

УДК 504.3.054

О.В. ШРУБКОВСЬКИЙ*

ОГЛЯД МЕТОДІВ АСИМІЛЯЦІЇ ДАНИХ ДЛЯ УТОЧНЕННЯ РЕЗУЛЬТАТІВ МОДЕЛЕЙ РОЗПОВСЮДЖЕННЯ ЗАБРУДНЕНЬ ПІСЛЯ АВАРІЙНИХ ВИКИДІВ

*Інститут проблем математичних машин і систем НАН України, м. Київ, Україна

Анотація. Асиміляція даних (DA) є важливою задачею в прогнозуванні забруднень, оскільки вона дозволяє інтегрувати спостереження з моделями, покращуючи точність прогнозів поширення забруднювачів в атмосфері, океані та на суші. Це особливо важливо для оцінки впливу аварійних викидів і управління екологічними ризиками. У статті здійснено огляд та порівняння чотирьох різних методів асиміляції даних у задачі розповсюдження забруднень після аварійних викидів. Методи містять у собі ансамблевий фільтр Калмана (EnKF), ансамблевий згладжувач (ES) та два нових підходи, що базуються на композиції з машинним навчанням (ML). Обґрунтовано математичну складову та проаналізовано переваги і недоліки кожного з них. Ансамблеві методи асиміляції даних, зокрема EnKF, є обчислювально ефективними та можуть давати хороші результати навіть у нелінійних моделях. Вони потребують менше ресурсів порівняно з класичними методами, зберігаючи важливу інформацію при обмеженому розмірі ансамблю. Однак їхнім недоліком є гаусове наближення, що може призвести до числових нестабільностей і нефізичних результатів у задачах із не-гаусівським розподілом. Крім того, необхідність численних перезапущів може збільшити обчислювальні витрати. Альтернативою є ES, який не потребує рекурсивних оновлень і скорочує час моделювання, але може давати гірші результати в деяких задачах, хоча його обчислювальна ефективність вища. Розглянуто два гібридних методи, що поєднують асиміляцію даних із машинним навчанням: перший полягає в корекції тенденцій або резольвенти, що має композиційну природу і показує кращі результати на типових задачах у порівнянні з аналогічними методами без корекцій, другий полягає в нейронній мережі з нестандартною функцією втрат. Обидва методи продемонстрували здатність коригувати модельні похибки та враховувати зміщення в прогнозах. Гібридні підходи, що комбінують традиційні методи асиміляції з машинним навчанням, забезпечують високу точність, знижуючи кількість циклів DA-ML.

Ключові слова: асиміляція даних, машинне навчання, нейронні мережі, ансамблевий фільтр Калмана.

Abstract. Data assimilation (DA) is a crucial task in pollution forecasting, as it enables the integration of observations with models, improving the accuracy of pollutant dispersion predictions in the atmosphere, ocean, and land. This is particularly important for assessing the impact of accidental emissions and managing environmental risks. This paper reviews and compares four different data assimilation methods for pollution dispersion following accidental releases. The methods include ensemble-based approaches: the ensemble Kalman filter (EnKF), Ensemble Smoother (ES), and two novel approaches based on machine learning (ML). The mathematical foundations of these methods are justified, and their advantages and drawbacks are analyzed. Ensemble data assimilation methods, particularly EnKF, are computationally efficient and can provide good results even in non-linear models. They require fewer resources compared to classical methods while preserving critical information with a limited ensemble size. However, their drawback is the Gaussian approximation, which can lead to numerical instabilities and non-physical results in problems with non-Gaussian distributions. Additionally, the need for multiple reinitializations can increase computational costs. An alternative is ES, which does not require recursive updates and reduces modeling time but may produce worse results in certain cases, despite its higher computational efficiency. Two hybrid methods that combine data assimilation with machine learning have been examined: the first involves correcting tendencies or resolvents, leveraging a compositional structure that outperforms similar methods without corrections in benchmark problems; the second utilizes a neural network with a non-standard loss function. Both methods have demonstrated the ability to correct model errors and account for biases in forecasts. Hybrid approaches that integrate traditional assimilation techniques with machine learning offer

high accuracy while reducing the number of DA-ML cycles.

Keywords: data assimilation, machine learning, neural networks, ensemble Kalman filter.

DOI: 10.34121/1028-9763-2025-1-113-123

1. Вступ

Асиміляція даних передбачає поєднання спостережень з апіорними знаннями для отримання оцінки істинного стану системи та пов'язаної з цією оцінкою невизначеності (наприклад, огляд у [1]). Хоча асиміляція даних потрібна в багатьох сферах, її витoki як наукової дисципліни виникли із задачі прогнозування погоди в геофізиці. У геофізичних задачах, які стали мотивацією для розробки асиміляції даних, розміри стану системи та спостережень є величезними, а оператори еволюції, пов'язані з чисельними моделями, надто нелінійні.

Важливою задачею прикладної геофізики є аналіз і прогнозування полів забруднення в атмосфері, океані та на суші. Моделювання забруднень вимагає розробки методів асиміляції даних (DA), щоб покращити прогностичні можливості моделей перенесення забруднювачів.

Зі статистичної точки зору, отримання оцінок істинного стану системи та його невизначеності в таких умовах можна здійснити, наприклад, за допомогою фільтрації — процесу, в якому послідовно визначається апостеріорний розподіл стану на поточний момент часу, спираючись на всі спостереження, зібрані до цього моменту [2]. Поєднання високої розмірності та нелінійності робить цю задачу дуже складною.

Важливою особливістю застосування методів DA до задач перенесення забруднень в геофізичній гідродинаміці є те, що ці забруднювачі зазвичай є пасивними трасерами, на відміну від інших асимільованих параметрів, які можуть впливати на динаміку перенесення [3]. Методи DA є необхідними для багатьох задач, пов'язаних з аварійними викидами забруднювачів в океані, таких як розливи нафти чи викиди радіоактивних речовин [3]. Конкретна проблема, пов'язана як баєсова асиміляція із застосуванням методів DA, полягає в тому, що розподіл концентрації в океані залишається дуже неоднорідним протягом тривалого часу після аварії (кілька років у разі масштабного викиду). Відповідні спостережувані концентрації забруднювача можуть охоплювати кілька порядків величин, що робить необхідним оцінку та, за потреби, адаптацію існуючих методів DA для розв'язання таких задач.

Мета цієї статті — аналіз порівняння різних підходів асиміляції даних для задачі розповсюдження забруднень після аварійних викидів. Зокрема, ми розглядаємо, як різні методи асиміляції даних можуть впливати на точність прогнозів концентрації забруднюючих речовин у середовищі та на оцінку невизначеності цих прогнозів.

2. Ансамблеві методи

Ансамблева асиміляція даних базується на ансамблі короткострокових прогнозів станів системи, які використовуються разом із даними вимірювань як вихідна інформація для визначення стану атмосфери. Ансамбль дозволяє обчислити невизначеність змінних на момент аналізу. В моделях розповсюдження забруднень, так само як в атмосферних, де навіть незначні відмінності між станами можуть із часом сильно зростати в обмеженому підпросторі, ансамблеві методи показують хорошу стабільність. Завдяки ним, можна покращити як початкові умови для прогнозів, наприклад, функцію джерела викиду та її розташування, так і точність самих прогнозів.

Розглядається така адвективно-дифузійна модель:

$$\frac{\partial c}{\partial t} + u \frac{\partial c}{\partial x} + v \frac{\partial c}{\partial y} - D \left(\frac{\partial^2 c}{\partial x^2} + \frac{\partial^2 c}{\partial y^2} \right) = q(t) \delta(x - x_s) \delta(y - y_s). \quad (1)$$

Тут c — концентрація забруднення, t — час, D — коефіцієнт дифузії, права частина

описує джерело забруднення, що знаходиться в (x_s, y_s) , q — інтенсивність джерела.

Ансамблеві методи можуть виявитись більш обчислювально ефективними в порівнянні зі стандартними методами та показують хороші результати навіть в нелінійних моделях. Щобільше помилки прогнозу зосереджені в невеликому підпросторі кількох конкретних параметрів, то більшість інформації міститься в ансамблі обмеженого розміру [4, 5]. Ще однією перевагою є те, що вони обчислювально здійснені і потребують зазвичай менше ресурсів за стандартні методи, як баєсовська асиміляція.

Найбільш популярні й ефективні ансамблеві методи містять у собі ансамблевий фільтр Калмана (EnKF), детерміністичний ансамблевий фільтр Калмана (DenKF) [6], ансамблеву згладжуючу асиміляцію даних (Ensemble Smother, ES). Ця концепція також зустрічається в машинному навчанні (ML): Bagging, Boosting, Stacking, Voting, Weighted Ensemble [7].

2.1. Ансамблевий фільтр Калмана (EnKF)

Ансамблевий фільтр Калмана (EnKF) — це наближений метод фільтрації, вперше запропонований у літературі з геофізики [8], який є Монте-Карло адаптацією стандартного фільтра Калмана. На відміну від стандартного фільтра Калмана [9], який працює з усім розподілом стану явно, EnKF зберігає, пропагує та оновлює ансамбль векторів, що наближає розподіл стану. Ця ансамблева репрезентація є формою зменшення розмірності, оскільки замість спільного розподілу, який включає повну матрицю коваріацій, пропагується лише малий ансамбль. Коли стають доступними нові спостереження, ансамбль оновлюється за допомогою лінійного зсуву на основі припущення про лінійно-гаусівську модель простору станів. Тому додаткові наближення вводяться, коли мова йде про негаусівську чи нелінійну природу моделі. Але метод себе добре показує і в високих розмірностях, і в нелінійних проблемах [2]. Одна з причин таких переваг ENKF полягає у тому, що він не страждає від виродження ваг (наприклад, коли всі ваги, окрім однієї, стають 0), як інші послідовні методи Монте-Карло [2, 10]. Важливо зазначити, що для лінійних моделей метод збігається до стандартного фільтра зі збільшенням числа членів ансамблю.

Одним із недоліків EnKF є гаусове наближення щодо розподілів похибок моделі та вимірів, що застосовується в схемах оновлення. Гаусове наближення може бути надто грубим для задач із розподілом, що суттєво відхиляється від нормального, що призводить до нефізичних рішень і створює числові нестабільності [11]. Ще одна проблема — необхідність повторних перезапусків моделювання, які вимагає послідовна асиміляція даних, що може призводити до перезапуску з нереалістичних станів моделі. Інші проблеми EnKF включають колапс ансамблю і розходження фільтра [12].

2.2. Математичне формулювання

Традиційний EnKF складається з двох кроків: прогнозу і аналізу. Під час прогнозу всі члени ансамблю пропагуються за допомогою фізичної моделі-резольвенти. Кількість членів ансамблю може варіюватись від декількох до тисяч. Схематична діаграма зображена на рис. 1: кожен крок часу члени ансамблю отримують корекцію згідно зі спостереженнями. Ці кроки повторюються постійно до найбільш актуальної історії (рис. 1, [12]).

Введемо такі позначення. Нехай m_n^s — вектор статичних параметрів моделі, що залежить від часу, тоді резольвента M визначається за допомогою рівняння

$$\begin{bmatrix} m_n^d \\ d_n^c \end{bmatrix} = M(m_{n-1}^s, m_{n-1}^d) \quad (2)$$

в задачі розповсюдження забруднень. Це, наприклад, параметри джерела забруднень: місце розташування і інтенсивність, m_d — вектор динамічних параметрів моделі — концентрації

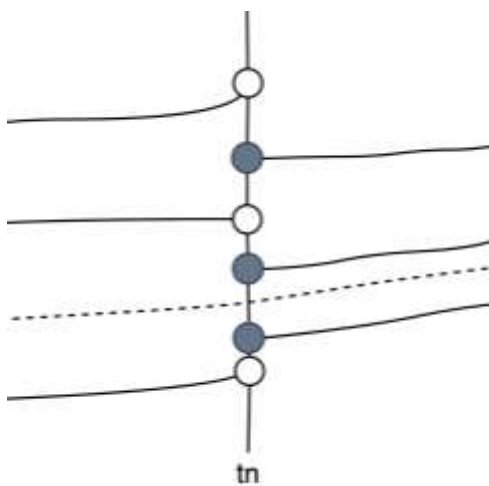


Рисунок 1 — Схема оновлення членів ансамблю, пунктирна лінія відповідає за справжній стан, чорні — члени ансамблю. Білими точками зображено прогнозовані значення, чорними після корекції

забруднень. Динамічні параметри знаходяться за допомогою моделі резольвенти, вони описуються диференціальними рівняннями в часткових похідних. Зокрема, концентрація є розв'язком рівняння дифузії-переносу.

Змінна d_n^{cal} у (2) — це вектор теоретичних(порахованих) спостережень. Разом вони формують вектор стану системи y^n . Для кожного члена ансамблю, яких N_e , ми формуємо ансамбль векторів стану системи ψ_n , який оновлює ЕпКФ. d_n^{obs} — вектор спостережень у момент часу t_n .

Відповідність між вектором обчислених спостережень і вектором стану встановлюється лінійним оператором H :

$$d_n^{cal} = H y_n. \quad (3)$$

$$H = [00I]. \quad (4)$$

Динамічні параметри моделі разом з обчисленими спостереженнями еволюціонують за допомогою резольвенти

$$\begin{bmatrix} m_n^d \\ d_n^c \end{bmatrix} = M(m_{n-1}^s, m_{n-1}^d). \quad (5)$$

Рівняння оновлення такі:

$$\psi_n^u = \psi_n^p + K(D_n - H\psi_n^p). \quad (6)$$

Індекс “ u ” означає оновлений стан ансамблів, індекс “ p ” — прогнозовані значення, D_n — ансамбль збурених спостережень, K — поправка фільтра:

$$K = C_{\psi,n}^p H^T (C_{\psi,n}^p H^T + C_D)^{-1}. \quad (7)$$

Тут $C_{\psi,n}^p$ — наближення матриці коваріації стану моделі, $C_D = \langle \varepsilon, \varepsilon^T \rangle$ — матриця коваріації похибок спостережень.

Збурені спостереження отримуються за допомогою додавання нормально розподіленої похибки з 0 середнім $\varepsilon_{n,i}$, індекс “ i ” відповідає за номер члена ансамблю. Оскільки помилки в спостереженнях у більшості випадків некорельовані, матрицю C_D можна вважати діагональною. Матрицю коваріації вектора стану $C_{\psi,n}^p$ апроксимуємо як

$$C_{\psi,n}^p = \frac{1}{N_e - 1} \sum_{j=1}^{N_e} (y_{n,j}^p - \langle y_{n,j}^p \rangle) (y_{n,j}^p - \langle y_{n,j}^p \rangle)^T. \quad (8)$$

Враховуючи визначення оператора H : $H C^p H^T = C_{d^{cal}}^p$.

Однією з переваг ЕпКФ є те, що нам не треба рахувати C^p повністю, адже вона множить на оператор H . Враховуючи рівняння поправки, отримуємо повну версію рівнянь:

$$K = C_{\psi,n}^p H^T (C_{\psi,n}^p H^T + C_D)^{-1}.$$

$$y_n^u = y_n^p + C_n^p H_n^T (H_n C_n^p H_n^T + C_{D,n})^{-1} (d_n^{obs} - H_n y_n^p). \quad (9)$$

Для розв'язання може бути застосований метод оптимальної інтерполяції [13] або ітераційний метод мінімізації [14].

На практиці метод застосовується в багатьох задачах геофізики. Зокрема, варто уваги дослідження в задачі розповсюдження забруднень в ідеалізованій 2D-моделі циркуляції океану [3]. В ньому було згенеровано забруднення в класичній моделі Стомеля [15], сформовано ансамбль концентрацій і проаналізовано EnKF в асиміляції даних із різними сценаріями вимірювання. Метод шукає покращення поля концентрації, щоб уникнути перенавчання, вільний параметр регуляризації було знайдено за допомогою евристичного методу L-кривої [16]. Наступний рис. 2 показує похибки і коефіцієнт кореляції в різних сценаріях асиміляції даних протягом 1 000 днів моделювання.

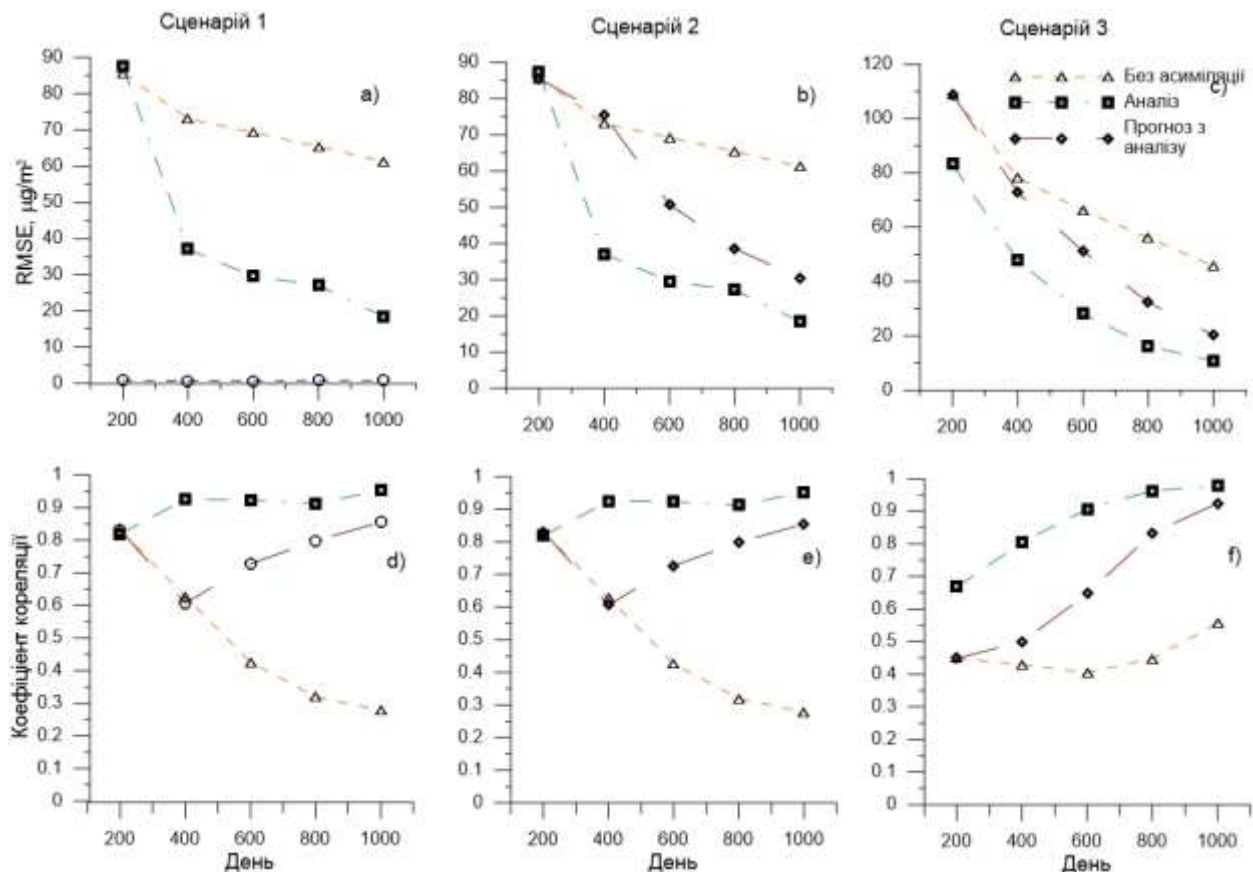


Рисунок 2 — Середньоквадратичні похибки (RMSE) (a, b, c) та коефіцієнти кореляції (d, e, f) між ансамблево усередненими змодельованими та реальними полями концентрації для різних днів, отримані в сценарії DA 1 (зліва), сценарії DA 2 (по центру) і сценарії DA 3 (праворуч), згідно з роботою [3]

Через суттєво нелінійну реакцію полів концентрації на поля швидкості розраховані концентрації, що відповідають заданому викиду у збурених полях течій, вони суттєво відрізнялися від полів, отриманих із незбуреною функцією течії.

Тести чутливості стосувалися впливу зменшення та збільшення кількості вимірювань вдвічі, а також впливу 10 % шуму у вимірюваннях. Статистичні показники похибок моделі виявилися не чутливими до проведених тестів і змінилися не більше ніж на 2 %. Отримані результати підтверджують, що запропонований метод можна вважати придатним для реальних задач, але він потребує подальшого тестування у відповідних умовах [3].

2.3. Ансамблевий згладжувач (ES)

Ансамблевий згладжувач (Ensemble smoother, ES) є життєздатною альтернативою EnKF, яка була впроваджена в 1996 р. [17]. На відміну від EnKF, ES здійснює глобальне оновлення,

одночасно асимілюючи всі доступні дані, що дозволяє значно скоротити час моделювання. Однак, як правило, цей метод дає гірший результат, ніж EnKF, особливо в задачах циркуляції океанів [17]. Основна перевага методу в тому, що він уникає рекурсивних оновлень і рестартів. У [18] показали, що метод потребує всього 10% обчислювальних ресурсів від EnKF.

У методі ES вектор стану визначений, як $\tilde{y} = \{y_n(t_0), \dots, y_n(t_n)\}$ — трирозмірна матриця замість двовимірної в EnKF. Матриця коваріації аналогічна до EnKF, лише вона по новому вектору станів:

$$\tilde{C}^p = \frac{1}{N_e - 1} \sum_{j=1}^{N_e} (\tilde{y}^p - \langle \tilde{y}^p \rangle) (\tilde{y}^p - \langle \tilde{y}^p \rangle)^T. \quad (10)$$

Наступна схема оновлення запропонована в [17]:

$$\tilde{y}_n^u = \tilde{y}_n^p + \tilde{C}_n^p \tilde{H}_n^T (\tilde{H}_n \tilde{C}_n^p \tilde{H}_n^T + \tilde{C}_{D,n})^{-1} D', \quad (11)$$

де $D' = (d_n^{obs} - H_n y_n^p)$, $\tilde{D}' = (D'_1, D'_m)^T$, індекс “m” відповідає за той час, коли спостереження доступні.

Похибка даних в ES вважається нормально розподіленою, як в EnKF, а матриця D' відповідає за ансамбль зміщених спостережень. Видно, що ансамбль усіх спостережень за всі точки часу оновлюється і до нього можна ставитись, як до комбінації ансамблю всіх прогнозів за часові кроки від 0 до n.

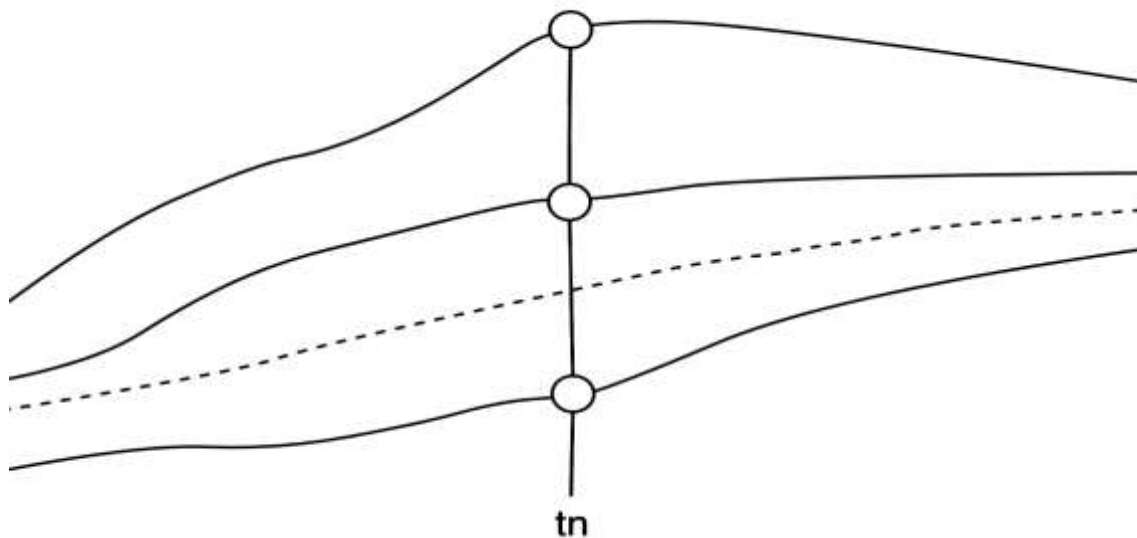


Рисунок 3 — Схема оновлення членів ансамблю, пунктирна лінія відповідає за справжній стан, чорні — члени ансамблю

Ми запускаємо весь ансамбль від початкового часу до кінцевого в прогнозуєму кроці та оновлюємо кожен ансамбль на етапі оновлення (рис. 3, [12]).

3. Асиміляція даних на основі машинного навчання (ML-DA)

Методи машинного навчання (ML) останнім часом набрали великої популярності у багатьох областях через вражаючі результати: від медицини, комп'ютерного зору та фінансів до екології та промисловості. Їх ефективність у вирішенні складних задач пов'язана з можливістю аналізувати великі обсяги даних, виявляти приховані закономірності та будувати прогнози з високою точністю. Підходи на основі штучного інтелекту можуть імітувати складні обчислення традиційних моделей із зниженими обчислювальними витратами, що

відкриває можливість значного прискорення та покращення аналізу й прогнозів, у тому числі в моделях прогнозування природи, що споріднені з досліджуваними [19]. ІІІ також грає важливу роль у вирішенні оптимізаційних задач [20].

Сурогатною моделлю називається модель, яка залежить лише від спостережень фізичної системи на відміну від справжньої динамічної. Оскільки ML не дуже добре працює з розрідженими даними, яких ще й зазвичай не так багато для якісного навчання, пропонуються гібридні методи з асиміляцією даних. Як показано в [21, 22], проблеми корекції помилок для навчання в загальному простіші і потребують менше даних.

Нехай еволюція стану задається рівнянням

$$y_{n+1} = M(y_n), \quad (12)$$

де M — динамічна модель резольвента.

Наша мета — знайти сурогат справжньої моделі, що буде застосований для прогнозування. Стандартний ML підхід складається з мінімізації функції втрати:

$$J(p) = \frac{1}{2} \sum_n \|y_{n+1} - M_n(p, y_n)\|_{Q_n^{-1}}^2 + L(p). \quad (13)$$

Тут p — набір параметрів моделі, наприклад, ваги і зсуви в нейронній мережі, $L(p)$ — параметр регуляризації, Q_n — матриця коваріації помилок моделі.

Оскільки стан системи в реальних задачах відомий через вектор спостережень, які пов'язані таким співвідношенням:

$$d_n^{obs} = H_n(y_n) + \varepsilon_n, \quad (14)$$

де H_n — оператор інтерполяції, то рівняння набуває вигляду [23–26]:

$$J(p) = \|y_0 - y_0^b\|_{B^{-1}}^2 + \frac{1}{2} \sum_{n=0}^{N_t} \|d_n^{obs} - H_n(y_n)\|_{C_D^{-1}}^2 + \frac{1}{2} \sum_{n=0}^{N_t-1} \|y_{n+1} - M_n(p, y_n)\|_{Q_n^{-1}}^2 + L(p), \quad (15)$$

де B — матриця коваріації помилки початкового наближення (між попереднім прогнозом і спостереженнями);

C_D — матриця коваріації помилок спостережень;

Q_n — матриця коваріації помилок моделі.

Якщо можливо, можна спробувати мінімізувати рівняння для параметрів і для вектора стану одночасно, але зазвичай це складно імплементувати в реальних задачах через високу розмірність, тому пропонується така стратегія сепарації задачі ДА:

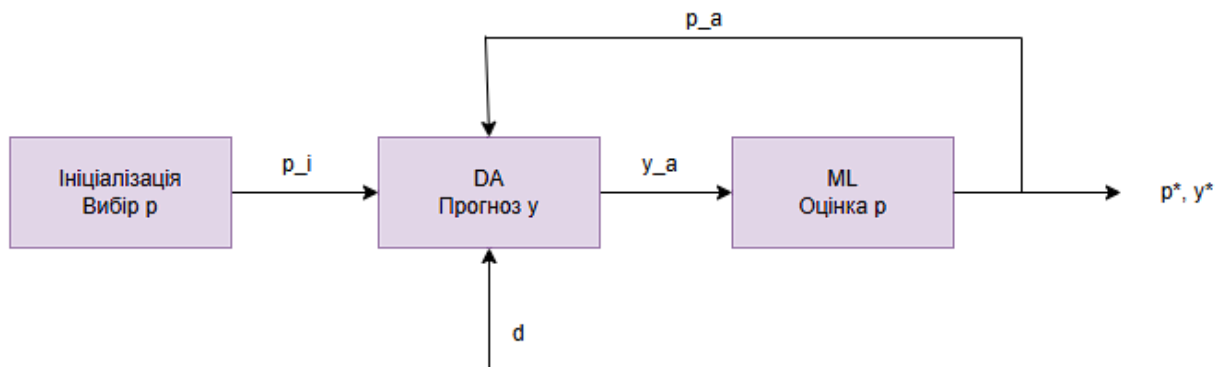


Рисунок 4 — Ілюстрація методу DA-ML для стратегії мінімізації цільової функції: чергування кроків DA з кроками ML для оцінки параметрів моделі p та траєкторії стану $y_0, \dots, y_n$ із підвищеною точністю

Починаємо з кроку DA з недосконалою фізичною моделлю для асиміляції спостережень. Потім виконуємо крок ML для навчання корекції похибки моделі на основі аналізу

кроку DA. Далі, розвиваючи метод DA–ML, можна ітеративно покращувати результати: використовувати скориговану модель для отримання точнішого аналізу в наступних кроках DA та навчатися на цьому аналізі, щоб покращити корекцію похибки моделі в подальших кроках ML (рис. 4, [27]).

Такий підхід гнучкий, адже кроки DA, ML повністю розділені. Проте застосування в реальних задачах має свої обмеження за двох причин: початковий вибір параметрів суттєво впливає для складних моделей і для сходження треба багато циклів DA-ML, що може бути складною обчислювальною задачею. Це може бути виправлено побудовою гібридної моделі на вже існуючій замість побудови сурогатної моделі з нуля. Алгоритм ML у цьому сценарії робитиме корекцію помилок, що зменшить кількість циклів і спростить обчислення [27].

Є два шляхи побудови гібридної моделі з сурогатом, один з яких полягає в корекції резольвенти, інший в корекції тенденцій (першої похідної по часу). Перший формулюється як

$$M(p, y) = M^{true}(p, y) + M^{ml}(p, y) \quad (16)$$

і потребує мало змін в оригінальній моделі через його простоту. Другий — замість корекції результату, ми коригуємо тенденцію:

$$\frac{dy}{dt} = F^{true}(y) + F^{ml}(p, y). \quad (17)$$

Це формулювання, по суті, не змінює функцію вартості, описану у рівнянні. Проте тенденції оригінальної моделі та навченої моделі стають складнішими через інтеграцію. На відміну від першого, цей підхід потребує значних змін у розрахунках, але він має кращий потенціал для прогнозування і кращу точність [28].

У [27] підхід протестовано на квазі-геострофічній моделі динаміки атмосфери, а корекція була застосована до резольвенти. В роботі було показано, що навіть малі нейронні мережі (NN) достатні для прогнозування значної частини помилки моделі і що гібридна модель має кращі властивості прогнозування, ніж оригінальна (DA): для невеликої кількості спостережень (10) результати RMSE $\approx 0,315$ (DA) проти $\approx 0,25$ (DA-ML), для 500 спостережень $\approx 0,225$ проти $\approx 0,17$ з інтервалом здійснення вимірів в один день. Із більшим інтервалом збільшується помилка через нелінійну природу помилок, при восьми днях помилка така ж сама, як в оригінальному сценарії. Було порівняно різні типи ML: у загальному більші моделі показують кращі результати, і нелінійні мережі показали себе значно краще лінійних.

Keller і Potthast [29] пропонують кардинально інший підхід, що заміняє класичну DA схему тренуванням нейронної мережі для асиміляції даних. Замість вивчення зв'язку між вхідними даними, спостереженнями і очікуваним результатом через тренування мережі на аналізованих (коригованих шляхом асиміляції вимірів) обчислювальних полях, ML тренується для знаходження співвідношення, завдяки якому використовується внутрішня функція втрат:

$$f(d^{cal}, d^{obs}) \rightarrow y. \quad (18)$$

Оскільки функція втрат зазвичай вимірює різницю між вхідними і вихідними даними, тут член, що відповідає за відхилення результату від першого припущення, замінено на член, який вимірює різницю між виходом моделі і наданими спостереженнями з вагами, визначеними матрицями коваріації помилок фону та спостережень відповідно. Функція втрат набуває такого вигляду, який відповідає функції асиміляції даних 3D-Var [13]:

$$l = (\hat{y} - y^b)^T B^{-1} (\hat{y} - y^b) + (\hat{d} - d)^T C_d^{-1} (\hat{d} - d). \quad (19)$$

Тут $\delta y = \hat{y} - y^b$ — інкремент аналізу, $\hat{d} - d$ — різниця між аналізом і спостереженнями.

При такому підході активно використовується факт, що функціонал мінімізується при тренуванні мережі. Така модель сильно залежить від першого наближення, проте коли

перше наближення якісне, вона може бути використана для тренування. Якщо перше наближення відсутнє, то як таке використовуються нульові значення.

Цей підхід протестовано на трьох задачах: одновимірних та двовимірних ідеалізованих задачах, на реальній задачі аналізу температури. Була використана нейронна мережа прямого поширення з трьома прихованими прошарками. Для перших двох задач підхід навіть із тривіальними початковими наближенням зміг практично ідеально відтворити стан. З динамічними локаціями спостережень результат лише трохи гірший за статичні, що може бути компенсовано більшою кількістю спостережень. Асиміляція температури на висоті двох метрів (T2M) в реальному сценарії показала, що якість результатів моделювання покращилась. Зокрема, це може виправляти зміщення у початкових полях, наприклад, нічне теплове зміщення у реаналізі COSMO-REA6 [30], і суттєво зменшувати помилки [29].

4. Висновки

У цій статті ми дослідили сучасні методи асиміляції даних, включаючи традиційні підходи на основі ансамблевого фільтра Калмана (EnKF), ансамблевого згладжувача (ES), та новітні гібридні методи, що базуються на машинному навчанні (ML-DA) у різноманітних задачах асиміляції. Представлені результати демонструють, що включення ML у процеси асиміляції даних дозволяє покращити прогнозування, зменшити обчислювальні витрати та підвищити ефективність аналізу, особливо у випадках з обмеженою кількістю спостережень.

Методи на основі ML показали здатність виправляти модельні похибки та враховувати зміщення у прогнозах, наприклад, нічне теплове зміщення у COSMO-REA6. Крім того, гібридні підходи, що коригують резольвенту або тенденцію моделі, дозволяють інтегрувати переваги ML із традиційними методами асиміляції даних, зменшуючи кількість циклів DA-ML і забезпечуючи високу точність.

Однак залишається ряд викликів, зокрема, залежність від початкових наближень і обчислювальна складність при великій кількості циклів. У цьому контексті ідея використання ML для корекції помилок, а не створення повністю нових сурогатних моделей, виглядає перспективною.

Крім зазначених складнощів, усі розглянуті гібридні методи на основі поєднання традиційних методів ДА та методів машинного навчання були перевірені на підставі моделювання геофізичних полів, які є більш-менш однорідними (наприклад, поле температури в океані, значення якої змінюються не більше ніж на $\approx 10\%$). У випадку моделювання розповсюдження забруднень, утворених аварійними викидами поля концентрації забруднення, є суттєво неоднорідними, а відповідні значення охоплюють декілька порядків величини. Це робить необхідним додаткову оцінку та, за потреби, адаптацію розглянутих методів для моделювання відповідних сценаріїв.

Перспективи майбутніх досліджень включають вдосконалення гібридних методів, розробку більш ефективних архітектур нейронних мереж і застосування ML-DA для складніших задач у реальних сценаріях. Отже, інтеграція методів машинного навчання в асиміляцію даних відкриває нові горизонти для прогнозування природних процесів і аналізу даних.

Подяки

Представлена робота була підтримана грантом Національного фонду досліджень України, N 2023.03/0021.

СПИСОК ДЖЕРЕЛ

1. Nychka D., Anderson, J.L. Data Assimilation. *Handbook of Spatial Statistics* / A. Gelfand, P. Diggle, P. Guttorp, M. Fuentes (eds.). New York, NY: Chapman & Hall/CRC, 2010. P. 477–492.
2. Katzfuss M., Stroudb J.R., Wikle C.K. Understanding the Ensemble Kalman Filter, *The American Statistician*, 2016. Vol. 70., P. 350–357. DOI: <https://doi.org/10.1080/00031305.2016.1141709>.
3. Kovalets I., Kim K.O., Shrubkovsky O., Maderich V. Ensemble Data Assimilation of Concentration Measurements Following the Accidental. Release of a Contaminant in the Ocean: Method Testing in an Idealized Setting, *Pure Appl. Geophys.*, 2022. DOI: <https://doi.org/10.1007/s00024-022-02990-5>.
4. Tippett M.K., Cohn S.E., Todling R., Marchesin D. Low-dimensional representation of error covariance. *Tellus*, 2000. Vol. 52 (A). P. 533–553.
5. Thomas M. Hamill, Ensemble-Based Data Assimilation. NOAA-CIRES Climate Diagnostics Center, Boulder, Colorado USA, 2002. P. 83–123.
6. Sakov P., Oke P.R. A deterministic formulation of the ensemble Kalman filter: an alternative to ensemble square root filters. *Tellus*, 2008. Vol. 60A. P. 361–371.
7. A comprehensive review on ensemble deep learning: Opportunities and challenges, Ammar Mohammed and Rania Kor. *Journal of King Saud University – Computer and Information Sciences*. 2023. P. 757–774. DOI: <https://doi.org/10.1016/j.jksuci.2023.01.014>.
8. Evensen G. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*. 1994. N 99 (C5). P. 10143–10162.
9. Kalman R.E. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*. 1960. Vol. 82 (1). P. 35–45. DOI: <https://doi.org/10.1115/1.3662552>.
10. Snyder C., Bengtsson T., Bickel P., Anderson J.L. Obstacles to High-dimensional Particle Filtering. *Monthly Weather Review*. 2008. N 136. P. 4629–4640.
11. Liu N., Oliver D.S. Critical Evaluation of the Ensemble Kalman Filter on History Matching of Geologic Facies. *SPE Journal*. 2005. N 6 (3). P. 344–367.
12. Xiaoyang Xia History-matching production data using ensemble smoother with multiple data assimilation: a comparative study, 2014. P. 1–78.
13. Daley R. Atmospheric data analysis. Cambridge University Press, 1991. P. 98–180.
14. Evensen G. Data Assimilation. The Ensemble Kalman Filter. Springer-Verlag, 2009. P. 71–92.
15. Stommel H. The westward intensification of wind-driven ocean currents. *Transactions American Geophysical Union*. 1948. Vol. 29 (2). P. 202–206.
16. Lawson C.R., Hanson R.J. Solving least squares problems. SIAM. 1995. P. 100–103.
17. Peter Jan Van Leeuwen, Data assimilation and inverse methods in terms of a probabilistic formulation. *Monthly Weather Review*. 1996. Vol. 124. P. 2898–2913.
18. Skjervheim J., Evensen G. An Ensemble Smoother for Assisted History Matching. *SPE Journal*. 2011. Vol. 5 (2). P. 412–423.
19. Pathak J., Subramanian S., Harrington P., Raja S., Chattopadhyay A., Mardani M., Kurth T., Hall D., Li Z., Azizzadenesheli K., Hassanzadeh P., Kashinath K., Anandkumar A. FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. 2022. URL: <https://arxiv.org/abs/2202.11214>.
20. Fan Z., Ghaddar B., Wang X., Xing L., Zhang Y., Zhou Z. Artificial intelligence for operations research: Revolutionizing the operations research process, 2024. P. 1–47. DOI: <https://doi.org/10.48550/arXiv.2401.03244>.
21. Jia X., Willard J., Karpatne A., Read J., Zwart J., Steinbach M., Kumar V. Physics guided RNNs for modeling dynamical systems: A case study in simulating lake temperature profiles. *Proc. of the 2019 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics*. Philadelphia, PA, 2019. P. 558, 566. DOI: <http://dx.doi.org/10.1137/1.9781611975673>.
22. Watson P.A.G. Applying machine learning to improve simulations of a chaotic dynamical system using empirical error correction. *Journal of Advances in Modeling Earth Systems*. 2019. Vol. 11 (5). P. 1402–1417. DOI: <https://doi.org/10.1029/2018MS001597>.
23. Hsieh W.W., Tang B. Applying neural network models to prediction and data analysis in meteorology and oceanography. *Bulletin of the American Meteorological Society*. 1998. Vol. 79 (9). P. 1855–1870. DOI: [https://doi.org/10.1175/1520-0477\(1998\)079<1855:ANNMTP>2.0.CO;2](https://doi.org/10.1175/1520-0477(1998)079<1855:ANNMTP>2.0.CO;2).

24. Abarbanel H.D.I., Rozdeba P.J., Shirman S. Machine learning: Deepest learning as statistical data assimilation problems. *Neural Computation*. 2018. Vol. 30 (8). P. 2025–2055. DOI: https://doi.org/10.1162/neco_a_01094.
25. Bocquet M., Brajard J., Carrassi A., Bertino L. Data assimilation as a learning tool to infer ordinary differential equation representations of dynamical models. *Nonlinear Processes in Geophysics*. 2019. Vol. 26 (3). P. 143–162. DOI: <https://doi.org/10.5194/npg-26-143-2019>.
26. Bocquet M., Brajard J., Carrassi A., Bertino L. Bayesian inference of chaotic dynamics by merging data assimilation, machine learning. arXiv:2001.06270, 2020. P. 1–27. DOI: <https://doi.org/10.48550/arXiv.2001.06270>.
27. Farchi A., Bocquet M., Laloyaux P., Bonavita M., Malartic Q. A comparison of combined data assimilation and machine learning methods for offline and online model error correction. *Journal of Computational Science*. 2021. Vol. 55. P. 101468. URL: <https://doi.org/10.1016/j.jocs.2021.101468>.
28. Farchi A., Laloyaux P., Bonavita M., Bocquet M. Using machine learning to correct model error in data assimilation and forecast applications. *Quarterly Journal of the Royal Meteorological Society*. 2021. Vol. 147 (739). P. 3067–3084. ISSN 1477-870X. DOI: <https://doi.org/10.1002/qj.4116>. URL: <http://dx.doi.org/10.1002/qj.4116>.
29. Keller J.D., Potthast R. AI-based data assimilation: Learning the functional of analysis estimation. arXiv:2406.00390 [physics.ao-ph], 2024. P. 1–16. DOI: <https://doi.org/10.48550/arXiv.2406.00390>.
30. Bollmeyer C., Keller J.D., Ohlwein C., Wahl S., Crewell S., Friederichs P., Hense A., Keune J., Kneifel S., Pscheidt I., Redl S., Steinke S. Towards a high-resolution regional reanalysis for the European CORDEX domain. *Quarterly Journal of the Royal Meteorological Society*. 2015. Vol. 141 (686). P. 1–15. DOI: <https://doi.org/10.1002/qj.2486>.

Стаття надійшла до редакції 27.01.2025