку 1.10 показана перша спроба використання регресійних методів для прогнозування еволюції поверхневої температури в Індіанаполісі (США). Ця модель використовує поле тиску атмосферної циркуляції за попередні 24 години як вхідні дані для прогнозування наступних 24 годин.

Незважаючи на оптимізм, висловлений Мелоуном на початку 1950-х років, реальність така, що застосування машинного навчання та статистичних методів стало доповненням NWP, а не як альтернатива йому. Сила фізичних рівнянь для моделювання еволюції атмосфери уздовж просторових і часових вимірів ще не зрівнялася з іншими методологіями.

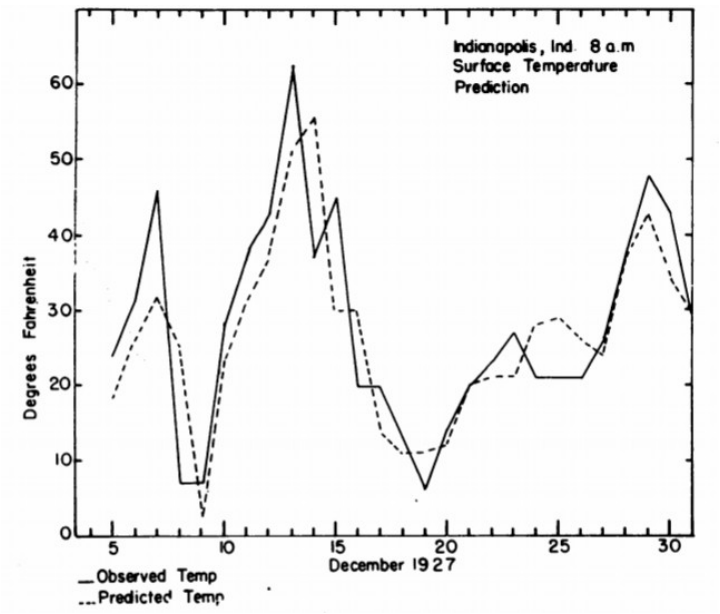



Рис. 1.10 - Порівняння між 24-годинним прогнозом середньодобової температури і спостережуваними значеннями температури в Індіанаполісі (США).

Залежно від розв'язуваної задачі і характеру вихідних даних можна виділити наступні категорії проблем, при яких різні машинні методики навчання були успішно застосовані для поліпшення нашого розуміння процесів в атмосфері.

- Виправлення систематичної похибки NWP: Ця категорія включає випадки, коли вихідні дані NWP обробляються для усунення зміщень, збільшення вихідної потужності або вирішення локальних топографічних ефектів з використанням спостережуваних даних.
- Оцінка передбачуваності: Через хаотичний характер атмосфери прогнози погоди мають внутрішню оцінку невизначеності, що обмежує її значення. Методи машинного навчання були використані для оцінки невизначеності та пов'язаних з нею показників впевненості ансамблевого прогнозування.
- Екстремальне виявлення: Ця категорія об'єднує класифікаційні задачі, в яких результатом є предикат або виявлення рідкісної події. Приклади таких застосувань можна знайти в методах, які передбачають такі явища, як град, пориви вітру або циклони.
- Параметри NWP: Моделі NWP генерують кілька змінних, використовуючи відповідні моделі або параметризації для опису процесів, які не можуть бути моделюється за допомогою явних фізичних рівнянь. Хоча історично

ці параметри ґрунтувалися на емпіричних моделях, використання машинного навчання починає зростати. Існують приклади методів машинного навчання, що застосовуються до модельних процесів, таких як радіаційне перенесення, конвективний і прикордонний шар або турбулентність.

Хоча чисто статистичні методології ще не змогли замінити моделювання NWP у прогнозуванні погоди, статистика відіграла важливу роль у підвищенні якості виходу ППП. Добре відомо, що моделі НВП мають певні дефекти, які можуть бути виправлені шляхом статистичної пост обробки їх результатів. Перша категорія задач, в яких моделі намагаються врахувати вихід NWP, є найбільш поширеним застосуванням машинного навчання до прогнозування погоди. Дані спостережень часто використовуються для виконання певної операції даних НВП з метою підвищення їх точності.

Статистичні моделі, що використовуються для пост обробки результатів NWP, еволюціонували в трьох загальних рамках: «Perfect Prog», Model Out-output Statistics (MOS) і «Повторний аналіз». Моделі «Perfect Prog» використовують регресію для представлення взаємозв'язку між початковим станом змінних NWP (аналізу), таких як temperature або precipitation, і спостереження за тими ж параметрами. Потім навчені моделі використовуються для корекції прогнозованих полів NWP, таких як температура або опади, виправляючи недоліки в прогнозі NWP. «Ідеальна прога» передбачає, що точність прогнозу не залежить від розміру вікна прогнозування. На противагу цьому, МОС відповідає лінійній регресійній моделі між обсягом виробництва ПНП у певний прогнозний момент часу зі спостереженнями в цей час. Оскільки МРС відповідає вихідному NWP, він може враховувати упередження та систематичність у моделі. При оновленні конфігурацій моделей NWP МРС необхідно перенавчати після того, як буде зібрано достатню кількість прогнозів нових моделей. Моделі «Perfect Prog», як правило, менш точні, ніж оптимізовані моделі MOS, але вони менш чутливі до змін конфігурації моделі і, як правило, є такими з часом стає більш міцним. При розробці всіх варіацій MOS і «Perfect Prog» обмеженням є обсяг даних, доступних для навчання моделей regression.

Лінійні регресійні моделі були використані в прогнозуванні погоди в більш широкому контексті, ніж усунення зміщення з полів, змодельованих NWP. Наприклад, були запропоновані регресійні методи для зменшення масштабу екстремальних опадів на основі даних NWP. Аналогічно пропонують форму параметричної регресії, засновану на використанні ядер для розв'язання локальних ефектів вітрів. Ми можемо знайти інші застосування регресії для прогнозування ймовірності суворих погодних умов або максимального розміру граду та його ймовірності. У випадках, коли вихід не є неперервною змінною, а скоріше бінарним результатом або набором категорій, логістична регресія може бути використана для прогнозувати настання певних подій, таких як конвективні процеси або відбір учасників ансамблю NWP.

Одним з основних обмежень моделей, заснованих на лінійній регресії, є те, що не всі взаємозв'язки між різними атмосферними процесами і їх змінними можуть бути моделюється за допомогою лінійних функцій. Крім того, методи на основі найменших квадратів, що використовуються для розв'язання лінійної регресії, які вимагають інверсії матриці, погано масштабуються, коли набір даних представляє велику кількість вхідних змінних. Для подолання цих обмежень в літературі були запропоновані різні методи, такі як багаторівнева регресія, що використовується для моделювання кліматичної мінливості для нелінійних процесів або квадратична регресія для покоління учасників ансамблю. У разі наявності великої кількості вхідних змінних або коли ці змінні тісно пов'язані між собою, процес навчання регресійних моделей може бути утруднений. Наприклад, методи, засновані на гребеневій регресії, були запропоновані як метод вибору змінних у багатомодельних сценаріях. Оператор найменшої абсолютної усадки та відбору (LASSO) є ще одним методом, подібним до регресії хребта, який був застосований для прогнозування погоди для Скорочення масштабування NWP або довгострокове сезонне прогнозування. Крім того, були запропоновані методи аналізу головних компонент (PCA) для параметризації процесів підсіткового масштабу в атмосфері шляхом визначення найбільш релевантних змінних використовується в регресіях.

Метод на основі дерева регресії та класифікації є популярним підходом до

моделювання нелінійних атмосферних процесів. До того, як деревні методи стали популярними в 1980-х роках, були введені дерева рішень для діагностики рівнів вологості верхніх рівнів в атмосфері. Дерева можуть бути легко масштабовані для навчання моделей з використанням великих наборів даних і з великою кількістю вхідних змінних. Легкість інтерпретації навчених моделей також сприяла популяризації цих методів у прогнозуванні погоди. Методи, засновані на дереві рішень, виявилися потужним інструментом у широкому спектрі погодних застосувань, таких як виявлення та діагностика турбулентності грози, екстремальні явища або для представлення кругової природи вітру. На рисунку 1.11 представлено дерево рішень для прогнозування градових опадів, яке чітко повідомляє про метод, який можуть дати люди-синоптики слідкуючі за прогнозом граду.

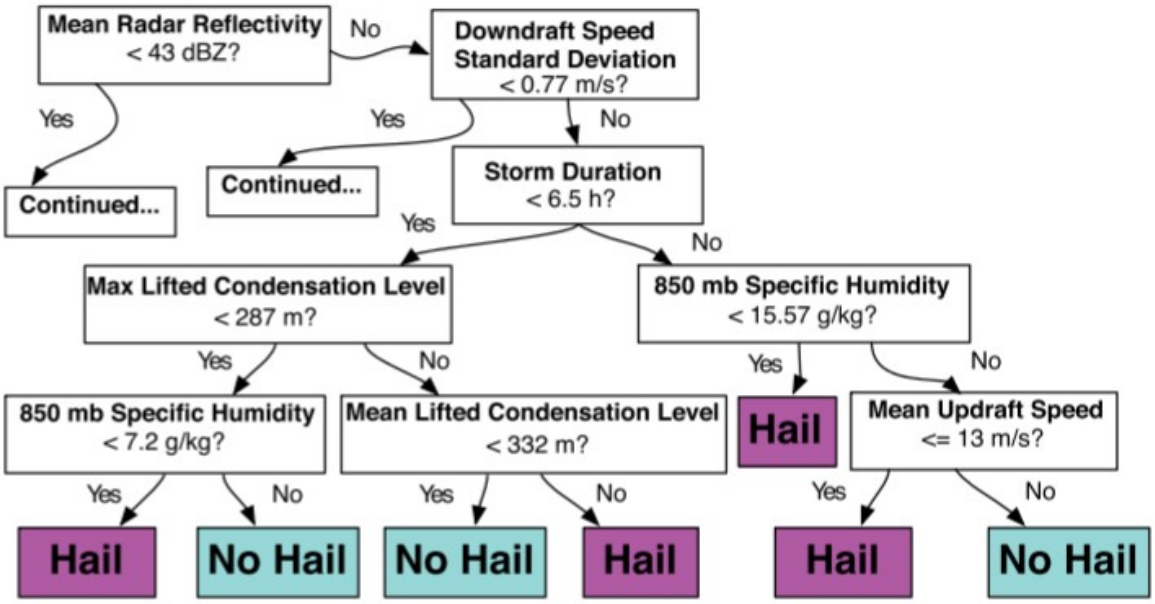



Рис 1.11 – Дерево рішень, що використовується для прогнозування події граду на основі порогових значень для різних спостережуваних і NWP параметрів.

Методи, засновані на дереві, здатні представляти нелінійності в даних за допомогою кускової апроксимації, рекурсивно розділяючи вхідний простір. Штучні нейронні мережі (ШНМ), підтримка векторних машин (SVM) або підтримка векторів (СЗР) забезпечують загальну, більш потужну альтернативу моделюванню нелінійних процесів. Як моделі ANN, так і SVM/SVR є гнучкими та

потужними, але створюють моделі, які часто важко інтерпретувати з точки зору основних фізичних концепцій, визначених моделлю. Ця характеристика обмежує застосування таких моделей до задач прогнозування погоди в тих областях, де вчені вимагають обґрунтування припущень, зроблених модель у зв'язку з необхідністю зв'язку з іншими моделями або необхідністю врахування взаємозалежностей між змінними. Методи, засновані на SVM і SVR, широко використовуються в метеорологічних додатках, наприклад, для виявлення і прогнозування торнадо або для прогнозування опадів.

ШНМ страждають від аналогічної проблеми: отримані моделі важко інтерпретувати за допомогою ваг і нелінійних функцій активації. ШНМ використовуються в широкому спектрі метеорологічних додатків з кінця 1980-х років, з такими додатками, як класифікація хмар, прогнозування та виявлення торнадо, висота хмар та видимість, класифікація опадів або корекція зміщення для опадів і температурних передбачень. Останнім часом ШНМ розширилися до методів глибокого навчання, застосування яких у сфері прогнозування погоди представлено в кінці цього розділу.

Моделі глибоких нейронних мереж ефективно та ефективно витягують складні закономірності з великих структурованих наборів даних. Специфіка методології вилучення просторової інформації з наборів даних зображень. Подібним чином, повторювані нейронні мережі (RNN), розроблені в галузі обробки природної мови, мають багато застосувань у широкому діапазоні набори даних, що містять часовий вимір. Комбінація обох моделей була вперше запропонована в контексті опадів з використанням радіолокаційних даних.

Ще однією проблемою прогнозування погоди є облік невизначеності, пов'язаної з метеорологічними ситуаціями. Будь-який даний прогноз має довірче значення, що вказує на його ймовірність. Хаотичний і нелінійний характер атмосфери накладає обмеження на нашу здатність прогнозувати її еволюцію. Спільнота прогнозистів погоди використовує термін «передбачуваність» для вираження здатності точно моделювати еволюцію певного параметра або ситуація. Загальним підходом до представлення мінливості атмосфери є використання

ансамблів числових моделей, які будуються за допомогою однієї і тієї ж моделі, збурюючи початкові умови або з урахуванням результатів різних моделей NWP. Інтерпретація ансамблевого НВП часто вимагає ідентифікації груп моделей або регіонів, які представляють схожі метеорологічні ситуації. Методології без вчителя, такі як кластеризація, використовуються для представлення ймовірності прогнозу шляхом агрегування членів ансамблю за аналогічними умовами.

Важливим фактором при застосуванні машинного навчання в області прогнозування погоди є великі обсяги даних. Хоча наявність великих наборів даних є перевагою для навчальних моделей, висока розмірність наборів даних стає проблемою для більшості методологій. Більшість прикладів застосування машинного навчання, передбачають різке спрощення або зменшення розмірності даних, шляхом обмеження вхідного простору до окремих точок сітки у просторі або часі.

Створення узагальнених моделей машинного навчання, які враховують ефекти багатьох вхідних змінних, зазвичай вимагає використання параметричних підходів. Ці параметричні моделі вимагають навчання декількох версій одних і тих же моделей для різних точок часу і простору. Такий підхід передбачає великі обчислювальні ресурси і часто його важко масштабувати. Крім того, для роботи призначені реалізації традиційних алгоритмів машинного навчання, таких як лінійна регресія, деревовидні методи або повністю пов'язані штучні нейронні мережі скорочення вхідних просторів і їх складність швидко стають некерованими в міру зростання числа вхідних змінних. Навчання моделей на високовимірних просторах залишається проблемою в галузі машинного навчання, що перешкоджає розробка суттєвих альтернатив NWP для симуляції погоди.

Останнім часом спільнота машинного навчання приділяє особливу увагу новим методологіям, які можна застосовувати до великих обсягів даних та/або висока розмірність. Зокрема, методи глибокого навчання продемонстрували найсучасніші результати з використанням великих наборів даних. Застосування цих методологій також нещодавно було досліджено в галузі прогнозування погоди з безпрецедентними результатами. Наприклад, на рисунку 1.12 показані результати

моделі глибокого навчання, навченої з використанням великої сукупності полів тиску NWP для виявлення струменевих потоків. Ця модель була навчена на великому кластері обчислювальних вузлів у суперкомп'ютерному центрі Лоуренса Берклі в США, продемонструвавши масштабованість і здатність CNN виділяти складні просторові закономірності в даних.

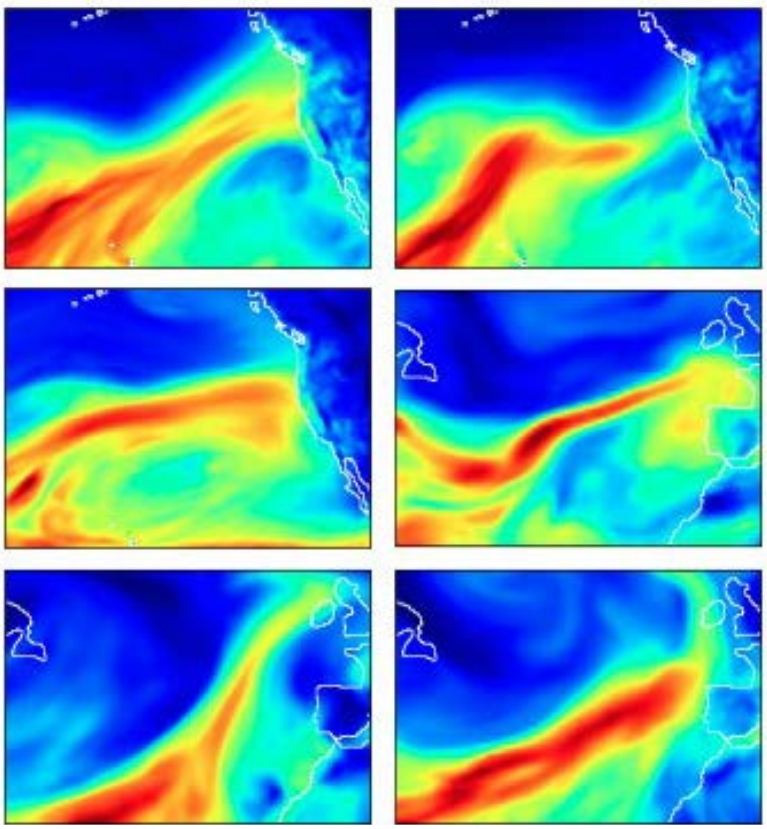


Рис. 1.12 - Зразки зображень атмосферних річок (струменевих потоків), відповідно класифіковані та витягнуті з багатотерабайтного набору даних NWP за допомогою глибокої моделі CNN.

Навіть якщо нові методології виявляться здатними вивчати моделі ПНП, що керують фізикою, можна уявити, що вони замінять NWP, на схід у короткострокові та середньострокові. Базові поля в NWP, такі як вітри точно моделюються за допомогою фізичних моделей, які розв'язуються за допомогою високооптимізованих чисельних методів. Код і компілятори, що працюють на цих моделях, постійно вдосконалювалися. Моделям машинного навчання все одно знадобиться деякий час, щоб досягти такого ж рівня ефективності в проведенні симуляцій при порівнянних змінах, ніж NWP.

Що стосується тенденції обсягу даних, що генеруються моделями ПНП і

системами спостережень протягом останніх десятиліть, ми можемо передбачити, що розмір і складність наборів даних збережеться зростає експоненціальними темпами. Інтерпретація та аналіз таких обсягів даних вимагатиме методологій, які здатні витягувати закономірності з великих і складних наборів даних, ефективно використовуючи наявні обчислювальні ресурси та ресурси зберігання.

Якби система була достатньо здатною проаналізувати всю сукупність історичних спостережень за атмосферою, виокремити закономірності та просторово-часові зв'язки між змінними, така система змогла б замінити НВП. У цьому сценарії не потрібно буде розуміти фізичні рівняння, які керують атмосферою, оскільки алгоритми зможуть витягти моделі, які симулюють. На рисунку 1.13 представлені підходи, що використовуються NWP і машинним навчанням при вирішенні проблеми прогнозування погоди.

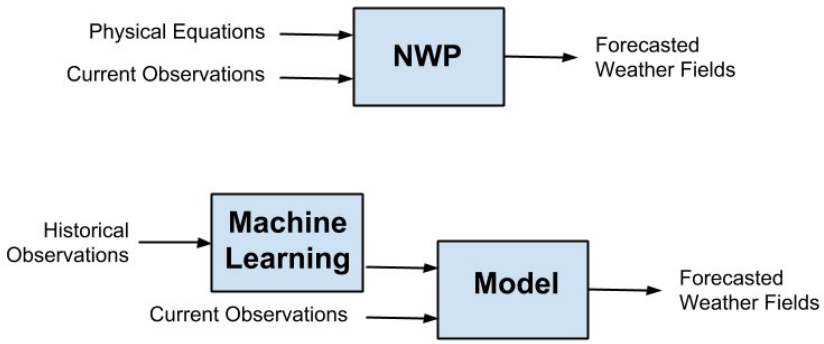



Рис. 1.13 - Порівняння традиційних підходів NWP та машинного навчання до прогнозування погоди.

NWP та спостережувані дані з наземних станцій та джерел дистанційного зондування забезпечують унікальний ресурс для дослідження структурованих проблем прогнозування погоди. Експерти, які працюють над моделюванням погоди та клімату, сходяться на думці, що ми побачимо поширення автоматизованих статистичних методів, які допоможуть в інтерпретації інформації про погоду в найближчому майбутньому. По-перше, як доповнення до NWP, машинне навчання на основі methods дозволить краще зрозуміти складні процеси, що відбуваються в атмосфері. На другому етапі стануть доступними більш комплексні методології, оскільки ефективність і потужність алгоритмів покращяться, а також з'явиться більше обчислювальної потужності.

# РОЗДІЛ 2. МЕТОДОЛОГІЇ ПОКРАЩЕННЯ ПРОГНОЗУВАННЯ ПОГОДИ

## 2.1. Використання кругової регресії ядра

Багатовимірна лінійна регресія є класичним методом моделювання взаємозв'язку між скалярною вихідною змінною та однією або декількома вхідними змінними. Цей метод у багатьох випадках служить базою для порівняння нових регресійних підходів.

Завдання в цьому випадку полягає в тому, щоб виконати регресію швидкості вітру NWP, що змінюється, використовуючи спостережувані дані. Вітер - це погодне явище, яке сильно залежить від топографу. Вітер, як правило, визначається за допомогою векторів, тоді як модуль pro визначає швидкість, а кут є напрямком вітру. Погодні моделі розкладають змінну швидкості вітру на декартові компоненти, щоб спростити її представлення. Проблема цього підходу полягає в тому, що виконання таких операцій, як регресія, не є простим через взаємозв'язок і залежність між обома компонентами.

Для цього етапу ми використовуємо модель NWP Global Forecasting System (GFS) та авіаційні рутинні метеорологічні звіти (METARs) з трьох аеропортів на півночі Іспанії: Віторія-Гастейс, Більбао і Сан-Себастьян-Доностія. Модель GFS представляє дані за допомогою регулярної сітки, яка охоплює весь світ, з просторовою роздільною здатністю приблизно 50 км і часовою роздільною здатністю 3 години. Цією моделлю керує Національне управління океанографії та атмосфери (NOAA), а глобальний набір даних за останні 15 років оприлюднюється. METARS – це високоякісні метеорологічні звіти, складені в більшості цивільних аеропортів світу. Ці звіти доступні кожні 30 хвилин і описують такі змінні, як вітер, температура, видимість і хмарність в аеропорту. METARC кодуються у вигляді текстових повідомлень і розповсюджуються по всьому світу, щоб диспетчери польотів і пілоти могли планувати процедури зльоту і посадки.

Для кожного з трьох аеропортів ми вибрали точки сітки NWP, найближчі до

точок спостереження, створивши тимчасовий ряд, який містить як змодельовані, так і спостережувані значення для різні погодні параметри. Ми порівнюємо різні методології лінійної та нелінійної регресії, використовуючи комбінації змінних NWP на вході та спостережувану швидкість вітру як вихідну змінну.

Причина вибору швидкості вітру в якості вихідної полягає в тому, що на цю змінну сильно впливає місцева топографія, що оточує аеропорти. Аеропорти, як правило, розташовані на відкритих просторах з постійним і вітровим режимом. Ці три аеропорти на півночі Іспанії знаходяться в гірському регіоні біля узбережжя Атлантичного океану і мають дуже специфічні вітрові показники. Розмір моделі NWP, використаної в цьому дослідженні, становить близько 75 км, що означає, що ці три аеропорти моделюються в суміжних осередки сітки. Робота зосереджена на вивченні потенціалу спостережуваних даних про швидкість вітру для корекції значень NWP для конкретних місць.

Непараметрична регресія - це категорія регресійного аналізу, в якій предиктор не приймає заздалегідь визначену форму, а будується кожен раз на основі даних. Непараметрична регресія вимагає більших розмірів вибірки, ніж регресія на основі параметричних моделей, оскільки дані повинні забезпечувати структуру моделі, а також оцінки моделі. Ця робота демонструє техніку, що використовує непараметричну регресію для покращення прогнозування швидкості вітру шляхом кластеризації даних про швидкість вітру навколо конкретних різних компонентів. Ця regression виконується dynamically, відбираючи історичні дані зі схожими характеристиками. Рівень подібності та відносні ваги для кожного елемента регресії контролюються за допомогою різних форм ядра.

Наприклад, на рисунку 2.1 представлений взаємозв'язок між спостережуваними та прогнозованими значеннями вітру, що генеруються для одного з досліджуваних аеропортів. В даному випадку вітер direction використовується в циклічному ядрі для динамічного зважування внеску кожної точки вхідних даних у регресію.

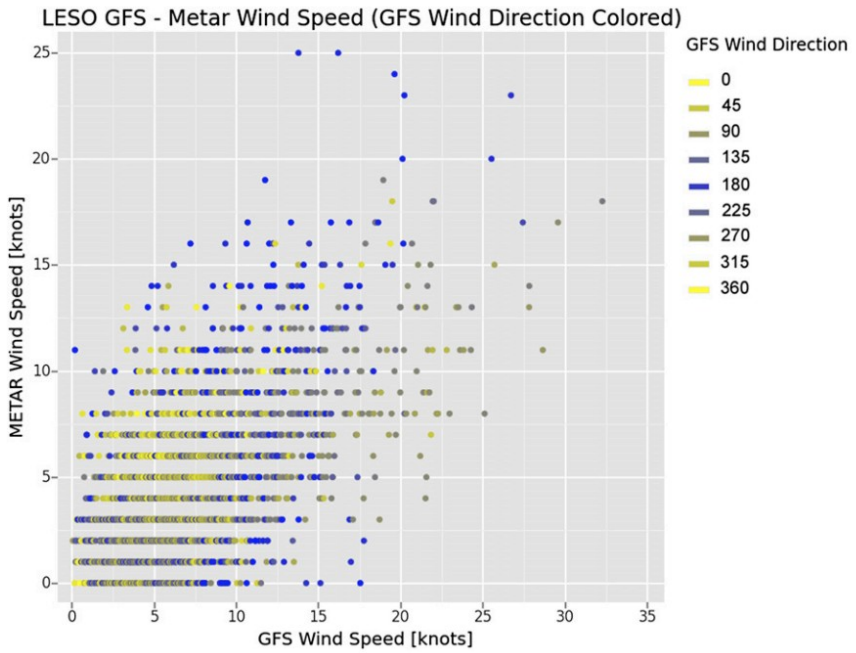



Рис 2.1 - Взаємозв'язок між показниками швидкості вітру GFS та METAR із Сан-Себастьяна. Напрямок вітру GFS представлено за допомогою кольорової шкали, де жовті кольори показують північні вітри, а кольори-підказки представляють південні вітри.

Ця робота вводить поняття циклічного ядра, яке забезпечує спосіб з'єднання елементів за допомогою циклічної змінної. Ми демонструємо, як циклічні ядра можуть успішно асимілювати спрямовані дані в регресійну модель. Ця концепція також може розширювати та покращувати інші циклічні або часові змінні та може бути використана для вилучення сезонних або щоденних моделей із даних.

## 2.2. Дерева кругової регресії

Більшість сучасних алгоритмів регресійного машинного навчання зосереджені на моделюванні взаємозв'язків між лінійними змінними. Кругові змінні мають відмінний характер від лінійних змінних, тому традиційні методології не в змозі представити їх пропорційно, що призводить до неоптимального результату в більшості випадків.

Продовжуючи вивчення методологій моделювання кругових змінних, ми вирішили розглянути regression trees для цього нового research study. Методи

ансамблю на основі дерев, такі як bagging або Random Forest, є універсальним, обчислювально ефективним і точним методом виконання регресія. Спираючись на новаторську концепцію кругового дерева регресії, досліджено альтернативні та більш ефективні методи введення циклічних змінних у дерева регресії.

Традиційні змінні **y**, **c** і **r** були використані в regression trees за допомогою двох ап-прохачів. Перший ігнорує кругову природу змінної, розглядаючи її як лінійну змінну. Проблема цього підходу полягає в тому, що 0 і 360 градусів представлені на кожному кінці змінного діапазону, але насправді це одне і те ж значення. Другий підхід полягає в тому, щоб розкласти кругову змінну на її декартові компоненти і використовувати їх як окремі змінні в дереві. Обидва ці підходи вносять обмеження і обмежують продуктивність дерев регресії. При першому підході дерева не можуть виконувати розщеплення, що перетинають початок координат, а в другому випадку розщеплення, виконані на декартових компонентах, незалежно ведуть до надмірного і неприродного поділу простору, визначеного круговими змінними, що негативно позначається на точності результуючих моделей.

Для цієї роботи ми використовуємо аналогічний набір даних, представлений у попередньому розділі, про нелінійну регресію ядра. Ми вирішили спрогнозувати спостережувану швидкість вітру в 5 різних місцях Європи. Дані з аеропортів METARS Берлін - Тегель, Лондон - Хітроу, Барселона - Ель-Прат, Париж - Шарль де Голль і Мілан - Мальпенса використовуються для навчання різних моделей дерев і проаналізовано отримані результати. Подібно до попередньої роботи, ми витягуємо дані часових рядів з моделі GFS для найближчих комірок сітки до кожного аеропорту. Моделі дерев навчаються з використанням тригодинних даних за 3 роки, в результаті чого в аеропорту було отримано приблизно 8760 зразків.

Ключова ідея кругових дерев полягає в тому, що кругові змінні можуть бути природним чином включені в дерева регресії, розглядаючи розщеплення, які перетинають початкову точку [0-360] у просторі пошуку. Оригінальна реалізація циклічних дерев виконує розбиття на несуміжні області змінного простору, що призводить до надмірної фрагментації набору даних. Удосконалення,

запропоноване в нашій роботі, обмежує простір пошуку суміжними областями змінного простору, що призводить до підвищення продуктивності обчислень і точності результатів. Рисунок 2.2 містить представлення поділів, виконаних нашим круговим деревом для набору даних, який містить одну циклічну та одну лінійну змінну. Розщеплення, породжені цим деревом, визначають суміжні області в просторі.

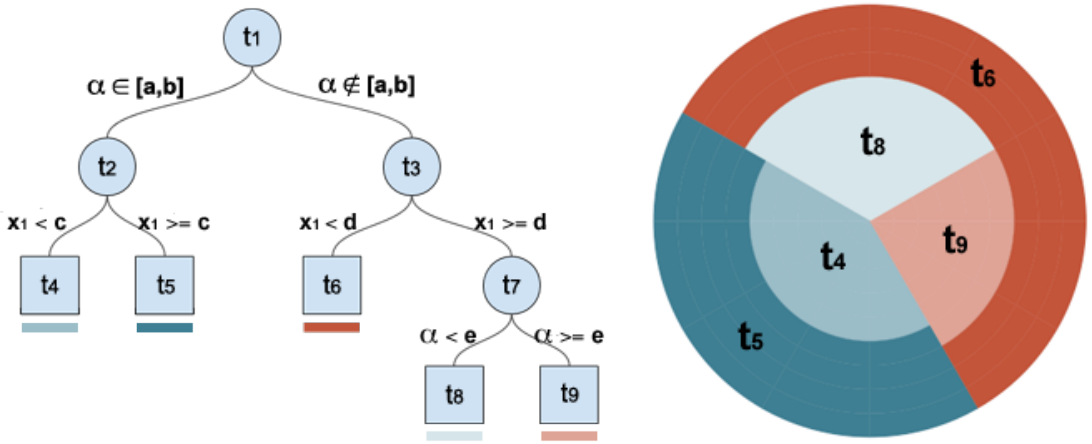



Рис. 2.2 - Приклад запропонованого дерева кругової регресії та представлення того, як простір ділиться на суміжні області.

У цій роботі представлено нову методологію, яка обмежує простір пошуку кожного розділу в дереві, що веде до створення суміжних розділів. Це обмеження, додане до початкової ідеї кругових дерев регресії в роботі Лунда, має наслідком генерування більш простих розділів, які покращують точність і ефективність потягів деревовидних моделей. Незважаючи на те, що простір пошуку оптимальних розбиття в кожному вузлі дерева більш обмежений, ніж у вихідній версії, загальна похибка дерева призводить до того, що нижче. Виявлено, що навіть якщо несуміжні поділи дають хороші результати у верхній частині дерева, ці розділи мають тенденцію вироджуватися як процес розщеплення еволюціонує, що призводить до низької продуктивності дерев.

Програмне забезпечення, розроблене в даній роботі, представлено у вигляді самодостатнього Python library. Ця бібліотека реалізує загальну версію regression tree, яку можна використовувати як з лінійними, так і з круговими змінними як на вході, так і на виході. Крім того, бібліотека пропонує можливість виконувати

несуміжні розділи, як у початковій пропозиції Лунда і суміжні, як у методології, яку ми запропонувати. Програмне забезпечення також можна використовувати як інтерфейс командного рядка, дозволяючи користувачам навчати та тестувати різні моделі дерева регресії та завантажувати набори даних для будь-якого аеропорту світу. Ці інструменти були розроблені, щоб користувачі могли експериментувати та досліджувати відмінності між моделями, вхідними змінними та аеропортами. Скрипти та бібліотеки написані простим способом, тому користувачі можуть читати код і розуміти структуру, а також змінювати або розширювати його Функцій. AeroCirTree поставляється з ліцензією GNU GPLv3, тому будь-хто може використовувати, модифікувати і спільно використовувати цю програму для будь-яких цілей.

### 2.3. Згорткові нейронні мережі

Працюючи над круговими регресійними деревами, проведено експерименти з використанням глибоких нейронних мереж для регресії даних про погоду. Хоча, наскільки нам відомо, в літературі з прогнозування погоди немає ніяких додатків, ми вирішили вивчити і застосувати ці методи до задач прогнозування погоди. Згорткові нейронні мережі, які використовуються в основному для задач класифікації зображень, здавалися хорошим кандидатом для моделювання сітчастих даних NWP.

До цього моменту нашого дослідження ми витягували дані з найближчої точки сітки NWP до бажаного спостережуваного набору даних, в результаті чого утворився часовий ряд. Згорткові нейронні мережі (CNN) пропонують можливість розглядати ціле зображення як вхідні дані в модель, і нейронна мережа може навчитися виявляти області зображення, які найбільше корелюють з результатом. Це була серйозна зміна в нашій методології дослідження. CNN не потребують локації окремих точок сітки як вхідних даних: вони можуть розглядати всі метеорологічні сітки (зображення) як єдиний вхід, уникаючи необхідність попереднього вилучення тимчасових рядів.

CNN в основному використовуються для класифікації та вилучення просторової інформації на зображеннях, вибудовуючи з дрібнозернистих деталей у структури вищого рівня. У цій роботі ми досліджували використання CNN для класифікації опадів (дощових/сухих) у різних містах Європи.

CNN вимагають великих наборів даних для навчання. Для цієї роботи ми використовували набір даних під назвою ERA-Interim, який містить дані, починаючи з 1979 року, з темпоральною роздільною здатністю 3 години. Він є загальнодоступним від Європейського центру середньострокових прогнозів погоди. Цей набір даних генерується за допомогою нумерної моделі погоди, яка моделює стан атмосфери для всієї планети з просторовою роздільною здатністю приблизно 80 Км. Вихідні дані представлені у вигляді звичайних числових сіток і доступної великої кількості фізичних параметрів, що представляють такі змінні, як температура, швидкість вітру та відносний вихор.

Для визначення дощових явищ у різних місцях ми використовуємо авіаційні планові метеорологічні звіти (METAR), подібно до наших попередніх робіт. Ми розглянемо 5 основних аеропортів, розташованих у різних містах Європи та періодом 5 років. Аеропорти: Гельсінкі-Ванта, Амстердам-Схіпхол, Дублін, Рим-Фьюмічіно та Відень.

Взявши з ERA-Interim розширену територію над Європою, створивши 3-годинну серію зображень, що складаються з 3 каналів, що відповідають висоті геопотенціалу **z** на 1000, 700 і 500 рівні тиску в атмосфері. Цей параметр являє собою висоту в атмосфері, при якій досягається певне значення тиску. Ці рівні відповідають, як правило, 100, 3000 і 5500 метрів над середнім рівнем моря **r**.

У цій роботі експериментуючи, додаючи часовий вимір до мережевої роботи CNN. Часова розмірність додається до цих мереж шляхом додавання четвертої осі до згорткових ядер (широта, довгота, висота, час). Використовуючи збір геопотенціальних полів як вхідних даних та умови опадів для різних локацій як вхідні дані, мережі CNN навчаються прогнозувати дощові умови в кожній точці. Незважаючи на те, що координати різних місць не передаються мережі, мережі вдається знайти зв'язок між полем тиску та даними про опади для конкретної

локації. Ця робота демонструє таким чином що CNN можуть бути використані для інтерпретації результатів NWP і автоматичного формування локальних прогнозів.

Також у рамках цієї роботи ми використовували техніку під назвою Class Activation Mapping (CAM) для інтроаналізу моделей CNN та візуальної оцінки регіонів вхідний простір, який має більший вплив на вихід. Використання CAM представлено на рисунку 2.3 з використанням теплової карти для представлення відносної ваги кожного пікселя в мережі. Мережа вчиться надавати більшу вагу пікселям, що оточують регіон, навіть коли цей параметр невідомий мережі.

Також у роботі було доведено, що 3D-згортки здатні природним чином включати часову складову даних у нейронні мережі, що призводить до значного підвищення точності результатів, якщо порівнювати з аналогічною мережею, навченою з окремими кадрами.

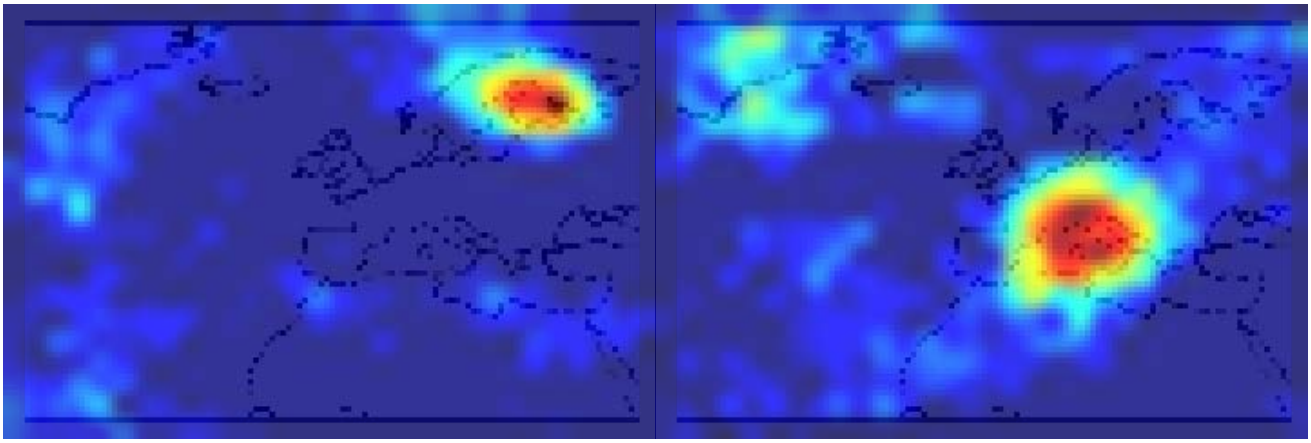


Рис 2.3 - Приклад отриманих карт активації класів для ERA-Interim CNN, навчених з використанням спостережуваних опадів в аеропортах Гельсінкі та Риму.

## 2.4. Згорткові кодери-декодери для регресії зображення в зображення

Для проведення експериментів у цій роботі використовується набір даних глобального аналізу клімату ERA-Interim, розроблений Європейським центром середньострокових прогнозів погоди (ECMWF). ERA-Interim надає велику кількість параметрів, з яких ми вибираємо геопотенціальну висоту і загальну кількість опадів. Ми обробляємо великі площі по всій Європі та витягуємо вибрані параметри протягом 1980-2016 років період з тимчасовою роздільною здатністю 6 годин, в результаті чого виходить набір даних з більш ніж 50 000 зразків.

Автоенкодери - загальні нейронні мережі, які відтворюють вхідні дані, виконуючи зменшення розмірності та подальше розширення вхідний простір. Ця техніка дозволяє вивчати стислі представлення даних без нагляду. Згорткові автокодери поєднуються з CNN з автоенкодерами для ефективного вивчення стиснених представлень зображень. Остання еволюція згорткових автокодерів демонструє, що подібні мережі пропонують ефективний спосіб виконання сегментації зображень, навчаючи модель на вибірках сегментованих зображень. На рисунку 2.4 показані перетворення, що здійснюються мережею кодера-декодера в геопотенційне поле, перетворюючи його в поле опадів для того самого регіону.

Ми розглянемо три сучасні мережі згорткових енкодерів-декодерів у сфері сегментації зображень: VGG-16, Seg-net та U-net. Ці мережі модифікуються для виконання завдань регресії зображень замість сегментації шляхом зміни функції втрат на MAE, і тестуються із завданням прогнозування опадів.

Ця робота демонструє, що нейронні мережі згорткового кодера-декодера можуть бути використані, як альтернатива параметризації, для навчання складних атмосферних процесів, використовуючи базові поля NWP як вводу. Наскільки нам відомо, це перша спроба забезпечити наскрізний підхід автоматизованого навчання для отримання нових параметрів з базових галузей НВП за допомогою глибини згорткові кодерно-декодувальні мережі. Ця робота також демонструє, як популярні мережі глибокого навчання в галузі сегментації зображень можуть бути адаптовані для інтерпретації та отримання нових параметрів погоди.

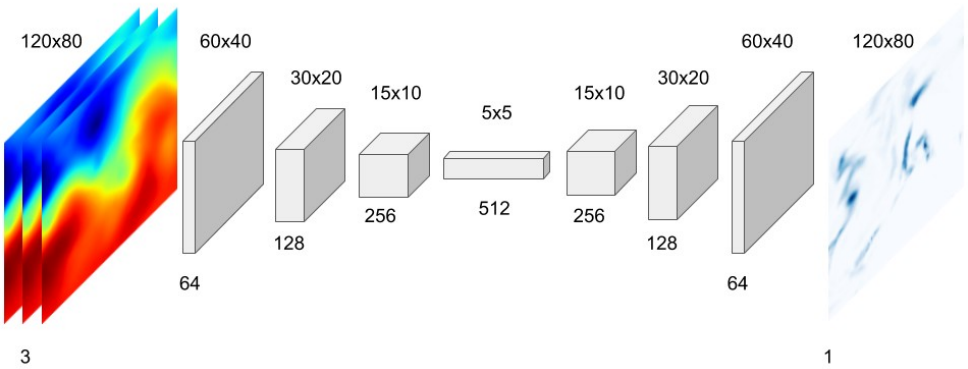



Рис 2.4 - Представлення перетворень, виконаних мережею кодер-декодер в поле геопотенціальної висоти і перетворення його в поле, що представляє сумарне поле опадів для того ж регіону.

# РОЗДІЛ 3 МЕТОДИ ПРОГНОЗУВАННЯ ШВИДКОСТІ ВІТРУ В АЕРОПОРТАХ

## 3.1. Прогнозування на основі непареметричної регресії

Вітер є одним із параметрів, який найкраще передбачити за допомогою чисельних моделей погоди, оскільки його можна безпосередньо розрахувати за допомогою фізичних рівнянь тиску, які керують його рухом. Однак місцеві вітри сильно залежать від топографії, яку глобальні числові моделі погоди через їх обмежену роздільну здатність не можуть відтворити. Щоб покращити навички чисельних моделей погоди, можна використовувати статистичні методи та методи аналізу даних. Методи машинного навчання можна застосувати для навчання моделі з даними, що надходять як з моделі, так і зі спостережень у цікавій області. У цій статті вивчається новий метод, заснований на непараметричній багатовимірній локально зваженій регресії, для покращення прогнозованої швидкості вітру в числовій моделі погоди. Дані про напрямок вітру використовуються для побудови різних регресійних моделей, як спосіб обліку впливу навколишнього рельєфу. Використання цієї методики забезпечує подібні рівні точності для прогнозів швидкості вітру порівняно з іншими алгоритмами машинного навчання з перевагою в тому, що вона більш інтуїтивно зрозуміла та проста для інтерпретації.

Моделі глобального числового прогнозування погоди (NWP) запускаються з просторовою роздільною здатністю, яка не може чітко відобразити вплив місцевої топографії. Було розроблено декілька інструментів і методологій для зменшення масштабу глобальних прогнозів NWP до регіонального чи місцевого масштабів. В основному всі вони можуть бути класифіковані як динамічні та статистичні підходи. Для методів динамічного масштабування метою є використання фізичної моделі високої роздільної здатності, вкладеної та ініціалізованої з граничними умовами моделі низької роздільної здатності, яка зазвичай охоплює більшу область. Статистичне зменшення базується на статистичному аналізі виходу NWP і даних спостережень для місця.

Статистичне зменшення масштабу також можна згрупувати в три категорії: а) класифікація/тип погоди визначає закономірності або синоптичні схеми погоди та аналізує дані відповідно до кожного випадку; б) Техніки функції регресії/передачі підбирають NWP і дані спостережень з використанням різних алгоритмів регресії та інших машинного навчання; і в) Генератори погоди засновані на ідеї створення стохастичного часового ряду як конвеєрного процесу для різних параметрів.

Методи непараметричного зменшення масштабу регресії базуються на ідеї, що предиктор не може бути сформульований за допомогою унікальної формули, а створюється під час виконання, розглядаючи весь набір даних і вибираючи його підмножину для побудови іншої регресійної моделі для кожного випадку. Непараметрична регресія потребує більших наборів даних, ніж регресія на основі параметричних моделей, оскільки лише обмежена частина даних використовується для побудови предиктора щоразу.

Непараметрична регресія вже була застосована в метеорологічних проблемах для зменшення масштабу моделей опадів. У цьому документі проста форма ядерної непараметричної регресії використовується для покращення прогнозу швидкості вітру, отриманого від NWP, враховуючи напрям вітру та змінні швидкості вітру для фільтрації даних. Ця форма регресії особливо підходить для прогнозування в реальному часі, оскільки її можна оновлювати, щоб включити найновіші дані, що робить її ідеальним кандидатом для оперативних додатків на вимогу.

Через циклічну природу напрямку вітру статистичний аналіз вітру неможливо виконати безпосередньо шляхом застосування до даних готових алгоритмів машинного навчання або зменшення масштабу. Запропонована регресійна модель використовує циклічний ядерний підхід для вибору схожих випадків вітру. Цей спосіб підгонки циклічних даних до моделі відрізняється від підходу, застосовуваного іншими методами спрямованої статистики, такими як кругова регресія або використання узагальнених адитивних моделей на окремих компонентах вітру.

Метою цієї методики є представлення методу вимірювання систематичної

похибки NWP при прогнозуванні швидкості вітру. Систематична похибка NWP в основному спричинена його обмеженою просторовою роздільною здатністю. Якщо цю похибку можна виміряти, враховуючи різницю в напрямку вітру та те, як вона впливає на швидкість вітру, можна буде відняти цю похибку з будь-якого лідируючого часу моделі, що вдосконалює свої навички. Запропонований метод створює непараметричну регресійну модель для оцінки зв'язку між прогнозованою швидкістю вітру NWP і спостережуваною швидкістю вітру, фільтруючи дані за напрямком вітру.

Аеропорти зазвичай розташовані на широких відкритих територіях з особливим вітровим режимом, сприятливим для повітряного руху. Навколишній рельєф впливає на поведінку вітру, блокуючи, посилюючи або змінюючи його напрямок під час руху, створюючи місцеві вітрові ефекти, які не вирішуються NWP. Якщо потрібно спрогнозувати швидкість вітру для певного напрямку, об'єднання подібних випадків вітру для побудови регресійної моделі є виправданим, оскільки на всі впливає одна і та ж топографічна конфігурація.

Щоб перевірити ефективність поточної методики зменшення масштабу, потрібно зібрати як NWP, так і дані спостережень. У запропонованій моделі регресії залежною змінною є спостережувана швидкість вітру, а незалежними змінними є швидкість вітру та напрямок вітру, що надходить від NWP. Ці дані мають бути представлені у вигляді часових рядів, використовуючи однакові одиниці вимірювання та часову роздільність для кожного з вибраних аеропортів.

Щоб перевірити продуктивність поточної методики зменшення масштабу, потрібно зібрати як NWP, так і дані спостережень. У запропонованій регресійній моделі залежною змінною є спостережувана швидкість вітру, а незалежними змінними є швидкість вітру та напрямок вітру, що надходить від NWP. Ці дані мають бути представлені у вигляді часових рядів, використовуючи однакові одиниці вимірювання та часову роздільність для кожного з вибраних аеропортів.

***Спостереження.*** Дані спостережень про погоду в цивільних аеропортах є загальнодоступними через метеорологічні звіти з аеродромів, відомі як METAR [Посібник ВМО з кодів], форму кодованих аеронавігаційних звітів про погоду, які

регулюються Міжнародною організацією цивільної авіації (ICAO). Ці повторні звіти зазвичай створюються щопівгодини та містять багато різних спостережень погодних умов, що впливають на аеропорт під час спостереження.

Кожен METAR містить таку інформацію, як ідентифікатор аеропорту, дата та час спостереження, вітер, хмарний покрив, температура, точка роси та тиск, серед інших параметрів, використовуючи кодований формат, визначений Всесвітньою метеорологічною організацією (ВМО). Для виконання випробувань витягується лише спостережуване значення швидкості вітру разом із відповідним значенням часової позначки для кожного з аеропортів. Швидкість і напрямок вітру METAR вказуються як виміряне або оцінене середнє значення за 10 хвилин до моменту випуску звіту. Поривчастий вітер, якщо він присутній, кодується як окрема змінна і не враховується.

ІКАО використовує код із чотирьох символів для ідентифікації кожного аеропорту. Коди ICAO для вибраних аеропортів: Форонда (LEVT), Лою (LEBB) і Гондаррібія (LESO).

***Дані NWP.*** Система Global Forecas t System (GFS) є глобальною числовою системою Wенаїї Модель, що працює в операційному режимів Національній Західній Європі США Служба з березня 2011 року, і всі її історичні файли NetCDF можуть бути доступні в Оперативній моделі Національного управління океанічних і атмосферних досліджень (NOAA) Archive та Distribution System (NOMADS), публічний репозиторій в Інтернеті. GFS має просторово роздільну здатність 0,5 градуса (приблизно 55 кілометрів) і часовою роздільною здатністю 3 години з новим пробігом в 6 годин. Для того, щоб оцінити систематичну похибку NWP.

Для цього дослідження використовується простий підхід для отримання часових рядів сайту. Найближча точка сітки GFS до кожного аеропорту вибирається без урахування будь-якої іншої форми просторової інтерполяції чи корекції. Для кожного місця U і V 10-метрові компоненти швидкості вітру витягуються в часовий ряд, що відповідає змінним GFS «компонент U висоти вітру над землею» і «V компонент висоти вітру над землею». Ці змінні містять миттєві значення, виміряні в метрах за секунду. Діючи так само, як і з даними

спостережень, дані вітру GFS збираються для найближчих точок сітки до аеропортів за 3 роки.

***Часовий ряд.*** NWP зазвичай представляють вітер через його декартові компоненти, що дуже зручно для обчислення середніх значень та іншого статистичного аналізу. Однак дані про вітер, що надходять від метеостанцій, зазвичай описуються за допомогою компонентів «напряму» та «швидкості». Для цього дослідження напрямок вітру використовується для фільтрації даних, включених до непараметричної регресії. Значення вітру, отримані з GFS, повинні бути перетворені з їх декартових компонентів у компоненти напрямку та швидкості перед включенням у часовий ряд.

Оскільки дані GFS були зібрані з роздільною здатністю 3 години, а METAR доступні кожні 30 хвилин, порівнювати можна лише частину значень. Комбінований часовий ряд для 00, 03, 06, 09, 12, 15, 18 і 21 години UTC створюється з використанням даних моделі GFS і звітів METAR, усі додаткові звіти METAR ігноруються.

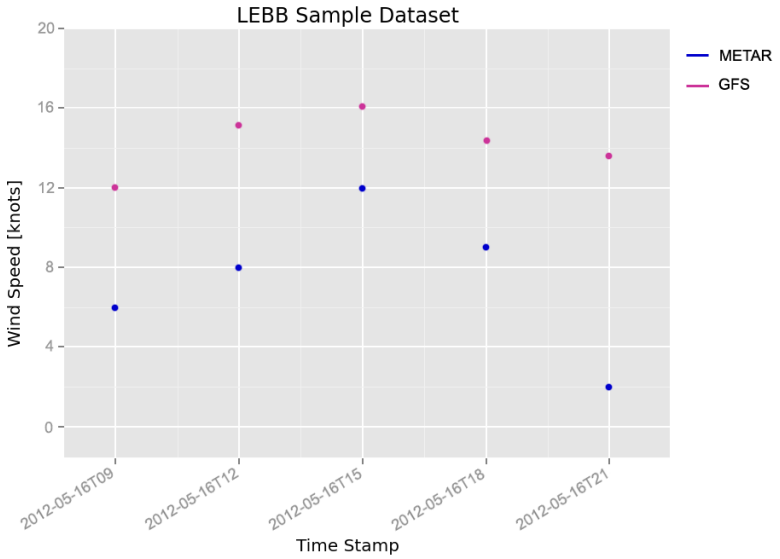



Рис. 3.1 - Вибірка даних часових рядів для швидкостей вітру METAR і GFS з аеропорту Лоіу (LEBB).

Форонда (LEVT) — єдиний аеропорт із трьох, що записує METAR 24 години на добу. Для цього аеропорту зібрано 5840 точок даних. Інші два аеропорти закриваються на кілька годин вночі. 00H і 03H METAR пропущено для аеропорту

Лойу (LEBB), а також 21H у випадку Хон-даррібії (LESO), даючи загальну кількість 4380 і 3650 рядків даних відповідно. Таблиця 1 показує зразок об'єднаних даних часового ряду для аеропорту Лою (LEBB), а (рис. 3.1) містить представлення значень швидкості вітру METAR і GFS.

**Основи методології.** Покращення помилки прогнозування швидкості вітру може бути досягнуто шляхом застосування простої однофакторної регресійної моделі, що поєднує значення швидкості вітру, отримані з METAR і моделі GFS. Однак NWP містить багато різних фізичних параметрів, які можна включити в метод регресії для покращення результатів. Методи непараметричної регресії вимагають перевірки великих наборів даних, оскільки вони використовують підмножини для обчислення регресії, що є причиною збору даних часових рядів за два роки в базі даних. Непараметрична регресія — це форма регресії, у якій предиктор не можна виразити через одну функцію, скоріше вона обчислює нову регресію для кожного прогнозованого значення. Цей вид алгоритму особливо підходить для прогнозування в реальному часі, оскільки регресійна модель побудована на основі часу виконання. Локально зважена регресія є формою непараметричної регресії, де значення, близькі до прогнозованої точки, мають сильніший вплив на регресію. У наступних розділах описано основні методи, які використовуються в цій моделі регресії.

***Регресійна модель.*** Метод зважених найменших квадратів — найпростіший спосіб підгонки даних до моделі, коли потрібне зважування. Його рівняння в матричній нотації можна виразити наступним чином:

$$(\mathbf{X}^T\mathbf{W}\mathbf{X})\beta = \mathbf{X}^T\mathbf{W}\mathbf{Y}$$

Де X — матриця незалежних змінних, W — діагональна матриця, що містить ваги кожної точки, Y — вектор, що містить значення залежної змінної, і вектор, що містить коефіцієнти для кожної змінної регресії.

Якщо визначено навчальний набір, де відомі значення як X, так і Y, значення можна визначити та використати для прогнозування нових значень залежної змінної.

Форма локально зваженої регресії може бути реалізована за допомогою

методу зважених найменших квадратів, що визначає вагову функцію для вибору локальних точок навколо точки регресії.

Порівняння між простою лінійною регресією та локально зваженою регресією для даних швидкості вітру в аеропорту Гондаррібія (LESO) (рис. 3.2) демонструє, як непараметрична регресія адаптується до нелінійності в даних за рахунок обчислювальних витрат на обчислення - створення нових параметрів регресії для кожної точки графіка. Функція зважування, яка використовується для створення локально зваженої регресії, показаної на цьому малюнку, пояснюється в наступному підрозділі.

***Функція ядра.*** Ядро — це добре відома вагова функція, яка використовується в методах непараметричного оцінювання для формування впливу різних точок даних, які беруть участь у непараметричній регресії. Є багато функцій, які можна використовувати як ядра, наприклад уніформа, трикутник, косинус або трикуб. Tricube є одним із найпопулярніших ядер, яке використовується в моделі, запропонованій у цьому дослідженні.

Таблиця 1. Зразок даних часового ряду для аеропорту Лою (LEBB), що поєднує дані з METAR і моделі GFS. Відображення лише перших трьох стовпців для економії місця (TimeStamp, Metar Wind Speed і GFS Wind Speed).

| *TimeStamp* | *MetarWindSpeed* | *GFSWindSpeed* | *GFSWindDir* |
|---|---|---|---|
| 2012-05-16T09:00 | 6.0 | 12.060 | 95 |
| 2012-05-16T12:00 | 8.0 | 15.114 | 81 |
| 2012-05-16T15:00 | 12.0 | 16.094 | 79 |
| 2012-05-16T18:00 | 9.0 | 14.291 | 79 |
| 2012-05-16T21:00 | 2.0 | 13.870 | 100 |

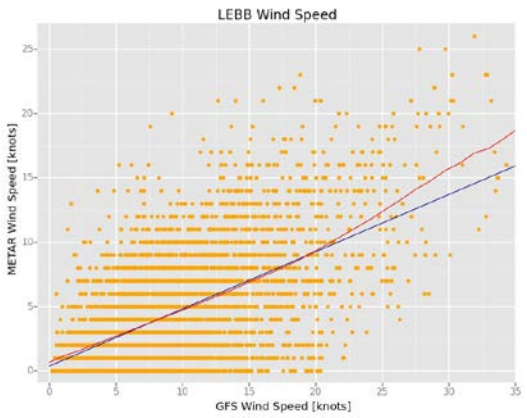



Рис. 3.2 - Порівняння простої лінійної регресії та локально зваженої регресії для даних швидкості вітру з аеропорту Лою (LEBB).

$$K(d) = \begin{cases} \frac{81}{90}(1 - |\frac{d}{d_{max}}|^3)^3, & \text{if } d \leq d_{max} \\ 0 & \text{if } d > d_{max} \end{cases}$$

Ця функція повертає відносну вагу кожної точки в моделі, де d представляє її відстань до прогнозованої точки, а dmax — максимальну відстань, починаючи з якої жодна точка не буде включена в регресійну модель. Це ядро визначає як кількість точок, так і їхню відносну вагу в регресійній моделі залежно від того, наскільки кожна точка далека від прогнозованого значення.

На рис. 3 показано, як ядро tricube зважує різні точки даних навколо значення 15 і 20 вузлів для створення регресійної моделі. У цій вибірці відносна вага кожної точки в регресії зменшується від свого максимального значення при 15 і 20 вузлах і стає нульовою для точок далі, ніж 10 вузлів з кожного боку, що є фіксованим значенням dmax.

***Перехресне розмірне зважування.*** У попередньому розділі наведено приклад того, як ядро формує вагу точок залежно від їх відстані від прогнозованої точки. Кожна точка, яка розглядається в регресії, містить багато інших параметрів, крім швидкості вітру. Ця техніка заснована на ідеї адаптації регресійних моделей, використовуючи історичні дані з характеристиками, подібними до дня, який ми намагаємося спрогнозувати. Наприклад, якщо модель NWP прогнозує вітер, що дме з півночі, кращі результати слід отримати, фільтруючи дані, щоб використовувати значення бази даних, де вітер дме з півночі. Однак дані також можна відфільтрувати за допомогою будь-якої іншої змінної, що міститься в наборі даних. Швидкість вітру можна спрогнозувати, відфільтрувавши точки даних, щоб включити ті, що демонструють характеристики, подібні до прогнозованого дня. У цьому документі напрямок вітру запропоновано як хорошу фільтруючу змінну, але також можна використовувати будь-яку іншу змінну.

Ядра використовуються для визначення вагової матриці W, яка використовується в регресії. W — квадратна матриця, яка визначається:

$$\mathbf{W} = \mathbf{I}\mathbf{K}_1\mathbf{K}_2 \ldots \mathbf{K}_n$$

де I — одинична матриця, а Ki — і-та матриця ядра. Ядерна матриця K — це діагональна матриця, у якій кожне значення головної діагоналі відповідає вазі

кожної точки даних у регресії. Таким чином, K є квадратною матрицею з розмірністю, що дорівнює кількості елементів у наборі даних. Вага кожного значення визначається ядерною функцією tricube. Оскільки всі матриці, які використовуються для обчислення вагової матриці W, є діагональними матрицями однакової розмірності, властивість комутативності може бути застосована, що означає, що порядок ядер не впливає на результат.

На рис. 3.4 показано дані про швидкість вітру для аеропорту Хондаррібія (LESO), де колір кожної точки відображає напрям вітру. На рис. 5 показано приклад такого перехресного зважування з використанням напрямку вітру для фільтрації даних. Ядро трикуба напрямку вітру вибирає підмножину вихідних точок із напрямком вітру приблизно від 0 до 180 градусів.

***Запропонована методологія.*** Ідея розгляду історичних подібних випадків природно виникає в діяльності прогнозування погоди. У попередньому розділі представлені деякі математичні інструменти та ідеї, які можуть допомогти відфільтрувати ці «схожі ситуації» з усього набору даних. У конкретному випадку прогнозування вітру топографія має великий вплив на визначення характеру вітрів у будь-якому місці.

Напрямок вітру класифікує вітри, що дмуть з різних місць, і його можна використовувати для представлення місцевих впливів на вітри. Групування даних вітру одного напрямку є способом неявного введення ефекту місцевої топографії.

У цьому розділі розглядається ідея використання значень напрямку вітру для прогнозування швидкості вітру. Ядерні матриці використовуються як спосіб зважування підмножин даних у моделі регресії. Однак функція ядра, представлена в попередньому розділі, призначена для роботи з лінійними змінними, а напрямок вітру є круговим.

***Циклічне ядро.*** Особливістю напрямку вітру є круговий простір замість лінійного. Вимірюється в градусах, напрямок вітру може приймати значення в діапазоні [0-360], де 0 і 360 представляють ту саму точку. Щоб обчислити відстань між двома кутами, необхідно вибрати мінімальну з двох можливих відстаней навколо циклу.

$$Dist(a,b) = min \begin{cases} b-a \\ a+360-b \end{cases} \quad \text{where } b \geq a$$

Ця відстань і визначена максимальна кутова відстань dmax використовуються в ядрі tricube для призначення вагових коефіцієнтів різним точкам даних і отримання найкращих оцінок швидкості вітру. Вибір оптимального значення для dmax є ключовим для побудови точної моделі регресії. Занадто малі значення dmax враховують лише невеликі сектори даних, що може спричинити переобладнання та погане узагальнення моделі, тоді як, з іншого боку, занадто великі значення дають надзвичайно загальні регресії, які не можуть розрізнити різні випадки.

***Побудова та валідація моделі.*** Для кожного аеропорту використовуються дані про швидкість і напрям вітру з моделі GFS і зі звітів METAR. Дані швидкості вітру GFS визначають залежну змінну матрицю X; Спостережувані значення швидкості вітру METAR утворюють матрицю пояснювальної змінної Y; Дані про напрямок вітру GFS використовуються для побудови вагової матриці W з використанням циклічних ядер.

Щоб підтвердити запропоновану модель, необхідно визначити методологію перевірки результатів. Необхідно визначити техніку перевірки моделі, щоб оцінити, як прогнози вітру узагальнюються в незалежному наборі даних. Затримка, перехресна перевірка та початкове завантаження — це різні методи випадкового поділу набору даних і перевірки статистичного методу. Затримка ділить набір даних на дві різні групи, одна з яких використовується для навчання статистичної моделі, а інша – для перевірки чи тестування продуктивності навченої моделі. Перехресна перевірка ділить набір даних на n різних груп і виконує перевірку. Одна група даних одночасно виключається з навчання, використовуючи цю виключену групу для проведення тестування. Потім цей процес повторюється, щоразу змінюючи групу, вибрану для виключення, доки не будуть охоплені всі групи. Бутстрапінг – це різновид утримування, коли кожна з підмножин отримується шляхом випадкової вибірки із заміною з вихідного набору даних.

Для перевірки запропонованої моделі обрано метод повторної витримки.

Рішення про використання методу повторного витримки замість перехресної перевірки або початкового завантаження приймається на основі розміру набору даних. Для великих наборів даних різниця між випадковим розміщенням точок даних або перехресною перевіркою фіксованих підмножин даних незначна.

Для кожного експерименту виконується повторна оцінка очікування, коли весь набір даних кожного аеропорту випадковим чином розбивається на два набори, один з яких використовується для навчання регресійної моделі, що містить 75% даних, а інший 25% використовується для перевірки. . У навчальному наборі дані зі спостережень швидкості вітру використовуються для навчання регресійної моделі, а ця сама змінна прихована та використовується для оцінки помилки в наборі перевірки. Середньоквадратична помилка (RMSE) швидкості вітру використовується як оцінка похибки моделі. Для кожного експерименту та аеропорту ця процедура виконується 10 разів, а її значення RMSE усереднюються.

**Результати.** Модель порівнюється з іншими найсучаснішими методами регресії, щоб оцінити її ефективність. Для перевірки дати кожного методу використовується 10-кратний процес оцінки затримки, як описано в попередньому розділі. Порівняння між різними методологіями завжди виконуються з використанням тих самих повторюваних розподілів. Значущість різниці між порівнюваними типами регресійних моделей статистично оцінюється за допомогою парного t-критерію. Використання параметричного тесту є виправданим, оскільки тест Шапіро-Вілкса забезпечив припущення про нормальність порівнюваних зразків RMSE.

Таблиця 3.2. Середнє середньоквадратичне значення швидкості вітру та результати прямого прогнозування спостережуваної швидкості вітру з використанням вихідних даних швидкості вітру моделі GFS у порівнянні з середньоквадратичним значенням середньоквадратичної середньої середньої середньоквадратичної швидкості, отриманої після застосування моделі одновимірної лінійної регресії, що містить ті самі дані.

$\boldsymbol{E}(RMSE)\pm\sigma(RMSE)$

| $Method$ | $LESO$ | $LEVT$ | $LEBB$ |
|---|---|---|---|
| GFS output | 4.361±0.067 | 4.015±0.074 | 6.863±0.107 |
| Linear Reg. | 2.765±0.110 | 3.929±0.083 | 3.560±0.068 |

Таблиця 3.3. Середнє середньоквадратичне значення швидкості вітру та результати для різних аеропортів із використанням різних значень dmax напрямку вітру в градусах у трикубічному ядрі для зважування даних у непараметричній регресії.

| | $\boldsymbol{E}(RMSE)\pm\sigma(RMSE)$ | | |
|---|---|---|---|
| $d_{max}$ | $LESO$ | $LEVT$ | $LEBB$ |
| 5 | 2.758±0.104 | 3.434±0.059 | 3.426±0.079 |
| 10 | 2.735±0.104 | 3.414±0.059 | 3.378±0.074 |
| 20 | 2.711±0.100 | 3.428±0.063 | 3.370±0.070 |
| 30 | 2.706±0.097 | 3.459±0.067 | 3.381±0.069 |
| 40 | 2.706±0.095 | 3.496±0.070 | 3.401±0.069 |
| 60 | 2.710±0.095 | 3.571±0.079 | 3.441±0.070 |
| 90 | 2.725±0.100 | 3.657±0.089 | 3.487±0.069 |
| 120 | 2.741±0.105 | 3.711±0.090 | 3.509±0.069 |
| 180 | 2.760±0.109 | 3.762±0.087 | 3.544±0.069 |

Перш за все, еталон встановлюється як еталон, щоб результати запропонованої регресійної моделі можна було порівняти з ним. Найпростіший і найменш точний метод прогнозування вітру полягає у використанні значення швидкості вітру з числової моделі погоди для прогнозування спостережуваної швидкості вітру в аеропорту. Цей результат можна легко покращити, застосовуючи просту лінійну регресію, щоб зв'язати значення швидкості вітру з NWP і даних спостережень із METAR. Таблиця 3.2 містить результати RMSE прямого порівняння між швидкостями вітру, прогнозованими GFS, і відповідними спостережуваними значеннями METAR. Таблиця 3.2 також містить результати RMSE, отримані за допомогою моделі однофакторної лінійної регресії для прогнозування швидкості вітру для різних аеропортів. Перший результат отримано з використанням усього набору даних, тоді як у випадку регресії методологія, пояснена в кінці попереднього розділу, використовується для розрахунку значень RMSE кожного аеропорту. Як показано в таблиці 3.2, лінійна регресія вносить помітне покращення в прогнозованій швидкості вітру. Запропонована непараметрична регресійна модель спрямована на покращення цих значень RMSE.

Таблиця 3.3 містить результати проведення експериментів з використанням різних значень напрямку вітру dmax для фільтрації даних, які використовуються в

регресії, як пояснюється в розділі методології. Як показано на малюнку 6, три аеропорти демонструють подібну поведінку, демонструючи середні значення RMSE мінімальної швидкості вітру з dmax в діапазоні від 10 до 30 градусів, але загальне досягнуте покращення є різним для кожного з них.

Як тільки значення напрямку вітру dmax, що мінімізує помилку, його можна виправити та ввести нове ядро для фільтрації даних за допомогою іншої змінної. Точки даних із подібним напрямком вітру, вибрані першим ядром, знову зважуються за допомогою другого ядра з новою змінною. Використання кількох ядер дозволяє нам фільтрувати дані відповідно до різних змінних і таким чином досягати кращих результатів у регресії. Наприклад, якщо значення напрямку вітру dmax зафіксовано на рівні 30 градусів в одному ядрі, вторинне ядро може використовуватися знову для зважування отриманої підмножини даних відповідно до іншої змінної.

Оскільки метою є покращення прогнозу швидкості вітру, отриманого з моделі, напрямок вітру та швидкість вітру GFS можна комбінувати, щоб вибрати точки даних, де вітер надходить з того самого напрямку та має однакову швидкість. Як пояснювалося в попередньому абзаці, можна застосувати два ядра, одне з яких використовує напрямок вітру, а інше швидкість вітру, щоб вибрати точки NWP із схожими характеристиками вітру (напрямком і швидкістю), як і прогнозовані. Використовуючи результати попереднього експерименту, який визначив оптимальне значення напрямку вітру dmax близько 30 градусів, запропоновано новий експеримент, який поєднує два ядра. Перше ядро фільтрує дані за напрямком вітру GFS, використовуючи оптимальне значення dmax, а друге ядро фільтрує дані за допомогою змінної швидкості вітру GFS. Як було виконано в попередньому експерименті, різні значення швидкості вітру перевіряються d’max, щоб знайти оптимальне значення, яке мінімізує помилку регресії. Щоб уникнути плутанини в позначенні параметрів двох ядер, посилання на ядро швидкості вітру використовують штрих. Таблиця 3.4 містить результати RMSE цього експерименту з використанням різних значень d’max у регресійній моделі.

Аналіз результатів, наведених у таблиці 3.4, показує, що запропонована

непараметрична регресія з використанням ядер статистично перевершує стандартну лінійну регресію (p-value=0,001). Можна спостерігати закономірність у середніх значеннях, коли продуктивність покращується, коли значення d'max збільшується до моменту, коли воно знову погіршується. Різні аеропорти показують різні оптимальні значення d'max, але всі вони мають мінімум у діапазоні від 2 до 4 вузлів.

Застосовуючи цю техніку, виникає велика кількість різних комбінацій для підгонки непараметричної моделі регресії, залежно від кількості та типу змінних, порядку їх застосування та форми використовуваного ядра.

Щоб оцінити продуктивність цієї моделі прогнозування вітру, виконується порівняння з популярною технікою регресійного машинного навчання. Випадковий ліс використовується як еталон. Випадковий ліс — це метод ансамблевого навчання для регресії. Він працює шляхом побудови безлічі дерев рішень під час навчання та виведення середнього результату з окремих дерев. Кожне дерево навчається з використанням випадкової підмножини всього набору даних.

Таблиця 3.4. Середнє RMSE та результати з використанням фіксованого значення напрямку вітру dmax та різних значень d'max швидкості вітру як параметрів ядра в непараметричній регресії.

| | | $\boldsymbol{E}(RMSE)\pm\sigma(RMSE)$ | | |
|---|---|---|---|---|
| $d_{max}$ | $d'_{max}$ | $LESO$ | $LEVT$ | $LEBB$ |
| 30 | 0.5 | 2.691±0.093 | 3.442±0.059 | 3.329±0.054 |
| 30 | 1.0 | 2.666±0.095 | 3.426±0.062 | 3.303±0.055 |
| 30 | 2.0 | 2.634±0.096 | 3.419±0.063 | 3.294±0.054 |
| 30 | 4.0 | 2.655±0.094 | 3.430±0.062 | 3.291±0.053 |
| 30 | 8.0 | 2.684±0.092 | 3.438±0.060 | 3.318±0.054 |

Таблиця 3.5. Результати RMSE для різних аеропортів із застосуванням оптимізованої непараметричної регресійної моделі (NP) і випадкового лісу (RF).

| | $\boldsymbol{E}(RMSE)\pm\sigma(RMSE)$ | | |
|---|---|---|---|
| $Model$ | $LESO$ | $LEVT$ | $LEBB$ |
| NP (30/2) | 2.764±0.096 | 3.433±0.063 | 3.328±0.054 |
| RF | 2.682±0.172 | 3.448±0.117 | 3.362±0.139 |

Для порівняння запропонованої моделі використовується модель Random Forest, яка містить 100 дерев і використовує значення швидкості та напрямку вітру з моделі GFS; і спостережні значення швидкості вітру з METAR. Як було виконано з непараметричною регресією, 75% даних використовуються для навчання випадкових лісів, а інші 25% використовуються для перевірки моделі, де значення швидкості вітру METAR використовуються для оцінки похибок. Для кожного аеропорту процес повторюється 10 разів, усереднюючи результати RMSE.

У таблиці 5 представлені дуже схожі результати продуктивності як запропонованої моделі, так і випадкового лісу. Парний t-тест не показує статистично значущих відмінностей між обома методами (p-value=0,36).

Одна з важливих переваг цієї моделі полягає в тому, що вона дуже інтуїтивно зрозуміла, і ядра можна налаштувати, щоб максимізувати продуктивність моделі для кожного аеропорту чи місця. Це велика перевага в порівнянні з результатами чорної скриньки інших методів, таких як випадковий ліс або нейронні мережі. Алгоритм випадкового лісу, використаний у цьому експерименті, не враховував круговий характер напрямку вітру. Таким чином, можна очікувати покращення його продуктивності, якщо використовувати версію, здатну працювати з даними спрямованості.

***Висновки.*** Для покращення продуктивності NWP можна використовувати різні види статистичної постобробки. У деяких випадках, коли потрібні локальні прогнози, статистичний аналіз і алгоритми машинного навчання можуть значно зменшити похибку прогнозованих змінних.

Непараметричні моделі регресії представляють простий і водночас ефективний спосіб представлення нелінійних зв’язків між змінними. Використання ядер у цих моделях дозволяє нам формувати вплив (вагу), який сусідні точки мають у регресії залежно від їх близькості до прогнозованої точки. У цьому дослідженні було доведено, що використання ядер у непараметричній регресії особливо добре працює для прогнозування швидкості вітру в аеропортах, де демонструються помітні місцеві вітрові режими. Ядра також пропонують простий спосіб підгонки спрямованих даних до регресійної моделі.

Як зазначалося в попередньому розділі, існує велика кількість різних можливостей підгонки даних до непараметричної моделі регресії залежно від кількості змінних, які використовуються як ядра для зважування даних, і порядку їх застосування. Наприклад, сонячне опромінення, яке доступне як змінна в більшості NWP, також призвело до надійного ядра фільтрації при прогнозуванні швидкості вітру. Його висока кореляція з турбулентністю вітру, що виникає через нагрівання поверхні землі сонцем, робить його хорошим способом обліку щоденних і сезонних турбулентних режимів вітру. Результати використання цієї змінної не включені в цю статтю, оскільки вона в основному зосереджена на факті використання напрямку вітру як основної змінної для визначення місцевого впливу рельєфу.

Було доведено, що циклічні ядра успішно асимілюють дані спрямованості в регресійну модель. Інші циклічні змінні, які зазвичай містяться в наборах даних із мітками часу, це «час доби» та «день року». Той самий метод непараметричної регресії з використанням циклічних ядер можна застосувати до цих змінних, виділяючи з даних сезонність і щоденні моделі. Аналіз часових рядів даних аеропорту, як сезонність або щоденна фільтрація шаблонів, не досліджувався в цій статті, але він заслуговує подальшого дослідження.

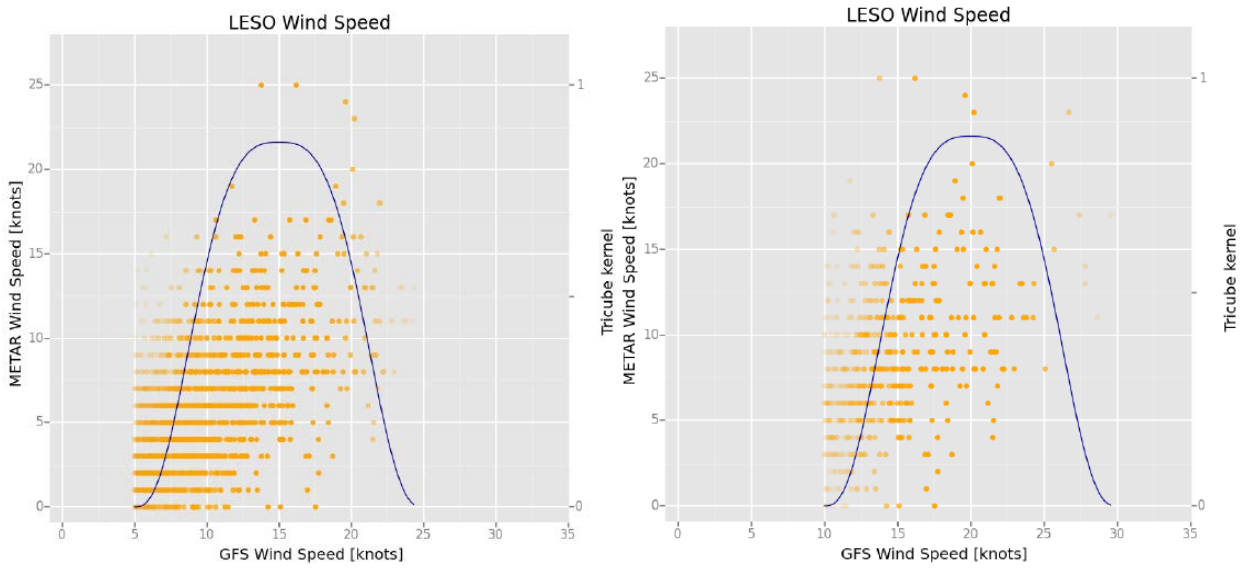



Рис. 3.3 - Вибір точок швидкості вітру для аеропорту Гондаррібія (LESO) з використанням двох трикубних ядер із центром 15 і 20 вузлів із значенням dmax 10 вузлів. Точки бліднуть під впливом ядра, інтенсивність кольору представляє їхню відповідну вагу в регресії.

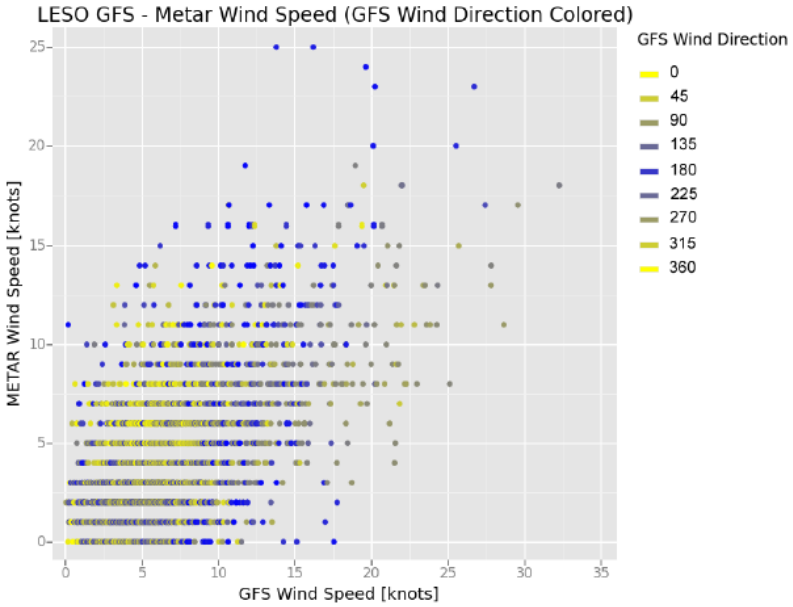



Рис. 3.4 - Зв'язок між значеннями швидкості вітру GFS і METAR для аеропорту Гондаррібія (LESO).

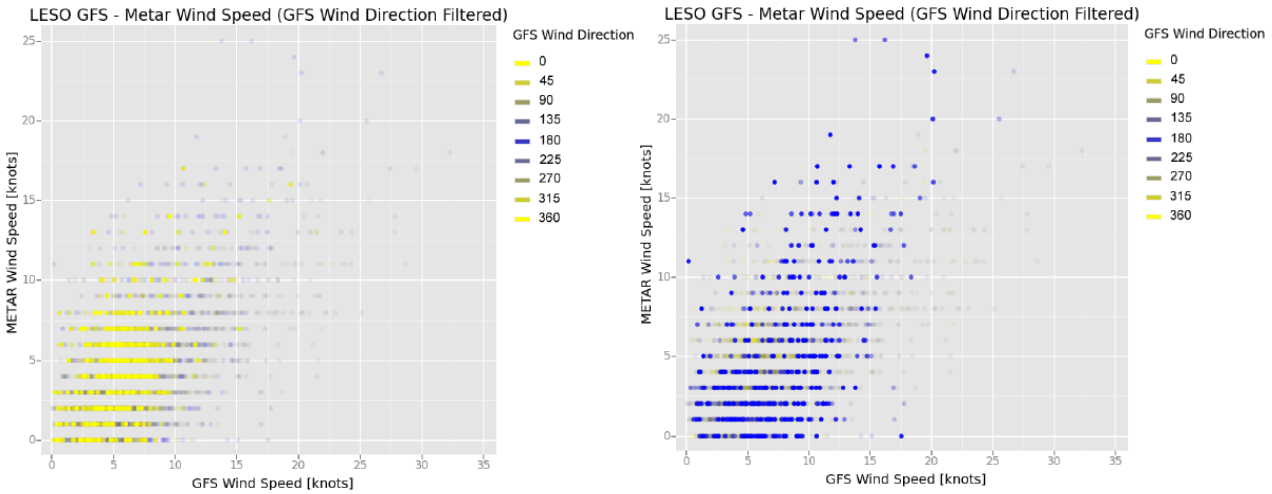



Рис. 3.5 - Вибір точок швидкості вітру для аеропорту Гондаррібія (LESO) з використанням трикубного ядра з двома напрямками вітру навколо 0 і 180 градусів.

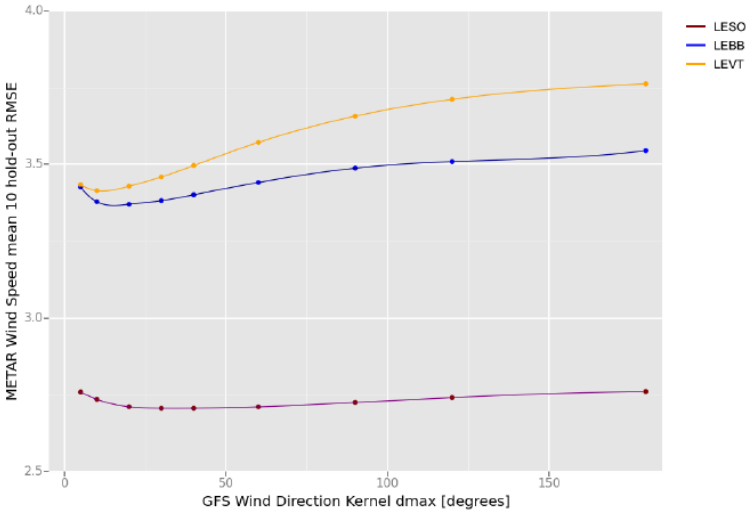



Рис. 3.6 - Еволюція середнього RMSE швидкості вітру для різних аеропортів як функція dmax значення ядра напрямку вітру.

## 3.2. Система прогнозування погоди в аеропорту на основі дерев кругової регресії

Сучасне прогнозування погоди здебільшого спирається на числові моделі, які імітують еволюцію атмосфери на основі рівнянь динаміки рідин і термодинаміки. Ці рівняння розв'язуються для дискретних точок регулярної сітки, що охоплює цікаву область. Моделі з вищою роздільною здатністю створюють більш детальні прогнози, але також вимагають великих обчислювальних ресурсів і більш тривалого часу роботи. Операційні моделі замінюють якість роздільної здатності на менший час обробки. Потреба в прогнозах з вищою роздільною здатністю спонукала до появи численних методологій для отримання більш детальних результатів, що називається зменшенням масштабу. Динамічне зменшення масштабу використовує вихід більш грубої моделі як початкову умову локальної моделі з вищою роздільною здатністю, яка краще вирішує процеси підсітки та топографію. Іншим підходом є статистичне зменшення, коли історичні дані спостережень використовуються для покращення результатів чисельної моделі. Існують численні методології статистичного масштабування, що базуються на різних принципах, таких як аналоги, інтерполяція або моделі машинного навчання.

Авіаційні операції сильно залежать від погодних умов і вимагають найякіснішої метеорологічної інформації для максимальної ефективності та безпеки. Міжнародна організація цивільної авіації (ICAO) і Всесвітня метеорологічна організація (WMO) встановили міжнародні стандарти для забезпечення високої якості метеорологічних звітів. Для створення цих звітів національні метеорологічні служби по всьому світу використовують висококваліфікований персонал, який постійно спостерігає та прогнозує умови навколо аеропорту, такі як видимість, напрямок і швидкість вітру або близькість штормових осередків. Авіаційні синоптики покладаються в основному на свої знання про аеропорт і якість використовуваного ЧПП.

Існує низка інструментів, які полегшують процес генерації прогнозів погоди в аеропорту, і на даний момент є областю активних досліджень. Аеропорти

зазвичай мають довгі та регулярні серії високоякісних історичних даних спостережень, які можна використовувати для створення статистичних моделей зменшення масштабу, щоб допомогти прогнозистам у їхній роботі. Вплив невирішених навколишніх гір, водойм чи місцевих кліматичних умов можна врахувати в цих моделях шляхом вивчення місцевих ефектів, спричинених погодними умовами в минулому.

Кругові змінні присутні в будь-якому спрямованому вимірюванні або змінній з притаманною періодичністю. Дані про погоду містять багато параметрів, які представлені як кругові змінні, наприклад напрямок вітру, географічні координати або мітки часу. Більшість сучасних алгоритмів машинного навчання регресії зосереджені на моделюванні зв'язків між лінійними змінними. Циркулярні змінні мають різну природу, ніж лінійні, тому традиційні методології не можуть повністю відобразити їхній зміст, що в більшості випадків призводить до неоптимальних результатів. Модель, представлена в цій роботі, базується на концепції кругових регресійних дерев, введених Лундом. Наша модель ефективніша з точки зору обчислень і генерує безперервні розбиття для циклічних змінних, що забезпечує кращу точність у порівнянні з її попередником.

Дерева кругової регресії можуть краще представляти циклічні змінні, оскільки вони враховують більше можливостей для розбиття простору, ніж дерева лінійної регресії. Дерева кругової регресії можуть визначати підмножини даних навколо початкової точки 0, 2↟ радіан. Наприклад, при прогнозуванні події, яка демонструє високу кореляцію із зимовими місяцями в північній півкулі, кругле дерево зможе виділити місяці з грудня по березень в одну групу. З іншого боку, лінійне дерево, швидше за все, розглядатиме розбиття, що починаються або закінчуються на початку року, не створюючи групи, що містить ці місяці.

У цьому документі представлено AeroCirTree, систему, засновану на описаній моделі дерева кругової регресії, яка здатна створювати покращені прогнози погоди для будь-якого аеропорту світу. Це програмне забезпечення представляє загальне рішення, де доступні всі необхідні інструменти, необхідні для отримання історичних даних про погоду, навчання моделей і створення нових

прогнозів. Ця система призначена для того, щоб допомогти авіаційним синоптикам створювати звіти кращої якості, а дослідникам машинного навчання створювати більш складні моделі.

***Методологія.*** Завдяки своїй простоті, швидкості навчання та продуктивності регресійні дерева є популярним і ефективним методом моделювання лінійних змінних. Дерева класифікації та регресії (CART) є однією з найпопулярніших версій дерев регресії.

Дерева лінійної регресії рекурсивно розділяють простір, знаходячи найкраще розбиття в кожному нетермінальному вузлі. Кожне розбиття ділить простір на два набори за допомогою функції вартості, яка зазвичай базується на метриці для мінімізації комбінованої дисперсії отриманих дочірніх вузлів.

Рисунок 3.7 містить приклад дерева регресії на основі двох лінійних змінних x1 і x2. З правого боку є графічне представлення того, як простір поділяється шляхом створення розділень на ці дві змінні.

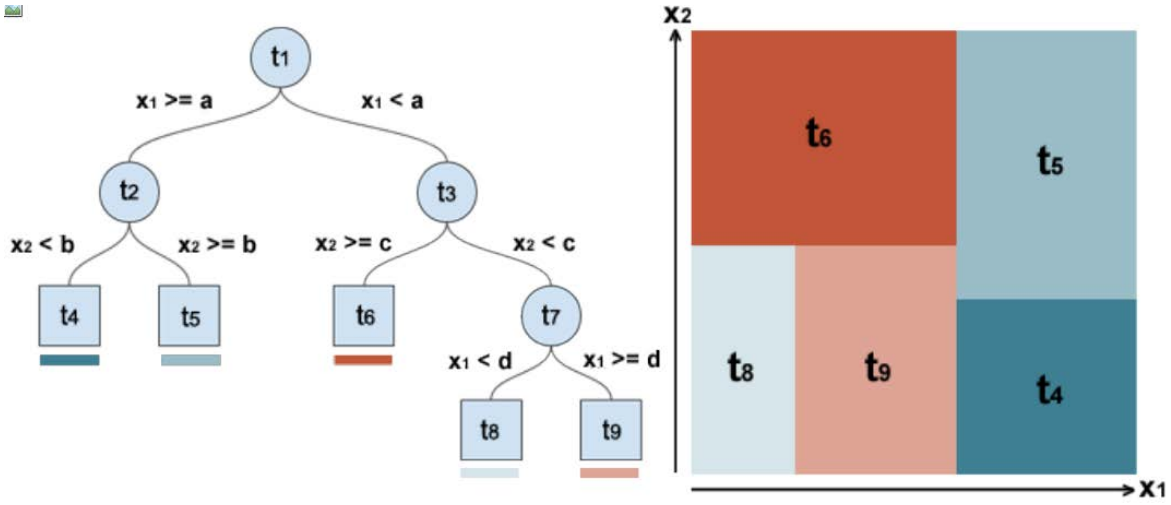



Рисунок 3.7 - Приклад класичного дерева лінійної регресії та представлення того, як розділено простір.

Кругові змінні — це числові змінні, значення яких обмежено циклічним простором. Наприклад, змінна, що вимірює кути в радіанах, має діапазон від 0 до 2↑, де обидва значення представляють одну й ту саму точку простору. Хоча ці змінні можуть бути включені в дерево лінійної регресії, їх потрібно розглядати як лінійні змінні, що є надмірним спрощенням і зазвичай призводить до

неоптимальних результатів.

Циркулярна змінна визначає круговий простір. Круговий простір є циклічним у тому сенсі, що він не обмежений; наприклад, поняття мінімального та максимального значення не застосовуються. Відстань між двома значеннями в просторі стає неоднозначним поняттям, оскільки її можна вимірювати за годинниковою стрілкою та проти неї, що дає різні результати. Крім того, цей простір не можна розділити на дві половини шляхом вибору значення, оскільки оператори «<» і «>» не застосовуються.

Для того, щоб розділити циклічну змінну, необхідно визначити принаймні два різних значення. Ці два значення описують два взаємодоповнюючі сектори, кожен з яких містить частину даних. Дерева циклічної регресії використовують цей підхід розщеплення для включення циклічних змінних у дерева регресії.

Є багато прикладів циклічних змінних. Будь-яка змінна, що представляє дані спрямованості або періодичну подію, є циклічною. Зокрема, у сфері прогнозування погоди в аеропорту прикладами циклічних змінних є напрямок вітру, час доби або дата.

Лунд пропонує методологію, яка дозволяє включати циклічні змінні в дерева регресії. Рисунок 3.8 містить представлення, подібне до попереднього прикладу, але з урахуванням однієї циклічної змінної ↲ та лінійної x1. Праворуч є діаграма, яка показує, як простір розділено за допомогою полярних координат.

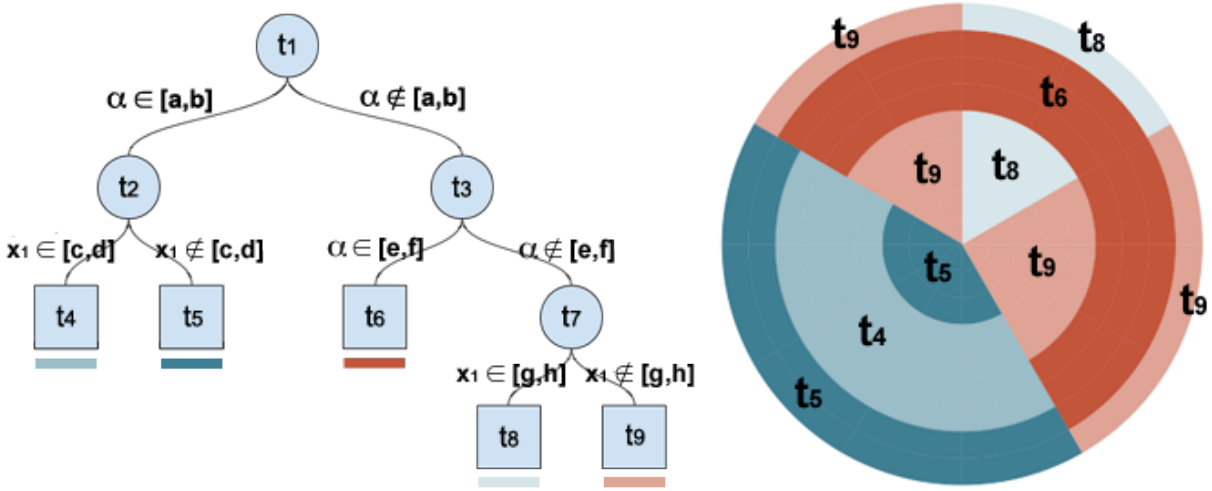



Рисунок 3.8 - Приклад оригінальної пропозиції Лунда про кругове регресійне дерево та представлення того, як поділяється простір.

Методологія, представлена в цій роботі, ґрунтується на концепції кругових дерев регресії, представляючи альтернативу, яка покращує продуктивність обчислень і точність їх результатів. На рисунку 3.9 показано, як відбувається поділ простору за запропонованою методикою.

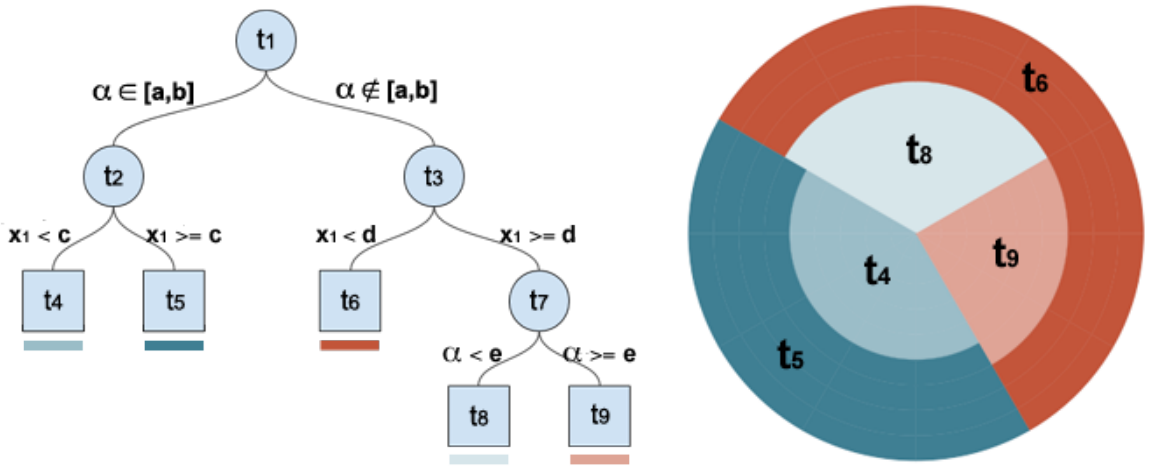



Рисунок 3.10 - Приклад запропонованого дерева кругової регресії та відображення того, як поділяється простір.

Візуально порівнюючи рисунок 3.9 і рисунок 3.10, видно, що регіони поділяються по-різному. Новизна цієї методології, порівняно з оригінальною версією, запропонованою Лундом, полягає в тому, що вона завжди генерує безперервні розбиття. Роблячи це, ми уникаємо надмірної фрагментації простору, а розбиття забезпечує краще узагальнення для його дочірніх вузлів. Оригінальна методологія використовує оператори «2» і «<» для генерації всіх розподілів для циклічних змінних. Це зазвичай генерує розділи, в яких підмножини, визначені реченням 2, оточені додатковим < підмножиною. Наша методологія використовує ці оператори для створення лише першого розбиття циклічної змінної, а потім використовує оператори «<» і «>», щоб створити наступні розбиття. Ця зміна також призводить до зменшення простору пошуку для можливих поділів. Запропонований алгоритм для генерації циклічних дерев має, як наслідок, O(n) вартість замість O(n2), порівняно з оригінальною пропозицією Лунда. Єдиним винятком є обчислення першого розбиття циклічної змінної, яке має обчислювальну вартість O(n2), оскільки воно має враховувати всі різні розбиття навколо кола.

***Програмне забезпечення та набори даних.*** AeroCirTree — це набір сценаріїв Python, який надає інструменти для навчання та тестування трьох раніше описаних методологій дерева регресії з використанням даних про погоду в аеропорту. Він використовує змінні NWP як вхідні дані та генерує більш точне значення для вибраної вихідної змінної, вивчаючи спостережувані значення для певного місця. Після навчання моделі її можна використовувати для підвищення точності прогнозованого вихідного значення, що забезпечується новими вхідними даними NWP.

Варто зазначити, що моделі дерев регресії представлені в цій роботі як метод статистичного зменшення результату NWP для конкретних місць. Вони використовуються не для прогнозування майбутніх значень часового ряду, а для покращення значень, отриманих NWP. Дані аналізу моделі NWP і дані спостережень використовуються для навчання дерев регресії. Ці дерева можуть пояснювати упередження та систематичні помилки моделі NWP. Навчені моделі можуть бути застосовані до будь-якого горизонту прогнозування, створеного NWP, для виправлення систематичних і випадкових помилок.

Програмне забезпечення AeroCirTree, представлене в цій роботі, пропонує загальну реалізацію дерева регресії. AeroCirTree дозволяє своїм користувачам тренувати дерева лінійної регресії, а також циклічні версії, використовуючи несуміжні або суміжні розбиття, як ми пропонуємо. Щоб визначити, яка методологія використовується, кожну змінну у вхідних або вихідних даних можна позначити як [лінійну, кругову] за допомогою файлу конфігурації. Додатковий тег, суміжний, який можна встановити на [true, false], вказує на методологію розділення, застосовану до циклічних змінних.Різні значення цих тегів вказують на різні версії дерев регресії. Наприклад, класичні дерева лінійної регресії можна створити, позначивши всі їхні вхідні змінні тегами linear і contiguous=true. Пропозиція Лунда щодо циклічного дерева вимагала б, щоб циклічні вхідні змінні були позначені тегами як circular і contiguous=false. Нарешті, запропонована нами методологія вимагатиме, щоб ті самі циклічні вхідні змінні були безперервними=істинними.

AeroCirTree використовує два набори погодних даних. Перший – це результат

глобальної NWP, яка називається моделлю Глобальної системи прогнозів (GFS) [11], якою оперативно керує Національне управління океанічних і атмосферних досліджень (NOAA). Другий використовує метеорологічні звіти з аеродрому (METAR) [6], які містять періодичні метеорологічні спостереження з аеропортів по всьому світу.

Кожен із цих наборів даних містить декілька змінних, що описують різні погодні параметри, такі як температура, вологість, швидкість вітру або хмарний покрив у різних місцях, які вони представляють. Модель GFS представляє дані за допомогою регулярної сітки, яка охоплює весь світ з просторовою роздільною здатністю приблизно 50 км і тимчасовою роздільною здатністю 3 години. NOAA підтримує архів операційної моделі та систему розповсюдження (NOMADS) для публікації даних GFS. Цей архів містить результати GFS за останні 10 років.

METAR – це текстові звіти про погоду, які кодують спостережувані метеорологічні параметри на злітно-посадкових смугах аеропорту за допомогою чітко визначеного коду.METAR виробляються з періодичністю щогодини або півгодини, а також є загальнодоступними через Глобальну телекомунікаційну систему ВМО (GTS). Національні центри прогнозування навколишнього середовища (NCEP) підтримують систему під назвою «Система збору метеорологічних даних» (MADIS), яка архівує всі звіти METAR, створені у світі за останні 10 років. Кожен звіт унікально ідентифікується своїм заголовком, який містить код аеропорту Міжнародної організації цивільної авіації (ІКАО) і часову позначку UTC.

Надане програмне забезпечення AeroCirTree містить утиліту командного рядка, яка витягує інформацію з цих двох наборів даних для будь-якого заданого аеропорту та діапазону дат. Вихідні дані представлені у вигляді зручного файлу CSV, що містить значення різних змінних у вигляді часових рядів. Усі операції, такі як визначення координат аеропорту в сітці GFS, розбір і вилучення METAR або уніфікація змінних одиниць, обробляються програмним забезпеченням, тому користувач може легко отримати чистий набір даних для потрібного аеропорту. Цей файл csv є вхідними даними для навчання нових моделей.

***Експерименти та результати.*** Гіпотеза цього дослідження полягає в тому, що запропонована нами методологія для створення регресійних дерев забезпечує краще узагальнення та точність, ніж попередні несуміжні кругові регресійні дерева при використанні циклічних змінних та еквівалентних класичних лінійних методологій.

У наступних розділах описані кроки, необхідні для вилучення необхідних даних, навчання моделей і створення прогнозів. Останній розділ містить аналіз запропонованої точності моделі та порівняння з результатами, наданими вихідними результатами GFS, методологією Лунда та класичними деревами лінійної регресії.

### *3.2.1. Вилучення даних і навчання моделі*

Щоб порівняти відмінності в продуктивності між методологіями, ми використовуємо дані про погоду, отримані з моделювання NWP, і дані спостережень з різних аеропортів. Дерева регресії навчаються з використанням NWP як вхідних даних і спостережуваної швидкості вітру як вихідної змінної. Варто зазначити, що моделі дерева регресії не використовуються для прогнозування швидкості вітру в майбутньому. Ці моделі використовуються для статистичного зменшення масштабу даних NWP, виправлення зміщень і систематичних помилок.

Ми вирішили спрогнозувати спостережувану швидкість вітру в 5 різних місцях Європи. Дані з аеропортів Берлін Тегель (EDDT), Лондон Хітроу (EGLL), Барселона Ель Прат (LEBL), Париж Шарль де Голль (LFPG) і Мілан Мальпенса (LIMC) використовуються для навчання різних моделей і аналізу результатів. Моделі тренуються з використанням тригодинних даних за 3 роки, що забезпечує приблизно 8760 зразків на аеропорт. Кожна модель генерує необхідні розділи для прогнозування спостережуваної швидкості вітру, використовуючи такі параметри GFS як вхідні змінні: відносна вологість, швидкість і напрямок 10-метрового вітру, а також час доби, пов’язаний із значеннями. Швидкість вітру є однією з найважливіших змінних погоди, що впливає на роботу аеропорту. Ця змінна також сильно залежить від іншої змінної, напрямку вітру, який є круговим. Причина включення цих двох змінних у наші експерименти полягає в тому, що разом вони

можуть представляти місцеві ефекти топографії, які не вирішуються погодними моделями. Відносна вологість поверхні використовується як індикатор таких явищ, як дощ або туман. Нарешті, час доби, також кругова змінна, сильно корелює з щоденними моделями вітру.

Критерій зупинки для всіх розглянутих дерев базується на кількості елементів у вузлі. Розбиття виконується рекурсивно, доки кількість записів даних у вузлі не впаде нижче певного значення. Потім процес розщеплення припиняється і вузол позначається як лист. Це значення отримує назву «максимальний розмір листа». Великі значення «максимального розміру листа» створюють неглибокі дерева, тоді як малі значення створюють глибокі дерева з більшою кількістю вузлів. Для кожного аеропорту генеруються різні версії моделі з використанням різних максимальних розмірів листя дерева. Максимальні значення розміру аркуша, які розглядаються в цьому експерименті, такі: 1000, 500, 250, 100 і 50.

```
   {"output":{"name":"metar_wind_spd",
"type":"linear"},"input":[
{"name":"gfs_wind_spd","type":"linear"},
{"name":"gfs_wind_dir","type":"circ"},
{"name":"gfs_rh","type":"linear"},
{"name":"time","type":"circ"}],
"contiguous":true
"max_leaf_size":100}
```

### *3.2.2. Експериментальний аналіз*

Для кожного аеропорту та значення максимального розміру аркуша генеруються три різні моделі: класичне дерево лінійної регресії (з використанням компонентів u, v швидкості вітру та часу доби), дерево Лунда та запропоноване нами дерево кругової регресії.

Щоб оцінити відмінності в точності між цими трьома методологіями, використовується процедура 5-кратної перехресної перевірки. Цей процес перевірки гарантує, що моделі тестуються з використанням даних, які не використовувалися під час навчання. Щоб уникнути відмінностей у результатах, викликаних різними розділеннями в процесі перевірки, той самий 5-кратний розподіл використовується

для перевірки всіх методологій для різних значень параметра «максимальний розмір листа». Помилка прогнозування визначається як різниця між швидкістю вітру, передбаченою деревом, яка є середнім цільових значень, що містяться у відповідному листі, та спостережуваним значенням швидкості вітру METAR. Уточнений індекс згоди (RIA) використовується для вимірювання відмінностей у точності між методологіями. Цей індекс забезпечує більший розподіл під час порівняння моделей, які працюють відносно добре, і є менш чутливим до помилок, зосереджених у викидах, порівняно з іншими методами, такими як абсолютна або середньоквадратична помилка. APB можна виразити як

$$RIA = 1 - \frac{\sum_{i=1}^{n} |P_i - O_i|}{2 \sum_{i=1}^{n} |O_i - \overline{O}|}$$

Де Oi представляє спостереження, а Pi – прогнози, створені моделлю. Таблиця 3.6 містить отримані значення RIA для кожної методології дерева, а також контрольне значення 10-метрової швидкості вітру, вироблене GFS в аеропортах, про які згадувалося раніше. Вищі значення RIA вказують на кращу точність результатів. Подібні результати з використанням різних комбінацій вхідних і вихідних змінних, що поєднують лінійні та кругові змінні, доступні у вигляді текстового файлу в основному сховищі коду.

Дивлячись на значення RIA, що містяться в таблиці 1, можна відзначити, що використання моделей дерева регресії значно покращує рівень точності вихідних даних моделі GFS. Рівень покращення сильно залежить від обраного аеропорту. Це може бути пов’язано з тим, що кожна точка сітки моделі GFS містить представлення погоди в області приблизно 50 квадратних кілометрів, і деякі місця та змінні краще представлені цим спрощенням, ніж інші. Наприклад, аеропорти, оточені горами, отримають більше переваг від статистичних моделей, ніж аеропорти, розташовані на великих рівнинах.

Порівнюючи різницю в точності між трьома моделями дерева регресії, показаними в таблиці 3.6, використання запропонованої моделі забезпечує кращі результати в більшості випадків. Рівень покращення також суттєво відрізняється в різних місцях розташування аеропорту. Результати аналізуються з огляду на дрібні

та глибокі дерева. Для неглибоких дерев дві кругові моделі показують дуже схожу поведінку, перевершуючи лінійний підхід. Оскільки максимальний параметр розміру стулки стає меншим, ми бачимо покращення точності для всіх трьох моделей. Глибші дерева все ще показують кращі результати для круглих моделей, але пропозиція Лунда починає демонструвати ознаки передчасного переобладнання порівняно з двома іншими моделями. У випадку найглибшого дерева (максимальний розмір листя дорівнює 50) усі три моделі демонструють погіршення продуктивності, при цьому Лунд є найбільш непомітним випадком.

У випадку Парижа Шарля де Голля (LFPG) невеликі круглі дерева демонструють покращення приблизно на 4–5% у порівнянні з класичною версією лінійного дерева. Це вдосконалення підтримується нашою запропонованою моделлю при розгляді глибших дерев. Однак модель Лунда не покращується з такою ж швидкістю. Більш систематичний аналіз результатів цього тесту пропонується в кінці розділу, надаючи статистичну значущість відмінностей між методологіями.

Таблиця 3.6 - Порівняння значень RIA під час прогнозування спостережуваної швидкості вітру METAR для різних аеропортів з використанням прямого результату GFS, класичного дерева лінійної регресії, кругового дерева Лунда та запропонованої моделі.

| *Airport* | *Method* | RIA per Max Leaf Size | | | | |
|---|---|---|---|---|---|---|
| | | 1000 | 500 | 250 | 100 | 50 |
| EDDT | GFS (ref.) | 0.669 | 0.669 | 0.669 | 0.669 | 0.669 |
| | Linear | 0.684 | 0.695 | 0.710 | 0.716 | 0.713 |
| | Lund | 0.700 | 0.713 | 0.720 | 0.715 | 0.702 |
| | AeroCirTree | 0.700 | 0.712 | 0.717 | 0.721 | 0.714 |
| EGLL | GFS (ref.) | 0.653 | 0.653 | 0.653 | 0.653 | 0.653 |
| | Linear | 0.687 | 0.703 | 0.716 | 0.728 | 0.730 |
| | Lund | 0.702 | 0.721 | 0.731 | 0.735 | 0.729 |
| | AeroCirTree | 0.702 | 0.720 | 0.730 | 0.737 | 0.737 |
| LEBL | GFS (ref.) | 0.362 | 0.362 | 0.362 | 0.362 | 0.362 |
| | Linear | 0.591 | 0.601 | 0.607 | 0.613 | 0.607 |
| | Lund | 0.602 | 0.608 | 0.615 | 0.606 | 0.590 |
| | AeroCirTree | 0.601 | 0.607 | 0.619 | 0.619 | 0.606 |
| LFPG | GFS (ref.) | 0.604 | 0.604 | 0.604 | 0.604 | 0.604 |
| | Linear | 0.674 | 0.691 | 0.702 | 0.711 | 0.707 |
| | Lund | 0.704 | 0.716 | 0.719 | 0.706 | 0.691 |
| | AeroCirTree | 0.704 | 0.712 | 0.715 | 0.714 | 0.707 |
| LIMC | GFS (ref.) | 0.401 | 0.401 | 0.401 | 0.401 | 0.401 |
| | Linear | 0.517 | 0.519 | 0.519 | 0.509 | 0.496 |
| | Lund | 0.521 | 0.520 | 0.518 | 0.500 | 0.482 |
| | AeroCirTree | 0.522 | 0.521 | 0.521 | 0.513 | 0.501 |

На рисунках 3.11 і 3.12 показано графічне представлення еволюції RIA при прогнозуванні швидкості вітру для аеропортів Лондон Хітроу (EGLL) і Барселона Ель Прат (LEBL) відповідно. Усі методології дерева регресії покращують свою точність, оскільки максимальний розмір аркуша зменшується, демонструючи ознаки надмірного підбору для випадку найменшого розміру аркуша. Значення швидкості вітру GFS у найближчій точці сітки показано як еталон для представлення відносного покращення, досягнутого кожною моделлю.

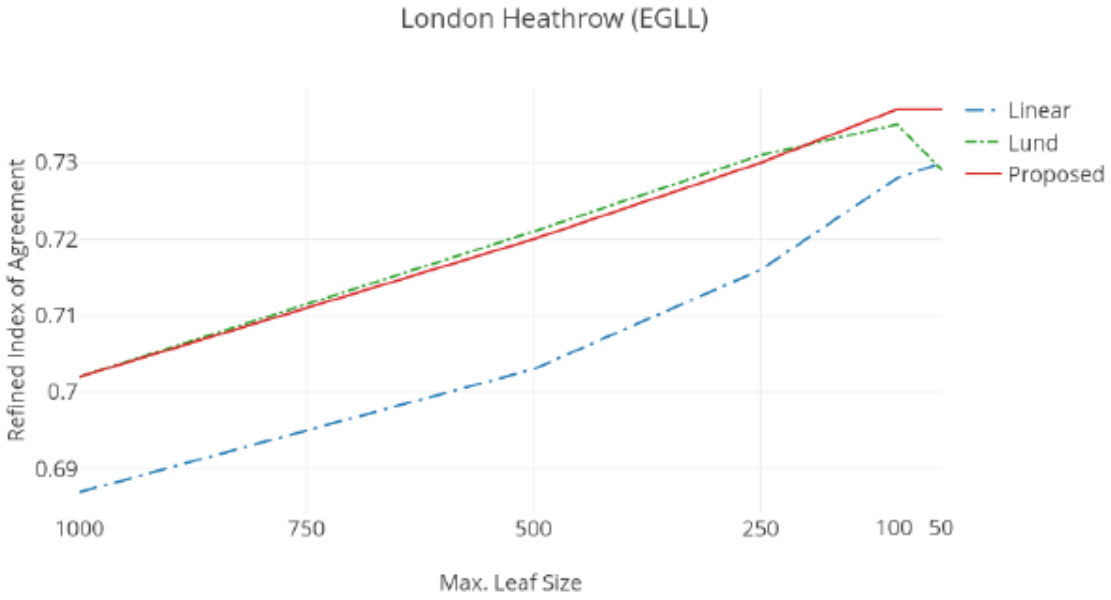



Рисунок 3.11 - Значення RIA для лондонського аеропорту Хітроу (EGLL), порівняння точності вихідних даних для різних максимальних розмірів стулок.

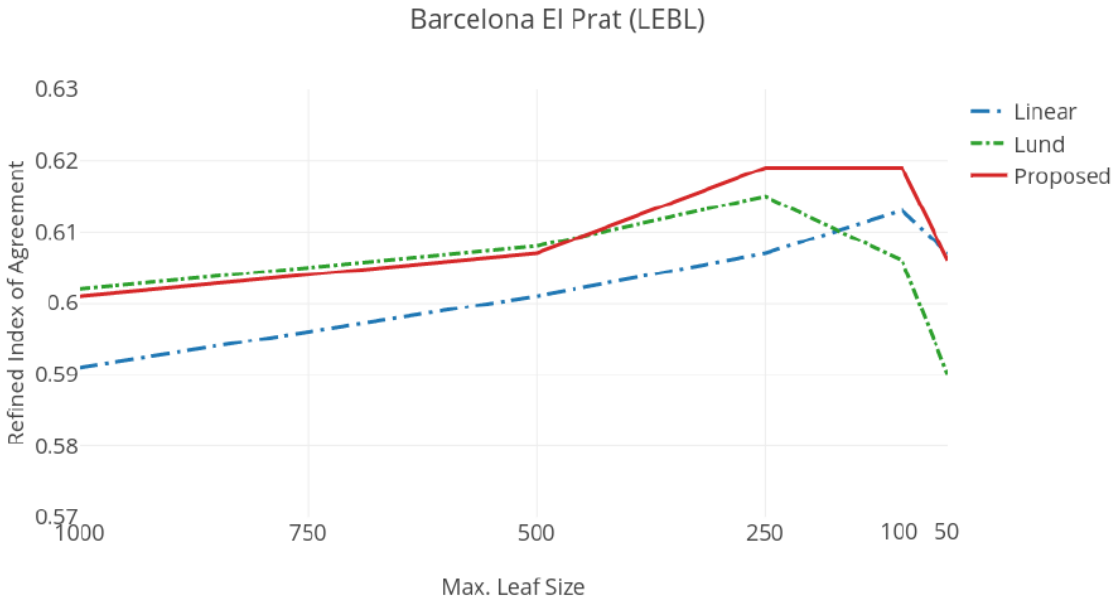



Рисунок 3.12 - Значення RIA для аеропорту Барселони Ель-Прат (LEBL), порівняння точності виходу для різних максимальних розмірів стулки.

Циклічні методології мають перевагу розгляду додаткових розділів для циклічних змінних, тих, які перетинають початок координат, у порівнянні з лінійними методами. Переваги використання круглих дерев більш помітні у випадку неглибоких дерев, тих із більшими значеннями максимального розміру листя. Перше розбиття циклічної змінної зазвичай відбувається в одному з перших вузлів дерева, поблизу кореневого вузла. Розбиття, яке відбувається у верхній частині дерева, сильно впливає на його продуктивність, оскільки вони ділять більшу пропорційну частину набору даних. Для неглибоких дерев критично важливо знайти хороший розділ на цих рівнях, тоді як глибші дерева можуть покращити погані розділи, створивши нові. Дерева несуміжної кругової регресії генерують розділи, які, здається, забезпечують гірше узагальнення для наступних поділів, ніж дві інші методології. Хороші результати, показані методом Лунда для неглибоких дерев, швидко погіршуються для більш глибоких дерев. Запропонована методологія, заснована на безперервних круглих деревах, досягає ефективності, подібної до методу Лунда для неглибоких дерев, а також кращих результатів, ніж дві інші методології для глибших. Крім того, як згадувалося в Розділі 2, запропонована методологія є більш ефективною з точки зору обчислень, ніж несуміжна версія.

Для оцінки результатів використовується методологія, запропонована Демсаром для оцінки статистичної значущості відмінностей між методами. Нульова гіпотеза подібності відхиляється для лінійної та обох дерев кругової регресії. Це виправдовує використання пост-хок двовимірних тестів, у нашому випадку Немені, які оцінюють статистичну різницю між парами алгоритмів. Результати цих тестів можна виразити графічно за допомогою діаграм критичної різниці (CD). Тест Немені попарно порівнює кожну методологію. Точність будь-яких двох методологій вважається суттєво різною, якщо відповідний середній ранг відрізняється принаймні на критичну різницю.

На рисунку 3.13 представлені результати RIA тесту Немені ($\alpha = 0,05$) з використанням діаграм CD для максимальних розмірів листків 1000, 100 і 50, оскільки вони представляють обидві крайні межі запропонованого діапазону.CD-

діаграми з’єднують групи алгоритмів, для яких не було виявлено значних відмінностей, або, іншими словами, ті, чия відстань менша за фіксовану критичну різницю, показану над графіком. Зверніть увагу, що алгоритми з нижчими значеннями на діаграмах CD означають вищі оцінки RIA. Ці тести були виконані за допомогою пакета scmamp R, який є загальнодоступним у Comprehensive R Archive Network (CRAN).

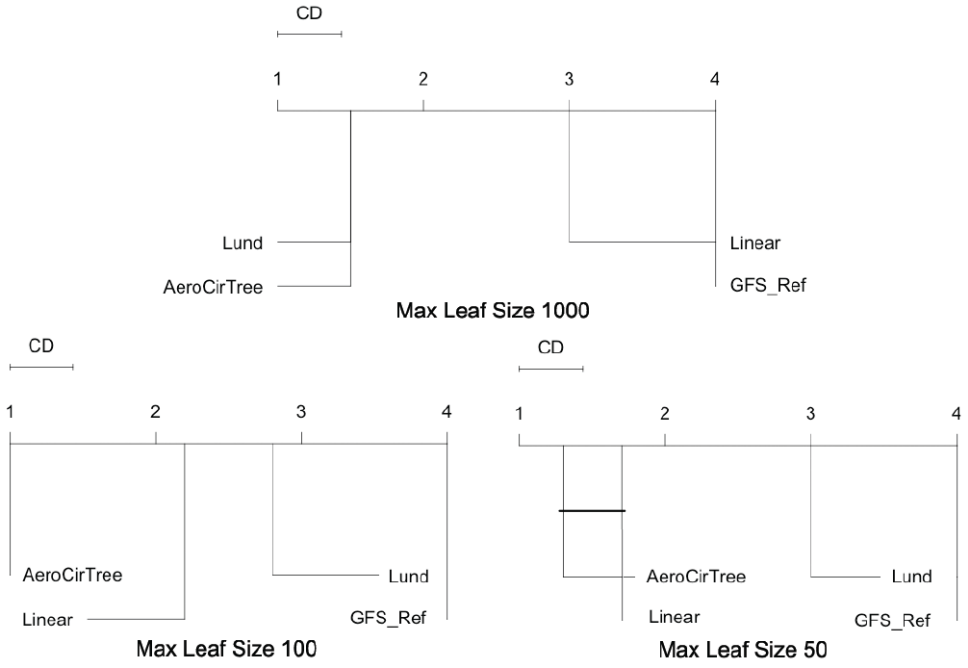



Рисунок 3.13 - Критичні відмінності порівняння трьох методологій для дрібних і глибоких дерев. α = 0,05

Як можна побачити на діаграмах CD на малюнку 3.13, для неглибоких дерев обидві циклічні методології перевершують лінійний підхід (максимальний розмір листя 1000). Коли експеримент просувається до глибших дерев (максимальний розмір листа 100), запропонована методологія статистично перевершує дві інші в розглянутих наборах даних. Навіть для випадку максимального розміру листа 50, коли всі методи демонструють погіршення точності, запропонована методика показує найкращі результати. Методологія Лунда, з іншого боку, виявляє значне погіршення точності для найменшого максимального розміру листа. Ці результати підтверджують нашу експериментальну гіпотезу: пропоноване дерево кругової регресії здатне генерувати моделі, які забезпечують кращі узагальнення для циклічних змінних.

***Розробка та використання програмного забезпечення.*** AeroCirTree — це пакет на Python 3, який реалізує дерева регресії та набір інструментів командного рядка для отримання даних про погоду та навчання моделей дерева для будь-якого аеропорту світу. Зазвичай користувачі використовують наданий пакет за допомогою трьох сценаріїв, які називаються aerocirtree extract, aerocirtree train і aerocirtree test, які отримують історичні часові ряди погодних даних, моделі поїздів і результати тестування відповідно для будь-якого аеропорту світу.

***Дизайн реалізації.*** Пропоноване дерево кругової регресії було реалізовано як пакет Python. Більшість його функціональних можливостей міститься в двох класах, які називаються Data та Node. Дерево моделюється як вкладена структура екземплярів Node. Кожен вузол у дереві містить екземпляр класу даних, який представляє підмножину набору даних, що міститься в цьому вузлі. Об’єкт Data побудований навколо класу Python Pandas DataFrame.

Node містить два атрибути класу типу Node, які називаються лівим дочірнім і правим дочірнім елементом, що визначає рекурсивну структуру. Кожен нетермінальний вузол у дереві містить два екземпляри Node, які складають його лівий і правий дочірні елементи. З іншого боку, кінцеві вузли або листи характеризуються тим, що для вмісту їхніх дочірніх елементів встановлено значення None.

Node також визначає метод .Split(), який створює розділення, генеруючи два нових екземпляри класу Node. Кожен із цих двох нових екземплярів Node містить одну частину вихідних даних і призначається лівому та правому дочірнім атрибутам. Дерево будується шляхом рекурсивного виклику методу .Split() для кожного з дочірніх вузлів, доки не буде виконано критерій зупинки.Критерії зупинки можуть бути налаштовані як мінімальна кількість елементів або значення дисперсії для вмісту даних вузла.

Кожен стовпець даних вузла має бути позначений як лінійний або циклічний, щоб визначити характер даних, які він представляє. Позначаючи стовпці тегами, ми можемо динамічно навчати різні версії дерева та порівнювати їхні результати. Класичні дерева регресії розглядають усі змінні як лінійні, тоді як запропонована

нами методологія дозволяє розглядати деякі змінні як циклічні. Наприклад, позначивши всі змінні як лінійні, ми отримаємо класичне дерево регресії.

Ця реалізація є загальною і може бути застосована до даних із будь-якого поля, якщо вона доступна у форматі csv.

***Керівництво користувача.*** AeroCirTree також надає серію сценаріїв для отримання даних про погоду, навчання та тестування моделей дерева регресії. Ці сценарії використовують описаний раніше пакет для навчання конкретних моделей для будь-якого аеропорту світу.

Ось приклад, який показує, як отримати дані для лондонського аеропорту Хітроу з 1 січня 2016 року до 1 червня 2016 року:

```
  $ ./aerocirtree_extract
--airport EGLL --start_date
20160101 --end_date 20160601
metar_press,metar_rh,metar_temp,
metar_wind_spd,gfs_press,gfs_rh,
gfs_temp,gfs_wind_dir,gfs_wind_spd,
time,date
1025.0,75.5,6.0,2.57,1016,92,3,280,3,
45.0,0
1024.0,80.92,5.0,4.12,1016,96,3,290,3,
90.0,0
1024.0,80.92,5.0,2.57,1015,97,4,300,3,
135.0,0
1024.0,86.99,6.0,2.57,1016,93,6,340,3,
180.0,0
...
```

Зауважте, що значення часу та дати перетворюються на числові значення як циклічні змінні, де початок [0-360] відповідає 00:00 та 1 січня відповідно. Вихідні дані цієї команди можна перенаправити до локального файлу. Ці файли використовуються як вхідні дані, необхідні для навчання моделей дерева.

Коли набір даних для даного аеропорту доступний, модель можна навчити, визначивши її вхідні та цільові змінні.

Вихідна змінна має бути однією зі спостережуваних змінних, отриманих зі звітів METAR, а вхідні змінні є прогнозованими змінними GFS або їхньою підмножиною.

Роблячи це таким чином, коли нові дані прогнозу з GFS доступні, модель може бути використана для створення розширеного прогнозу цільової змінної. Різні параметри для створення моделі вказуються у файлі конфігурації. Цей файл конфігурації містить об'єкт JSON із трьома полями: «вихід», «вхід» і «максимальний розмір аркуша». Ім'я цільової змінної, створене деревом, вказується у «виводі». Вхідні змінні перераховані в полі «введення» разом із тегом, що дозволяє розглядати їх як кругові чи лінійні. Параметр максимального розміру аркуша вказує значення для керування глибиною результуючого дерева. Наприклад, щоб визначити модель для прогнозування температури з використанням відносної вологості GFS, напрямку вітру як кругової змінної та максимального розміру листя 100, необхідно вказати файл із таким вмістом:

```
  {"output":{"name":"metar_temp",
"type":"linear"},"input":
[{"name":"gfs_wind_dir","type":
"circular"},{"name":"gfs_rh",
"type":"linear"}], "contiguous":true,
"max_leaf_size":100}
```

Для навчання моделі ми використовуємо потяг aerocirtree, який отримує як аргументи шляхи до файлу, що містить дані, і файл конфігурації.Припустимо, що вихідні дані, отримані в попередньому розділі, були збережені у файлі з іменем EGLL.csv, а представлений файл конфігурації збережено як модель A.json, модель можна навчити, виконавши:

```
 $ ./aerocirtree_train --data
EGLL.csv --config Model_A.json
```

Ця команда вивчає вказану модель і зберігає її, використовуючи назву, яка поєднує в собі імена вхідних файлів і використання розширення .mod. Попередню модель буде збережено на диску під назвою файлу EGLL Model A.mod.

Нарешті, тест aerocirtree можна використовувати для запуску моделі на нових даних. Цей сценарій отримує шлях до збереженого файлу моделі та вхідний файл csv як аргументи. Сценарій повертає результуючі результати моделі для кожного рядка вхідного файлу.

Наприклад, припустимо, що ми хочемо перевірити нашу попередньо навчену

модель EGLL Model A.mod з новими даними, що містяться у файлі EGLL.csv, ми можемо запустити:

```
$ ./aerocirtree_test --data
EGLL.csv --model EGLL_Model_A.mod
```

Ця команда обчислює кінцеві значення температури для кожного з введених значень в аеропорту Лондон-дон Хітроу.

***Висновки.*** У цій роботі представлено програмне забезпечення для прогнозування погоди в будь-якому аеропорту світу. Він також пропонує нову методологію дерева кругової регресії, яка пропонує кращу точність у порівнянні з класичними лінійними методами, а також кращу точність і обчислювальну ефективність, ніж оригінальна пропозиція Лунда щодо дерев кругової регресії.

Це програмне забезпечення містить бібліотеку, яка реалізує загальну версію дерев регресії, а також інструменти командного рядка для навчання, тестування та завантаження нових наборів даних аеропорту. Ці інструменти розроблено для того, щоб користувачі могли створювати власні прогнози, а також щоб вони могли експериментувати та досліджувати відмінності між моделями, вхідними змінними та аеропортами. Сценарії та бібліотеки написані простим способом, щоб користувачі могли читати код, щоб зрозуміти, що робить програма, а також змінювати її частини. AeroCirTree постачається з ліцензією GNU GPLv3, тому будь-хто може використовувати, змінювати та ділитися цією програмою з будь-якою метою.

Модель, запропонована в цій роботі, базується на новій методології побудови базового дерева кругової регресії. Дерева регресії розвинулися завдяки впровадженню багатьох різних методів, які підвищують їхню точність і ефективність. Добре відомі методи модифікації стандартних дерев регресії, такі як обрізка, балансування, згладжування або випадкові ліси і ансамблі, також можуть бути застосовані до дерев кругової регресії та можуть підвищити точність результатів. у порівнянні з базовими деревами регресії. Майбутня робота могла б реалізувати ідеї, представлені у згаданих публікаціях, пропонуючи більш просунуті моделі.

### 3.3. Автоматизація прогнозів погоди на основі згорткових мереж

Прогнозування погоди базується на числових прогнозах погоди (NWP), які фіксують стан атмосфери та моделюють її еволюцію на основі фізичних і хімічних моделей. Глобальні моделі NWP зазвичай забезпечують велику кількість параметрів, що представляють різні фізичні змінні в просторі та часі. Через відсутність просторової та часової роздільної здатності ці поля потребують інтерпретації висококваліфікованим персоналом для створення прогнозів для будь-якого конкретного регіону. Це все ще сьогодні процес, заснований на людині, який покладається на спеціально підготовлених і досвідчених професіоналів для інтерпретації змодельованих і спостережених даних. Змінні NWP визначають стан атмосфери та її зміни в просторі та часі. NWP визначають високоструктурований набір даних, у якому зв'язки між його змінними визначаються фізичними рівняннями, такими як збереження маси, імпульсу та енергії.

Останні досягнення в області нейронних мереж довели, що за рахунок збільшення кількості загальних прихованих шарів можна досягти безпрецедентних результатів у багатьох різних сферах.

Зокрема, дослідження навколо згорткових нейронних мереж (CNN) довели свою ефективність у вирішенні проблем класифікації та сегментації зображень.

Машинне навчання було застосовано до різних областей прогнозування погоди, таких як зменшення масштабу або поточна прогнозування. Основним недоліком традиційних методологій є їх нездатність включати як просторові, так і часові компоненти, присутні в даних. Більшість існуючих досліджень у цій галузі ґрунтуються на ручному вилученні точок у моделі, що представляють певне місце, та навчанні моделей за допомогою отриманих даних. Проблема цього підходу полягає в тому, що погода є динамічною системою, і аналіз окремих точок окремо втрачає важливу інформацію, що міститься в синоптичному та мезомасштабі.

CNN дозволяють аналізувати та витягувати просторову інформацію в зображеннях, перетворюючи дрібні деталі на структури вищого рівня. Тимчасовий вимір можна додати до цих мереж шляхом додавання третьої осі до згорткових

ядер. Ця робота демонструє, як CNN можна використовувати для автоматичної інтерпретації результатів числового прогнозу погоди (NWP) для створення місцевих прогнозів. Основними результатами цієї роботи є:

CNN можуть надати модель для безпосередньої інтерпретації полів числової моделі погоди та створення місцевих прогнозів погоди.

Class Activation Mapping (CAM) надає цінний механізм для візуальної оцінки просторових і часових кореляцій різних полів, допомагаючи аналізувати та розробляти нові моделі.

Тривимірні згортки можуть природним чином включати часовий компонент у нейронні мережі, значно підвищуючи точність результатів.

***Набори даних.*** Для цієї роботи ми пропонуємо використовувати NWP і дані спостережень за опадами з різних місць, щоб експериментувати з різними конфігураціями моделей CNN. Мета полягає в тому, щоб навчити модель, яка передбачає дощ в конкретному місці, використовуючи як вхідні дані числової моделі погоди. У цьому розділі ми описуємо, як були згенеровані ці набори даних.

ERA-Interim — це загальнодоступний набір даних для повторного аналізу метеорологічних даних від Європейського центру середньострокових прогнозів погоди (ECMWF). Цей набір даних створено за допомогою числової моделі погоди, яка імітує стан атмосфери всієї планети з просторовою роздільною здатністю приблизно 80 км. Дані доступні з 1979 року з роздільною здатністю 3 години. Вихідні дані представлені у формі регулярних числових сіток і є велика кількість доступних фізичних параметрів, що представляють такі змінні, як температура, швидкість вітру та відносна завихреність.

Звичайні авіаційні метеорологічні звіти (METAR) — це оперативні текстові авіаційні метеорологічні звіти, які кодують спостережувані метеорологічні змінні для кожного комерційного аеропорту світу. METAR створюються з періодичністю щогодини або півгодини, а також оприлюднюються через систему зв’язку Всесвітньої метеорологічної організації (ВМО). Кожен звіт унікально ідентифікується своїм заголовком, який містить код аеропорту Міжнародної організації цивільної авіації (ІКАО) і мітку часу UTC.

Для проведення наших експериментів ми розглянули 5 основних аеропортів, розташованих у різних містах Європи, і період у 5 років (2012-2017). Аеропорти та їхні відповідні коди ICAO: Гельсінкі-Ванта (EFHK), Амстердам-Схіпхол (EHAM), Дублін (EIDW), Рим-Фьюмічіно (LIRF) і Відень (LOWW).Ми витягли розширену територію над Європою з ERA-Interim, створивши 3-годинну серію зображень, складених 3 смугами, що відповідають геопотенційній висоті z при рівнях атмосферного тиску 1000, 700 і 500. Цей параметр являє собою висоту в атмосфері, на якій досягається певне значення тиску, і рівні зазвичай відповідають 100, 3000 і 5500 метрів над середнім рівнем моря відповідно.

Причина вибору цих полів полягає в тому, що синоптики зазвичай ґрунтують свої прогнози на них. Вони містять інформацію про форму, розташування та еволюцію систем тиску в атмосфері.

Використовуючи дані METAR, умови опадів [дощ, суха] були виділені для кожного аеропорту за той самий період часу та частоту. Отриманий часовий ряд набору даних містить понад 12000 зразків. На рисунку 3.14 представлено зразок розглянутих полів ERA-Interim із їхнім розміром і географічним розмахом ліворуч. Права сторона показує, як дані ERA-Interim узгоджуються зі спостережуваними опадами для місця вибірки.

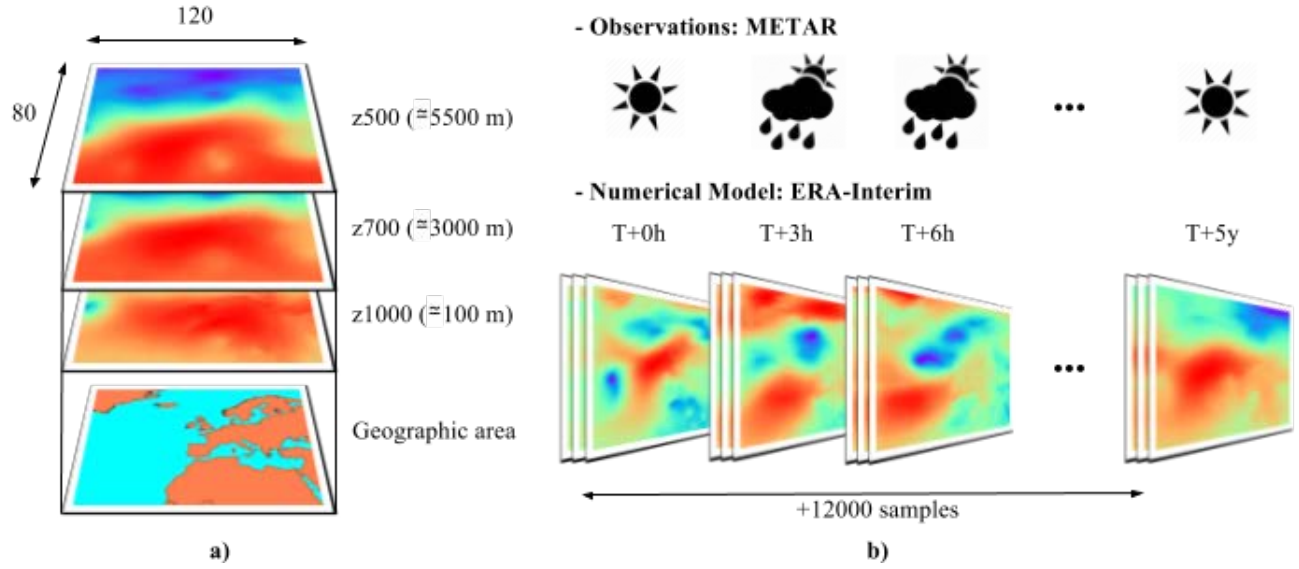



Рис. 3.14 a) представляє 3 геопотенціальні підмножини, отримані з ERA-Interim, що відповідають різним висотам атмосфери, складені на карті для представлення просторового протяжності. b) Представляє весь витягнутий часовий ряд і вирівнювання обох наборів даних.

***Експерименти та результати.*** Метою запропонованого експерименту є прогнозування подій перед опадами для розглянутих аеропортів, використовуючи геопотенціальні дані ERA-Interim як вхідні дані та спостереження METAR для анотації зразків [дощ, сухо]. Використовуються дві різні CNN. Перша модель виконує 2D згортки, а друга включає часовий вимір на основі 3D згорток. Ми прагнемо довести, що ці моделі можуть охопити частину розумового та інтуїтивного процесу, якому дотримуються прогнозисти під час інтерпретації числових даних про погоду.

***Архітектура CNN.*** Для проведення експериментів використовується 2-шаровий CNN. Кожен згортковий рівень використовує ядро 3x3, за яким слідує максимальний рівень об'єднання 2x2. Після операцій згортання повністю підключений рівень використовується для підключення виходу [дощ, посушливий] за допомогою функції активації «softmax».

Таблиця 3.7. Точність прогнозування дощу для різних місць, порівнюючи 2D і 3D CNN з еталонною точністю кліматології.

| AIRPORT | RAIN CLIM. | 2D CNN | 3D CNN |
|---|---|---|---|
| EFHK | 60.8 | 73.6 | 75.4 |
| EHAM | 74.2 | 77.8 | 79.3 |
| EIDW | 61.2 | 70.7 | 72.6 |
| LIRF | 83.1 | 87.3 | 88.2 |
| LOWW | 75.7 | 77.1 | 78.8 |

Для 3D CNN конфігурація подібна до попередньої версії, але ядра в обох рівнях згортки та максимального об'єднання мають додатковий вимір із розмірами 3x3x3 та 2x2x2 відповідно. 3D CNN навчається шляхом об'єднання вхідного набору даних у групи з 8 послідовних зображень. Ця сукупність представляє еволюцію атмосфери протягом 24 годин. Потім нейронна мережа може витягти інформацію з часового виміру, використовуючи спостереження, що відповідає останньому зображенню серії, як вихід.

2D і 3D CNN були реалізовані в TensorFlow і тренувалися для кожного аеропорту протягом 300 епох, використовуючи 80% даних. Решта 20% використовувалися як підтвердження для перевірки точності моделей.

***Результати.*** Таблиця 3.7 містить результати, отримані різними моделями з

використанням набору даних перевірки. Значення точності представляють відсоток успіху моделі під час прогнозування дощу чи сухості для кожного місця. Кліматологія для кожного місця, кількість спостережень за дощем над загальною кількістю спостережень, також включено до результатів як еталон. Модель, результат якої завжди «сухий», матиме такий показник успіху.

На рисунку 3.15 представлено результати за допомогою гістограми з накопиченням. Відносне покращення в порівнянні з кліматологією, досягнуте за допомогою 2D і 3D згорткових моделей, представлено зеленою та червоною частинами смужки.

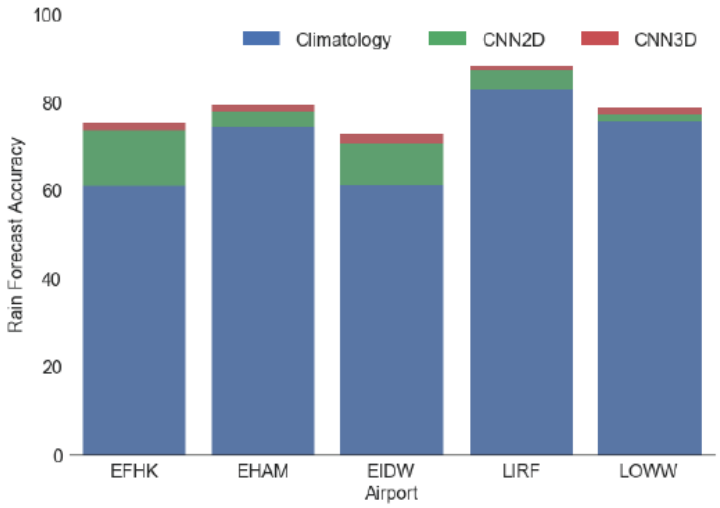



Рис. 3.15 - Результати точності, отримані для різних аеропортів і методології.

***Відображення активації класу.*** Class Activation Mapping (CAM) — це техніка, яка локалізує специфічні для класу області зображення в навченій CNN.

Ця техніка використовує останній рівень CNN для створення графічного представлення для певного результату на основі його ваг. Отримане зображення — це теплова карта, яка показує, які частини зображення мають більший вплив на результат.

Наприклад, на рисунку 3.16 зображено два різних представлення CAM для 2D моделей CNN, навчених з використанням даних про кількість опадів у Гельсінкі та Римі. Тепліші кольори на зображенні представляють вищі значення ваги, тому мережа приймає рішення переважно на основі об’єктів, розташованих у цих областях. Зображення на рисунку 3.16 підтверджують інтуїтивну ідею про те, що

місцеві погодні умови мають більший вплив, ніж віддалені, на прогнозування погоди в певному місці. Зображення накладено на карту узбережжя, щоб служити орієнтиром відносного розташування структур на теплових картах.

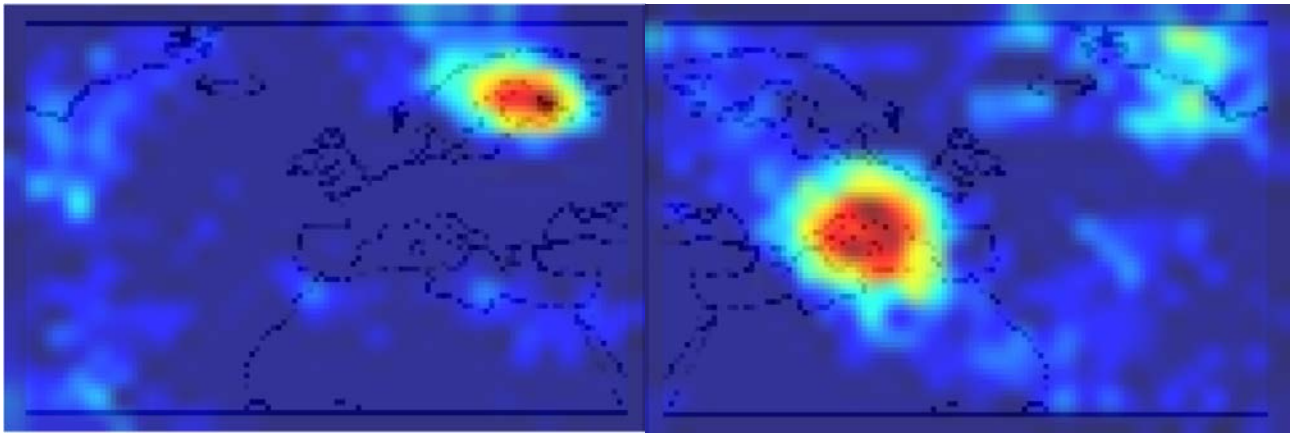


Рис. 3.16 - Приклад отриманих карт активації класу для ERA-Interim CNN, навченого з використанням спостережених опадів в аеропорту Гельсінкі-Вантаа, EFHK, (з ліва) та римському аеропорту Фьюмічіно, LIRF, (з права). Берегові лінії були накладені як орієнтир для читачів.

Ця техніка виявилася дуже корисною для візуальної оцінки правильності моделі CNN. Інше використання може бути для вибору вхідної змінної, ідентифікації параметрів NWP, які демонструють більш високу кореляцію щодо класу, який потрібно передбачити.

Ця робота демонструє, як CNN можна безпосередньо застосувати до вихідних даних числових моделей погоди, використовуючи дані спостережень для анотації зразків. Конструкція CNN, що використовується в наших експериментах, дуже проста порівняно з деякими найсучаснішими архітектурами. Незважаючи на їхню простоту, результати показують, що згорткові шари можна використовувати для інтерпретації результатів погодних моделей.

Параметри NWP, що використовуються в експериментах, прямо не корелюють зі змінною виходу опадів. NWP має багато інших змінних, таких як вологість, завихреність або навіть загальна кількість опадів, які можна використовувати для прогнозування моделей опадів з кращою точністю. Мета цього початкового експерименту полягала в тому, щоб продемонструвати, що CNN можуть вивчати певні конфігурації систем атмосферного тиску та пов’язувати їх із подіями випадання опадів (фронти, конвекція тощо).

## 3.4. Керований даними підхід до параметризації опадів за допомогою згорткових нейронних мереж кодера-декодера

Чисельне прогнозування погоди (NWP) в даний час лежить в основі більшості операцій з прогнозування погоди по всьому світу. Постійне зростання навичок NWP протягом останніх 40 років пов'язане зі зростанням обчислювальних потужностей, збільшенням доступності даних спостережень і розвитком кращих методів засвоєння даних. Однак, незважаючи на технологічний прогрес, деяким важливим фізичним процесам, таким як конвекція, тертя, турбулентність або радіація, вдається уникнути адекватної репрезентації при підсіткові масштаби для більшості систем NWP. NWP використовує параметри для моделювання цих фізичних процесів, які покладаються на наявність явно визначених параметрів. Параметризація часто ґрунтується на емпіричних припущеннях, які використовуються для представлення процесів, що відбуваються в обраних масштабах і є основним джерелом невизначеності в NWP.

Більшість параметризацій базується на рівняннях детермінованої моделі, що представляють спрощення фізичних процесів. Є також деякі, які базуються на статистичних або імовірнісних підходах. Ці параметризовані процеси можуть взаємодіяти один з одним, що призводить до складної поведінки моделі. Часто невелика модифікація в одному з його компонентів може призвести до неузгодженості з іншими параметрами та, зрештою, призвести до нестабільності в оцінці NWP. Таким чином, процес проектування та підтримки окремих параметризацій та зв’язків між ними є трудомістким і вимагає високого рівня знань у галузі.

Легкий доступ до великих обсягів даних дозволяє використовувати керовані даними методи для отримання параметризації. Було запропоновано машинне навчання для отримання параметризації за допомогою таких методів, як дерева регресії або нейронні мережі. У метеорології питання про те, чи моделі глибоких нейронних мереж, наприклад, навчені на атмосферних даних,

можуть конкурувати з моделями NWP на основі фізики, нещодавно було розглянуто, продемонструвавши багатообіцяючі результати. Подібні мережі глибокого навчання також досліджувалися в контексті параметризації фізичних процесів підсітки в NWP. Наприклад, глибокі нейронні мережі використовувалися для виконання прогностичного моделювання або для вивчення конвективних і радіаційних процесів за моделями розділення хмар з використанням одностовпцевих моделей.

***Згорткові нейронні мережі в контексті прогнозу погоди.*** Результати NWP виражаються у вигляді двовимірних (2d) числових масивів чисел для будь-якого заданого часу на окремих рівнях атмосфери. Зазвичай ці двовимірні поля погоди представлені у вигляді цифрових зображень, що містять розміри широти та довготи. У цьому розділі ми представляємо згорточні нейронні мережі (CNN), популярну методологію аналізу зображень у сфері комп'ютерного зору. Ми описуємо їх застосування до NWP і проблем параметризації, посилаючись на 2d масив числових даних як зображення.

***Згортка зображення.*** Згортка в контексті обробки зображень — це математична операція, яка виконується над околицею пікселів зображення, зважених двовимірною матрицею, яка називається ядром. Результатом операції згортки є інше трансформоване зображення з тими самими розмірами, що й вихідне зображення (без урахування ефектів меж). Виконання згортки зображення вимагає повторного застосування тієї самої операції шляхом переміщення ядра по всьому зображенню. Результат кожної операції згортки призначається пікселю в новому зображенні в позиції, визначеній центром ядра. Наступне рівняння показує операцію декомпозиції згортки між різними точками сітки в області поля погоди F, ядро згортки K і те, як результати призначаються вихідному зображенню C:

$$c_{ij} = \begin{bmatrix} f_{i-1,j-1} & f_{i-1,j} & f_{i-1,j+1} \\ f_{i,j-1} & f_{i,j} & f_{i,j+1} \\ f_{i+1,j-1} & f_{i+1,j} & f_{i+1,j+1} \end{bmatrix} \circledast \begin{bmatrix} k_{11} & k_{12} & k_{13} \\ k_{21} & k_{22} & k_{23} \\ k_{31} & k_{32} & k_{33} \end{bmatrix}$$

На рисунку 3.17 показано, як застосовується операція згортки з використанням ядра 3 × 3 над областю поля геопотенціалу NWP 850 гПа з координатною сіткою. Використовуються два ядра: ядро згладжування (верхнє) і ядро визначення країв (нижнє). На малюнку показано вплив двох різних ядер згортки на поле геопотенціалу висоти 850 гПа у вигляді зображень у градаціях сірого. Ядро Right-Sobel, наприклад, виявляє краї з орієнтацією вправо або, іншими словами, робить помітними області зображення, які містять градієнти, що зменшуються справа наліво. Ми спостерігаємо, що це ядро визначило особливості в геопотенціалі, які нагадують ті, які навчений синоптик міг би визначити як фронтальні системи. Цей автоматизований метод ідентифікації ключових характеристик забезпечує цілеспрямований підхід до ідентифікації погодних явищ на величезних територіях і великих обсягах числових даних.

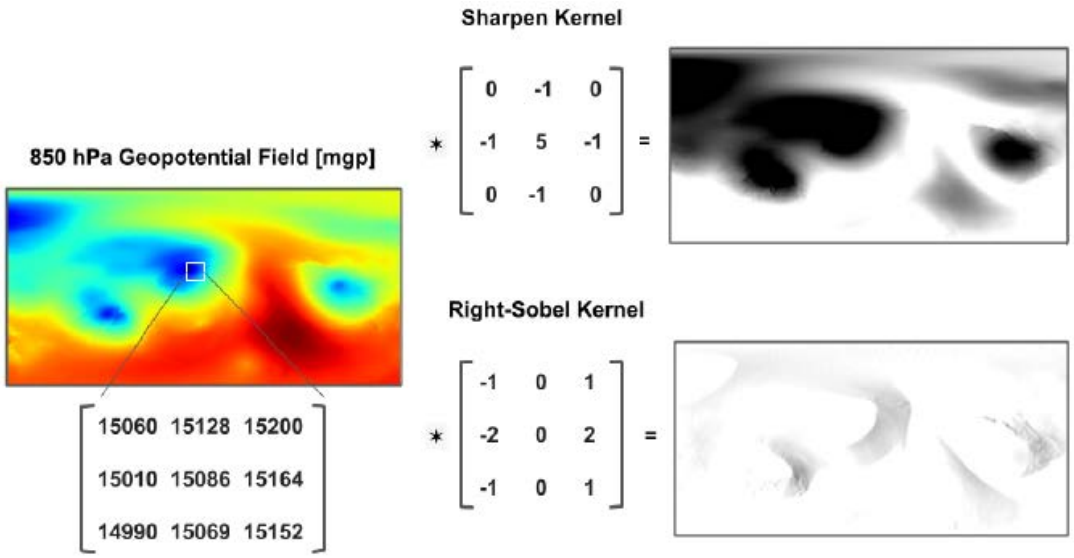



Рис. 3.17. Ілюстрація операції згортання, застосованої до зображення географічних висот 850 гПа. Використовуються два ядра: ядро різкості та ядро визначення країв. Підсумковий результат операції відображається як зображення у градаціях сірого.

***Згорткові нейронні мережі.*** Згорткові нейронні мережі (CNN) містять серію шарів, які виконують операції згортання, витягуючи просторову інформацію вхідних зображень. Замість використання попередньо визначених ядер, CNN використовує метод градієнтного спуску для пошуку оптимальних значень які зменшують функцію втрат. Ця функція втрат вимірює точність

завдання під час навчання. Під час процесу навчання мережа вивчає ваги, які перетворюють набір вхідних зображень у близьке представлення відповідних виходів. Згортки на кожному рівні CNN зазвичай супроводжуються операцією, яка спричиняє зменшення розміру зображення, і функцією активації для введення нелінійності в модель. Зменшення розмірності зазвичай досягається шляхом застосування спеціального методу підвибірки, що називається об'єднанням, або виконанням операції згортання покроково по зображенню. Функція згортання, об'єднання та нелінійної активації, що виконується на кожному рівні CNN, призводить до поступового зменшення розміру вхідного зображення, що означає, що ядра можуть охоплювати дедалі більші області відносно вихідного зображення. Ця характеристика дозволяє CNN ідентифікувати елементи в різних масштабах зображення. На рис. 3.18 показано зменшення розміру поля висот геопотенціалу NWP за допомогою 4-рівневої CNN. Кожна операція об'єднання в цю мережу зменшує розмір зображення вдвічі. У міру того, як розміри зображення зменшуються, кожна точка сітки представляє більшу область, і ядра можуть витягувати мезомасштабні та синоптичні масштабні особливості початкового зображення.

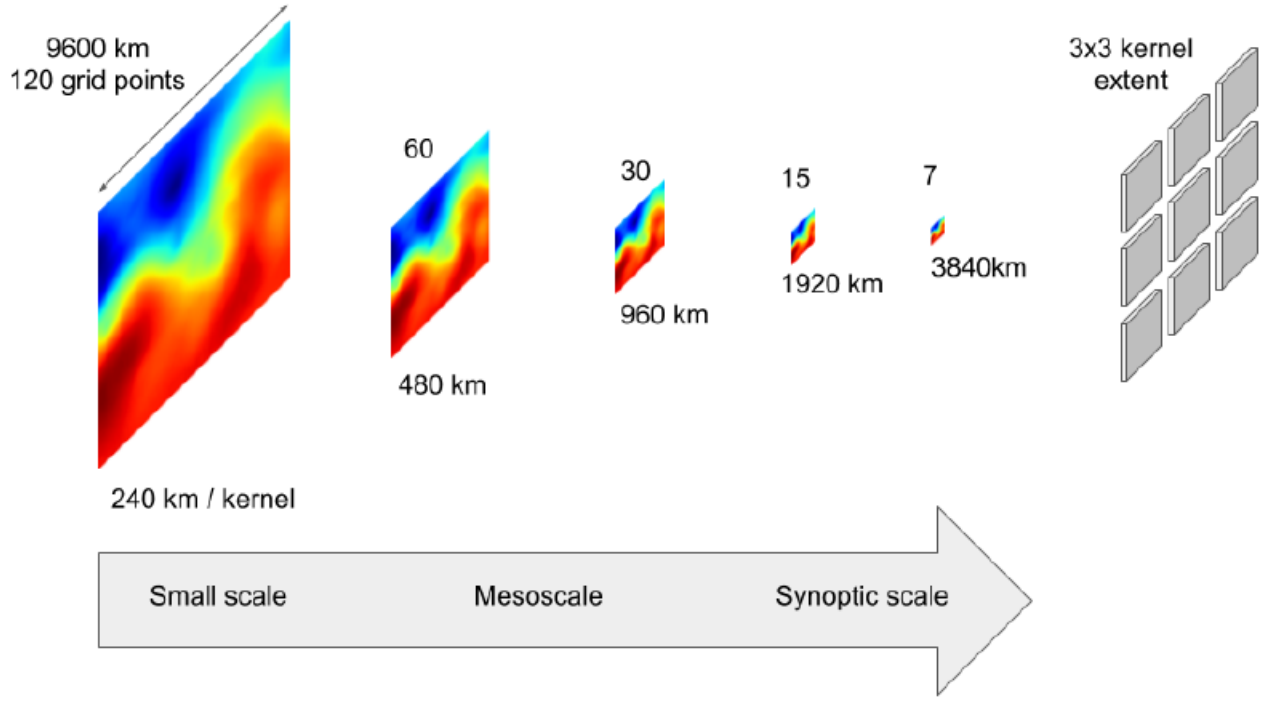



Рис. 3.18 - Графічне представлення конвеєра експерименту, що включає процеси вибору змінної та порівняння мережі кодера-декодера.

Кожен шар у CNN перетворює розмірність вхідного зображення, зменшуючи його просторову протяжність і розширюючи кількість функцій уздовж осі третього виміру. Це розширення глибини зображень досягається за допомогою 3-вимірних ядер шляхом укладання 2-вимірних ядер, кожне з яких виявляє конкретні шаблони або особливості вхідного зображення.

CNN є надзвичайно популярним і ефективним методом вирішення завдань класифікації та регресії. Хоча ці мережі спочатку були розроблені в області комп'ютерного зору, вони знайшли застосування в інших галузях, таких як аналіз медичних зображень, фізика високих енергій або астрономія.

***Набір даних та методологія.*** Тут ми використовуємо глобальний набір даних повторного аналізу клімату NWP ERA-Interim. ERA-Interim містить дані повторного аналізу з 1979 року до теперішнього часу з 6-годинною тимчасовою роздільністю. Просторова роздільна здатність набору даних становить приблизно 80 км (скорочена сітка Гауса N128) на 60 вертикальних рівнях від поверхні до рівня тиску 0,1 гПа. Дані ERA-Interim є загальнодоступними через веб-інтерфейс публічних наборів даних ECMWF. З великого обсягу вихідних змінних ми вибираємо геопотенційну висоту (Z) і загальну кількість опадів (P). Крім того, ми зосереджуємося на регіоні середніх широт, що визначає прямокутну територію над Європою, обмежену (широта: [75, 15], довгота = [-50, 40]). На рисунку 3.19 зображено географічну зону, а також відповідність між геопотенціальною висотою та часовим рядом загального поля опадів. Зауважте, що ми вибираємо підмножину доступних геопотенційних висот (Z), які відповідають таким рівням тиску атмосфери: {1000, 900, 800, 700, 600, 500, 400, 300, 200, 100} гПа. Отримані геопотенційні дані про висоту зберігаються як 4-вимірний числовий масив із формою [54.023, 10, 80, 120] для відповідних розмірів [час, висота, широта, довгота]. ERA-Interim (tp) представляє накопичену суму за 3-годинний період для кожної точки сітки. Це поле додатково агрегується, щоб відповідати 6-годинній частоті поля геопотенційної висоти. Результатом є тривимірний числовий масив із формою [54.023, 80, 120] для відповідних розмірів [час, широта, довгота].

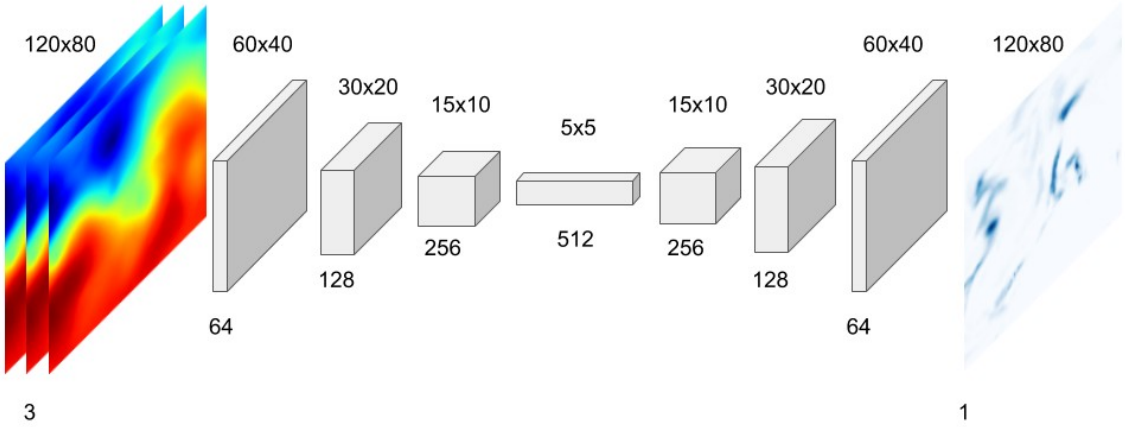



Рис. 3.19 - Перетворення розмірності даних, що виконуються кодером-декодером VGG-16 для відображення між вхідним і вихідним просторами.

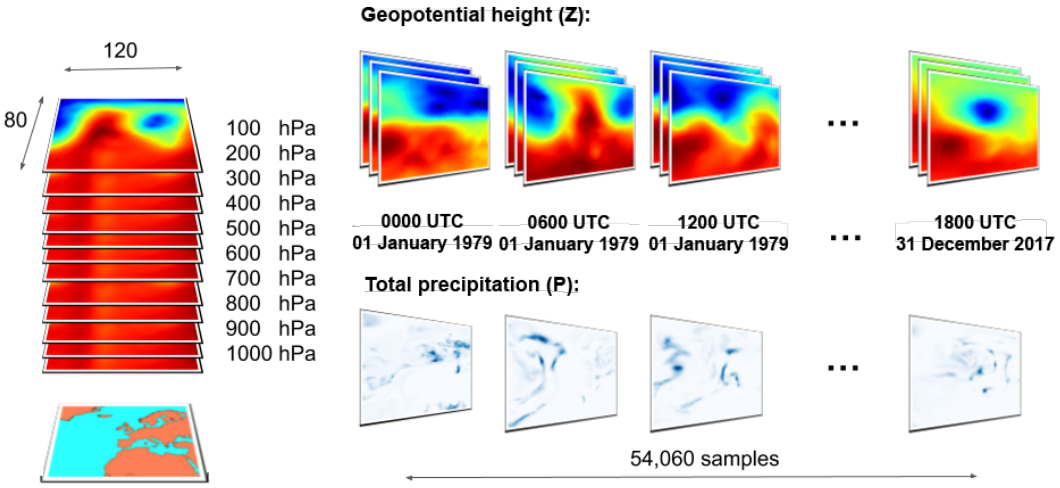



Рис. 3.20 - Відображення географічної досліджуваної області (широта: [75, 15], довгота = [-50, 40]) і тимчасової межі, охопленої геопотенційною висотою ERA-Interim, і загальною кількістю опадів.

***Експериментальний дизайн.*** Мета експериментів, описаних тут, полягає в тому, щоб продемонструвати, що згорткові нейральні мережі енкодер-декодер можуть навчитися робити висновки про складні атмосферні процеси, такі як опади, використовуючи базові Поля НВП як вхідні дані. Ми розглянемо 3 сучасні нейронні мережі згорткового енкодера-декодера для знаходження взаємозв'язків між висотою геопотенціалу та загальною кількістю опадів, порівнюючи їх продуктивність. Для полегшення порівняння різних моделей енкодер-декодер ми пропонуємо використовувати конвеєр, що складається з двох етапів. На першому етапі виконується процес варіативного вибору для

визначення висоти геопотенціалу, що мінімізує похибку при прогнозуванні загального поля опадів. На другому етапі порівнюються результати трьох сучасних архітектур згорткового енкодера-декодера в області сегментації зображень при навчанні прогнозуванню ERA-проміжне загальне поле опадів.

***Вибір змінної.*** Вхідний набір даних містить десять рівнів геопотенційної висоти; однак через лінійне збільшення кількості параметрів, які можна навчити, і вимог до апаратного забезпечення для навчання цих згорткових мереж кодера-декодера, ми обмежуємо кількість вхідних рівнів до трьох. Було запропоновано різні методи для відбору змінних (Saeys et al. 2007). Ці методи зазвичай зменшують простір пошуку вхідних змінних і оптимізують побудову точних предикторів. Для нашого експерименту ми будуємо спрощену мережу згорткового кодера-декодера та виконуємо вичерпний пошук за всіма можливими комбінаціями розглянутих геопотенційних рівнів висоти. Ця спрощена мережа має архітектуру, схожу на глибші мережі кодера-декодера, але її складність зменшується за рахунок обмеження кількості шарів і глибини операцій згортання. Виконання подібного вичерпного пошуку вхідних змінних із глибшими, складнішими мережами, такими як ті, що представлені в наступній частині експериментального процесу (Розділ X.Y), було б обчислювально неможливим. Однак ця спрощена мережа дозволяє швидко повторювати весь простір ознак, щоб ідентифікувати підмножину геопотенційних висот, що мінімізує помилку поля перед опадами в навчальному наборі. Для визначення рівнів геопотенційної висоти, які дають найкращі результати опадів, спрощена мережа згорткового кодера-декодера навчається за 1-, 2-, 3- комбінаціями з набору 10 рівнів тиску, таким чином отримуємо 175 комбінацій. Результати цього процесу вибору змінних використовуються для порівняння точності трьох найсучасніших згорткових мереж глибокого кодера-декодера на другому етапі конвеєра. Таким чином, остаточна перевірка, що порівнює точність різних мереж, повинна виконуватися з використанням іншого розділу набору даних, який залишається невидимим під час процесу вибору змінної.

Весь конвеєр можна розглядати як процес вибору змінних, за яким слідує

процес порівняння моделі нейронної мережі. Початковий набір даних, який містить 54 060 зразків, випадковим чином розбивається на набори даних для навчання та перевірки, що містять 80% і 20% даних відповідно, щоб різні метеорологічні ситуації були рівномірно представлені в обох розділах. Ці розділи використовуються для оцінки відмінностей у точності між порівнюваними глибокими архітектурами. Процес вибору змінної виконується внутрішньо за допомогою навчального набору даних, який далі розділений на 80% і 20% внутрішніх розділів для навчання та перевірки різних підмножин вхідних змінних. Остаточне порівняння між архітектурами виконується з використанням початкового зовнішнього 20% розподілу перевірки, який не втручається ні в процес вибору змінних, ні в навчання різних архітектур. На малюнку 5 представлено запропонований експеримент і зв'язок між вибором змінної та процесами оцінки моделі.

Мережа, яка використовується в процесі вибору змінної, є спрощенням мережі згорткового кодера-декодера VGG-16, який продемонстрував найсучасніші результати в задачах сегментації зображень. На малюнку 3 представлено перетворення розмірності, виконане мережею VGG-16 для вхідного простору з використанням згорток 3x3 із використанням значення кроку 2 для виконання зменшення просторового розміру. Спрощення, запропоноване для цієї частини експериментального процесу, полягає у видаленні останньої операції згортки в секції кодера. Інформація стискається на глибину 256 каналів замість 512 у повній мережі VGG-16. Ця спрощена мережа значно зменшує загальну кількість параметрів у порівнянні з повним VGG-16, що призводить до значного зменшення кількості обчислювальних ресурсів, необхідних для її навчання. Кожна мережа навчається протягом 20 епох (ітерації внутрішнього вхідного тренувального розподілу), а їхні результати чесно обчислюються за допомогою метрики MAE, яка порівнює прогнозовані вихідні дані із загальним полем опадів у внутрішньому наборі даних перевірки. Таким чином, результати цього першого експерименту використовуються для визначення геопотенційних рівнів висоти, які мінімізують помилку при прогнозуванні загальної кількості опадів.

***Вибір моделі CNN.*** Другий крок конвеєра зосереджений на пошуку того, яка згорточна мережа глибокого кодера-декодера є більш точною для прогнозування загального поля опадів ERA-Interim. Для цієї частини ми використовуємо рівні геопотенційної висоти, обчислені раніше, щоб порівняти три різні найсучасніші мережі сегментації.

Ми розглядаємо три різні найсучасніші мережі згорткового кодера-декодера в області сегментації зображення: VGG-16, Segnet та U -net. Ці мережі модифіковано для виконання завдань регресії замість класифікації шляхом зміни функції втрат на MAE. Щоб здійснити чесне порівняння між цими трьома мережами, ми будуємо їх, використовуючи однакову кількість шарів і глибину операцій згортання. Базова структура для всіх трьох мереж представлена на рисунку 3.21.

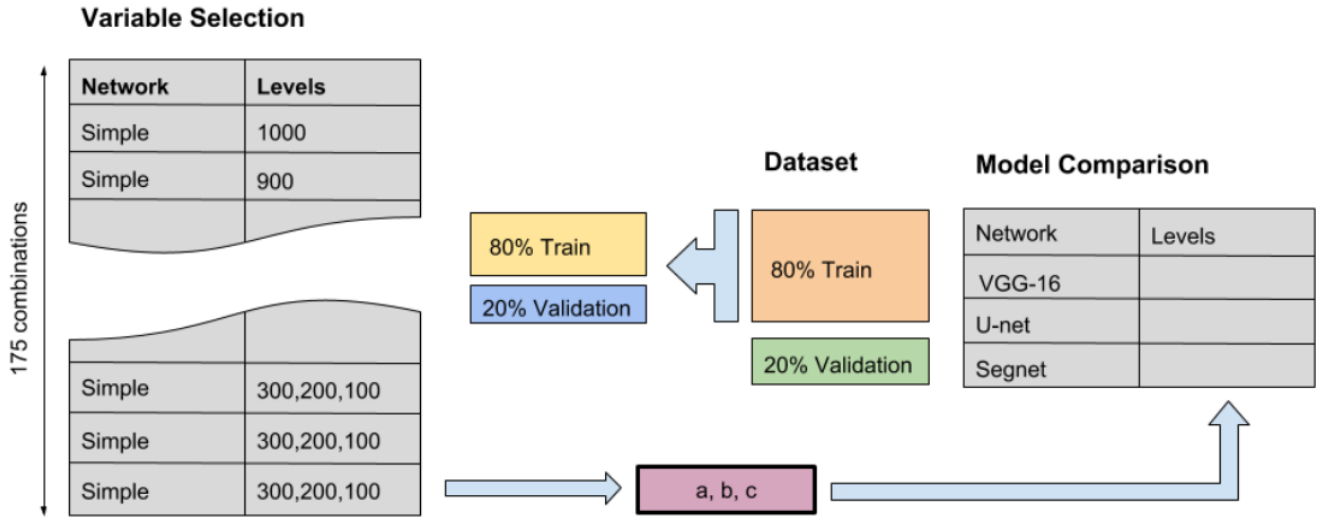



Рис. 3.21 - Графічне представлення конвеєра експерименту, що включає процеси вибору змінної та процеси порівняння мережі кодера-декодера.

Таким чином, усі три мережі виконують однакові перетворення розмірності при оцінці загальної кількості опадів на основі вхідної геопотенційної висоти. Різниця між цими мережами полягає в кількості операцій згортання, які виконуються на кожному рівні, і в конфігурації з’єднань між шарами. VGG-16 представляє лінійну архітектуру, в якій кожен шар пов'язаний лише з сусідніми. Seg-net обчислює індекс максимальної операції об’єднання на кожному з рівнів кодера та передає це значення своєму симетрику в декодері, щоб їх можна було

використовувати на етапі підвищення дискретизації. Декодер U-net, з іншого боку, об'єднує вагові коефіцієнти, які використовуються в частині кодера, щоб реконструювати просторову інформацію. Метою цієї другої половини експериментального процесу є визначення мережі, яка забезпечує найкращу точність при прогнозуванні загальної кількості опадів. Усі три мережі навчаються з використанням початкового зовнішнього розподілу 80/20, визначеного на початку експерименту.

Таким чином, розподіл перевірки, який використовується для порівняння точності різних мереж, залишався непоміченим під час вибору змінних і навчання різних мереж. Цей метод забезпечує незалежність і справедливість результатів обох експериментів.

Мережі навчаються протягом 50 епох, використовуючи підмножину геопотенційних висот, вибраних у першій частині експерименту як вхідні дані, і загальне поле опадів як вихідні дані. Для навчання трьох мереж використовується той самий оптимізатор (стохастичний градієнтний спад), швидкість навчання (0,01) і функція втрат (MAE), що й у процесі вибору змінної. Ці мережі потім порівнюються з полем загальної кількості опадів у розділеному наборі даних перевірки, створеному ERA-Interim, щоб визначити помилку.

Моделі реалізовано за допомогою фреймворку Keras і бек-енду TensorFlow.

**Результати.**

***Процес відбору змінних.*** У першій частині експериментального процесу ми визначаємо підмножину геопотенційних рівнів висоти, які дають кращу оцінку в навчальному наборі загального поля опадів ERA-Interim. Ми навчаємо просту мережу кодера-декодера 175 разів, як описано в попередньому розділі, використовуючи навчальний спліт.

Отримані моделі потім порівнюються за допомогою внутрішнього розподілу перевірки, який формується з решти 20% початкового розподілу навчання. Показник MAE використовується для порівняння похибки прогнозування загальної кількості опадів у кожній точці сітки. На малюнку 6 наведені оцінки MAE, отримані для кожної комбінації будь-яких двох геопотенційних рівнів висоти відносно

загального поля опадів, створеного NWP. Результати показують, що комбінації нижчих рівнів атмосфери дають кращі оцінки поля опадів, ніж вищих рівнів. Головна діагональ матриці на малюнку 6 представляє результуючі помилки при використанні одного рівня геопотенціалу для навчання мережі. Індивідуально нижні рівні атмосфери мають нижчі значення MAE при прогнозуванні опадів, становлячи 900 гПа з найменшою похибкою. У разі використання двох вхідних даних найменші похибки виявляються при поєднанні низького та середнього рівнів геопотенціалу, оскільки комбінація рівнів 1000 та 500 гПа представляє найменшу похибку.

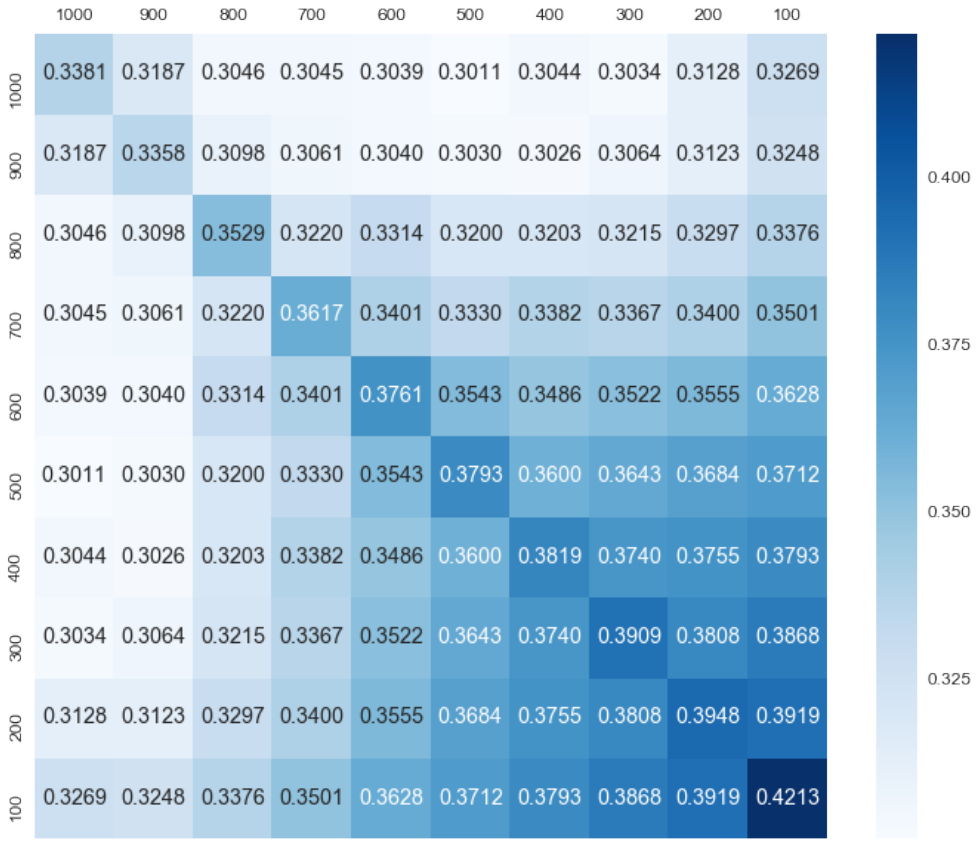



Рис. 3.22 - Матриця, що представляє середні результати валідації MAE для кожної простої мережі енкодер-декодер, навченої з кожною можливою комбінацією двох геопотенційних рівнів.

Навчання мережі кодера-декодера з трьома рівнями геопотенціалу призводить до значного покращення продуктивності. Таблиця 3.8 містить п’ять найнижчих результатів помилок і відповідні їм атмосферні рівні. На жаль, результати для трьох рівнів не можна легко відобразити графічно. Порівняно з

попередніми результатами, є помітне покращення продуктивності, коли третій рівень додається як вхідний сигнал до мережі кодера-декодера. Це означає, що нейронна мережа здатна знаходити внутрішні зв'язки між різними рівнями атмосфери та пов'язувати їх із опадами. Поєднання геопотенційних висот 1000, 800 і 400 гПа призводить до найменшої похибки загального поля опадів у навчальному розділі. Цей результат напрочуд схожий на традиційну практику прогнозування погоди з використанням геопотенційних полів 850 і 500 гПа (у поєднанні з іншими, такими як температура і вітер) для визначення розташування погодних фронтів і, отже, опадів.

ТАБЛИЦЯ 3.8. Топ-5 середніх результатів MAE під час навчання простої мережі кодера-декодера з кожною комбінацією трьох геопотенційних рівнів висоти для прогнозування ERA-Interim загального поля опадів.

| *z levels* [hPa] | *MAE* [mm] |
|---|---|
| 1000, 800, 400 | 0.2895 |
| 1000, 800, 500 | 0.2897 |
| 1000, 900, 500 | 0.2897 |
| 1000, 900, 400 | 0.2901 |
| 1000, 700, 400 | 0.2927 |

Значення середньої абсолютної похибки (MAE), представлені в таблиці 3.8, розраховуються з використанням середнього значення результатів MAE для області сітки 120 × 80 і для всіх часових записів у розділі перевірки. Враховуючи, що загальне поле опадів виражається в міліметрах рідинного еквівалента опадів, похибка цих мереж при прогнозуванні загальної кількості опадів у середньому становить менше 1/3 літра на квадратний метр за 6-годинний період у порівнянні з значення, створені NWP.

***Порівняння глибоких згорткових мереж.*** Для цієї частини експериментального процесу ми вибираємо підмножину геопотенційних висот 1000, 800 і 400 гПа, щоб оцінити продуктивність раніше представлених глибших, найсучасніших мереж сегментації кодера-декодера, адаптованих для виконання завдань регресії. . Кількість параметрів і глибина цих мереж значно перевищують спрощену мережу, яка раніше використовувалася для вибору рівнів геопотенціалу.

Таблиця 3.9 представляє загальну кількість параметрів, які можна навчити, для кожної з цих мереж і загальну кількість часу, необхідного для навчання різних мереж протягом 50 ітерацій (епох) протягом початкового спліт-навчання. Загальна кількість параметрів, які можна навчати, вказує на розмір кожної мережі, а значення часу в цій таблиці вказує на час, необхідний для навчання кожної мережі за допомогою TensorFlow, бібліотеки з відкритим кодом. повторно орендований Google, і вузли GPU P-100.

ТАБЛИЦЯ 3.9. Кількість параметрів для кожної мережевої архітектури кодера-декодера та кінцевий час навчання для кожної мережі (50 епох).

| *Network* | *Parameters* | *Time [hours]* |
|---|---|---|
| Simple (ref.) | 745,000 | 0.6 |
| VGG-16 | 16,467,469 | 4.7 |
| U-net | 7,858,445 | 2.4 |
| Segnet | 29,458,957 | 8.6 |

На рисунку 3.23 показано процес навчання чотирьох різних мереж кодера-декодера протягом 50 епох у навчальному наборі даних. Наприкінці кожної епохи під час навчання набір даних перевірки використовується для оцінки похибки моделі, і вдосконалення різних моделей можна чесно порівняти за допомогою невидимих даних. На початку процесу навчання мережа швидко навчається і сповільнюється по мірі навчання. Зменшення помилки перевірки різне для кожної мережі, яка вирівнюється в різних точках і з різними темпами.

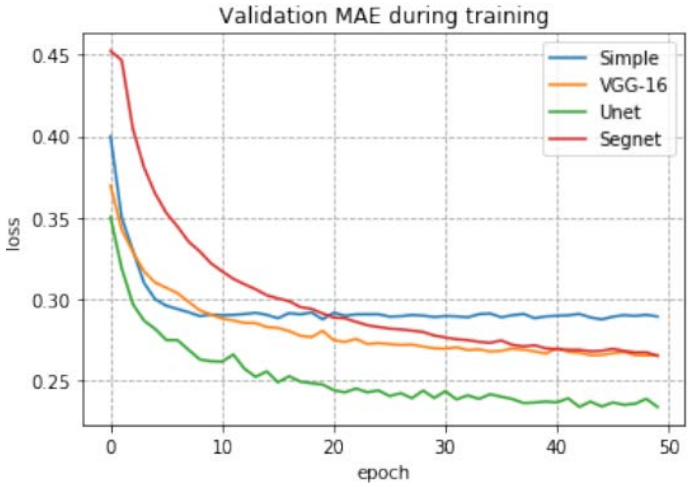



Рис. 3.23 Порівняння еволюції помилки перевірки під час навчання для чотирьох згорткових мереж кодера-декодера протягом 50 епох.

Беручи до уваги результати перевірки на рисунку 3.23 і розмір кожної мережі в таблиці 3.9, ми виділяємо поведінку U-net, яка демонструє найменшу помилку перевірки, а також має найменшу кількість параметрів з трьох розглянутих мереж глибокого навчання.

На рисунку 3.24 представлено візуальне порівняння результатів, отриманих кожною моделлю для нетипової атмосферної ситуації близько 0000 UTC 04 червня 1983 року. Перші два стовпчики зліва представляють геопотенціальну висоту 1000 гПа і загальну кількість опадів, як це зроблено за моделлю ERA-Interim. Загальна кількість опадів – це загальна кількість опадів, накопичена за 6-годинний період, наступний за вказаним часом. Аналогічним чином, використовуючи ту саму колірну шкалу, наступні 4 стовпці представляють опади, що генеруються різними мережами кодерів-декодерів.

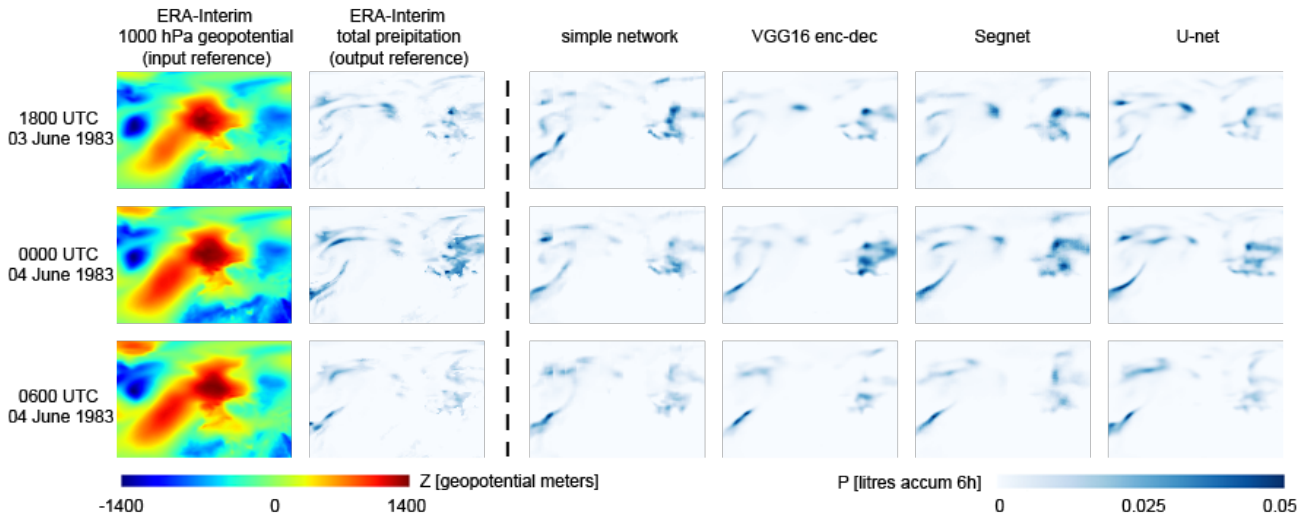



Рис. 3.24 - Візуальне порівняння сумарного поля опадів, створюваного різними мережами. ERA-Interim поля геопотенціальної висоти 1000 гПа та загальної кількості опадів включені для довідки.

Просторова структура та інтенсивність поля опадів представлена кожною мережею по-різному, з незначними відмінностями щодо ERA-проміжного еталонного виходу. Різні згорткові мережі енкодер-декодер використовують різні методи для реконструкції просторової інформації, втраченої на етапі кодування. Крім фіксації просторової структури поля опадів, різні сітки дають точні результати для інтенсивності опадів на кожній сітці точка.

***Статистичний аналіз результатів.*** Щоб порівняти продуктивність різних архітектур згорткового кодера-декодера, ми використовуємо розподіл зовнішньої перевірки, щоб отримати загальну кількість опадів у найближчій точці сітки для дев'яти різних міст, створених кожною мережею. На рисунку 3.25 показано географічне розташування цих дев'яти міст у регіоні нашого набору даних. Ці міста розташовані в різних кліматичних зонах і мають різний характер опадів.

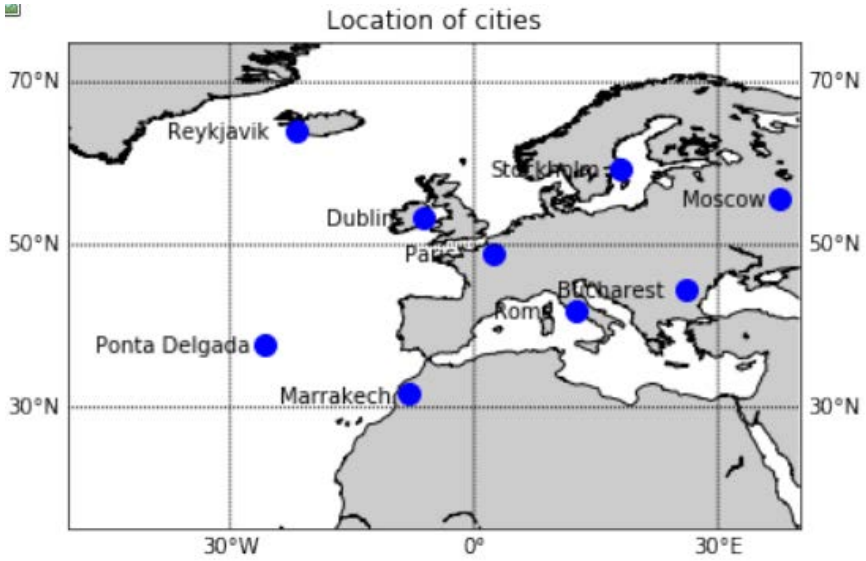



Рис. 3.25 - Розташування дев'яти різних міст у складі регіону.

Результати оцінюються з використанням поля загальної кількості опадів ERA-Interim як еталону для тих самих точок сітки з використанням метрики помилки зі знаком або зсуву. Цей показник надає інформацію про можливі зміщення та розподіл помилки на відміну від раніше використовуваного MAE, який не надає інформації про знак помилки. Для кожного міста і моменту часу розраховується похибка прогнозу загальної кількості опадів. Потім ці результати агрегуються за містом і типом мережі. На рисунку 3.26 використовується скрипковий графік для представлення результатів помилок у кожному місці для різних архітектур. Скрипковий графік пропонує модифікацію коробчатих графіків, додаючи інформацію про розподіл щільності до основної сумарної статистики, притаманної коробчатим діаграмам. Горизонтальна блакитна смуга до центру кожної зі скрипок на рисунку 3.26 представляє середнє значення. Нижня частина кожного графіка показує середнє (μ) і стандартне відхилення (s) значень похибок

для кожної мережі та місця. Форма скрипки дає візуальну індикацію виконання кожної моделі. Ширші та гостріші форми скрипки навколо значення 0 є показником хорошої продуктивності мережі.

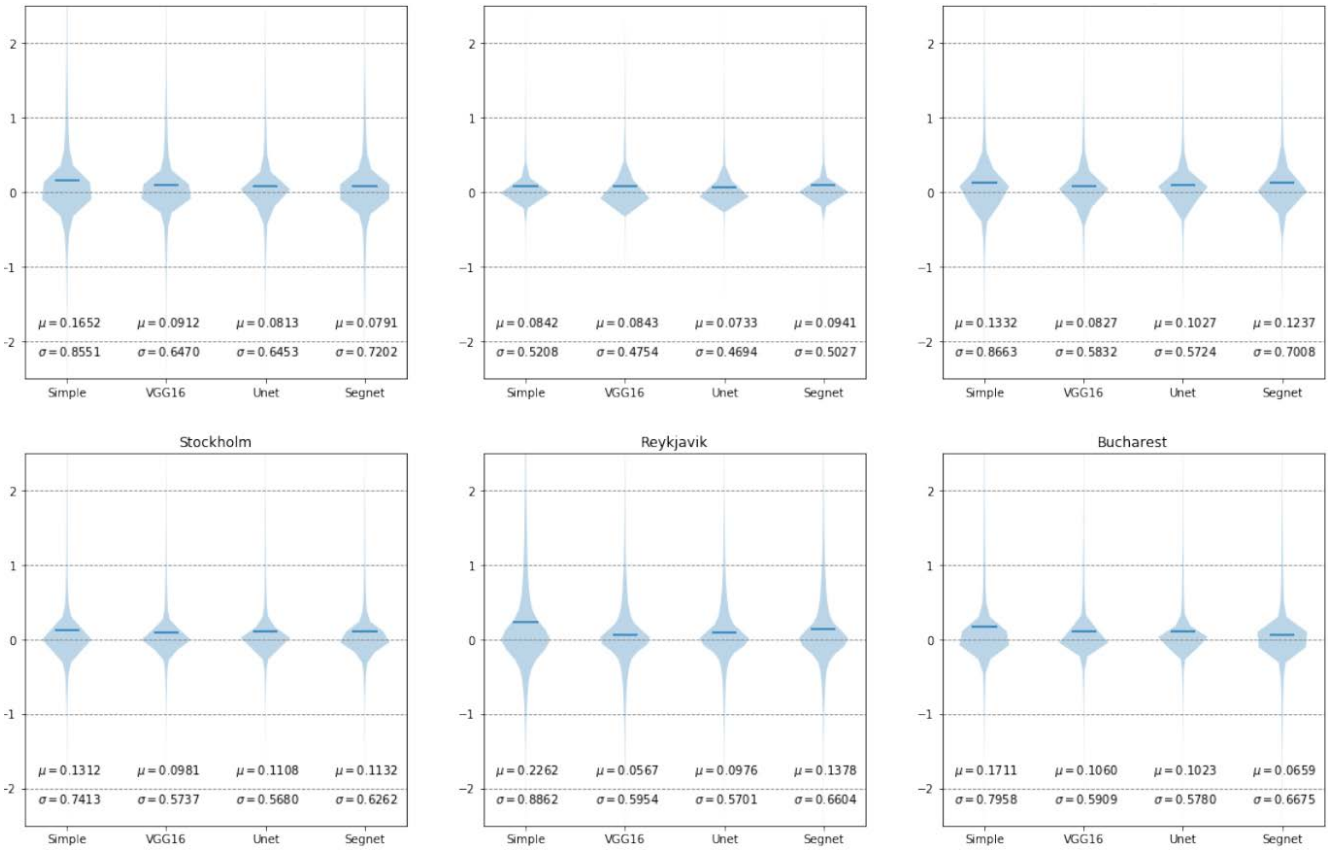



Рис. 3.26 - Відображення функції щільності помилок і середнього (μ) і стандартного відхилення (s) значень для різних архітектур нейронної мережі в кожному місті.

Щоб статистично порівняти результати, ми використовуємо методологію, запропоновану, щоб оцінити статистичну значущість відмінностей між результатами помилок кожної мережі в дев'яти місцях. Початковий тест Фрідмана відхиляє нульову гіпотезу подібності між 4 мережами згорткового кодера-декодера. Це виправдовує використання пост-хок двовимірних тестів, Nemenyi у нашому випадку, щоб оцінити значущість відмінностей між різними парами мереж кодера-декодера.

Результати цих тестів виражаються графічно за допомогою діаграм критичної різниці (CD). Тест Немені виконує попарне порівняння результатів помилок для будь-яких двох архітектур. Відмінності вважаються суттєвими, якщо відповідний середній ранг відрізняється принаймні однією критичною різницею.

На рисунку 3.27 показано CD-діаграму, що представляє результати тесту Немені (α = 0,05) з використанням значень помилок у дев'яти місцях для кожної мережі згорткового кодера-декодера. Діаграми CD здійснюють попарне порівняння між методами, з'єднуючи архітектури, для яких не виявлено значущих статистичних відмінностей, або, іншими словами, ті, чия відстань менша за фіксовану критичну різницю, показану як еталон у верхній частині рисунка 3.28. Мережі ранжовані з нижчими значеннями на діаграмах CD означають вищі значення помилок. Ці тести було виконано за допомогою пакета scmamp R, який є загальнодоступним у Comprehensive R Archive Network (CRAN).

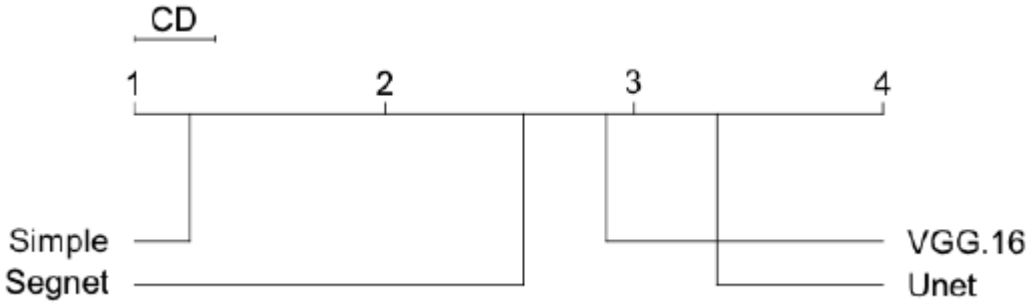



Рис. 3.27 Критичні відмінності порівняння 3 архітектур згорткового кодера-декодера разом із еталонною спрощеною моделлю. α = 0,05.

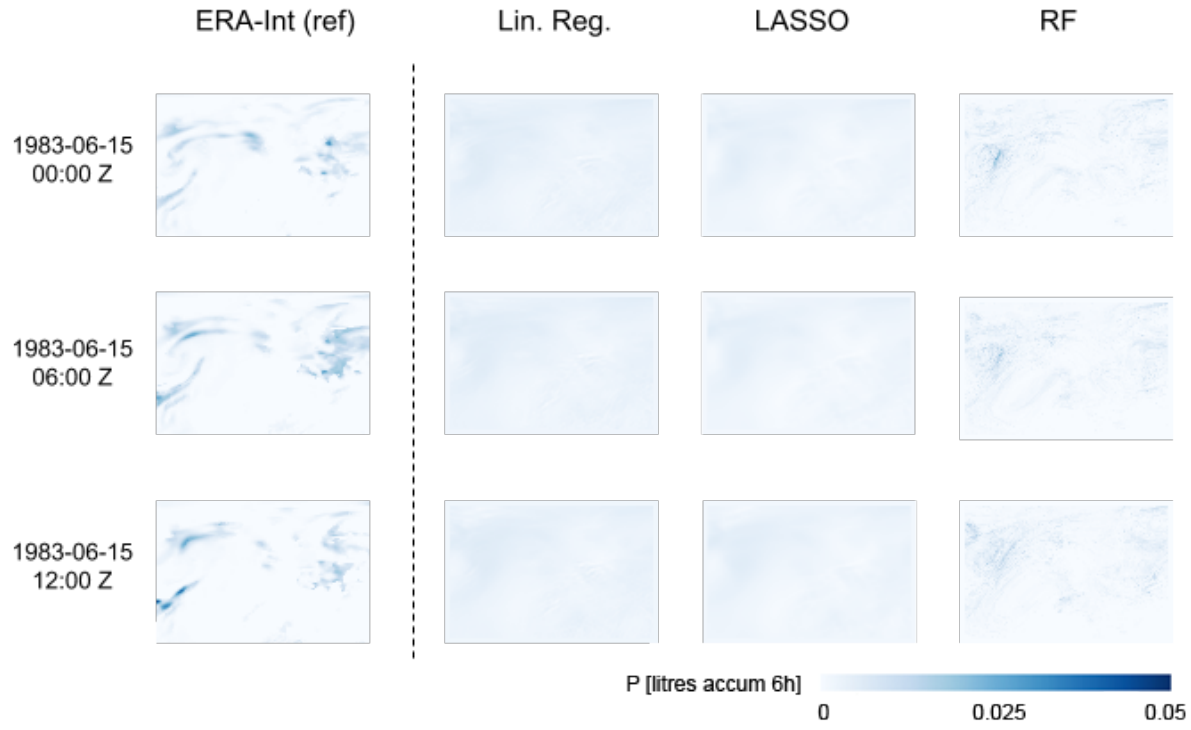



Рис. 3.28 Виведення трьох сучасних методів регресії (розмір латки = 7) з використанням тих самих дат, показаних на рисунку 3.24.

На рисунку 3.28 показано вихідні дані, отримані кожним алгоритмом регресії для тієї самої метеорологічної ситуації, представленої на рисунку 3.24. Моделі не можуть забезпечити чіткість, необхідну для представлення поля опадів. Можна побачити, що результати, згенеровані випадковим лісом, забезпечують незначне покращення у визначенні положення областей опадів, можливо, тому, що це єдиний нелінійний метод. Однак потужності цієї методики недостатньо для точного вирішення цієї проблеми.

У цій роботі ми представляємо серію моделей нейронної мережі, які навчені генерувати оцінки опадів на основі геопотенціального поля. Таким чином, якість цих моделей обмежена якістю базової параметризації NWP, яка використовується для представлення загального поля опадів. Ту саму мережу кодера-декодера в ідеалі можна навчити, використовуючи дані спостережених опадів замість параметризованих змінних. Останні досягнення в технологіях спостереження за Землею, такі як Глобальний моніторинг опадів (GPM), пропонують високоякісні набори даних про глобальні опади, які можна використовувати в поєднанні з NWP для вивчення кращих моделей параметризації.

Еволюція методології, представлена в цій роботі, полягає у введенні часового виміру в згорткові мережі енкодер-декодер з використанням рекурентних конфігурацій. Рекурентні нейронні мережі продемонстрували видатні досягнення в області аналізу часових рядів і розпізнавання мови, а також відкрили новий цікавий напрямок досліджень для моделей, які можуть вивчити як просторові, так і часові виміри даних NWP.

# ВИСНОВКИ

В магістерській кваліфікаційній роботі досягнуто наступних результатів:

1. Було розглянуто розвиток прогнозування погоди від спостережень за природою до використання наукових методів, винаходу приладів і сучасних технологій, таких як комп'ютерні моделі та супутники.

2. Було розглянуто порівняння методів прогнозування погоди, включаючи чисельні моделі та традиційні підходи, а також методології їх удосконалення. Чисельні методи, які базуються на математичних моделях, забезпечують високу точність, проте залежать від якості вхідних даних і обчислювальних ресурсів, тоді як традиційні підходи корисні для швидких оцінок. Для покращення прогнозів рекомендовано інтеграцію сучасних технологій, таких як машинне навчання, удосконалення моделей і розширення джерел даних через міжнародну співпрацю й автоматизовані системи.

3. У межах дипломної роботи розроблено методику для аналізу та прогнозування погодних умов із використанням нейронних мереж. Проведено огляд сучасних методів прогнозування погоди, зокрема фізико-математичних моделей і машинного навчання. Була створена та навчена нейронна мережа для передбачення швидкості вітру в аеропортах.

## СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ


1. Annie Jin SA (2012) The potential of social media for luxury brand management. Mark Intell Plan 30(7):687–699.
2. Bastien V, Kapferer J-N (2013) More on luxury anti-laws of marketing. In: Wiedmann KP, Hennigs N (eds) Luxury Marketing. Gabler Verlag, Wiesbaden, pp 19–34.
3. Batat W (2019) The new luxury experience. Springer, Berlin.
4. Bilge HA (2015) Luxury consumption: literature review. Khazar J of Humanities and Social Sciences. https://doi.org/10.5782/2223-2621.2014.18.1.35.
5. Chen Y, Fay S, Wang Q (2011) The role of marketing in social media: how online consumer reviews evolve. J Interact Mark 25(2):85–94.
6. Chu S-C, Kim Y (2011) Determinants of consumer engagement in electronic word-of-mouth (eWOM) in social networking sites. Int J Advert 30(1):47–75.
7. Cuomo MT, Foroudi P, Tortora D, Hussain S, Melewar T (2019) Celebrity endorsement and the attitude towards luxury brands for sustainable consumption. Sustainability 11(23):6791.
8. Cvijikj IP, Michahelles F (2013) Online engagement factors on Facebook brand pages. Soc Netw Anal 3(4):843–861.
9. De Bruyn A, Lilien GL (2008) A multi-stage model of word-of-mouth influence through viral marketing. Int J Res Mark 25(3):151–163.
10. De Vries L, Gensler S, Leeflang PS (2012) Popularity of brand posts on brand fan pages: an investigation of the effects of social media marketing. J Interact Mark 26(2):83–91.
11. Dubois B, Laurent G, Czellar S (2001) Consumer rapport to luxury: analyzing complex and ambivalent attitudes, vol 736. Groupe HEC, Jouy-en-Josa.s
12. Erkan I, Evans C (2016) The influence of eWOM in social media on consumers' purchase intentions: an extended approach to information adoption. Comput Hum Behav 61:47–55.

13. Erkan I, Evans C (2018) Social media or shopping websites? The influence of eWOM on consumers' online purchase intentions. J Mark Commun 24(6):617–632.
14. Evans D (2010) Social media marketing: an hour a day. Wiley, New York
15. Fionda AM, Moore CM (2009) The anatomy of the luxury fashion brand. J Brand Manag 16(5–6):347–363
16. Freire NA (2014) When luxury advertising adds the identitary values of luxury: a semiotic analysis. J Bus Res 67(12):2666–2675
17. Gibson J (2017) Purchase funnel—definition and introduction. Retrieved February 07, 2018, from http://marketing-made-simple.com
18. Godey B, Manthiou A, Pederzoli D, Rokka J, Aiello G, Donvito R, Singh R (2016) Social media marketing efforts of luxury brands: influence on brand equity and consumer behavior. J Bus Res 69(12):5833–5841
19. Heine K (2012) The concept of luxury brands. Luxury Brand Management, Berlin
20. Heine K, Phan M (2011) Trading-up mass-market goods to luxury products. Austr Mark J 19(2):108–114
21. Hennig-Thurau T, Gwinner KP, Walsh G, Gremler DD (2004) Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the Internet? J Interact Mark 18(1):38–52
22. Hennig-Thurau T, Malthouse EC, Friege C, Gensler S, Lobschat L, Rangaswamy A, Skiera B (2010) The impact of new media on customer relationships. J Serv Res 13(3):311–330
23. Hutter K, Hautz J, Dennhardt S, Fьller J (2013) The impact of user interactions in social media on brand awareness and purchase intention: the case of MINI on Facebook. J Prod Brand Manag 22(5/6):342–351.
24. Chen Y, Fay S, Wang Q (2011) The role of marketing in social media: how online consumer reviews evolve. J Interact Mark 25(2):85–94.
25. Chu S-C, Kim Y (2011) Determinants of consumer engagement in electronic word-of-mouth (eWOM) in social networking sites. Int J Advert 30(1):47–75.

26. Cuomo MT, Foroudi P, Tortora D, Hussain S, Melewar T (2019) Celebrity endorsement and the attitude towards luxury brands for sustainable consumption. Sustainability 11(23):6791.
27. Cvijikj IP, Michahelles F (2013) Online engagement factors on Facebook brand pages. Soc Netw Anal 3(4):843–861.
28. De Bruyn A, Lilien GL (2008) A multi-stage model of word-of-mouth influence through viral marketing. Int J Res Mark 25(3):151–163.
29. De Vries L, Gensler S, Leeflang PS (2012) Popularity of brand posts on brand fan pages: an investigation of the effects of social media marketing. J Interact Mark 26(2):83–91.
30. Dubois B, Laurent G, Czellar S (2001) Consumer rapport to luxury: analyzing complex and ambivalent attitudes, vol 736. Groupe HEC, Jouy-en-Josa.s

# ДЕМОНСТРАЦІЙНІ МАТЕРІАЛИ

ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ
НАВЧАЛЬНО–НАУКОВИЙ ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

КВАЛІФІКАЦІЙНА РОБОТА
на тему:

# «МЕТОДИКА ДЛЯ АНАЛІЗУ І ПЕРЕДБАЧЕННЯ ПОГОДНИХ УМОВ ЗА ДОПОМОГОЮ НЕЙРОННОЇ МЕРЕЖІ»

***виконав студент: Ляшук М.А.***
***керівник: Шкапа В.В. к.ф.м.н, доцент.***



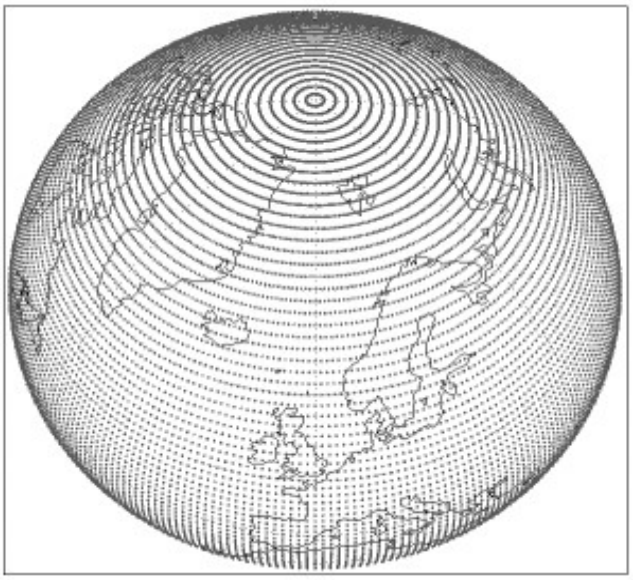
# *Мета роботи*

***Актуальність*** – сучасні погодні умови мають значний вплив на різні сфери життя, включаючи сільське господарство, транспорт, енергетику та урбаністику, військові цілі. Актуальність прогнозування погоди у військових цілях обумовлена тим, що погодні умови безпосередньо впливають на ефективність виконання бойових завдань, операційної діяльності та логістики. Наприклад, несприятливі погодні явища можуть вплинути на авіаційні операції, точність артилерійського вогню, мобільність наземної техніки та функціонування систем зв'язку. Традиційні методи прогнозування погоди не завжди дають достатню точність, особливо в умовах глобальних кліматичних змін. Використання нейронних мереж дозволяє обробляти великі обсяги даних і виявляти складні закономірності, що забезпечує більш точне та адаптивне прогнозування. Це сприяє прийняттю ефективних рішень та мінімізації ризиків, пов'язаних із несприятливими погодними явищами.

***Мета роботи*** – дослідити методики аналізу і прогнозування погодних умов з використанням сучасних методів машинного навчання, зокрема нейронних мереж, для підвищення точності прогнозіві покращення ефективності роботи систем моніторингу погоди.

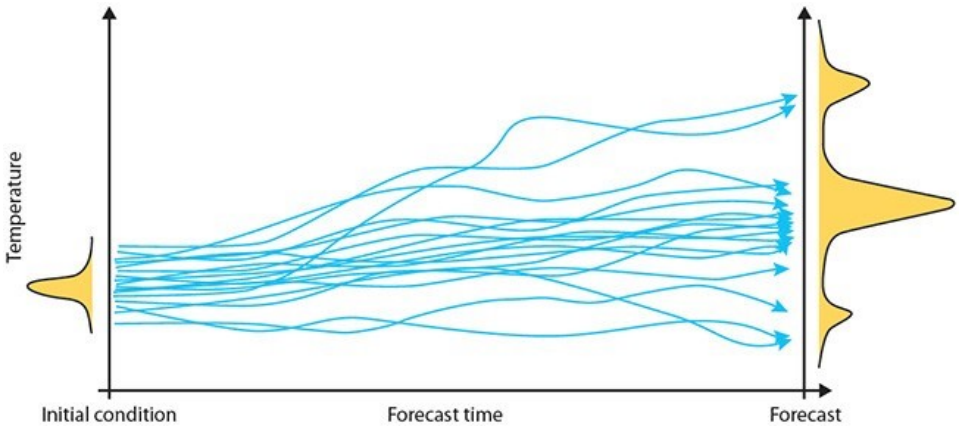
***Об'єкт дослідження*** – процеси аналізу і прогнозування погодних умов на основі метеорологічних даних.

***Предмет дослідження*** – методи і алгоритми машинного навчання, зокрема нейронні мережі, які використовуються для обробки та аналізу метеорологічних даних з метою прогнозування погодних умов.



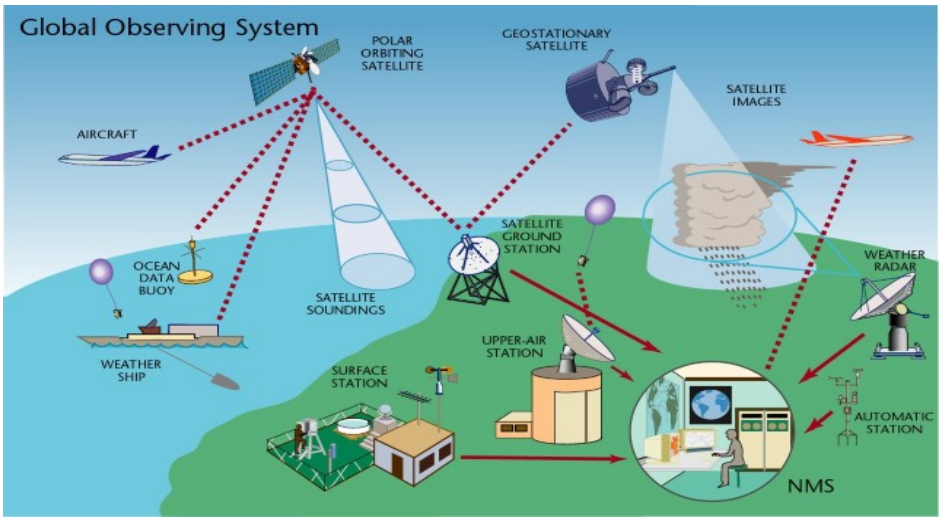
***Приклад регулярної гауссової сітки F80, що використовується деякими NWP для представлення атмосфери Землі***

3



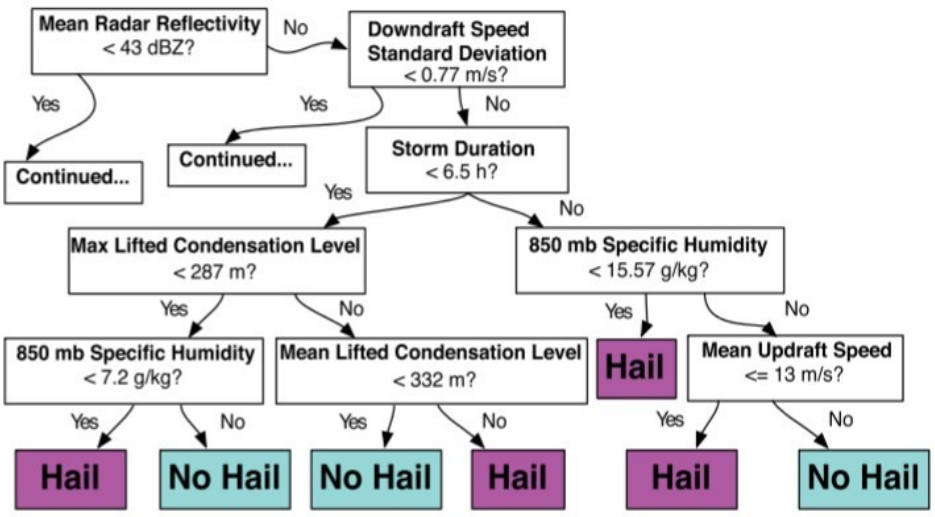

***Ансамбль прогнозів виробляє ряд можливих сценаріїв з урахуванням початкового розподілу ймовірностей прогнозованого параметра.***

4

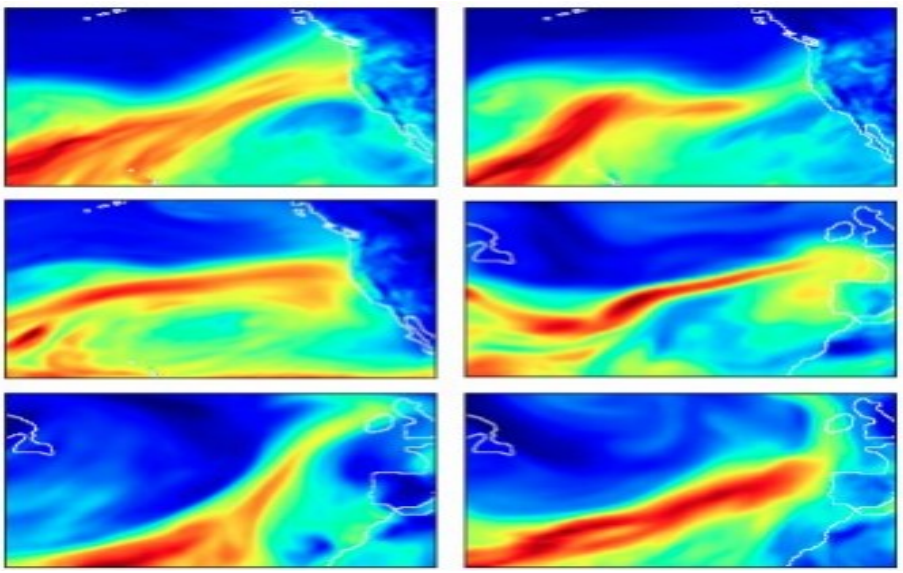



***Глобальна система спостережень (GOS)***



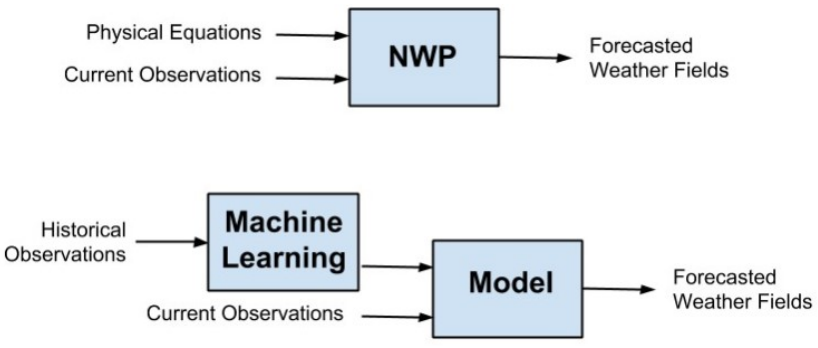



***Дерево рішень, що використовується для прогнозування події граду на основі порогових значень для різних спостережуваних і NWP параметрів.***



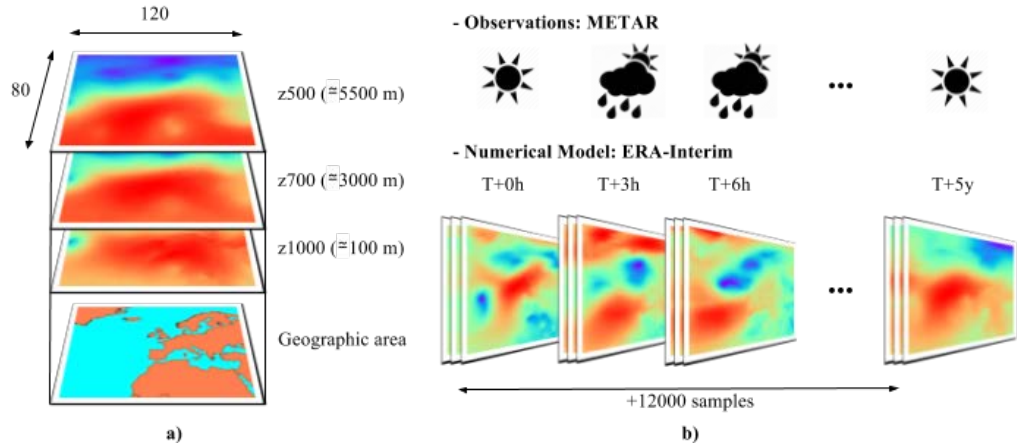
***Зразки зображень атмосферних річок (струменевих потоків), відповідно класифіковані та витягнуті з багато терабайтного набору даних NWP за допомогою глибокої моделі CNN***

7



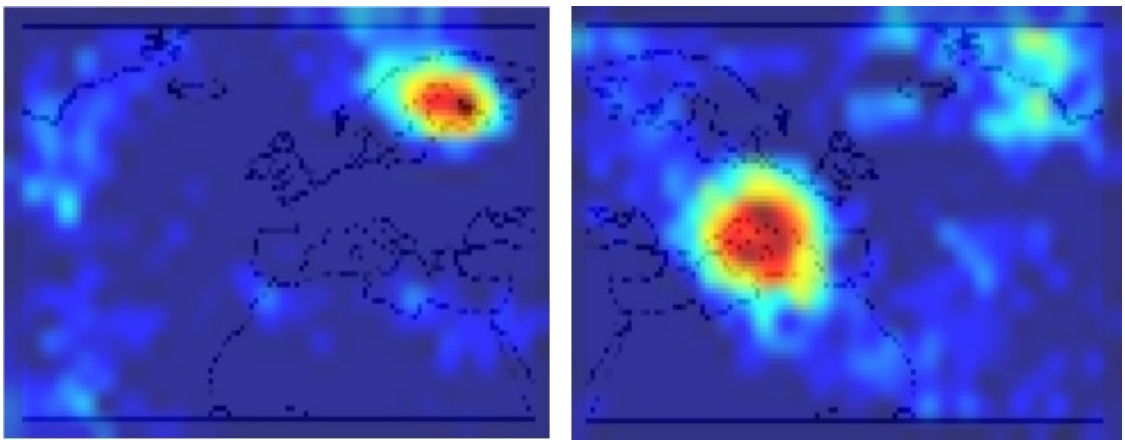

***Порівняння традиційних підходів NWP та машинного навчання до прогнозування погоди***

8

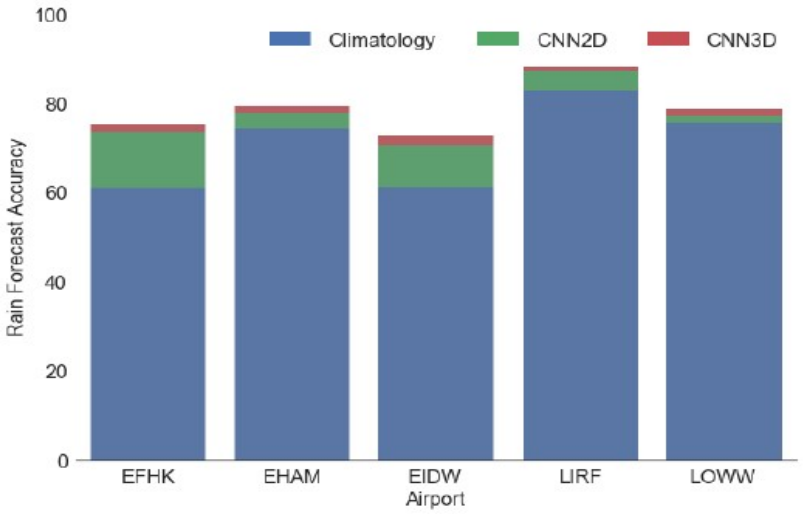



***Дані від ERA-Interim***





***Приклад отриманих карт активації класу для ERAInterim CNN, навченого з використанням спостережених опадів***




***Результати точності, отримані для різних аеропортів і методології***

11

ЗАКЛАД ВИЩОЇ ОСВІТИ
«МІЖНАРОДНИЙ НАУКОВО-ТЕХНІЧНИЙ
УНІВЕРСИТЕТ ІМЕНІ АКАДЕМІКА ЮРІЯ БУГАЯ»

**СЕРТИФІКАТ**

засвідчує, що

*Ляшук Микола*

взяв(ла) участь у VIII Всеукраїнської студентської науково-практичної конференції

«ВИЩА ОСВІТА – СТУДЕНТСЬКА НАУКА – СУЧАСНЕ СУСПІЛЬСТВО: НАПРЯМИ РОЗВИТКУ»

РЕКТОР ВЕРОНІКА ХУДОЛЕЙ

11 грудня 2024 року, м. Київ

***Апробація результатів дослідження***

12

## *Висновки*

В магістерській роботі досягнуто наступних результатів :

1. Було розглянуто розвиток прогнозування погоди від спостережень за природою до використання наукових методів, винаходу приладів і сучасних технологій, таких як комп'ютерні моделі та супутники.

2. Було розглянуто порівняння методів прогнозування погоди, включаючи чисельні моделі та традиційні підходи, а також методології їх удосконалення . Чисельні методи, які базуються на математичних моделях, забезпечують високу точність, проте залежать від якості вхідних даних і обчислювальних ресурсів, тоді як традиційні підходи корисні для швидких оцінок. Для покращення прогнозів рекомендовано інтеграцію сучасних технологій, таких як машинне навчання, удосконалення моделей і розширення джерел даних через міжнародну співпрацю й автоматизовані системи.

3. У межах дипломної роботи розроблено методику для аналізу та прогнозування погодних умов із використанням нейронних мереж. Проведено огляд сучасних методів прогнозування погоди, зокрема фізико-математичних моделей і машинного навчання. Була створена та навчена нейронна мережа для передбачення швидкості вітру в аеропортах.



# ДЯКУЮ ЗА УВАГУ!