DOI 10.31891/2307-5732-2024-333-2-35
УДК 004.8, 004.94

СИДОР ПЕТРО
Чернівецький національний університет ім. Юрія Федьковича, м. Чернівці, Україна
https://orcid.org/0009-0004-4237-7866
e-mail: sydor.pete@gmail.com
ВИКЛЮК ЯРОСЛАВ
Національний університет «Львівська Політехніка», м. Львів, Україна
https://orcid.org/0000-0003-4766-4659
e-mail: vyklyuk@ukr.net

АНСАМБЛЕВІ МОДЕЛІ ПРОГНОЗУВАННЯ ПОВЕНЕЙ У ВЕЛИКІЙ БРИТАНІЇ НА ОСНОВІ СОНЯЧНОЇ АКТИВНОСТІ

Це дослідження має на меті з'ясувати приховані залежності між потоками сонячних частинок, які надходять від Сонця, та повенями, спричиненими опадами, в регіоні Великої Британії. Унікальність цієї роботи полягає у застосуванні даних про сонячну активність як вхідних параметрів для машинного навчання та їх кореляції з реальними даними про інтенсивність опадів та повеней у Великій Британії, що слугують вихідними даними. За період з жовтня 2001 року по грудень 2019 року було проаналізовано 20 значущих повеней. Аналіз було виконано на денній основі з урахуванням часового зсуву між сонячними випромінюваннями та їх впливом на погодні умови, що варіювався від 0 до 9 днів. Дослідження застосовує кореляційний аналіз для визначення ступеня взаємозв'язку між часовими рядами сонячної активності та повенями, створюючи основу для більш глибокого розуміння можливих причинно-наслідкових зв'язків. На наступному етапі, використовуючи методи прогностичного моделювання, зокрема алгоритми дерева рішень та ансамблі класифікаційних моделей, було досліджено потенційний зв'язок між змінами в сонячній активності та повенями. Було створено ансамбль моделей за методом жорсткого голосування, який враховує різні показники, такі як щільність протонів, диференціальний потік протонів і температуру іонів, з часовим зсувом до 9 днів. Аналіз показав, що ці ключові параметри можуть ефективно передбачати настання повеней з точністю до 92% за вказаний період.Результати підкреслюють важливість інтеграції космічної погоди та сонячної активності в моделі прогнозування земних погодних умов, зокрема, у контексті повеней в Англії. Це дозволяє не тільки краще розуміти причини природних катаклізмів, але й підвищує ефективність планування заходів з надзвичайних ситуацій і зменшення можливих збитків від них. Таким чином, ця робота вносить важливий вклад у поле досліджень космічної погоди та її впливу на Землю, зокрема у сфері гідрометеорології та керування ризиками повеней.

Ключові слова: захист від повеней, опади, повені, машинне навчання, класифікація, математичне моделювання, аналіз даних, ансамбль класифікаційних моделей

SYDOR PETRO
Chernivtsi National University named after Yuriy Fedkovycha
VYKLYUK YAROSLAV
Lviv Polytechnic National University

ENSEMBLE MODELS FOR FORECASTING FLOODS IN GREAT BRITAIN BASED ON SOLAR ACTIVITY

This study aims to uncover hidden dependencies between streams of solar particles emanating from the Sun and flood events caused by precipitation in the UK. The uniqueness of this work lies in applying solar activity data as input parameters for machine learning and correlating them with real data on precipitation intensity and floods in the UK, which serve as output data. The analysis covered 20 significant floods from October 2001 to December 2019. The analysis was performed on a daily basis, taking into account the time shift between solar emissions and their impact on terrestrial weather conditions, which ranged from 0 to 9 days. The study employs correlation analysis to determine the degree of interrelation between time series of solar activity and floods, laying the foundation for a deeper understanding of possible cause-and-effect relationships. Subsequently, using predictive modeling methods, including decision tree algorithms and ensemble classification models, the potential connection between changes in solar activity and terrestrial floods was explored. An ensemble of models was created using the hard voting method, which takes into account various indicators such as proton density, differential proton flux, and ion temperature, with a time shift of up to 9 days. The analysis showed that these key parameters could effectively predict the occurrence of floods with up to 92% accuracy for the specified period. The results underscore the importance of integrating space weather and solar activity into the models for predicting terrestrial weather conditions, particularly in the context of floods in England. This not only allows for a better understanding of the causes of natural disasters but also enhances the effectiveness of emergency planning and reduces potential damages from them. Thus, this work makes a significant contribution to the field of space weather research and its impact on Earth, particularly in the areas of hydrometeorology and flood risk management.

Keywords: flood protection, precipitation, floods, machine learning, classification, mathematical modeling, data mining, ensemble of classification models

Вступ

Екстремальні погодні умови, такі як інтенсивні опади, які спричиняють повені, визнані як одна з найбільших природних загроз з серйозними соціальними, економічними та екологічними наслідками [1]. Повені можуть спричинити втрати життя, руйнування майна, знищення врожаю та худоби. Довгострокові ефекти включають збої у постачанні питної води та електроенергії, руйнування транспортної та комунікаційної інфраструктури, а також негативний вплив на фізичне та психічне здоров'я людей через переміщення населення. Попри прогрес у розумінні процесів, що призводять до сильних опадів та можливих

повеней, потреба у вдосконаленні прогнозування екстремальних погодних та гідрологічних явищ збільшується через їхні значні негативні наслідки.

Зв'язок між сонячною активністю та кліматом Землі досліджується вже понад 200 років [2]. Сонячна енергія, що досягає Землі, варіюється на різних часових масштабах і корелює з атмосферними параметрами [3], проте оцінити її вплив на кліматичні та екологічні процеси складно. Незважаючи на те, що остаточного визнання зв'язку між сонячним вітром та тропосферою досягнуто не було, численні дослідження вказують на вплив сонячної активності на клімат, включно з атмосферною циркуляцією, температурою, опадами та екстремальними погодними умовами [4–11].

Основні досліджені механізми включають прямий нагрів Землі сонячним випромінюванням і вплив УФ-випромінювання на озоновий шар стратосфери, що веде до змін в атмосферній циркуляції та кліматі [12]. Інший важливий механізм стосується впливу галактичних космічних променів, які можуть сприяти утворенню ядер конденсації хмар і таким чином впливати на хмарність [13–16].

Сонячна активність та збурення міжпланетного середовища мають прогностичне значення для розвитку екстраполічних циклонів, що є ключовими для погоди в середніх широтах. Дослідження показують зміни в циклонічній активності відповідно до сонячної активності, а також вплив сонячних протонних подій на інтенсивність опадів [17–19].

Ця стаття зосереджена на аналізі залежностей між опадами та повенями у Великій Британії спричиненими потоками частинок від Сонця на основі 20 повеней у період з жовтня 2001 р. по грудень 2019 р. З використанням машинного навчання і класифікаційного прогностичного моделювання, були встановлені приховані залежності між цими явищами та розроблена прогнозна модель.

Матеріали та методи

Аналіз предметної області

Територія Сполученого Королівства має густу дренажну мережу з приблизно 200,000 км водотоків, що дренують близько 1500 окремих басейнів [20, 21]. Ці численні водотоки переважно короткі, мілководні та чутливі до значних антропогенних змін. Режим річок визначається кліматичними умовами (особливо опадами, температурою повітря, інсоляцією), геологічними особливостями кожного водозбірного басейну (такими як їхня проникність), морфологією території та антропогенним фактором (змінами в руслах річок, використанням води, змінами у землекористуванні тощо).

В роботі [22] зазначили, що зимовий NAO (North Atlantic Oscillation) впливає на річкові потоки, контролюючи перенос вологи та тепла над Великою Британією. Laizé та Hannah [23] підкреслили, що вищий NAO індекс збільшує західні повітряні потоки через Велику Британію, що призводить до вищих, ніж в середньому, рівнів опадів і температур, а отже, і до більших річкових потоків. В той час як високогірні басейни отримують значні кількості опадів, низинні райони випробовують менші обсяги, тому на режим стоку впливають також інші фактори, такі як проникність, висота і фізичні характеристики басейну.

Водотоки Великої Британії різноманітні: від гірських потоків, що отримують до п'яти метрів опадів на рік, до низинних річок з підземним живленням на південному сході, де рівень опадів нижчий [21]. Опади у Великій Британії розподілені відносно рівномірно протягом року, але із схильністю до осінньо-зимового піку, особливо в західних басейнах. Сезонні коливання температури повітря та сонячного світла спричиняють високу випаровуваність в літній період (квітень-вересень), що впливає на внутрішньорічний розподіл стоків у річках з природними режимами. Зазвичай максимальні витрати води реєструються взимку, а мінімальні - влітку або восени. Варто зауважити, що міські водотоки були суттєво змінені та не завжди слідують цьому взірцю. Наприклад, низькі витрати можуть штучно збільшуватись через переливання резервуарів або перекидання води між басейнами.

Вивчаючи тенденції річкових потоків за чотири стандартні сезони у період з 1969 по 2008 рік у 89 басейнах з майже природними режимами стоку у Великій Британії, Hannaford і Buys [24] зробили висновок, що спостерігається загальне збільшення зимових річкових потоків (із найбільшим зростанням у північних та західних високогірних басейнах, тоді як низькі потоки зменшились у деяких західних басейнах); регіонально послідовне зниження весняних потоків; збільшення літніх потоків (у північних та західних басейнах); і в основному слабкі позитивні та негативні тенденції (в англійських низинах); збільшення осінніх потоків (особливо для високих потоків у центральній і південно-західній Британії та на північному сході Шотландії). Спостережувані тенденції, такі як збільшення зимового стоку та зниження весняного стоку, можуть впливати на управління водними ресурсами і вказувати на збільшення ризику повеней.

Опис набору даних

Щоб перевірити можливий зв’язок між повінню, спричиненою опадами, та сонячною активністю, ми використали кілька наборів даних і джерел даних. Для аналізу були використані дані з 20 незалежних блоків даних для різних повеней (r). Кожен блок даних складався з окремих наборів даних:

- Flood (F): $DS_F^r =< Date, precipitations, days\ from\ the\ beginning\ of\ the\ flood >$
- INTEGRAL PROTON FLUX (IPF, p/cs2-sec-ster) : $DS_{IPF}^r =< Date, (IPF > 10\,MeV), (IPF > 30\,MeV) >$
- DIFFERENTIAL ELECTRON AND PROTON FLUX (DF, p/cs2-sec-ster). Ці блоки містили різні характеристики сонячної енергії для різних періодів під час різних повеней. Виміряні діапазони для диференціального потоку електронів становили 38-53 кеВ і 175-315 кеВ для всіх досліджених повеней, тоді

як виміряні діапазони для диференціального потоку протонів змінювалися залежно від періоду, в якому відбулася повінь. Диференціальні потоки протонів були виміряні в наступних діапазонах: 47-65 кеВ, 47-68 кеВ, 65-112 кеВ, 112-187 кеВ, 115-195 кеВ, 310-580 кеВ, 761-1220 кеВ, 795-1193 кеВ, 1060-1900 кеВ і 1060-1910 кеВ, але єдиним загальним діапазоном для всіх повеней був 310-580 кеВ. Ми використовували лише наступні ознаки, які були загальними для всіх повеней: $DS^r_{DF} =< Date,\ 38\ \text{keV} \leq \text{DF} \leq 53\ \text{keV},\ 175\ \text{keV} \leq \text{DF} \leq 315\ \text{keV}, 310\ \text{keV} \leq \text{DF} \leq 580\ \text{keV} >$. (табл 1)

- SOLAR WIND (SW):

$$DS^r_{SW} =< Date, Proton\ Density\left(\frac{\text{particles}}{\text{cc}}\right), Bulk\ Speed\left(\frac{\text{km}}{\text{s}}\right), Ion\ Temperature(\text{degrees K}) >$$

- 10.7 CM RADIO FLUX (RF, solar flux units): $DS^r_{RF} =< Date,\ Radio\ Flux >$

Дані вибирались в діапазоні 10 днів до і тиждень після паводку.

Таблиця 1

Поля DIFFERENTIAL FLUX (DS^r_{IPF}) для різних річок

River	DIFFERENTIAL FLUX						
2001_0645_GBR	38-53	175-315	47-65	112-187	310-580	761-1220	060-1910
2002_0463_GBR_1	38-53	175-315	65-112	112-187	310-580	761-1220	060-1910
2002_0463_GBR_2	38-53	175-315	65-112	112-187	310-580	761-1220	060-1910
2002_0488_GBR	38-53	175-315	65-112	112-187	310-580	761-1220	060-1910
2002_0774_GBR	38-53	175-315	47-68	115-195	310-580	761-1220	060-1910
2004_0423_GBR	38-53	175-315	47-68	115-195	310-580	761-1220	060-1910
2007_0201_GBR	38-53	175-315	47-68	115-195	310-580	761-1220	060-1910
2007_0247_GBR	38-53	175-315	47-68	115-195	310-580	761-1220	060-1910
2007_0278_GBR	38-53	175-315	47-68	115-195	310-580	761-1220	060-1910
2008_0055_GBR	38-53	175-315	47-68	115-195	310-580	761-1220	060-1910
2008_0381_GBR	38-53	175-315	47-68	115-195	310-580	761-1220	060-1910
2009_0497_GBR	38-53	175-315	47-68	115-195	310-580	761-1220	060-1910
2012_0446_GBR	38-53	175-315	47-68	115-195	310-580	795-1193	1060-1900
2012_0488_GBR	38-53	175-315	47-68	115-195	310-580	795-1193	1060-1900
2012_0548_GBR	38-53	175-315	47-68	115-195	310-580	795-1193	1060-1900
2012_0549_GBR	38-53	175-315	47-68	115-195	310-580	795-1193	1060-1900
2012_0552_GBR	38-53	175-315	47-68	115-195	310-580	761-1220	060-1910
2013_0572_GBR	38-53	175-315	47-68	115-195	310-580	761-1220	060-1910
2015_0561_GBR	38-53	175-315	47-68	115-195	310-580	761-1220	060-1910
2017-0490-GBR	38-53	175-315	47-68	115-195	310-580	761-1220	060-1910
2019_0568_GBR	38-53	175-315	47-68	115-195	310-580	761-1220	060-1910

Попередня обробка вхідних даних

Часова трансформація

Слід зазначити, що дані сонячної активності та дані по паводкам фіксувались з різним часовим інтервалом (табл.2.)

Таблиця 2

Часові інтервали вхідних та вихідного наборів даних

Набори даних	Часовий інтервал
DS^r_{IPF} , DS^r_{IPF}	5 хв
DS^r_{SW}	1 хв
DS^r_{RF}	1 або 3 рази в день
DS^r_F	1 раз в день

Для подальшого аналізу блоки даних для кожної річки були згруповані в окремі набори даних (DS^r), що згруповані до максимального інтервалу в 1 день.

$$DS^r = DS^r_{IPF} \bigcup DS^r_{IPF} \bigcup DS^r_{SW} \bigcup DS^r_{RF} \bigcup DS^r_F \quad (1)$$

Так як згідно гіпотези ми досліджували вплив спалахів сонячної активності на повені, то в якості групування використовувались як максимальне значення max() так і відносний спалах протягом дня:

$$\frac{\max(X_i) - \min(X_i)}{\min(X_i)}$$

Кореляційний аналіз

Так, як вхідні набори даних для різних річок були різними, спочатку була проведена спроба знаходження незалежних функціональних залежностей для кожної річки окремо. Для цього спочатку був проведений кореляційний аналіз між вхідними факторами та опадами з врахуванням часових (лагових) затримок. Для врахування лагової затримки кожен часовий ряд вхідного параметру зсувався по вертикалі вниз на необхідну кількість днів (лагів). Записи, в яких при цьому з'являлись пропущені дані, вилучались. Результати приведені в таблиці 3. Як видно з таблиці, для всіх повеней не існує рівномірних лінійних залежностей між факторами, навіть з урахуванням лагів. Тобто, якщо в якійсь події повені існує висока кореляція для одного з факторів (наприклад, 2001_0645, R = 0,87, Lag = 2), але вона повністю відсутня для інших подій повені. Це свідчить про випадковість цієї залежності.

Таблиця 3

Максимальні значення коефіцієнтів кореляції між вхідними факторами та опадами для лага 0-3

	DS^r_{IPF}		DS^r_{DF} (electron)		DS^r_{DF} (proton)										DS^r_{SW}			DS^r_F
flood events	**IPF** > 10 *MeV*	**IPF** > 30 *MeV*	38 **keV** $\le$ **DF** $\le$ 53 **keV**	175 **keV** $\le$ **DF** $\le$ 31 **keV**	47 **keV** $\le$ **DF** $\le$ 65 **keV**	47 **keV** $\le$ **DF** $\le$ 68 **keV**	65**keV** $\le$ **DF** $\le$ 112 **keV**	112 **keV** $\le$ **DF** $\le$ 187**keV**	115 **keV** $\le$ **DF** $\le$ 195**keV**	310 **keV** $\le$ **DF** $\le$ 580 **keV**	761 **keV** $\le$ **DF** $\le$ 1220	795 **keV** $\le$ **DF** $\le$ 1193 **keV**	1060 **keV** $\le$ **DF** $\le$ 1910 **keV**	1060 **keV** $\le$ **DF** $\le$ 1900 **keV**	**BULK SPEED**	**ION TEMPERATURE**	**PROTON DENSITY**	**10.7 cm Radio Flux**
2001_0645	0.79	0.87	0.69	0.58	0.64	-		0.76	-	0.73	0.77	-	0.75	-	0.48	0.51	0.38	0.64
2002_0463	0.03	0.88	0.96	0.97	-	-	0.93	0.94	-	0.94	0.94	-	0.94	-	0.77	0.90	0.63	0.81
2002_0488	0.31	0.41	0.03	-0.03	-	-	0.59	0.33	-	0.04	0.07	-	0.07	-	0.24	0.37	0.43	0.39
2002_0774	0.05	0.04	0.10	0.06	-	0.30	-	-	0.38	0.36	0.25	-	0.27	-	0.24	0.25	0.53	0.05
2004_0423	0.27	0.27	0.82	0.65	-	-0.15	-	-	0.76	0.85	0.85		0.85		-0.15	0.19	0.06	0.57
2007_0201	-0.22	-0.21	0.05	0.68	-	0.16	-	-	0.11	-0.15	-0.26	-	-0.14	-	0.68	0.72	0.06	-0.39
2007_0247	0.05	0.06	-0.01	0.09	-	0.01	-	-	0.04	0.17	0.55	-	0.37	-	0.38	0.28	0.29	-0.22
2007_0278	0.63	0.73	0.06	0.74	-	-0.02	-		0.08	-0.07	-0.06	-	-0.06	-	0.01	0.83	0.77	-0.05
2008_0055	-0.10	-0.10	0.21	0.07	-	0.69	-	-	0.71	0.49	0.15	-	0.17	-	0.29	0.21	0.08	0.37
2008_0381	0.21	0.76	0.65	0.49	-	0.18	-	-	0.41	0.75	0.69		0.61		0.48	0.72	0.88	-0.02
2009_0497	0.33	0.33	0.04	-0.02	-	0.38	-	-	0.39	0.29	0.19	-	0.21	-	0.28	0.30	0.11	0.68
2012_0446	0.12	0.12	-0.08	-0.02	-	0.86		-	0.87	0.87	-	0.09	-	0.08	0.64	0.40	0.12	0.45
2012_0488	-0.26	-0.26	0.36	-0.21	-	0.39	-		0.08	-0.27	-	0.63	-	0.50	0.76	0.34	-0.27	-0.15
2012_0548	0.31	0.31	0.26	0.34	-	0.29	-	-	0.27	0.31	-	0.19	-	0.19	0.38	0.51	0.21	0.55
2012_0549	0.32	0.28	0.01	0.17	-	0.29	-	-	0.35	0.48	-	0.26	-	0.26	0.45	0.74	0.47	-0.07
2012_0552	0.22	0.14	0.35	0.16	-	-0.09	-	-	-0.10	-0.07	0.05	-	0.17	-	0.21	0.37	0.20	0.03
2013_0572	0.36	0.38	0.12	0.29	-	0.06		-	0.29	0.16	0.18	-	0.61	-	-0.14	-0.09	0.47	-0.05
2015_0561	-0.14	-0.20	-0.18	-0.05	-	0.24	-		0.13	0.09	0.84	-	0.84	-	0.29	0.17	0.52	0.71
2017-0490	0.51	0.53	0.06	0.17	-	0.90	-	-	0.36	-0.12	0.54	-	-0.12	-	0.65	-0.03	0.17	0.37
2019_0568	0.32	0.32	0.22	0.03	-	0.23	-	-	0.09	0.06	-0.07		-0.07		0.96	0.71	0.65	0.19

Модель прогнозу настання повені для кожної річки можна формалізувати так:

$$\text{Precipitations}_r = F\left(X_{r1},\ldots,X_{rm},X_{1,t-1},\ldots,X_{m,t-1},\ldots,X_{1,t-n},\ldots,X_{m,t-n}\right), \quad (2)$$

де r - індекс річки, m - кількість вхідних параметрів, n - максимальний лаг. Слід зазначити, що врахування лагу призводить як до збільшення вхідних параметрів (якщо модель враховує значення певного фактора за декілька днів, а не просто зміщені на лаг), так і до зменшення записів. Оскільки кількість записів для кожної повені коливається від 11 до 38, навіть лаг 2 призводить до того, що кількість вхідних параметрів перевищує кількість записів, що унеможливлює використання як нелінійного, так і лінійного методів. Отже, цю задачу можна вирішити, об'єднавши всі дані в один набір даних:

$$DS = \sum_{r=1}^{20} DS^r \quad (3)$$

Для подальшого аналізу були залишені лише ті характеристики сонячної активності, які присутні у всіх наборах даних DS^r, а саме: $IPF > 10\,MeV$ (X_1), $IPF > 30\,MeV$ (X_2), $38\ \text{keV} \le \text{DF} \le 53\ \text{keV}$ (X_3), $175\ \text{keV} \le \text{DF} \le 315\ \text{keV}$ (X_4), $310\ \text{keV} \le \text{DF} \le 580\ \text{keV}$ (X_5), PROTON DENSITY (X_6), BULK SPEED (X_7), ION TEMPERATURE (X_8), 10.7 cm Radio Flux (X_9).

Проблема дисперсії

Об'єднання наборів даних спричинює появу іншої проблеми – різна дисперсія вихідних даних, адже дані по опадам були отримані для різних річок. Тому подавати на вихід таке поле як опади не є коректним, адже різні річки по різному реагують на кількість опадів і кількість опадів також залежить від географічного розташування річки. Тому було запропоновано в якості вхідних параметрів використовувати не абсолютні значення піків, а їх час настання (положення) на графіках сонячної активності (рис.1). В якості вихідного поля слугувала дата початку паводку. Дні паводку позначались як True. Дні без паводку – False. Фактично задача тоді зводиться до задачі бінарної класифікації де на вхід подаються значення True/False по кожному полю з врахуванням лагу. А на виході теж бінарне поле. Для отримання фінального набору даних для кожної річки розраховувались положення піків для полів сонячної активності, а також фіксувався початок та закінчення паводку. Положення піків визначалось програмно з подальшою ручною верифікацією. Після бінаризації проводилась лагова трансформація цих бінарних наборів даних. Для цього кожне вхідне поле дублювалось та проводився зсув по вертикалі на необхідну кількість лагів. Записи, в яких при цьому з'являлись пропуски – вилучались. Фрагмент результуючого набору даних представлено в таблиці 4.

Таблиця 4

Фрагмент результуючого набору даних для лагу від 0 до 9

> 10 MeV(t-0)	> 30 MeV(t-0)	38-53(t-0)	175-315(t-0)	310-580(t-0)	PROTON DENSITY(t-0)	...	310-580(t-9)	PROTON DENSITY(t-9)	BULK SPEED(t-9)	ION TEMPERATURE(t-9)	10.7 cm Radio Flux(t-9)	days from the beginning of the flood
False	True	False	True	True	False	...	False	False	False	False	False	False
True	False	True	False	False	True	...	True	False	False	False	False	True
False	False	False	False	False	False	...	False	True	False	False	False	True
False	False	False	False	False	False	...	False	False	False	True	True	True
False	False	False	False	False	False	...	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...
False	False	False	False	True	False	...	False	False	False	False	False	False
True	True	False	False	False	False	...	False	True	False	True	False	True
False	False	False	False	False	False	...	True	False	False	False	False	True
False	False	False	False	True	False	...	False	False	False	False	False	False
False	False	False	False	False	False	...	False	False	True	True	False	True

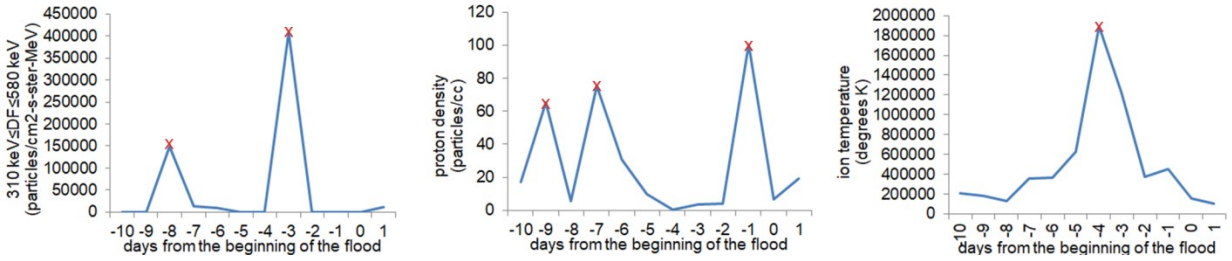

Рис. 1. Приклад встановлення піків

Класифікація та прогноз паводків

Метрики оцінювання

Для оцінки якості навчання моделі існують чотири різні метрики, а саме: 'accuracy', 'precision', 'recall', 'f1'. В нашому випадку є важливим передбачення саме паводку. Ситуація, коли модель помилково передбачає паводок, а на справді його не було є не важливим в нашому випадку. Помилка ж моделі, коли згідно прогнозу паводку не має, а він насправді є – є критичною. Для оцінки саме таких ситуацій служить метрика 'recall'. Саме вона оцінює точність позитивних прогнозів. Інші ж метрики враховують прогноз як настання паводку так і його відсутність. Тому ці метрики апріорі будуть мати вищі значення точності, але вони не є адекватними в нашому випадку.

$$\text{recall} = \text{tp} / (\text{tp} + \text{fn}), \quad (4)$$

де tp — кількість істинних позитивних результатів, а fn — кількість помилкових негативних результатів.

Вибір моделей

На сьогоднішній день існує величезна кількість класифікаційних моделей та не існує однозначного правила їх вибору. Можна зазначити, що переважна кількість класифікаційних моделей не дозволяє описати правила класифікації, чи побудувати дерево рішень. Тому в роботі розглядались два підходи:

1. Побудова прогнозу на основі одного класифікатора - дерева рішень
2. Побудова прогнозу на основі ансамблю моделей.

В першому випадку можна наглядно зрозуміти та обґрунтувати прийняте рішення по класифікації. В другому можна досягнути значно більшої точності прийняття рішення.

Точність моделі визначалась шляхом кроссвалідації, згідно якого навчальний набір ділився випадковим чином на 3 частини. Кожна з цих частин по черзі виступала як тестова. Тобто класифікатор навчався тричі на 3-х різних наборах даних. Для кожного випадку розраховувалась точність тестового та навчального наборів та усереднювалась. Аналіз величин цих метрик дав змогу оцінити точність, адекватність та наявність перенавчання.

Дерево рішення

Для визначення правил класифікації та візуалізації результатів було використано метод дерева рішень [25]. Це непараметричний метод навчання з учителем і один із широко використовуваних алгоритмів класифікації. Алгоритм дерева рішень будує гілки дерева за ієрархічним підходом. Кожна гілка використовує правило «що-якщо» та ділить набір даних на підмножини на основі найважливіших характеристик. Основна ідея дерева рішень полягає в тому, щоб визначити характеристики, які містять найбільше інформації про цільову функцію, а потім розділити набір даних разом із значеннями цих характеристик, щоб значення цільових характеристик у результуючих вузлах були максимально чистими. як можна. Правила вивчаються послідовно, використовуючи навчальні дані одне за одним. Кожного разу, коли вивчається правило, кортежі, які покривають правила, видаляються. Мета полягає в тому, щоб створити модель, яка передбачає значення цільової змінної шляхом вивчення простих правил прийняття рішень, отриманих на основі характеристик даних. Дерево можна розглядати як кусково-постійне наближення. Основною перевагою вибору цього методу є простота розуміння та можливість візуалізації результату, тоді як недоліком є складність роботи з відсутніми даними та те, що він може створювати складні дерева, які можуть бути неефективно класифіковані. Індекс Джині був обраний як критерій для вимірювання порогу розщеплення [26]. Це показник нерівності розподілу деякого значення чисел, який розраховує ймовірність конкретної ознаки, яка класифікується неправильно при виборі. випадковим чином. Стратегія, яка використовується для вибору розподілу в кожному вузлі, полягає в пошуку найкращого розподілу.

Ансамбль моделей

Ансамблеві методи поєднують передбачення з кількох моделей, щоб отримати кращу ефективність прогнозування, ніж можна було б отримати за допомогою будь-якого із складових алгоритмів навчання окремо. Існує три різні способи побудови модельних ансамблів, бегінг, стекінг та бустінг [27, 28]. У цьому дослідженні було використано 25 різних алгоритмів машинного навчання (таблиця 5) з різними параметрами та 3 ансамблі. Також ми протестували ансамблі моделей, заснованих на бустингу (Ada Boost Classifier і Gradient Boosting Classifier) і бегінгу (Bagging Classifier). Після цього ми об’єднали їх усіх в одну остаточну модель ансамблю шляхом жорсткого голосування (таблиця 4):

Таблиця 4

Перелік класифікаторів та ансамблів, які використовувалися в розрахунках

#	**класифікатори**
1	2
1.	DecisionTreeClassifier()
2.	LogisticRegression(random_state=1)
3.	QuadraticDiscriminantAnalysis()
4.	GaussianNB()
5.	RandomForestClassifier(max_depth=5, max_features=1, n_estimators=10)
6.	SVC(decision_function_shape='ovo')
7.	SGDClassifier()
8.	MLPClassifier(alpha=1e-05, hidden_layer_sizes=(20, 10), random_state=1,solver='lbfgs')
9.	ExtraTreesClassifier(random_state=0)
10.	KNeighborsClassifier(n_neighbors=3)
11.	OutputCodeClassifier(estimator=RandomForestClassifier(random_state=0),random_state=0)
12.	OneVsOneClassifier(estimator=LinearSVC(random_state=0))
13.	OneVsRestClassifier(estimator=SVC())
14.	RidgeClassifier()
15.	PassiveAggressiveClassifier(random_state=0)
16.	GaussianProcessClassifier(kernel=1**2 * RBF(length_scale=1), random_state=0)
17.	BernoulliNB()
18.	LabelPropagation()
19.	LabelSpreading()
20.	LinearDiscriminantAnalysis()
21.	LinearSVC(random_state=0, tol=1e-05)
22.	MultinomialNB()

Продовження табл. 4

1	2
23.	NearestCentroid()
24.	Perceptron()
25.	SVC(gamma='auto')
	Ансамблі
26.	AdaBoostClassifier (n_estimators =100, random_state =0)
27.	GradientBoostingClassifier (Learning_rate =1,0, max_depth =1, random_state =0)
28.	BaggingClassifier (base_estimator =SVC(), random_state =0)
29.	VotingClassifier (…,voting='hard')

Результати та дискусія

Лаговий аналіз

Згідно нашої гіпотези, період затримки між паводком та спалахом сонячної активності може сягати 10 днів. Для перевірки цієї гіпотези був проведений наступний експеримент. Всі класифікаційні моделі та ансамблі моделей були навчені та протестовані для вхідних даних, що не містили часової затримки. Далі до вхідних параметрів додавались дані, що містили часову затримку в один день. Після чого моделі заново навчались та розраховувалась метрика recall. Ці ітерації продовжувались до 9 лагів. Формально задачі класифікації зводились до вигляду:

$$lag(0): Flood = F(X_1,\ldots,X_9)$$

$$lag(1): Flood = F(X_1,\ldots,X_9,X_{1,t-1},\ldots,X_{9,t-1})$$

…

$$lag(9): Flood = F(X_1,\ldots,X_9,X_{1,t-1},\ldots,X_{9,t-1},\ldots,X_{1,t-9},\ldots,X_{9,t-9})$$

Отримані результати наведені в таблиці 5 та 6.

Таблиця 5

Точність recall для навчального набору при послідовному додаванні лагів

classifier	Lag 0	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	Lag 7	Lag 8	Lag 9
1	2	3	4	5	6	7	8	9	10	11
DecisionTreeClassifier	0,88	0,97	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
LogisticRegression	0,60	0,65	0,65	0,65	0,72	0,72	0,80	0,85	0,93	1,00
QuadraticDiscriminantAnalysis	0,68	0,70	0,83	0,93	0,97	1,00	1,00	1,00	1,00	1,00
GaussianNB	0,41	0,49	0,58	0,61	0,63	0,72	0,73	0,80	0,88	1,00
RandomForestClassifier	0,75	0,86	0,81	0,77	0,84	0,74	0,74	0,70	0,75	0,86
SVC	1,00	0,80	0,92	0,92	0,93	0,94	0,87	0,87	0,89	0,92
SGDClassifier	0,37	0,42	0,48	0,67	0,64	0,75	0,73	0,91	0,97	1,00
MLPClassifier	0,75	0,93	0,97	0,98	1,00	1,00	1,00	1,00	1,00	1,00
ExtraTreesClassifier	0,88	0,97	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
RadiusNeighborsClassifier	0,67	0,85	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
KNeighborsClassifier	0,63	0,64	0,69	0,80	0,81	0,83	0,87	0,85	0,85	0,90
OutputCodeClassifier	0,79	0,93	0,99	0,99	1,00	1,00	1,00	1,00	1,00	1,00
OneVsOneClassifier	0,58	0,67	0,61	0,65	0,71	0,73	0,82	0,89	0,97	1,00
OneVsRestClassifier	1,00	0,80	0,92	0,92	0,93	0,94	0,87	0,87	0,89	0,92
RidgeClassifier	0,55	0,67	0,62	0,66	0,70	0,74	0,82	0,87	0,96	1,00
PassiveAggressiveClassifier	0,00	0,55	0,60	0,65	0,64	0,66	0,82	0,96	0,98	1,00
GaussianProcessClassifier	0,00	0,00	0,00	0,91	1,00	1,00	0,61	0,65	0,74	0,88
AdaBoostClassifier	0,53	0,63	0,60	0,63	0,70	0,73	0,80	0,85	0,93	1,00
GradientBoostingClassifier	0,50	0,64	0,62	0,66	0,71	0,73	0,82	0,86	0,94	1,00
BaggingClassifier	0,95	0,80	0,93	0,95	0,94	0,94	0,91	0,93	0,90	0,91
BernoulliNB	0,00	0,57	0,59	0,64	0,65	0,71	0,73	0,78	0,89	0,97
LabelPropagation	0,88	0,95	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
LabelSpreading()	0,82	0,94	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
LinearDiscriminantAnalysis	0,50	0,66	0,61	0,66	0,71	0,75	0,81	0,89	0,97	1,00
LinearSVC	0,58	0,67	0,61	0,65	0,71	0,73	0,82	0,89	0,97	1,00
MultinomialNB	0,00	0,58	0,57	0,63	0,65	0,69	0,72	0,76	0,88	0,97
NearestCentroid	0,40	0,47	0,53	0,58	0,65	0,72	0,75	0,84	0,90	0,98
Perceptron	0,27	0,44	0,66	0,53	0,70	0,70	0,66	0,77	0,91	1,00
SVC	0,00	0,00	1,00	0,78	0,73	0,59	0,59	0,65	0,74	0,85
GaussianMixture	0,29	0,39	0,43	0,45	0,50	0,53	0,55	0,67	0,76	0,91

Таблиця 6

Точність recall для тестового набору при послідовному додаванні лагів

classifier	Lag 0	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	Lag 7	Lag 8	Lag 9
DecisionTreeClassifier	0,16	0,37	0,40	0,43	0,45	0,52	0,59	0,64	0,70	0,84
LogisticRegression	0,03	0,29	0,44	0,48	0,53	0,55	0,64	0,74	0,84	0,97
QuadraticDiscriminantAnalysis	0,29	0,32	0,30	0,35	0,38	0,70	1,00	1,00	0,20	0,74
GaussianNB	0,30	0,45	0,51	0,51	0,54	0,61	0,65	0,68	0,75	0,90
RandomForestClassifier	0,08	0,09	0,18	0,38	0,51	0,55	0,92	0,97	1,00	1,00
SVC	0,06	0,20	0,20	0,36	0,49	0,74	0,92	1,00	1,00	1,00
SGDClassifier	0,33	0,35	0,43	0,67	0,49	0,49	0,62	0,64	0,80	0,97
MLPClassifier	0,14	0,34	0,45	0,43	0,53	0,53	0,65	0,68	0,85	0,93
ExtraTreesClassifier	0,15	0,33	0,30	0,41	0,48	0,62	0,80	0,85	0,97	1,00
RadiusNeighborsClassifier										
KNeighborsClassifier	0,22	0,30	0,34	0,32	0,36	0,48	0,54	0,69	0,80	0,95
OutputCodeClassifier	0,14	0,29	0,26	0,34	0,45	0,63	0,81	0,93	0,98	1,00
OneVsOneClassifier	0,07	0,34	0,45	0,51	0,54	0,57	0,53	0,65	0,76	0,92
OneVsRestClassifier	0,06	0,20	0,20	0,36	0,49	0,74	0,92	1,00	1,00	1,00
RidgeClassifier	0,05	0,32	0,46	0,50	0,53	0,53	0,57	0,58	0,69	0,87
PassiveAggressiveClassifier	0,23	0,39	0,53	0,49	0,35	0,55	0,64	0,60	0,78	0,93
GaussianProcessClassifier	0,00	0,00	0,00	0,00	0,30	0,59	1,00	1,00	1,00	1,00
AdaBoostClassifier	0,12	0,34	0,44	0,50	0,54	0,58	0,63	0,65	0,76	0,91
GradientBoostingClassifier	0,08	0,36	0,47	0,52	0,57	0,55	0,64	0,68	0,81	0,93
BaggingClassifier	0,04	0,19	0,17	0,29	0,48	0,59	0,85	0,99	1,00	1,00
BernoulliNB	0,02	0,36	0,42	0,46	0,52	0,62	0,67	0,75	0,81	0,96
LabelPropagation	0,10	0,30	0,34	0,30	0,46	0,51	0,56	0,64	0,75	0,84
LabelSpreading()	0,14	0,31	0,36	0,36	0,49	0,52	0,58	0,63	0,77	0,86
LinearDiscriminantAnalysis	0,07	0,36	0,46	0,51	0,53	0,50	0,53	0,54	0,59	0,59
LinearSVC	0,07	0,34	0,45	0,51	0,54	0,57	0,53	0,65	0,76	0,92
MultinomialNB	0,00	0,23	0,35	0,42	0,53	0,65	0,69	0,75	0,84	0,97
NearestCentroid	0,63	0,54	0,59	0,58	0,55	0,56	0,62	0,62	0,73	0,88
Perceptron	0,33	0,33	0,39	0,34	0,51	0,56	0,58	0,60	0,78	0,92
SVC	0,00	0,00	0,00	0,09	0,25	0,99	1,00	1,00	1,00	1,00
GaussianMixture	0,30	0,43	0,54	0,43	0,47	0,28	0,33	0,67	0,37	0,48

Ці таблиці дають змогу проаналізувати динаміку зміни метрики при послідовному врахуванні нових лагів до вхідних параметрів. Значення recall порівнювалось для тестового та навчального наборів. Аналіз адекватності моделей оцінювався за такими ознаками: якщо помилка тестового та навчального наборів є близькою (маленька дисперсія) – це свідчить про те що модель добре навчилась та прогнозує невідомі значення на рівні відомих. А абсолютне значення свідчить наскільки точною є така модель. Якщо ж точність на навчальному наборі сягає 1, а на тестовому близька до 0.5 – це явна ознака перенавчання. Тобто ідеально прогнозуються відомі дані, а невідомі вгадуються 50/50 – абсолютна нездатність прогнозувати. Такі моделі мають бути усунені з аналізу. (таблиця 7)

Згідно з цією таблицею можна бачити лише один класифікатор не справився із задачею – RadiusNeighborsClassifier. У всіх інших точність прогнозу зростає із збільшенням лагу. Це означає, що дійсно є присутньою суттєва часова затримка між спалахом на сонці та настанням паводку. Для аналізу, які фактори важливі при такому прогнозуванні, побудуємо дерево рішень рис 2. (recall = 0.84)

Таблиця 7

Дисперсія похибки між тестовим та навчальним наборами даних

classifier	Lag 0	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	Lag 7	Lag 8	Lag 9
1	2	3	4	5	6	7	8	9	10	11
DecisionTreeClassifier	0,82	0,62	0,60	0,57	0,55	0,48	0,41	0,36	0,30	0,16
LogisticRegression	0,96	0,56	0,32	0,26	0,27	0,24	0,20	0,14	0,09	0,03
QuadraticDiscriminantAnalysis	0,58	0,54	0,63	0,62	0,61	0,30	0,00	0,00	0,80	0,26
GaussianNB	0,27	0,09	0,12	0,16	0,14	0,15	0,11	0,15	0,14	0,10
RandomForestClassifier	0,90	0,89	0,78	0,50	0,39	0,26	-0,24	-0,38	-0,33	-0,16
SVC	0,94	0,75	0,78	0,61	0,47	0,21	-0,06	-0,14	-0,12	-0,08
SGDClassifier	0,11	0,17	0,10	-0,01	0,23	0,36	0,16	0,30	0,18	0,03
MLPClassifier	0,81	0,64	0,54	0,56	0,47	0,47	0,35	0,32	0,15	0,07
ExtraTreesClassifier	0,83	0,66	0,70	0,59	0,52	0,38	0,20	0,15	0,03	0,00
RadiusNeighborsClassifier	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
KNeighborsClassifier	0,65	0,54	0,51	0,60	0,55	0,42	0,38	0,18	0,06	-0,06

Продовження табл. 7

1	2	3	4	5	6	7	8	9	10	11
OutputCodeClassifier	0,82	0,69	0,74	0,66	0,55	0,37	0,19	0,07	0,02	0,00
OneVsOneClassifier	0,88	0,49	0,26	0,22	0,24	0,22	0,36	0,26	0,21	0,08
OneVsRestClassifier	0,94	0,75	0,78	0,61	0,47	0,21	-0,06	-0,14	-0,12	-0,08
RidgeClassifier	0,91	0,52	0,26	0,24	0,24	0,29	0,30	0,34	0,28	0,13
PassiveAggressiveClassifier		0,30	0,12	0,25	0,46	0,16	0,22	0,37	0,21	0,07
GaussianProcessClassifier				1,00	0,70	0,41	-0,64	-0,53	-0,36	-0,14
AdaBoostClassifier	0,78	0,46	0,26	0,20	0,22	0,21	0,21	0,23	0,18	0,09
GradientBoostingClassifier	0,83	0,45	0,25	0,22	0,20	0,24	0,22	0,21	0,15	0,07
BaggingClassifier	0,96	0,77	0,82	0,70	0,49	0,37	0,07	-0,07	-0,11	-0,09
BernoulliNB		0,36	0,30	0,28	0,20	0,13	0,08	0,04	0,09	0,02
LabelPropagation	0,89	0,69	0,66	0,70	0,54	0,49	0,44	0,36	0,25	0,16
LabelSpreading()	0,84	0,67	0,64	0,64	0,51	0,48	0,42	0,37	0,23	0,14
LinearDiscriminantAnalysis	0,86	0,46	0,25	0,23	0,25	0,33	0,35	0,39	0,39	0,41
LinearSVC	0,88	0,49	0,26	0,22	0,24	0,22	0,36	0,26	0,21	0,08
MultinomialNB		0,60	0,39	0,34	0,17	0,06	0,05	0,00	0,04	0,01
NearestCentroid	-0,55	-0,15	-0,12	0,01	0,15	0,22	0,17	0,26	0,19	0,10
Perceptron	-0,23	0,26	0,40	0,35	0,27	0,20	0,12	0,21	0,15	0,08
SVC			1,00	0,89	0,66	-0,68	-0,70	-0,53	-0,36	-0,18
GaussianMixture	-0,02	-0,10	-0,27	0,06	0,06	0,47	0,39	0,01	0,52	0,47

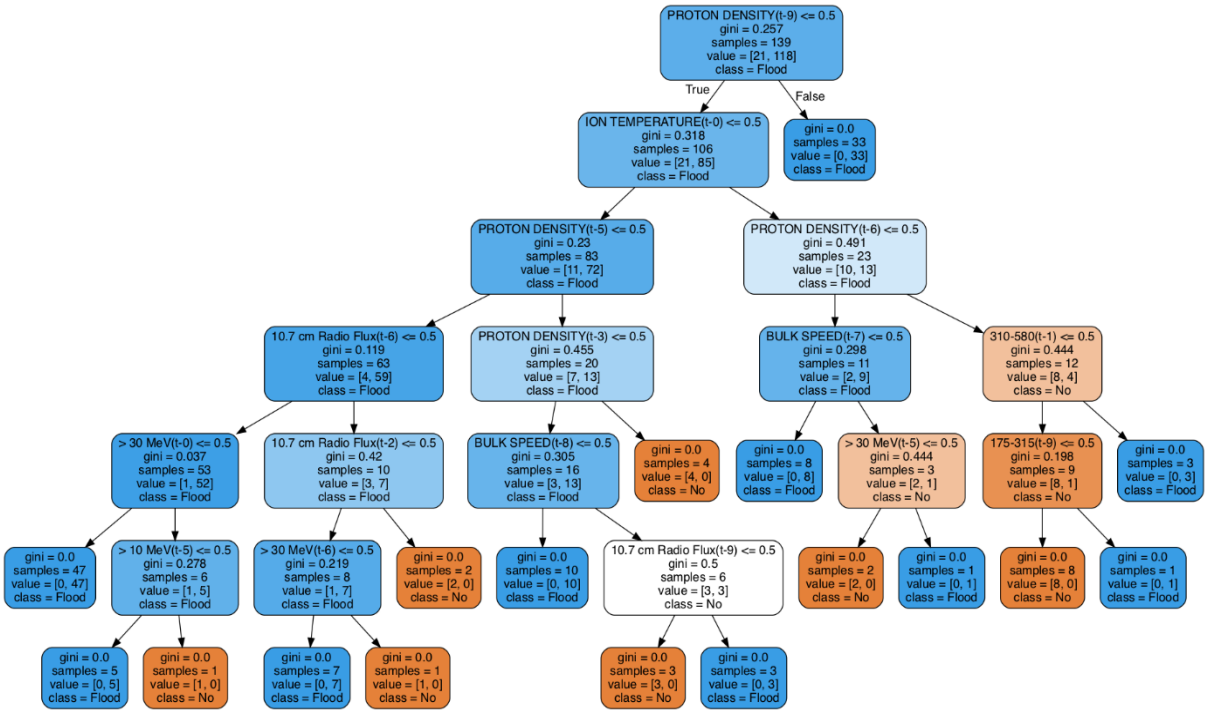

Рис. 2. Дерево рішень прогнозу паводку при врахування лагової затримки від 0 до 9 днів

Важливим при класифікації є Індекс Джині, також відомий як домішка Джіні, обчислює кількість імовірності певної ознаки, яка неправильно класифікована при випадковому виборі. Якщо всі елементи пов'язані з одним класом, то його можна назвати чистим. Як видно з рисунку, для встановлення паводку першою перевіркою є Proton Density із затримкою в 9 днів. Якщо спалаху інтенсивності цього фактору в цей день не спостерігалося тоді перевіряється Ion Temperature із затримкою 0 днів. Якщо ж спостерігається спалах на Proton Density, то із 100% вірогідністю має настати паводок. На основі цього дерева рішень можна визначити і важливість факторів (таблиця 8):

Як видно з таблиці, не дивлячись на те, що перша перевірка стосується PROTON DENSITY(t-9), найбільш вагомими факторами є PROTON DENSITY(t-3), 310-580(t-1) та ION TEMPERATURE(t-0). Видно також, що різні фактори впливають на настання паводку з різними часовими затримками. Також видно, що спалах такого фактору як PROTON DENSITY може призводити до повеней з різними часовими затримками. Або необхідно декілька спалахів, щоб це призвело до паводку. Слід зазначити, що точність цього класифікатора на навчальному наборі становить 1, а на тестовому 0.84.

Таблиця 8

Найбільш вагомі фактори при класифікації:

Фактор (лаг)	**Важливість**
PROTON DENSITY(t-3)	0.19
310-580(t-1)	0.10
ION TEMPERATURE(t-0)	0.09
10.7 cm Radio Flux(t-9)	0.08
PROTON DENSITY(t-6)	0.08
PROTON DENSITY(t-5)	0.07
10.7 cm Radio Flux(t-2)	0.07
PROTON DENSITY(t-9)	0.06
BULK SPEED(t-7)	0.05

Побудова прогнозних моделей

Для побудови прогнозу паводків на n днів наперед необхідно вилучити із вхідних параметрів дані з лагами [0-(n-1)]:

$$Forecast\left(1\,day\right): Flood = F\left(X_{1,t-1},\ldots,X_{9,t-1},\ ,\ldots,X_{9,t-1},\ldots,X_{1,t-9},\ldots,X_{9,t-9}\right)$$

…

$$Forecast\left(9\,days\right): Flood = F\left(X_{1,t-9},\ldots,X_{9,t-9}\right)$$

Таблиця 9

Точність recall для прогнозних моделей для тестового набору даних

classifier	Lag 0	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	Lag 7	Lag 8	Lag 9
DecisionTreeClassifier	0,85	0,81	0,83	0,83	0,82	0,76	0,75	0,79	0,82	0,84
LogisticRegression	0,97	0,97	0,97	0,97	0,97	0,96	0,94	0,95	0,99	1,00
QuadraticDiscriminantAnalysis	0,68	0,44	1,00	1,00	1,00	1,00	1,00	1,00	0,00	0,00
GaussianNB	0,90	0,90	0,90	0,90	0,90	0,86	0,84	0,84	0,84	0,84
RandomForestClassifier	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,99	1,00
SVC	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,99	1,00	0,92
SGDClassifier	0,93	0,91	0,91	0,91	0,91	0,90	0,91	0,92	0,89	0,88
MLPClassifier	0,94	0,92	0,94	0,96	0,92	0,91	0,89	0,91	0,92	0,87
ExtraTreesClassifier	1,00	0,99	1,00	0,99	0,99	0,97	0,95	0,91	0,90	0,87
RadiusNeighborsClassifier	0,95	0,91	0,89	0,88	0,86	0,92	0,84	0,78	0,78	0,80
KNeighborsClassifier	1,00	0,99	1,00	1,00	1,00	1,00	0,97	0,96	0,91	0,92
OutputCodeClassifier	0,90	0,91	0,89	0,87	0,86	0,90	0,89	0,92	0,91	0,92
OneVsOneClassifier	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,99	1,00	0,92
OneVsRestClassifier	0,85	0,83	0,84	0,84	0,85	0,86	0,90	0,94	0,98	1,00
RidgeClassifier	0,94	0,93	0,93	0,91	0,88	0,90	0,87	0,91	0,91	1,00
PassiveAggressiveClassifier	1,00	0,99	1,00	1,00	1,00	0,98	1,00	1,00	1,00	1,00
GaussianProcessClassifier	0,95	0,92	0,91	0,92	0,91	0,91	0,89	0,90	0,85	0,84
AdaBoostClassifier	0,86	0,85	0,80	0,78	0,80	0,83	0,79	0,76	0,75	0,91
GradientBoostingClassifier	0,87	0,88	0,83	0,80	0,82	0,86	0,84	0,79	0,75	0,91
BaggingClassifier	0,62	0,62	0,65	0,68	0,74	0,84	0,85	0,89	0,91	1,00
BernoulliNB	0,90	0,91	0,89	0,87	0,86	0,90	0,89	0,92	0,91	0,92
LabelPropagation	0,96	0,93	0,95	0,93	0,94	0,93	0,91	0,92	0,91	1,00
LabelSpreading()	0,90	0,87	0,85	0,80	0,81	0,80	0,76	0,71	0,72	0,67
LinearDiscriminantAnalysis	0,93	0,95	0,92	0,91	0,93	0,86	0,86	0,86	0,84	0,87
LinearSVC	1,00	1,00	1,00	1,00	1,00	1,00	1,00	.00	1,00	1,00
MultinomialNB	0,91	0,94	0,91	0,90	0,93	0,89	0,91	0,91	0,88	0,86
NearestCentroid	0,91	0,91	0,89	0,85	0,86	0,88	0,86	0,88	0,91	0,91
Perceptron	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,99	0,94	0,87
VotingClassifier	0,97	0,97	0,97	0,97	0,97	0,94	0,92	0,93	0,92	0,92

Як можна бачити, кількість вхідних параметрів буде зменшуватись, а отже це має призвести до зменшення точності прогнозу. В роботі були проаналізовані точності прогнозних класифікаційних моделей від 0 до 9 днів наперед. Також був побудований ансамбль моделей, що поєднував всі моделі шляхом hard voting та проаналізована динаміка його точності в залежності від затримки пронозу. Результати представлені в таблиці 9 та на рисунку 3.

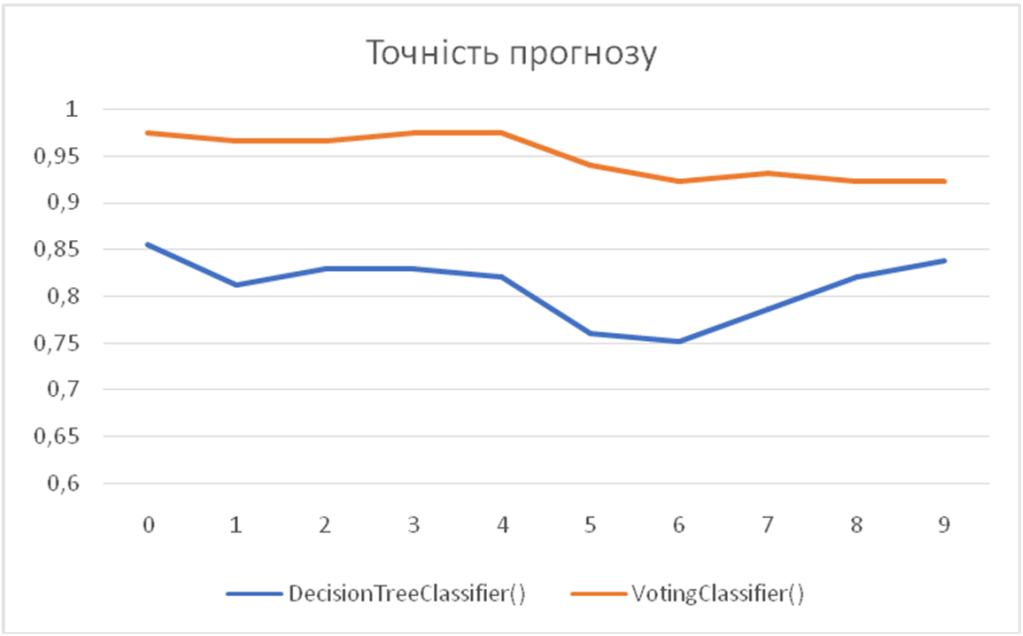

Рис. 3. Зміна точності класифікаційних моделей Decision Tree та VotingClassifier залежно від дальності пронозу

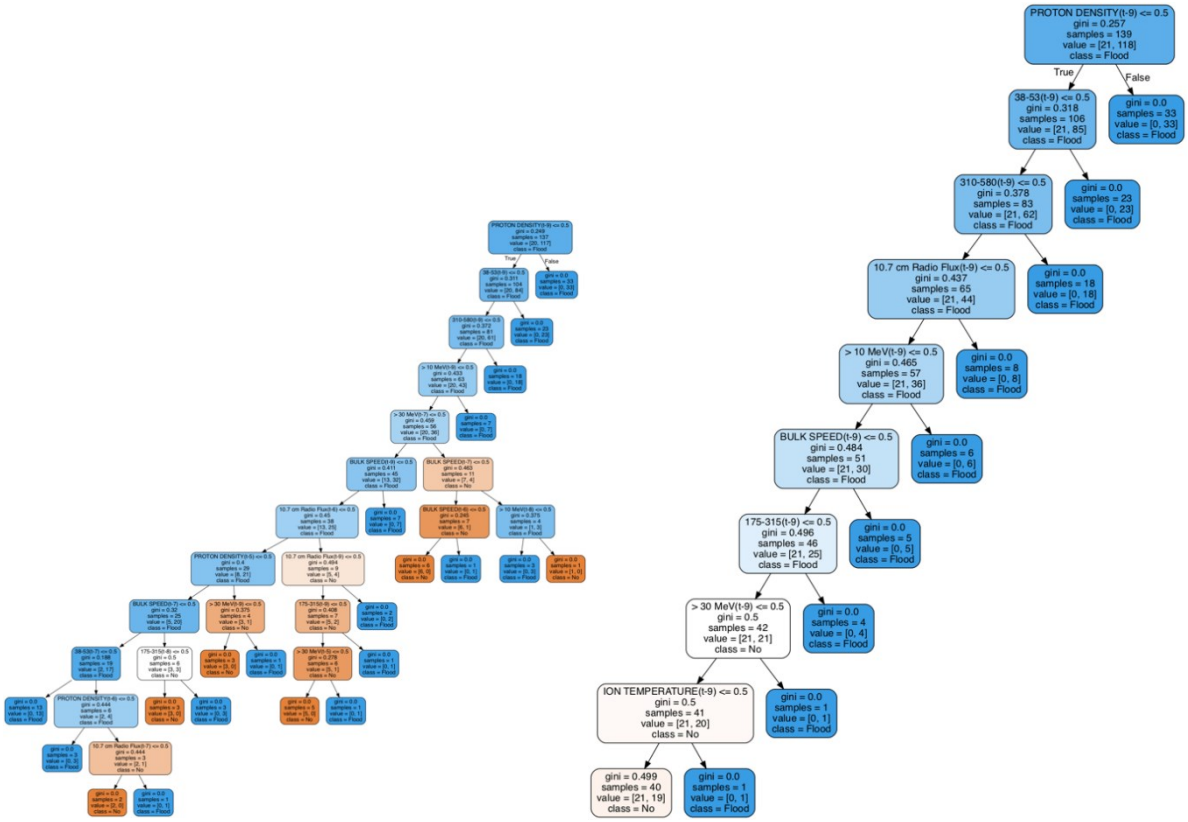

Рис. 4. Дерево рішень на 9 днів вперед

Як видно з таблиці та рисунку точність ансамблю моделей VotingClassifier є найвищою та поступово спадає із збільшенням дальності пронозу. Слід зазначити, що точність на тестовому наборі 0.97 свідчить про високу точність та відсутність перенавчання. Тобто такий ансамбль моделей може бути використаний для прогнозу паводків до 9 днів наперед. Недоліком є те, що на основі нього не можна побудувати дерево рішення. Точність Decision Tree коливається на одному рівні в межах похибки моделі. Це пояснюється тим, що перший критерій перевірки це PROTON DENSITY з лагом 9. Тому і точність моделі практично не залежить від вилучення факторів з малими лагами. Такий підхід дозволяє побудувати дерево рішень для прогнозу на будь який лаг. Тобто в такому підході прогноз для лагів від 0 до 9 вимагає побудови 10 різних дерев рішень. Для випадку 0-9 лагів дерево рішень побудоване на рис 3. Побудуємо для прикладу дерево для прогнозу на 5 та 9 днів наперед: (рис. 4)

Розрахована важливість факторів наведена в таблиці 10.

Таблиця 10

Найбільш вагомі фактори при прогнозуванні на 5 та 9 днів наперед:

Прогноз на 5 днів		**Прогноз на 9 днів**	
Фактор (лаг)	Важливість	Фактор (лаг)	Важливість
BULK SPEED (t-7)	0,10	$310\ \text{keV} \le DF \le 580\ \text{keV}$ (t-9)	0,19
175-315(т-8)	0,09	$38\ \text{keV} \le DF \le 53\ \text{keV}$ (t-9)	0,15
310-580(t-9)	0,08	PROTON DENSITY (t-9)	0,13
38-53(t-9)	0,06	10,7 см Radio Flux (t-9)	0,12
> 30 МеВ(t-7)	0,06	BULK SPEED (t-9)	0,12
PROTON DENSITY (t-5)	0,06	$175\ \text{keV} \le DF \le 315\ \text{keV}$ (t-9)	0,12
PROTON DENSITY (t-9)	0,05	*IPF* > 10 МеВ(t-9)	0,12
BULK SPEED (t-6)	0,05	ION TEMPERATURE (t-9)	0,03
> 30 МеВ(t-5)	0,05	*IPF* > 30 МеВ(t-9)	0,03

Як видно з рисунка 4, перший тест для PROTON DENSITY залишається, але найважливішими факторами для прогнозування повені за 9 днів є $310\ \text{keV} \le DF \le 580\ \text{keV}$, і $38\ \text{keV} \le DF \le 53\ \text{keV}$. Таким чином, на основі отриманого ансамблю класифікаційних моделей можна прогнозувати до 9 днів наперед з точністю 92%. А за допомогою дерева рішень можна обгрунтовувати та будувати рекомендації для прогнозування паводків.

Потенційне теоретичне (фізичне) пояснення механізму, який міг би пояснити розглянуту взаємодію в цій статті, було представлено кількома авторами. Відповідно до [29, 30], високоенергетичні частинки від Сонця захоплюють повітряні маси гідродинамічним тиском і безпосередньо впливають на атмосферні процеси. Якщо в місці контакту з повітряними масами відбувається насичення вологою, то можуть утворюватися хмари і випадати опади, при цьому механізм утворення опадів пояснюється принципом валентності електронів. Автори стверджують, що поява хмар і опадів, а також поява спекотних хвиль і сухих періодів зумовлена насамперед електромагнітними характеристиками сонячного вітру, розташуванням Сонця, від якого він випромінюється, і його хімічна структура. Вищевказаний механізм пояснюється циркуляцією векторів міжпланетних магнітних полів. Prikryl [31] обговорили дві раптові повені в Словаччині, що відбулися після прибуття двох високошвидкісних потоків сонячного вітру з корональних дір. В наступній праці ці автори довели, що сильні опади, що призводять до повеней і раптових повеней у Японії, Австралії та континентальній частині США, як правило, слідують за надходженням високошвидкісних потоків сонячного вітру з корональних дір. Вони припустили, що низхідні атмосферні гравітаційні хвилі можуть спровокувати утворення серії конвективних осередків, які спричинили сильні опади та повені. Відповідно до цих раніше опублікованих результатів, статистичні результати, представлені в цьому дослідженні, показують, що повені, спричинені опадами, мають тенденцію слідувати за раптовими надходженнями потоків сонячно заряджених частинок.

Висновки

Навіть за відсутності пояснення фізичних механізмів, встановлення відповідного прихованого зв'язку залежності, що дозволяє розглянути вплив сонячної активності на процеси навколишнього середовища, такі як повені, спричинені опадами, є внеском у цю область досліджень. Відповідно до раніше опублікованих результатів, це дослідження з використанням ансамблевого підходу прогнозного моделювання класифікації машинного навчання надало докази того, що повені, спричинені опадами, у Великобританії, як правило, слідують за спалахом сонячного вітру. Доведено, що на основі виявлення раптових потоків сонячно заряджених частинок можна очікувати появу опадів, які призведуть до повеней, на кілька днів вперед. Встановлено, що найважливішими факторами для прогнозування повеней є густина протонів із затримкою 9 днів, диференціальний потік протонів у діапазоні 310-580 кеВ та температура іонів. Дослідження в цій статті показали, що ансамбль моделей класифікації є точним і адекватним і може бути використаний для прогнозування появи спричинених опадами повеней до 9 днів наперед з точністю 92%. Враховуючи широкі можливості його практичного застосування, можна констатувати, що значення геліоцентричного електромагнітного довгострокового прогнозу погоди надзвичайно велике, якби можна було з набагато більшою точністю обробляти інформацію про те, яку погоду слід очікувати протягом наступних кількох місяців [33].

Література

1. Gaume, E.; Bain, V.; Bernardara, P.; Newinger, O.; Barbuc, M.; Bateman, A.; Blaškovi˘cová, L.; Bloschl, G.; Borga, M.; Dumitrescu, A.; et al. A compilation of data on European flash floods. J. Hydrol. 2009, 367, 70–78. [CrossRef]

2. Weronika Nitka, Krzysztof Burnecki 2019 Impact of solar activity on precipitation in the United States Physica A 527 (2019) 121387

3. Milovanović, B., Radovanović, M. (2009). Повезаност Сунчеве активности и циркулације атмосфере у периоду 1891-2004, *Journal of the Geographical Institute "Jovan Cvijić" SASA, 59/1, 35-48.*

4. Ma, L.H., Y.B. Han, Z.Q. Yin, Possible influence of the 11-year solar cycle on precipitation in huashan mountain of China over the last 300 years, Earth Moon Planets 107 (2010) 219–224, http://dx.doi.org/10.1007/s11038-010-9367-y.

5. Wilcox, J.M.; Scherrer, P.H.; Svalgaard, L.; Roberts,W.O.; Olson, R.H. Solar magnetic sector structure: Relation to circulation of the Earth's atmosphere. Science 1973, 180, 185–186. [CrossRef]

6. Wilcox, J.M.; Scherrer, P.H.; Svalgaard, L.; Roberts, W.O.; Olson, R.H.; Jenne, R.L. Influence of solar magnetic sector structure on terrestrial atmospheric vorticity. J. Atmos. Sci. 1974, 31, 581–588. [CrossRef]

7. Maliniemi, V., T. Asikainen, K. Mursula, Decadal variability in the Northern Hemisphere winter circulation: Effects of different solar and terrestrialdrivers, J. Atmos. Sol.-Terr. Phys. 179 (2018) 40–54, http://dx.doi.org/10.1016/j.jastp.2018.06.012.

8. Vyklyuk, Y., Radovanovic, M., Milovanovic, B., Leko, T., Milenkovic, M., Milo_sevi_c, Z., Milanovic Pesic, A., Jakovljevic, D., 2016. Hurricane genesis modelling based on the relationship between solar activity and hurricanes. Nat. Hazards. http://dx.doi.org/10.1007/s11069-016-2620-6.

9. Vyklyuk, Y., Radovanovi´c, M.M., Stanojevi´c, G.B., Milovanovi´c, B., Leko, T., Milenkovi´c, M., Petrovi´c, M., Yamashkin, A.A., Milanovi´c Peši´c, A., Jakovljevi´c, D., Malinovic Milicevic, S.: J. Atmos. Sol.-Terr. Phys. 180, 159–164 (2017) Hurricane genesis modelling based on the relationship between solar activity and hurricanes II. https://doi.org/10.1016/j.jastp.2017.09.008

10. Srećković, V., Šulić, M., Vujčić, V., Jevremović, D., Vyklyuk, V. (2017). The Effects of Solar Activity: Electrons In the Terrestrial Lower Ionosphere. *J. Geogr. Inst. Cvijic. 67(3) (221–233).*

11. Nina., A, Čadež, V.M., Bajčetić, J., Andrić, M., Jovanović, G. (2017). Responses Of The Ionospheric D-Rregion To Periodic And Transient Variations Of The Ioniying Solar Lyα Radiation. *J. Geogr. Inst. Cvijic. 67(3) (235–248)*

12. Haigh, J. D.: The impact of solar variability on climate, Science, 272, 981–984, 1996.

13. Svensmark H, Friis-Christensen E (1997) Variation of cosmic ray flux and global cloud coverage—A missing link in solar–climate relationships. J Atmos Sol-Terr Phys 59:1225–1232.

14. Carslaw KS, Harrison RG, Kirkby J (2002) Cosmic rays, clouds, and climate. Science 298:1732–1737.

15. Gray, L.J. Beer, M., Geller,J. D. Haigh,M. Lockwood, K. Matthes, U. Cubasch, D. Fleitmann,G. Harrison, L. Hood, J. Luterbacher,G. A. Meehl,D. Shindell, B. van Geel, and W. White., Solar influences on climate, Rev. Geophys. 48 (2010) RG4001, http://dx.doi.org/10.1029/2009RG000282.

16. Solheim, J.-E., K. Stordahl, O. Humlum, The long sunspot cycle 23 predicts a significant temperature decrease in cycle 24, J. Atmos. Sol.-Terr. Phys. 80 (2012) 267–284, http://dx.doi.org/10.1016/j.jastp.2012.02.008.

17. Veretenenko, S., Thejll, P., 2013. Influence of energetic Solar Proton Events on the development of cyclonic processes at extratropical latitudes. J. Phys. Conf. Ser. 409, 012237. http://dx.doi.org/10.1088/1742-6596/409/1/012237

18. Veretenenko,S.,Thejll,P.,2004.Effects of energetic solar proton events on the cyclone development in the North Atlantic.J.Atmos.Sol.-Terr.Phys.66(5), 393–405.

19. Bhattacharyya S, Narasimha R. (2005): Possible association between Indian monsoon rainfall and solar activity. *Geophysical Research Letters,* vol. 32, L05813, AGU.

20. Kirby, C. & Marsh, T. J. (eds). 1990. Water Quality in the Environment. Natural Environment Research Council: Swindon; 34.

21. UK Centre for Ecology and Hydrology. National River Flow Archive (2022). UK River and Flow Regime. https://nrfa.ceh.ac.uk/uk-river-flow-regimes accessed: 02-15-2022

22. Kingston, D.G., McGregor, G.R., Hannah, D.M. & Lawler, D.M. (2007). Large-scale climatic controls on New England river flow. *Journal of Hydrometeorology* 8(3), 367–379. doi: DOI: 10.1175/JHM584.1

23. Laizé, C.L:R. & Hannah, D.M. (2010). Modification of climate–river flow associations by basin properties. *Journal of Hydrology* 389, 186–204. DOI:10.1016/j.jhydrol.2010.05.048

24. Hannaford, J. & Buys, G. (2012). Trends in seasonal river flow regimes in the UK. *Journal of Hydrology* 475, 158–174.

25. Safavian, S.R.; Landgrebe D. A survey of decision tree classifier methodology. Systems Man and Cybernetics IEEE Transactions 1991, vol. 21, no. 3, pp. 660-674.].

26. Laura Elena Raileanu & Kilian Stoffel. Theoretical Comparison between the Gini Index and Information Gain Criteria. Annals of Mathematics and Artificial Intelligence 2004, vol. 41, pp. 77–93. doi: 10.1023/B:AMAI.0000018580.96245.c6.

27. Graczyk M., Lasota T., Trawiński B., Trawiński K. (2010) Comparison of Bagging, Boosting and Stacking Ensembles Applied to Real Estate Appraisal. In: Nguyen N.T., Le M.T., Świątek J. (eds) Intelligent Information and Database Systems. ACIIDS 2010. Lecture Notes in Computer Science, vol 5991. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-12101-2_35

28. Trostianchyn, A., Duriagina, Z., Izonin, I., Tkachenko, R., Kulyk, V., & Pavliuk, O. (2021). Sm-Co ALLOYS COERCIVITY PREDICTION USING STACKING HETEROGENEOUS ENSEMBLE MODEL. *Acta Metallurgica Slovaca*, *27*(4), 195-202. https://doi.org/10.36547/ams.27.4.1173

29. Stevancevic M, Radovanovic M, Todorovic N. (2004): The Possibility of Application of Electromagnetic Method in Mid term Weather Forecasting. Collection of Papers EkoIst'04 Ecological Truth, 30. 05. – 02. 06. 2004, Bor, p. 396- 399, (in Serbian).

30. Stevancevic M, Radovanovic M, Todorovic N. (2006): Analysis of characteristic mistakes in the heliocentric electromagnetic long-term forecast. “Tourist Valorisation of Tara”, Theme Collection of Geographic Institute “Jovan Cvijic” Serbian Academy for Sciences and Art and Sport-Recreative Center Bajina Bašta, Belgrade, p. 101-110, (in Serbian).

31. Prikryl, P.; Bruntz, R.; Tsukijihara, T.; Iwao, K.; Muldrew, D.B.; Rušin, V.; Rybanský, M.; Turˇ na, M.; Št’astný, P. Tropospheric weather influenced by solar wind through atmospheric vertical coupling downward control. J. Atmos. Sol. Terr. Phys. 2018, 171, 94–110. [CrossRef]

32. Prikryl, P.; Rušin, V.; Prikryl, E.A.; Št’astný, P.; Turˇ na, M.; Zele ˇnáková, M. Heavy rainfall, floods, and flash floods influenced by high-speed solar wind coupling to the magnetosphere-ionosphere-atmosphere system. Ann. Geophys. 2021, 39, 769–793.

33. Radovanovic, M., Ducic, V., Mukherjee, S., 2014. Climate changes instead of global warming. Therm. Sci. 18 (3), 1055–1061.