https://doi.org/10.36296/1819-8058.2026.2(85).140-156

СТАТИСТИКА ЧАСОВИХ РЯДІВ ІНСОЛЯЦІЇ М. КИЄВА. І. ЧАСТОТНИЙ ТА КОРЕЛЯЦІЙНИЙ АНАЛІЗИ ДЛЯ ФОТОЕЛЕКТРИЧНИХ СИСТЕМ

Отримано 22 квіт. 2026 р.; рекомендовано до публікації 26 чер. 2026 р.
Доступно онлайн 30 чер. 2026 р.

Гаєвський О. Ю.[1], Гаєвська Г. М.[2]

Автор для кореспонденції: Гаєвський Олександр,
e-mail: a.gaevskii@kpi.ua

[1] д-р. фіз.-мат. наук, професор
https://orcid.org/0000-0001-6144-2441

[2] ст. викладач
https://orcid.org/0000-0001-7760-6789

[1, 2] НТУУ «Київський політехнічний інститут ім. Ігоря Сікорського» , м. Київ, Україна

[1] Інститут відновлюваної енергетики НАН України, м. Київ, Україна

Анотація. *До сучасних фотоелектричних систем ставляться жорсткі вимоги під час проєктування та прогнозування вироблення електроенергії, що робить дослідження регіональних статистичних властивостей сонячного випромінювання дедалі важливішим. У цій статті представлено статистичний аналіз щоденних рядів сонячного опромінення для Києва за останні 10 років, виражених через індекс прозорості Kt та розкладених на сезонні та стохастичні компоненти. Показано, що розподіл частоти Kt має бімодальну структуру з яскраво вираженими піками, що відповідають станам «хмарно» та «ясне небо». «Сідлова» зона між цими піками вказує на те, що проміжні стани змінної хмарності трапляються рідше та за своєю суттю менш стабільні. Цю бімодальність необхідно враховувати під час проєктування ФЕ систем, оскільки розрахунки, засновані виключно на довгострокових місячних середніх значеннях, призводять до значних систематичних помилок, що потенційно переоцінює виробіток енергії протягом тривалих періодів хмарності. Основна увага приділяється стохастичним залишкам X ряду Kt, які є важливими для прогнозування виробництва та визначення розмірів систем акумуляторного накопичення енергії. Стаціонарність залишкового часового ряду, який зберігає бімодальний розподіл, була підтверджена за допомогою розширеного тесту Дікі Фуллера (ADF). Аналіз функцій автокореляції (ACF) та часткової автокореляції (PACF) виявив короткострокові залежності, які ефективно враховуються моделями ARMA(p,q). Оцінка параметрів у цьому сімействі моделей визначила модель ARMA(1,1) як найадекватнішу погляду точності та економного набору параметрів, що підтверджено інформаційним критерієм Акайке (AIC). Результати цих частотних та кореляційних аналізів є важливими для тестування автономних і резервних фотоелектричних систем шляхом генерації різноманітних погодних сценаріїв, включно з найгіршими умовами.*

Ключові слова*: сонячна радіація, Київ, індекс прозорості, стохастичні залишки, часові ряди, стаціонарність, розподіл частот, KDE-апроксимація, бімодальність, автокореляційні функції ACF/PACF, моделі ARMA, фотоелектричні системи, енергетична надійність.*

Абревіатури

ACF(Autocorrelation Function) – автокореляційна функція

ADF (Augmented Dickey-Fuller (test)) – розширений тест Дікі – Фуллера

ADS (Atmosphere Data Store) – сховище атмосферних даних

AIC (Akaike Information Criterion) – інформаційний критерій Акайке

ARMA (Autoregressive Moving Average) – авторегресійна ковзна середня

BESS (Battery Energy Storage System) – акумуляторна система накопичення енергії

CAMS (Copernicus Atmosphere Monitoring Service) – служба моніторингу атмосфери «Copernicus»

CDF (Cumulative Distribution Function) – кумулятивна функція розподілу

ERA5 (European Centre for Medium-Range Weather Forecasts Reanalysis) – система реаналізу Європейського центру середньострокових прогнозів погоді

GHI (Global Horizontal Irradiation) – глобальне горизонтальне опромінення

KDE (Kernel Density Estimation) – ядерна оцінка щільності

LCOE (Levelized Cost of Energy) – усереднена вартість енергії

PACF (Partial Autocorrelation Function) – частинна автокореляційна функція

PDF (Probability Density Function) – функція щільності ймовірності

PV (Photovoltaic) – фотоелектричний

RMSE (Root Mean Square Error) – середньоквадратична похибка

Вступ. Стрімкий розвиток фотоелектричних систем (ФЕС) та їх інтеграція до сучасних енергосистем висувають жорсткі вимоги до точності проєктування та прогнозування вироблення електроенергії. На відміну від традиційної генерації сонячна енергія характеризується високою стохастичністю, обумовленою складною динамікою атмосферних процесів. В умовах початку децентралізованого енергопостачання та використання автономних (резервних) ФЕС стандартних методів розрахунку за середньомісячними значеннями стає недостатньо. Тому важливі дослідження статистичних властивостей метеорологічних факторів (освітленості, температури навколишнього середовища, швидкості вітру) для прогнозування та оцінок надійності роботи установок на ВДЕ [1–3]. Вивчення статистичних властивостей сонячного опромінення в конкретних регіонах актуальне також під час проектування та оцінок ефективності роботи гібридних систем [4].

Ефективність і надійність енергооб'єктів безпосередньо залежить від розуміння структури часових рядів інсоляції. Статистичний аналіз дає змогу виділити детерміновані сезонні тренди та глибоко дослідити природу стохастичних залишків – відхилень, спричинених локальними метеорологічними факторами. Саме ці відхилення визначають імовірність дефіциту енергії, відповідність критеріям надійності енергопостачання та диктують вимоги до ємності акумулювальних систем енергозбереження (BESS).

Особливий інтерес являють дані з інсоляції для умов м. Києва. Клімат цього регіону характеризується значною мінливістю хмарного покриву, нгасамперед у перехідні осінньо-весняні періоди. Висока амплітуда стохастичних коливань інсоляції в умовах помірно континентального клімату потребує застосування інструментів математичної статистики, як-от частотний та кореляційний аналізи (ACF/PACF), авторегресійні моделі ARMA (Autoregressive and Moving Average Model) [5, 6].

Мета роботи і постановка задачі. Метою цієї роботи є розробка й верифікація статистичної моделі, яка описує внутрішню структуру та закономірності мінливості часових рядів сонячної інсоляції для умов м. Києва на основі актуальних даних за останнє десятиріччя. У процесі дослідження вирішувались такі задачі:

- декомпозиція часового ряду: виділення довгострокової сезонної складової та отримання ряду стохастичних залишків інсоляції;
- частотний (імовірнісний) аналіз: дослідження функцій щільності ймовірності (PDF) та інтегральних функцій розподілу (CDF) індексу прозорості Kt та його залишків X;
- кореляційна діагностика: обчислення та аналіз автокореляційних (ACF) та частинних автокореляційних (PACF) функцій для оцінки «пам'яті» досліджуваного процесу;
- ідентифікація та параметризація моделі: вибір оптимальної моделі ARMA, розрахунок її коефіцієнтів та оцінка статистичної значущості параметрів;
- верифікація моделі: перевірка адекватності моделі шляхом аналізу залишкового шуму на відповідність критеріям стаціонарності та відсутності серійної залежності (тести ADF, Льюнг – Бокс [6, 7]).

Для забезпечення достовірності статистичних висновків у роботі використано масив даних добової інсоляції за останній десятирічний період, 2026–2025 рр., у м. Києві. Багаторічна ретроспектива дає змогу врахувати не лише сезонні цикли, а й варіативність синоптичних процесів, за винятком впливу випадкових екстремальних погодних аномалій одного конкретного року. Дослідження багаторічних рядів актуально в задачах оптимізації компонентного складу ФЕС, довгостроковому прогнозуванні й отриманні поточної діагностики з ефективності шляхом генерування синтетичних рядів інсоляції [8–10].

Вихідними даними для дослідження слугували добові ряди сумарної сонячної інсоляції GHI (Global Horizontal Insolation) з сервісу Atmosphere Data Store (ADS) бази даних Copernicus Atmosphere Monitoring Service (CAMS) [11]. Цей сервіс надає результати моделювання як для умов ясного неба (Cloud-free / Clear-sky), так і для фактичних метеоумов (Actual weather). У цій роботі порівняння зазначених двох сценаріїв дало змогу виділити чистий внесок хмарності в стохастичну структуру часового ряду. Використання сервісу ADS, в основі якого лежить глобальний набір даних реаналізу ERA5-Land, забезпечує високу часову і просторову (9 км) дискретизацію, а також статистичну достовірність.

Процеси обробки та аналізу рядів виконувались у цій роботі в середовищі Python/Spyder існуючими інструментами бібліотек numpy, pandas, scipy, statsmodels, sklearn, tslearn та інших [12, 13].

Структура добових рядів сонячного опромінення м. Києва. Для аналізу статистичних закономірностей сонячного опромінення земної поверхні ми переходимо до індексу прозорості K_t – безрозмірного показника, що дорівнює відношенню глобального сонячного опромінення (GHI) на поверхні Землі до позаземного опромінення G_0:

$$K_t = \frac{G}{G_0} \quad (1)$$

де G – глобальна горизонтальна радіація на земній поверхні (GHI), G_0 – позаземна горизонтальна радіація. Перехід до K_t у подальшому дасть змогу перетворити вихідні ряди до стаціонарної форми. Необхідною умовою при цьому є усунення впливу детермінованої астрономічної складової G_0. При розрахунках K_t (1) фільтруються фізично неможливі значення поза діапазоном 0.05 ÷ 0.085.

Під час аналізу поведінки сонячної радіації слід враховувати, що часовий ряд не є однорідним, а являє собою суперпозицію різних за природою процесів, тому ряд K_t прийнято поділяти на його сезонну складову та залишки:

$$K_t = K_{t,\text{season}} + X \quad (2)$$

Перша складова $K_{t,\text{season}}$ являє собою детерміновану компоненту, яка визначається циклічними коливаннями інсоляції, зумовленими астрономічними факторами: обертанням Землі навколо своєї осі (добовий цикл) і навколо Сонця (річний цикл). Друга складова X стохастичні залишки, які є різницею між реальними вимірами та сезонним трендом. Залишки виникають через динаміку атмосферних процесів: рухи хмарності, зміни вологості та запиленості повітря.

Якщо аналізувати ряд (2) цілком, потужні сезонні коливання маскуватимуть локальні швидкі короткочасні зміни. Відділення сезонності дає змогу сфокусувати математичний апарат на аналізі випадкових відхилень. Це важливо для:

- мінімізації відхилень модельних рядів від реальних значень (наприклад, за критерієм RMSE), оскільки модель, яка навчена на залишках, точніше відображає короткострокові зміни погоди;
- оптимізації обчислень завдяки фокусуванню саме на аналізі стохастичних процесів без втрати ресурсівна сезонну складову.

Знання часової залежності сезонної складової $K_{t,\text{season}}$ дає змогу визначати теоретичний максимум генерації ФЕС для конкретної географічної точки в конкретний час. Вона приблизно описується моделлю чистого неба (Clear Sky Models) і сама вона не несе невизначеності, яку потрібно знаходити статистичними методами. Невизначеність міститься саме в залишках X, де прихована основна волатильність, що ускладнює прогнозування. Залишки акумулюють у собі вплив метеорологічної турбулентності, випадкових змін хмарності та аерозольного складу атмосфери. Метою виділення залишків є досягнення стаціонарності часового ряду: під час дослідження переходимо від нестаціонарного ряду інсоляції до стаціонарного ряду залишків. Стаціонарність є обов'язковою умовою для застосування моделей ARMA, а також для забезпечення збіжності алгоритмів машинного навчання [5, 6].

Ряди добових значень коефіцієнта прозорості K_t і залишків X на інтервалі 10 років показані на графіках на рис. 1 – зверху і знизу, відповідно. На графік K_t накладена сезонна складова, яка має чітко виражений періодичний характер. Приведемо результати статистичних оцінок отриманих рядів. Середні значення 0.466 – для K_t, 0.00018 – для X (має бути ≈ 0). Стандартне відхилення залишків – 0.1832. Перевірка отриманого часового ряду залишків X розширеним тестом Дікі – Фуллера ADF [6, 7] показала його стаціонарність: *p-value* << 0.001 (має бути < 0.05). Тест Льюнга – Бокса LB [5, 6], який перевіряє, чи є автокореляції часового ряду (у нашому випадку – залишків X), показав їх наявність на лагах до 20 днів (p <0.05). Це свідчить про доцільність переходу в подальшому до застосування та ідентифікації авторегресійної моделі з ковзним середнім ARMA(p,q).

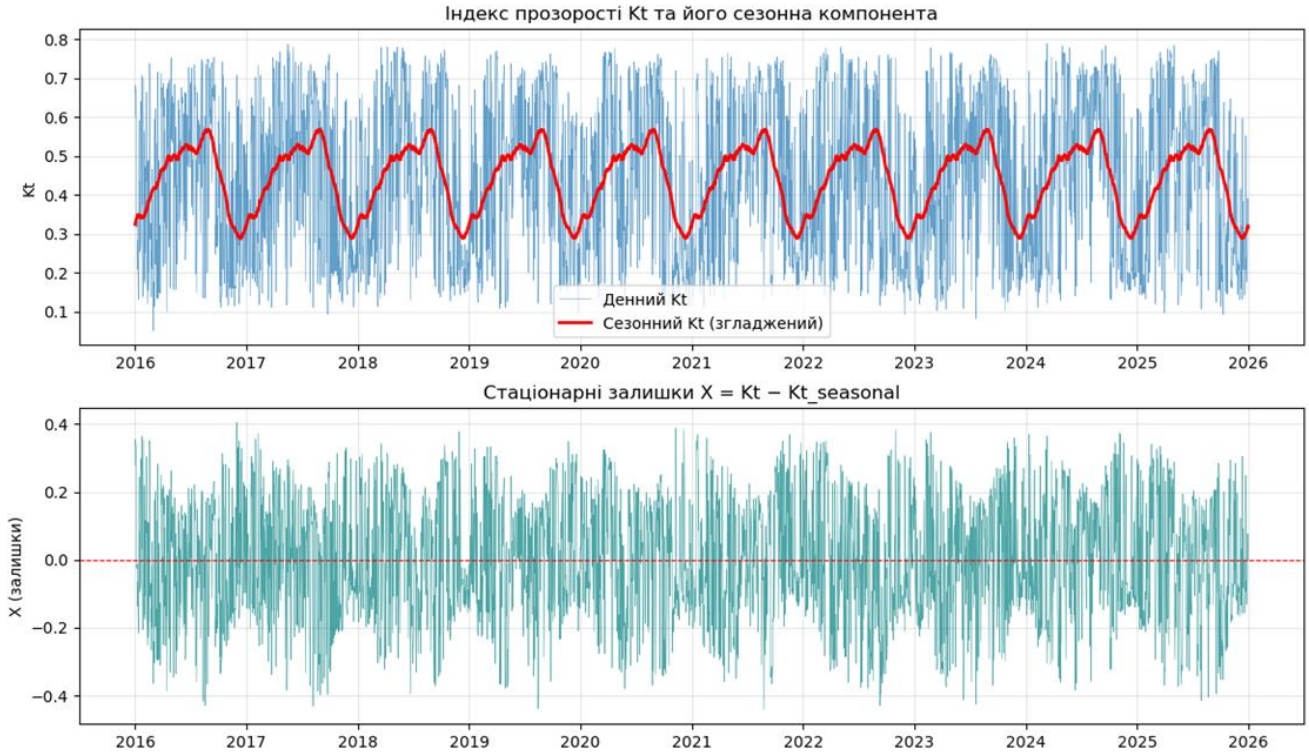


Рис. 1. Часові ряди періоду 2016–2025 рр. для м. Києва: зверху – добові значення індексу прозорості K_t з сезонною складовою (червона крива); знизу – стохастичні залишки X

Частотні та кумулятивні розподіли індексу K_t та залишків. Основою для розуміння динамічної поведінки сонячного ресурсу, що виходить за межі передбачуваної сезонної циклічності, є аналіз щільності ймовірності (PDF) та функції розподілу (CDF) індексу прозорості та стохастичних залишків. Теоретичний потенціал фото-

електричної системи визначається детермінованою компоненто – інсоляцією при ясному небі. Але для визначення реальної експлуатаційної надійності ФЕС потрібний аналіз статистичних властивостей фактичної інсоляції – розкиду, асиметрії та форми розподілу. З погляду проєктування та оптимізації ФЕС дослідження розподілів K_t, X важливе для рішення таких завдань фотоенергетики:

– *Прогнозування екстремальних режимів*. Аналіз «хвостів» розподілу дає змогу оцінити ймовірність і тривалість глибоких провалів генерації, спричинених аномальними погодними умовами, що недоступно при використанні середньомісячних даних.

– *Розрахунок критеріїв надійності (LLP – Loss of Load Probability)* [9, 14]. Точне знання функції розподілу залишків необхідне для обчислення ймовірності дефіциту енергії, тобто для обґрунтування вибору потужності панелей та ємності накопичувачів.

– *Стохастична оптимізація*. Форма щільності ймовірності слугує основою для моделювання реалістичних сценаріїв роботи системи в умовах невизначеності, що необхідно для оцінок економічної окупності (LCOE, CAPE*X*) [15, 16] у довгостроковій перспективі.

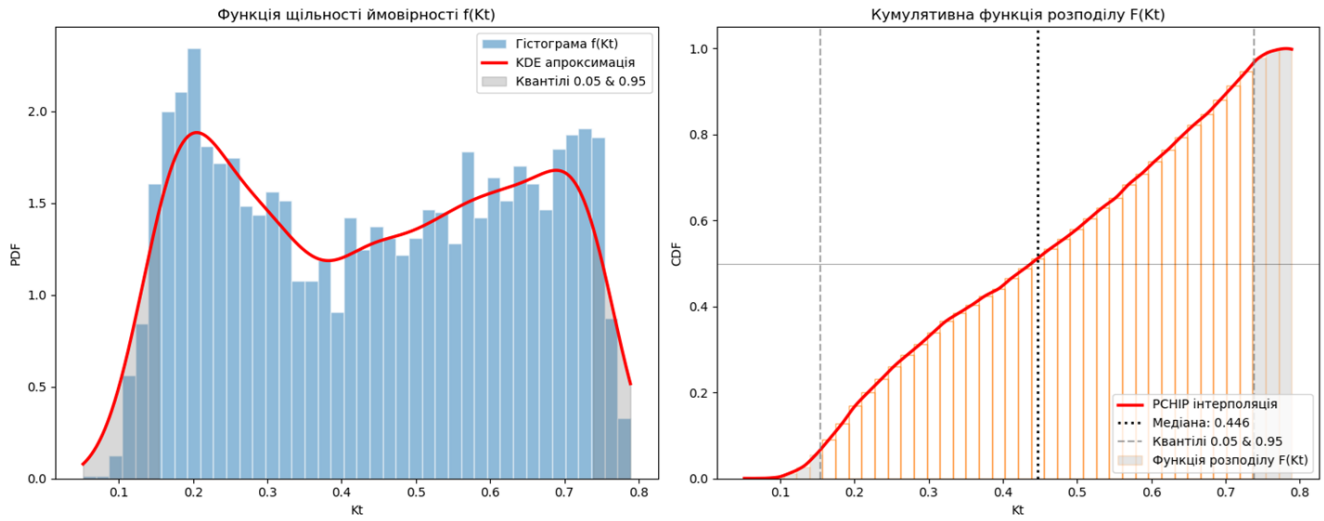
Результати статистичного аналізу ряду добових значень K_t показані на рис. 2 у формі гістограми частот та її непараметричної апроксимації методом ядерної оцінки щільності KDE (Kernel Density Estimation) [17] і кумулятивної функції розподілу (CDF) з її PCHIP-інтерполяцією (Piecewise Cubic Hermite Interpolating Polynomial). Область визначення апроксимуючої функції щільності ймовірності була встановлена за емпіричними межами накопиченого масиву даних. При цьому обрізаються теоретично нескінченні «хвости» апроксимуючих функцій, щоб у подальшому забезпечити достовірність синтетичних рядів для тестування моделей ФЕС.

На графіках частотного розподілу індексу прозорості (рис. 2, *а*) бачимо бімодальну структуру з двома вираженими локальними піками. Перший з них (0.20) відповідає режиму «хмарного неба» (overcast). Висока щільність імовірності в цій зоні характерна для клімату Києва, особливо в осінньо-зимовий період. Другий пик (0.73) відповідає режиму «ясного неба» (clear sky). Зона «сідла» з відносно низькою щільністю в центрі вказує на те, що проміжні стани (мінлива хмарність) є менш стійкими – атмосфера прагне «звалитися» в один з двох означених станів.

Графік накопиченої ймовірності CDF (рис. 2, *б*) підтверджує неоднорідність статистичної структури K_t: медіана (0.446) перебуває якраз у зоні «сідла» PDF. Цей важливий факт означає, що середнє значення індексу прозорості для Києва описує стан, який насправді зустрічається найрідше. Можна вважати це аргументом проти використання найпростіших середніх моделей сонячної радіації м. Києва в задачах фотоенергетики.

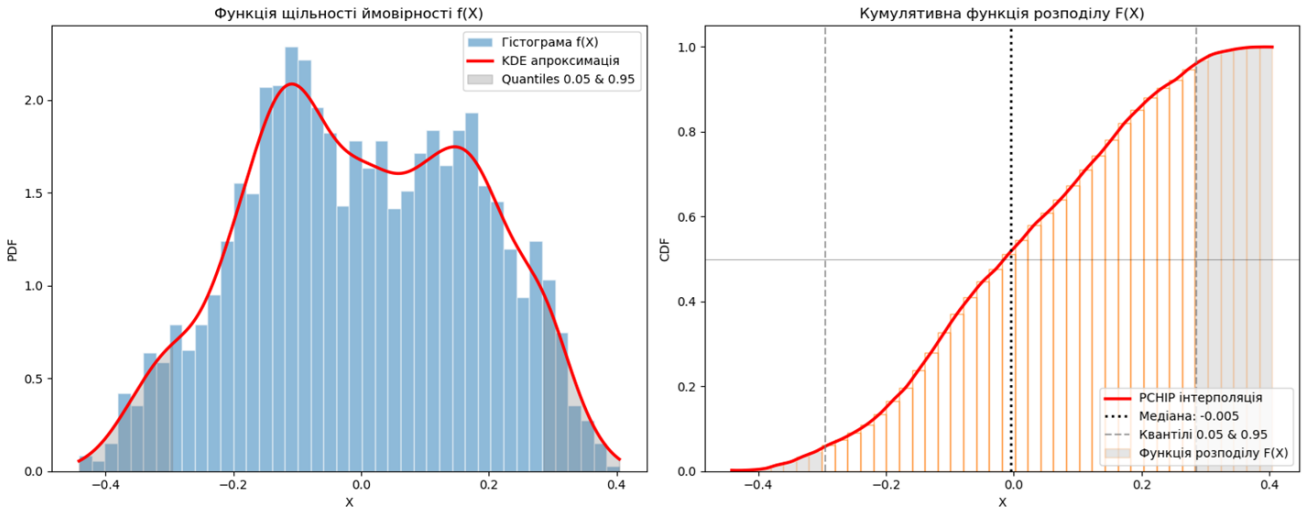
На графіках (див. рис 2, *а*, *б*) вказаний інтервал 0.05–0.95 квантілів (довірчий інтервал), що обмежує діапазон індексу прозорості K_t (приблизно від 0.15 до 0.74), за межами якого перебувають екстремальні, але малоймовірні викиди. Використання KDE-апроксимації (червона лінія на PDF) дає змогу згладити шум гістограми і візуалізувати локальні максимуми розподілу. PCHIP-інтерполяція на графіку CDF забезпечує монотонність кривої, що важливо для коректного синтезу рядів методом зворотної інтегральної трансформації [9, 18].

На частотному розподілі стохастичних залишків X (рис. 3, *а*) також спостерігаємо бімодальність, що виражена слабше. Це очікуваний, але цікавий результат, який вказує на те, що бімодальність не лише сезонний ефект (вплив сезонів був виключений при розрахунку X), а й фундаментальна властивість сонячного опромінення в регіоні.



а *б*

Рис. 2. Результати статистичного аналізу індексу прозорості K_t: а – гістограма частот та її непараметрична KDE апроксимація; б – кумулятивна функція розподілу (CDF) та її PCHIP-інтерполяція. Тінню показані 95%-ні довірчі інтервали



а *б*

Рис. 3. Результати статистичного аналізу залишків X: а – гістограма частот та її непараметрична KDE апроксимація; б – кумулятивна функція розподілу (CDF) та її PCHIP-інтерполяція. Тінню показані 95%-ні довірчі інтервали

Отже, графік PDF (див. рис. 3, *а*) для залишків показує збереження бімодальності, але її характер якісно змінився. Два максимуми зблизилися, а зона «сідла» стала менш глибокою порівняно з PDF(K_t). Моди залишків зміщені – тепер вони центровані щодо нуля (приблизно -0.12 і 0.13). Негативний пік – це «нестача» радіації через хмари, позитивний – «надлишок» радіації щодо середнього тренду в ясні періоди. Форма розподілу стала більш симетричною та компактною, що значно полегшує завдання статистичного синтезу рядів, оскільки екстремальні хвости стали менш вираженими. Бачимо також характерне центрування (медіана (-0.005), що необхідно для роботи ARMA-моделей.

На графіку накопиченої ймовірності CDF (рис. 3, *б*) крива стала плавнішою, майже наближаючись до S-подібної форми (притаманної нормальному розподілу), але з характерним «зламом» у районі медіани (зміна знаку другої похідної). Цей злам – прямий наслідок бімодальності та аргумент на користь того, що нормальний розподіл Гауса не підходить для моделювання залишків інсоляції – він повністю ігнорує наявність двох погодних режимів. Визначені особливості розподілу частот залишків та емпіричної функції розподілу є фундаментом для подальшої генерації ансамблів синтетичних часових рядів, необхідних для верифікації проєктних рішень ФЕС [8–10].

Аналіз кореляцій. Попередні дослідження статистики рядів залишків показали наявність у рядах внутрішньої структури та зв'язків між елементами X_t (значеннями в різні моменти t). Для отримання інформації про кореляційну структуру застосуємо апарат функції автокореляції (ACF – Autocorrelation Function) та часткової автокореляції (PACF – Partial Autocorrelation Function) [5, 6]. Саме ці функції дають змогу виявити приховану періодичність і визначити структуру стохастичної моделі (наприклад, ARMA/ARIMA).

Функція ACF, що містить інформацію про кореляцію ряду з самим собою, зрушеним на *k* кроків, допомагає знайти сезонність (добову, місячну чи річну). Коефіцієнти ACF ρ_k вказують на кореляційну залежність між поточним значенням X_t і зсунутим на лаг *k* минулим значеннями X_{t-k}. Для стаціонарних рядів ці коефіцієнти визначаються на основі експериментальних даних або результатів моделювання як

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{E[(X-\mu)(X_{t-k}-\mu)]}{E[(X_{t-k}-\mu)^2]} \quad (3)$$

де γ_k – автоковаріація на лазі *k*, γ_0 – дисперсія ряду, μ – математичне очікування. ACF вказує на інерційність атмосфери та відображає тривалість інтервалів схожої погоди. Обчислення ACF дає змогу визначити порядок ковзного середнього (MA). Якщо ACF обривається після лага *q*, це свідчить на користь моделі MA(*q*).

Друга функція – PACF – показує кореляцію зі зсувом *k*, виключаючи вплив всіх проміжних. Це дає змогу зрозуміти, чи обумовлена інсоляція в день *i* безпосередньо інсоляцією в день *i–k*, чи цей зв'язок є лише наслідком ланцюжка проміжних метеорологічних станів. PACF – ключовий інструмент для визначення порядку авторегресії (AR). Різке згасання PACF після лага *p* вказує на модель AR(*p*).

Для фотоенергетики спільний кореляційний аналіз ACF/PACF важливий з таких причин. По-перше, він визначає тип процесу: за видом згасання функцій можемо судити, чи маємо ми справу з процесом AR, MA або змішаним типом ARMA. По-друге, цей аналіз дає змогу визначити горизонт прогнозу за ознакою виходу автокореляції за межі довірчого інтервалу (кореляція стає статистично незначною). При кластеризації ряду *X* це буде визначати межу передбачуваності конкретного кластера інсоляції.

Для отриманих внаслідок передобробки рядів інсоляції були розраховані коефіцієнти ACF і PACF для лагів від 1 до 12 у двох сценаріях: фактичні метеодані (*а*, *б*) і модель чистого неба (*в*, *г*). Результати наведені на рис. 4.

Для отриманих даних характерне швидке згасання обох функцій: ACF стає незначною (менше 0.05) на 9-му лазі, а PACF – вже на 3-му або 4-му. Для фактичних метеоданих (за показаннями датчиків) значення першого лага PACF(1) = 0.354, для ясного неба PACF(1) = 0.458. Відношення PACF(1)/ PACF(2) становить для цих випадків 4.72 і 7.75, відповідно. Це може говорити на користь вибору моделі ковзного середнього першого порядку для умов фактичної погоди та ясного неба. Враховуючи ці значення та поведінку корелограм, можна назвати найімовірнішими кандидатами для опису динаміки *X* моделі AR(1), AR(2) чи ARMA(1,1).

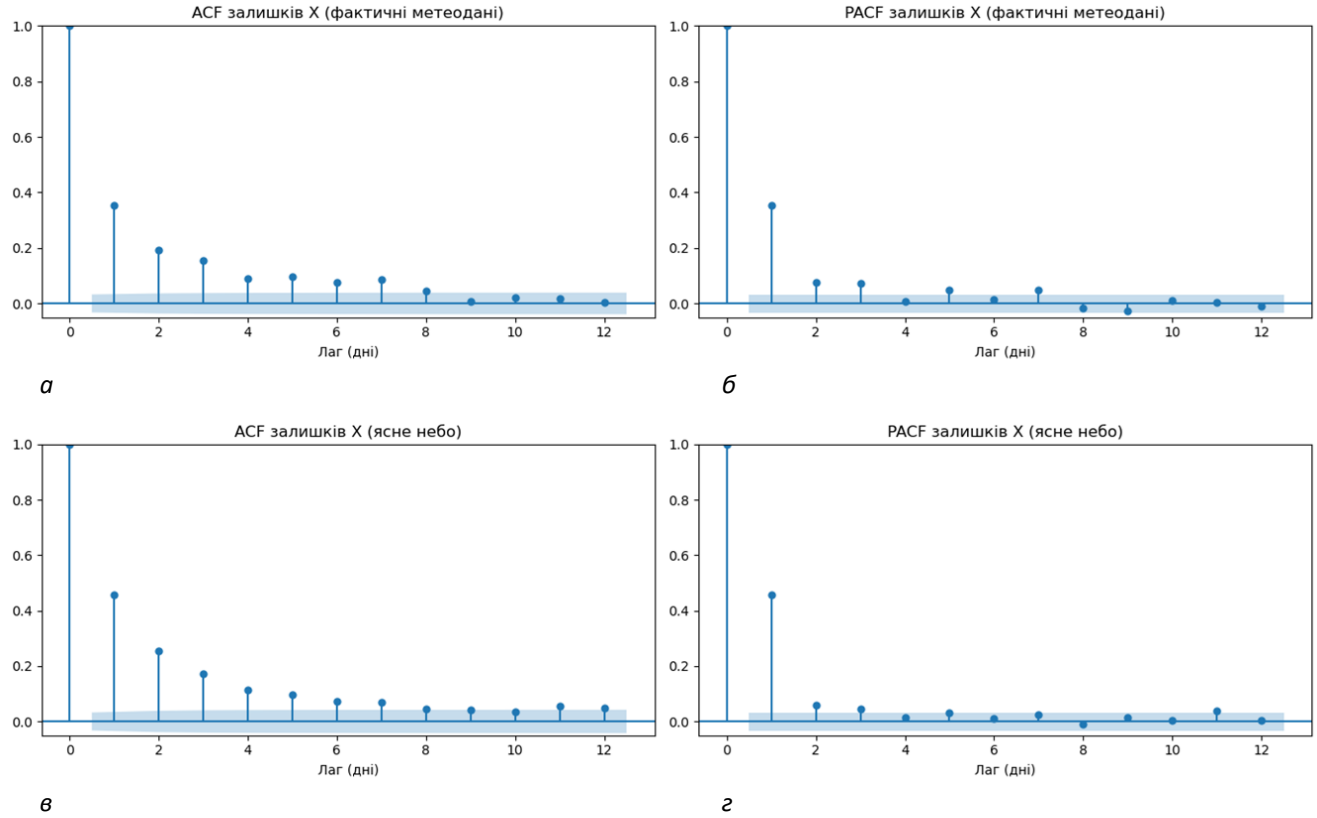


Рис. 4. Коефіцієнти автокореляції (ACF) (а, в) та частинної автокореляції (PACF) (бb, г) залишків X для фактичних метеоданих (а, б) та для моделі ясного (в, г). Блакитною смужкою показаний діапазон за межами довірчого інтервалу

Вибір та верифікація моделі ARMA. У моделі ARMA(p,q) часовий ряд X_t представлений як комбінація своїх попередніх значень X_{t-i} (авторегресія) і попередніх помилок (ковзне середнє). Рівняння для поточного X_t має вигляд

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \phi_i X_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \qquad (4)$$

де ϕ_i – параметри авторегресії (AR), θ_j – параметри ковзного середнього (MA), ε_t – білий шум (помилки) у момент часу *t*, *c* – константа.

Параметри ϕ_i показують вагу безпосереднього впливу минулого значення X_{t-i} на поточне X_t, коли вплив проміжних лагів $X_{t-i+1},\dots X_{t-1}$ вже врахований. Розрахунок ϕ_i здійснюється через коефіцієнти ρ_k (значення ACF) за допомогою систем лінійних рівнянь Юла – Уокера [5, 6]. Параметри θ_j часто описують інерційність атмосферних процесів: якщо θ велике, це означає, що «помилка» (випадкова зміна хмарності минулого часу) продовжує сильно впливати на поточне значення радіації. На графіку ACF це виглядає як короткий сплеск на перших 1–3 лагах, який швидко зникає. Взагалі з поведінки корелограм можна робити такі висновки: якщо ACF обривається, а PACF загасає – у ряді домінує MA-компонента; якщо PACF обривається, а ACF загасає – домінує AR-компонента.

Ідентифікація адекватної моделі з усього класу ARMA, тобто вибір значень p та q може бути здійснений двома способами.

1. Розрахунок параметрів ϕ_i, θ_j, виходячи з **корелограми ACF, PACF**. Для знаходження кожного конкретного параметра ϕ_k лага k потрібно розв'язати згадану систему *k* лінійних рівнянь Юла –Уокера, у якої ρ_k слугують коефіцієнтами. Параметри θ_j обчислюються за складнішою процедурою, оскільки потрібно розв'язувати систему рівнянь, які являють собою вирази ρ_k через нелінійні функції від множини змінних $\{\theta_j\}$. Розв'язання здійснюється послідовними ітераціями в межах методу моментів або за допомогою рекурсивного алгоритму інновацій [5].

2. Знаходження параметрів ϕ_i, θ_j, виходячи з наявного **емпіричного часового ряду X_t**. При цьому здійснюється підганяння ϕ_i, θ_j під ряд залишків методом максимальної правдоподібності [6].

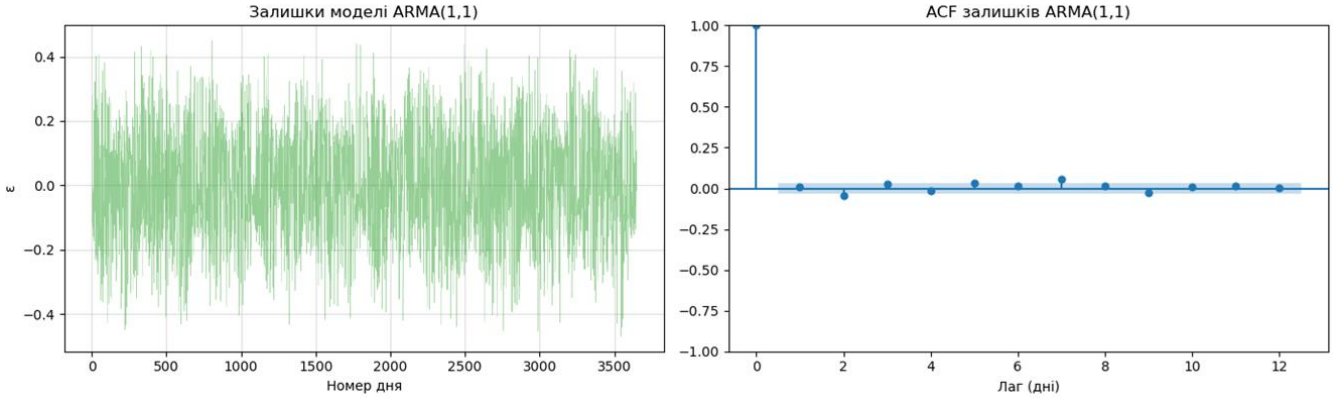
Для надійності ідентифікації ми використовували обидва способи. Для автоматичного перебору параметрів із заданого набору (у другому способі) застосовувався метод пошуку через сітку (Grid Search). Відбір оптимальної комбінації параметрів здійснювався за умовою мінімальності інформаційного критерію Акаїки (AIC) [19]: чим менше AIC, тим краще модель описує дані за мінімальної кількості параметрів.

Внаслідок розрахунку трьох моделей ARMA, що актуальні для опису інсоляції в м. Києві, були знайдені значення коефіцієнтів ϕ_i, θ_j у двох сценаріях: фактичної погоди та ясного неба (таблиця). Оскільки внесок ковзних середніх дуже помітний та має демпфуючий характер ($\theta_j < 0$) для подальших розрахунків вибираємо модель ARMA(1,1).

Таблиця. Кореляційні коефіцієнти моделей AR(1), AR(2) і ARMA(1,1)

Модель	Коефіцієнт	Фактична погода	Ясне небо
AR(1)	ϕ_1	0.354	0.457
AR(2)	ϕ_1	0.327	0.430
	ϕ_2	0.075	0.059
ARMA(1,1)	θ_1	-0.308	-0.139

Після видалення ряду ARMA з *X* отримані залишки та їх корелограма мають вигляд, показаний на рис. 5 (для фактичних метеоумов). На цьому етапі необхідно переконатися, що ці залишки мають характеристики, наближені до ідеальних шумових. Доведення моделі до стану «білого шуму» в залишках мінімізує ризик недооцінки екстремальних періодів (хмарно – ясно). Це критично важливе для автономних систем, де помилка з розрахунку «хвостів» розподілу веде до відмови системи електропостачання.



a *b*

Рис. 5. Результати для залишків ARMA(1,1): а – часовий ряд залишкі ARMA; б – корелограма ACF

Графік ACF (рис. 5, *б*) візуально нагадує білий шум: практично всі значення виходять за межи довірчого інтервалу – перебувають у синій смужці, крім лагу 7, який трохи «вистрибує» зі смужки. Тест Льюнга – Бокса показує такі значення критерію p-value: 0.001 – для 10-го лага, 0.022 – для 20-го та 0.063 – для 30-го. Це означає, що на коротких та середніх інтервалах (до 20 днів) у залишках все ще простежується слабка автокореляція. Саме лаг 7, який показує остаточний ефект тижневої циклічності, може впливати на показники короткострокових інтервалів. Навіть після застосування ARMA(1,1), складні метеорологічні процеси (наприклад, проходження серії фронтів або затяжні блокувальні антициклони) можуть залишати в даних «мікрозалежності», які проста лінійна модель не може повністю прибрати. Це може призводити до того, що навіть незначний сплеск на лазі 7 викликає формальне зниження *p-value* в тесті Льюнга – Бокса до 0.001. Хоча впевнено можна казати, що це не має вирішального значення для точності моделювання. Важливе те, що загальні амплітуди автокореляцій ACF на перших лагах знижені в десятки разів порівняно з вихідним ACF рядом залишків *X*.

Слід також зауважити, що при великому об'ємі вибірки (> 3650 відліків) тест Льюнга – Бокса стає дуже чутливим і виявляє навіть незначні відхилення від ідеального білого шуму, які практично не впливають на точність прогнозу. Таким чином, модель ARMA(1,1) можна визнати адекватною для синтетичної генерації рядів та їх подальшого використання під час моделювання фотоелектричних систем.

Обговорення та висновки. У цьому дослідженні представлено комплексний статистичний аналіз довгострокових щоденних рядів сонячного опромінення для Києва. Використовуючи дані глобального горизонтального опромінення (GHI) як для фактичних метеорологічних умов, так і для умов ясного неба, отримали індекс ясності K_t, який було розкладено на сезонну та стохастичну компоненти.

Результати показують, що розподіл частоти K_t має чітку бімодальну структуру з яскраво вираженими піками, що відповідають станам «хмарно» та «ясне небо». «Сідлова» зона між цими піками вказує на те, що проміжні стани змінної хмарності трапляються рідше та за своєю

суттю менш стабільні. Ця статистична неоднорідність додатково підтверджується кумулятивною функцією розподілу (CDF), де медіана перебцває в межах «сідлової» зони. Отже, ключовим висновком цього дослідження є те, що середній індекс прозорості K_t для Київської області являє собою «фізичний артефакт» – стан, який виникає з найнижчою статистичною ймовірністю. Ця проблема бімодального розподілу означає, що атмосфера переважно займає режими «хмарно» або «ясне небо». Отже, проєктування фотоелектричних систем, яке засноване виключно на довгострокових щомісячних середніх значеннях, призводить до значних систематичних помилок, потенційно переоцінюючи генерацію або не враховуючи критичні запаси акумульованої енергії, що необхідні протягом тривалих періодів похмурості.

Основна увага цього дослідження була приділена стохастичним залишкам *X* рядів *Kt*, які є важливими для прогнозування виробітку енергії та визначення розміру ємності систем зберігання. Стаціонарність часового ряду залишків була підтверджена за допомогою тесту ADF. Хоча ці залишки зберігають бімодальний розподіл частот, їхні максимуми на кривій PDF(X) розташовані ближче один до одного та зосереджені навколо медіани (нуля). «Перегин», що спостерігається поблизу медіани на графіку CDF(*X*), є прямим наслідком цієї бімодальності. Зміна знаку другої похідної та наявність двох різних мод надають вагомі докази проти використання гаусового (нормального) розподілу для моделювання залишків. Запропонована стохастична модель, що містить емпіричні характеристики PDF та CDF, забезпечує міцну основу для створення ансамблів синтетичних часових рядів, необхідних для оцінки показників ефективності та надійності фотоелектричних систем.

Для з'ясування внутрішньої структури та залежностей у рядах залишків функції ACF та PACF були обчислені як для фактичних метеорологічних сценаріїв, так і для сценаріїв ясного неба. Хоча тест Люнга – Бокса вказав на наявність автокореляції, подальший аналіз показав, що вона швидко спадає з часом. Для обох наборів даних коефіцієнти PACF(1) мали суттєві значення, тоді як наступні коефіцієнти PACF(2) були значно нижчими, що вказує на процес з короткочасною пам'яттю.

Оцінка параметрів для сімейства моделей ARMA(*p,q*) показала, що модель ARMA(1,1) є найадекватнфшою з погляду як точності, так і економного набору параметрів, що підтверджено критерієм AIC. Застосування цієї моделі ефективно перетворило складну бімодальну структуру відхилень *X* на стаціонарний процес. Отримані залишки ARMA наближаються до білого шуму, попри те, що тест Люнга – Бокса виявив слабку автокореляцію залишків. Незначний сплеск на лазі 7 відображає, імовірно, фундаментальну синоптичну періодичність – типовий цикл трансформації повітряних мас у помірно континентальному кліматі. Хоча модель ARMA(1,1) ефективно усуває первинну стохастичну інерцію, наявність цієї 7-денної «мікрозалежності» свідчить про те, що майбутні дослідження можуть отримати користь від інтеграції сезонних компонентів, таких як моделі SARIMA, для подальшого вдосконалення методу генерації синтетичних рядів.

Виявлені бімодальні особливості та визначені коефіцієнти ARMA забезпечують надійний стохастичний «генератор погоди», спеціально розроблений для Київської області. Ці результати можна безпосередньо інтегрувати в програмне забезпечення з моделювання генерації енергії для оптимізації ФЕС у мікромережах, що дасть змогу проводити точніші оцінки LCOE та забезпечить стабільність децентралізованих систем електропостачання за нестабільних метеорологічних умов.

ПОСИЛАННЯ

1. Ding Y. Data Science for Wind Energy. Boca Raton, FL: CRC Press, Taylor & Francis Group, 2020. 387 p. DOI: 10.1080/00401706.2020.1744901.
2. Zobaa A. F., Bihl T. J. (Eds). Big Data Analytics in Future Power Systems. Boca Raton: CRC Press, 2020. 188 p.
3. Unni A., Channi H. K. Data Analytics for Performance Optimization in Renewable Energy. In: Optimization in Sustainable Energy. Eds: P.Chatterjee, A.Khosla, A.Kumar, G.Demir Wiley, 2026. Ch. 9. https://doi.org/10.1002/9781394242139.ch9.
4. Кузнєцов М. П. Особливості комбінованих енергосистем з відновлюваними джерелами енергії: монографія. Київ: IBE, 2022. 142 с.
5. Brockwell P. J., Davis R. A. Introduction to Time Series and Forecasting. Third Edition. Springer, 2016. 425 p.
6. Box G. E. P., Jenkins G. M., Reinsel G. C., Ljung G. M. Time Series Analysis: Forecasting and Control. 5th Edition. Hoboken, New Jersey: John Wiley and Sons Inc., 2015. 712 p. https://doi.org/10.1111/jtsa.12194.
7. Said S. E., Dickey D. A. Testing for unit roots in autoregressive-moving average models of unknown order. Biometrika, 1984. Vol. 71. Issue 3. P. 599–607.
8. Larrañeta M., Fernandez-Peruchena C., Silva-Pérez M. A., Lillo-bravo I., Grantham A., Boland J. Generation of synthetic solar datasets for risk analysis. Solar Energy.– 2019. Vol. 187. P. 212–225. https://doi.org/10.1016/j.solener.2019.05.042.
9. Гаєвський О. Ю. Гаєвська Г. М. Синтетичні ряди інсоляції при розрахунках фотоелектричних станцій. Відновлювана енергетика. 2025. №. 3. С. 97–106. https://doi.org/10.36296/1819-8058.2025.3(82).97-106
10. Gaevskii O., Gaevska H. Two clustering approaches for generating synthetic insolation series. XXVII Conference "Renewable energy and energy efficiency in the XXI century". Kyiv, 2026. May 20–22.

11. Atmosphere Data Store. [Электронный ресурс]. URL: https://ads.atmosphere.copernicus.eu/.

12. Peixeiro M. Time Series Forecasting in Python. O'Reilly, 2022. 456 p. ISBN 9781617299889.

13. Time-series analysis and forecasting with Python. Tiger Data. [Електронний ресурс]. URL: https://www.tigerdata.com/learn/time-series-analysis-and-forecasting-with-python.

14. Khatib T., Ibrahim I.A., Mohamed A. A review on sizing methodologies of photovoltaic array and storage battery in a standalone photovoltaic system. Energy Conversion and Management, — 2016. Vol. 120. P. 430–448. https://doi.org/10.1016/j.enconman.2016.05.011.

15. Cristea M., Cristea C., Tîrnovan R. A., Șerban F. M. Levelized Cost of Energy (LCOE) of Different Photovoltaic Technologies. Appl. Sci. 2025. Vol. 15. Issue 12. 6710. https://doi.org/10.3390/app15126710.

16. Pillai D. S., Bayindir A. B., Thiruchutan A., Lopez Garcia J. et al. A comprehensive review of CAPEX-driven LCOE optimization strategies for utility-scale PV systems. Solar Energy. 2026. Vol. 306. 114296. https://doi.org/10.1016/j.solener.2025.114296.

17. Chen Y. C. A Tutorial on Kernel Density Estimation and Recent Advances. Biostatistics & Epidemiology. 2017. Vol. 1. Issue 1. DOI: 10.1080/24709360.2017.1396742.

18. Amato U., Andretta A., Bartoli B. et al. Markov processes and Fourier Analysis as a tool to describe and simulate daily solar irradiance. Solar Energy. 1986. Vol. 37. No. 3. Pp. 179–194. https://doi.org/10.1016/0038-092X(86)90075-7

19. Konishi S., Kitagawa G. Information Criteria and Statistical Modeling. Springer Nature. 2008. 274 p.

https://doi.org/10.36296/1819-8058.2026.2(85).140-156

STATISTICS OF SOLAR IRRADIANCE TIME SERIES FOR KYIV. I. FREQUENCY AND CORRELATION ANALYSES FOR PHOTOVOLTAIC SYSTEM APPLICATIONS

Received Apr. 22, 2026; accepted Jun. 26, 2026
Available online June. 30, 2026

Gaevskii O.[1], Gaevska H.[2]

Author for correspondence: Oleksandr Gaevskii,
e-mail: a.gaevskii@kpi.ua

[1] Doctor of Phys. Math. Sci., Prof.
https://orcid.org/0000-0001-6144-2441

[2] Senior teacher
https://orcid.org/0000-0001-7760-6789

[1, 2] NTUU «Igor Sikorsky Kyiv Polytechnic Institute», Kyiv, Ukraine

[1] Institute of Renewable Energy, NAS Ukraine, Kyiv, Ukraine

Abstract. *Modern photovoltaic (PV) systems demand precise design and power generation forecasting, making the investigation of regional statistical properties of solar radiation increasingly critical. This paper presents a statistical analysis of daily solar irradiation series for Kyiv over the past 10 years, expressed via the clearness index Kt and decomposed into seasonal and stochastic components. It is demonstrated that the frequency distribution of Kt exhibits a bimodal structure with pronounced peaks corresponding to "overcast" and "clear-sky" states. The "saddle" zone between these peaks indicates that intermediate states of variable cloudiness are less frequent and inherently less stable. This bimodality must be accounted for in PV system design, as calculations based solely on long-term monthly averages lead to significant systematic errors, potentially overestimating energy yield during prolonged overcast periods. The primary focus is placed on the stochastic residuals X of the Kt series, which are essential for generation forecasting and sizing battery energy storage systems (BESS). The stationarity of the residual time series, which retains a bimodal distribution, was confirmed via the Augmented Dickey-Fuller (ADF) test. Analysis of the autocorrelation (ACF) and partial autocorrelation (PACF) functions revealed short-term dependencies effectively captured by ARMA (p,q) models. Parameter estimation across this model family identified the ARMA (1,1) model as the most adequate in terms of both accuracy and parsimony, as confirmed by the Akaike Information Criterion (AIC). The results of these frequency and correlation analyses are essential for testing autonomous and backup PV systems through the generation of diverse weather scenarios, including worst-case conditions.*

Keywords: *solar radiation, clearness index, stochastic residuals, time series, Kyiv, stationarity, frequency distribution, KDE approximation, bimodality, ACF/PACF, ARMA models, photovoltaic systems, energy reliability.*

Abbreviations

ACF – Autocorrelation Function
ADF – Augmented Dickey-Fuller (test)
ADS – Atmosphere Data Store
AIC – Akaike Information Criterion
ARMA – Autoregressive Moving Average
BESS – Battery Energy Storage System
CAMS – Copernicus Atmosphere Monitoring Service
CDF – Cumulative Distribution Function
ERA5 – ECMWF Reanalysis v5
GHI – Global Horizontal Irradiation
KDE – Kernel Density Estimation
LCOE – Levelized Cost of Energy
PACF – Partial Autocorrelation Function
PDF – Probability Density Function
PV – Photovoltaic
RMSE – Root Mean Square Error

Introduction. The rapid development of photovoltaic (PV) systems and their integration into power grids demand high accuracy in design and generation forecasting. Unlike conventional power plants, solar energy is highly stochastic due to complex atmospheric dynamics. With the rise of decentralized power and autonomous (backup) PV systems, standard methods based on monthly averages are insufficient. Consequently, investigating the statistical properties of meteorological factors (irradiance, ambient temperature, wind speed) is essential for assessing the reliability of renewable energy installations [1–3]. Furthermore, regional solar radiation analysis is vital for evaluating hybrid system performance [4].

The efficiency and reliability of energy facilities depend directly on understanding the structure of insolation time series. Statistical analysis allows for isolating deterministic seasonal trends and investigating stochastic residuals—

deviations caused by local weather. These fluctuations determine the probability of energy deficits, compliance with reliability criteria, and the required capacity for battery energy storage systems (BESS).

The Kyiv region is of particular interest due to its significant cloud cover variability, especially during transitional autumn–spring seasons. The high amplitude of stochastic fluctuations in this temperate continental climate necessitates advanced mathematical tools, including frequency and correlation analyses (ACF/PACF) and Autoregressive Moving Average (ARMA) modeling [5, 6].

Objective and Problem Statement. The objective of this work is to develop and verify a statistical model describing the internal structure and variability patterns of solar insolation time series in Kyiv, based on data from the last decade. To achieve this, the following tasks were addressed:

- *Time series decomposition*: Extracting the long-term seasonal component and deriving the stochastic insolation residual series;
- *Frequency (probabilistic) analysis***:** Investigating the probability density functions (PDF) and cumulative distribution functions (CDF) of the clearness index Kt and its residuals X;
- *Correlation diagnostics:* Computing and analyzing autocorrelation (ACF) and partial autocorrelation (PACF) functions to assess the "memory" of the process;
- *Model identification and parameterization:* Selecting the optimal ARMA model, calculating its coefficients, and assessing the statistical significance of parameters;
- *Model verification:* Evaluating model adequacy by testing residual noise for stationarity and the absence of serial dependence (using ADF and Ljung–Box tests [6, 7]).

To ensure the reliability of the statistical conclusions, th study utilizes a daily insolation dataset for Kyiv covering th most recent ten-year period (2016–2025). This retrospe tive approach accounts for both seasonal cycles and syno tic variability, minimizing the impact of single-year weathe anomalies. Such long-term analysis is critical for optimizin PV system components, long-term forecasting, and generating synthetic insolation series for performance diagnostics [8–10].

The primary data consists of daily global horizontal irradiation (GHI) series obtained from the Atmosphere Data Store (ADS) of the Copernicus Atmosphere Monitoring Service (CAMS) [11]. This service provides modeled data for both clear-sky and actual weather conditions. Comparing these scenarios allowed for isolating the specific contribution of cloudiness to the stochastic structure of the time series. The ADS service, based on the global ERA5-Land reanalysis, ensures high temporal and spatial (9 km) resolution and statistical robustness.

Data processing and analysis were performed in the Python/Spyder environment using the *NumPy*, *Pandas*, *SciPy*, *Statsmodels*, *Scikit-learn*, and *Tslearn* libraries [12, 13].

Structure of Daily Solar Irradiation Series for Kyiv. To analyze the statistical patterns of surface solar irradiation, we employ the clearness index K_t—a dimensionless parameter defined as the ratio of global horizontal irradiation (GHI) at the Earth's surface to the extraterrestrial irradiation G_0:

$$K_t = G/G_0\,, \tag{1}$$

where G is the global horizontal irradiation at the Earth's surface (GHI) and G_0 is the extraterrestrial horizontal irradiation. Transitioning to K_t facilitates the conversion of the original series into a stationary form. A necessary condition for this is the elimination of the deterministic astronomical component G_0. During the calculation of K_t (1), improbable values outside the range of 0.05 to 0.85 are filtered out.

When analyzing solar radiation behavior, it should be noted that the time series is not homogeneous but represents a superposition of processes of different natures. Therefore, the K_t series is conventionally decomposed into a seasonal component and residuals:

$$K_t = K_{t,\mathrm{season}} + X. \tag{2}$$

The first component $K_{t,\mathrm{season}}$ represents the deterministic part governed by cyclic fluctuations driven by astronomical factors: the Earth's rotation (diurnal cycle) and its orbit around the Sun (annual cycle). The second component, X, represents stochastic residuals—the difference between actual measurements and the seasonal trend. These residuals arise from atmospheric dynamics, including cloud movement, humidity changes, and atmospheric aerosol content.

Analyzing the series (2) as a whole would result in pronounced seasonality masking rapid, short-term local changes. Isolating the seasonality allows the mathematical apparatus to focus on random deviations. This is crucial for:

Minimizing modeling errors: (e.g., by the RMSE criterion), as a model trained on residuals more accurately reflects short-term weather variability;

Optimizing computations: by focusing specifically on stochastic processes without the redundancy of the seasonal component.

The temporal dependence of $K_{t,season}$ defines the theoretical maximum PV generation for a specific location and time. This is typically described by Clear Sky Models and contains no inherent uncertainty requiring statistical modeling. The uncertainty resides within the residuals X, where the volatility that complicates forecasting is concentrated. Residuals aggregate the effects of meteorological turbulence, random cloud cover changes, and aerosol composition. Extracting residuals ensures time series stationarity, a mandatory condition for applying ARMA models and ensuring the convergence of machine learning algorithms [5, 6].

The time series of daily clearness index K_t values and residuals X over the 10-year interval are shown in Fig. 1 (top and bottom, respectively). The seasonal component, exhibiting a clear periodic character, is superimposed on the K_t plot. Statistical estimates for the obtained series are as follows:

the mean values are 0.466 for K_t, and 0.00018 for X (which should ideally be zero). The standard deviation of the residuals is 0.1832. Verification of the residual series *X* using the Augmented Dickey–Fuller (ADF) test [6, 7] confirmed its stationarity (*p-value* << 0.001). The Ljung–Box (LB) test [5, 6] confirmed the presence of autocorrelation in the residuals *X* at lags up to 20 days ($p < 0.05$). This justifies the application and identification of an autoregressive moving average ARMA(p,q) model.

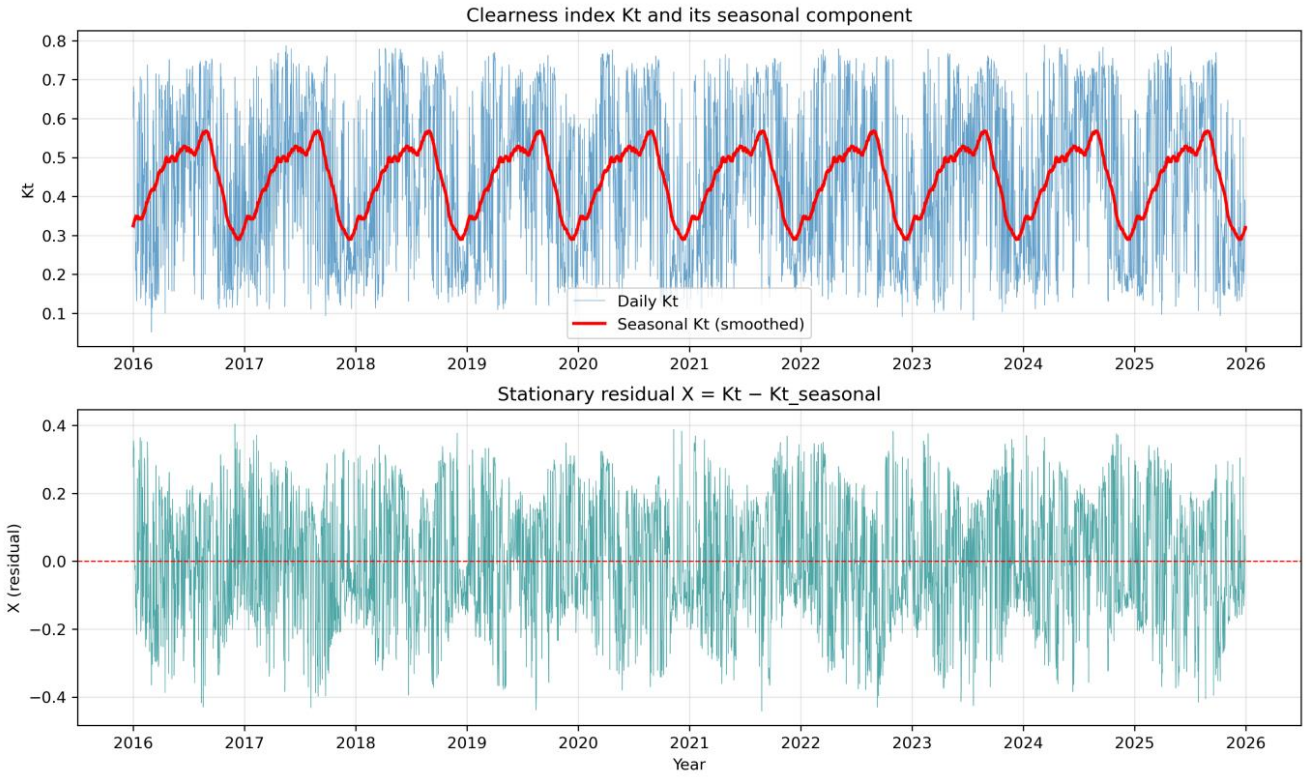


Fig. 1. Time series for the 2016–2025 period for Kyiv: top—daily clearness index K_t values with the seasonal component (red curve); bottom—stochastic residuals X

Frequency and Cumulative Distributions of the Index and Residuals. Analyzing the probability density function (PDF) and cumulative distribution function (CDF) of the clearness index and stochastic residuals is fundamental to understanding solar resource dynamics beyond predictable seasonal cycles. While a PV system's theoretical potential is dictated by the deterministic clear-sky component, assessing actual operational reliability requires a statistical analysis of real-world insolation, including its variance, skewness, and distribution shape. From a PV engineering perspective, investigating K_t and *X* distributions is essential for the following tasks:

- *Forecasting extreme regimes:* Analyzing distribution "tails" enables the estimation of the probability and duration of significant generation deficits caused by anomalous weather, which monthly average data fail to capture.
- *Calculating reliability criteria:* (e.g., Loss of Load Probability—LLP) [9, 14]. Accurate knowledge of the residual distribution is necessary to compute energy deficit probabilities and justify the selection of PV array and storage capacities.
- *Stochastic optimization:* The PDF shape provides a basis for modeling realistic operational scenarios under uncertainty, essential for long-term economic viability assessments (LCOE, CAPEX) [15, 16].

The statistical results for the daily K_t series are presented in Fig. 2 as a frequency histogram with a nonparametric Kernel Density Estimation (KDE) approximation [17], and a CDF with PCHIP (Piecewise Cubic Hermite Interpolating Polynomial) interpolation. The domain of the approximating PDF was defined based on the empirical boundaries of the collected data. This approach truncates the theoretically infinite "tails" of the approximation functions, ensuring the reliability of synthetic series generated for PV system testing.

The frequency distribution of the clearness index (Fig. 2a) reveals a bimodal structure with two pronounced local peaks. The first peak (0.20) corresponds to the "overcast" regime, characteristic of Kyiv's climate, particularly during autumn and winter. The second peak (0.73) corresponds to the "clear-sky" regime. The "saddle" zone between these peaks, characterized by low density, indicates that intermediate states (variable cloudiness) are less stable; the atmosphere tends to transition toward one of the two dominant states.

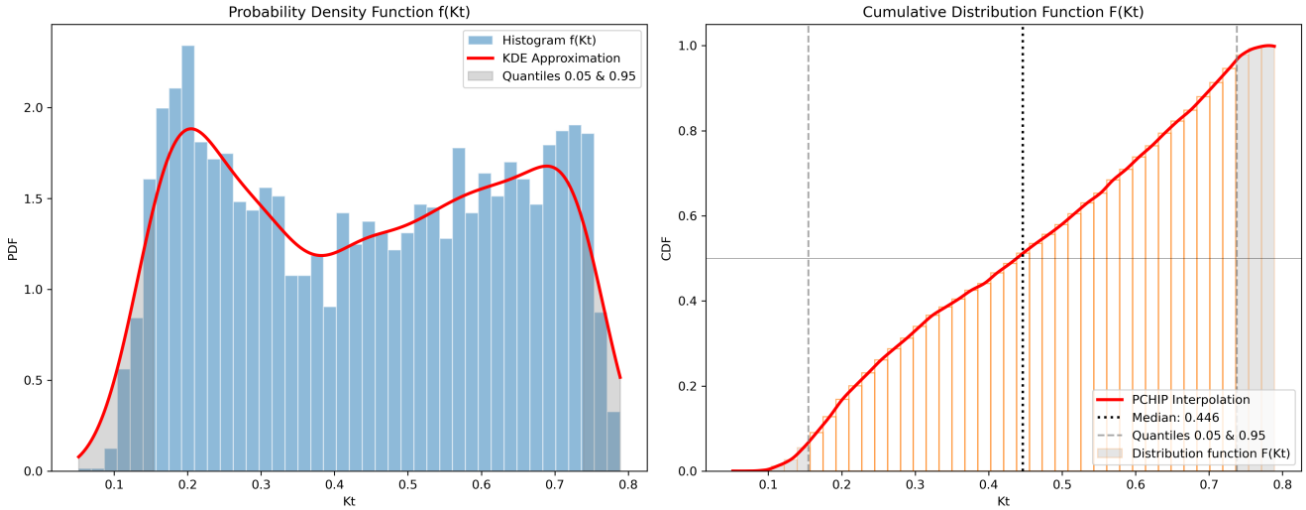


Fig. 2. Statistical analysis of the clearness index K_t: (a) frequency histogram and its nonparametric KDE approximation; (b) cumulative distribution function (CDF) with PCHIP interpolation. Shading indicates 95% confidence intervals

The CDF plot (Fig. 2b) confirms the statistical inhomogeneity of K_t: the median (0.446) falls precisely within the PDF's "saddle" zone. This signifies that the mean clearness index for Kyiv describes a state that occurs least frequently in reality. This finding strongly supports the argument against relying on simplistic average solar radiation models for PV engineering applications in this region.

The plots (Fig. 2a, 2b) illustrate the 0.05–0.95 quantile interval (confidence interval), which constrains the clearness index K_t range (approximately from 0.15 to 0.74), beyond which low-probability extreme outliers occur. The KDE approximation (red line on the PDF) smooths histogram noise and highlights local distribution maxima. PCHIP interpolation on the CDF plot ensures curve monotonicity, which is critical for accurate series synthesis using the inverse transform sampling method [9, 18].

The frequency distribution of stochastic residuals X (Fig. 3a) also exhibits bimodality, though less pronounced. This noteworthy result indicates that bimodality is not merely a seasonal artifact—since seasonal effects were removed during the calculation of X—but a fundamental property of solar radiation in the region.

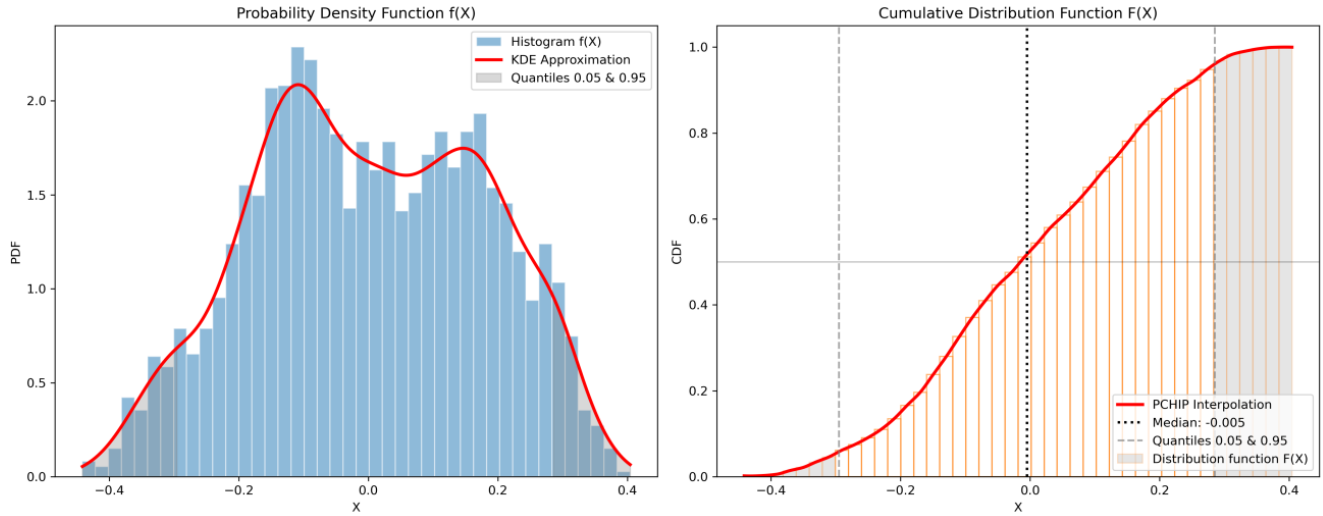


Fig. 3. Statistical analysis of residuals X: (a) frequency histogram and its nonparametric KDE approximation; (b) cumulative distribution function (CDF) with PCHIP interpolation. Shading indicates 95% confidence intervals

Although the PDF of the residuals (Fig. 3a) retains a bimodal structure, its character changes qualitatively. The two maxima converge, and the "saddle" zone becomes shallower compared to the PDF of K_t. The residual modes have shifted and are now centered around zero (approximately -0.12 and 0.13). The negative peak represents a radiation "deficit" due to cloud cover, while the positive peak reflects a radiation "surplus" relative to the seasonal trend during clear periods. The distribution has become more symmetric and compact, facilitating statistical series synthesis as the

extreme tails are less pronounced. Furthermore, the residuals exhibit the necessary centering (median of -0.005) required for ARMA modeling.

On the CDF plot (Fig. 3b), the curve is smoother, approaching an S-shape (typical of a normal distribution) but with a distinct inflection near the median (a change in the sign of the second derivative). This inflection is a direct consequence of bimodality and serves as a strong argument that the Gaussian distribution is unsuitable for modeling insolation residuals, as it fails to account for the two distinct weather regimes. These identified features of the residual frequency distribution and the empirical CDF form the basis for generating ensembles of synthetic time series required to verify PV system design solutions [8–10].

Correlation Analysis. The statistical investigation of the residual series revealed an internal structure and dependencies between elements X_t at different time steps t. To characterize this, we utilize autocorrelation (ACF) and partial autocorrelation (PACF) functions [5, 6]. These tools enable the detection of hidden dependencies and help determine the structure of the stochastic model (e.g., ARMA/ARIMA).

The ACF, which measures the correlation of a series with its own values shifted by k steps, helps identify potential periodicities. The ACF coefficients ρ_k quantify the correlation between the current value X_t and the past value X_{t-k} at lag k. For stationary series, these coefficients are determined as:

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{E[(X-\mu)(X_{t-k}-\mu)]}{E[(X_{t-k}-\mu)^2]} \quad (3)$$

where γ_k is the autocovariance at lag k, γ_0 is the series variance, and μ is the expected value. The ACF indicates atmospheric inertia and reflects the persistence of similar weather conditions. It is used to determine the moving average (MA) order; if the ACF cuts off after lag q, an MA(q) model is suggested.

The PACF measures the correlation at lag k while excluding the influence of all intermediate lags. This helps determine whether insolation on day i-k is directly influenced by day i or if the relationship is merely a consequence of a chain of intermediate meteorological states. The PACF is the primary tool for determining the autoregression (AR) order, where a sharp decay after lag p indicates an AR(p) model.

For PV engineering, joint ACF/PACF analysis is crucial for two reasons. First, it identifies the process type (AR, MA, or mixed ARMA) based on the decay patterns. Second, it defines the forecasting horizon; once the autocorrelation falls within the confidence interval, it becomes statistically insignificant, marking the predictability boundary for a specific insolation cluster.

For the preprocessed insolation series, ACF and PACF coefficients were calculated for lags 1 to 12 across two scenarios: actual meteorological data and the clear-sky model. The results are presented in Fig. 4.

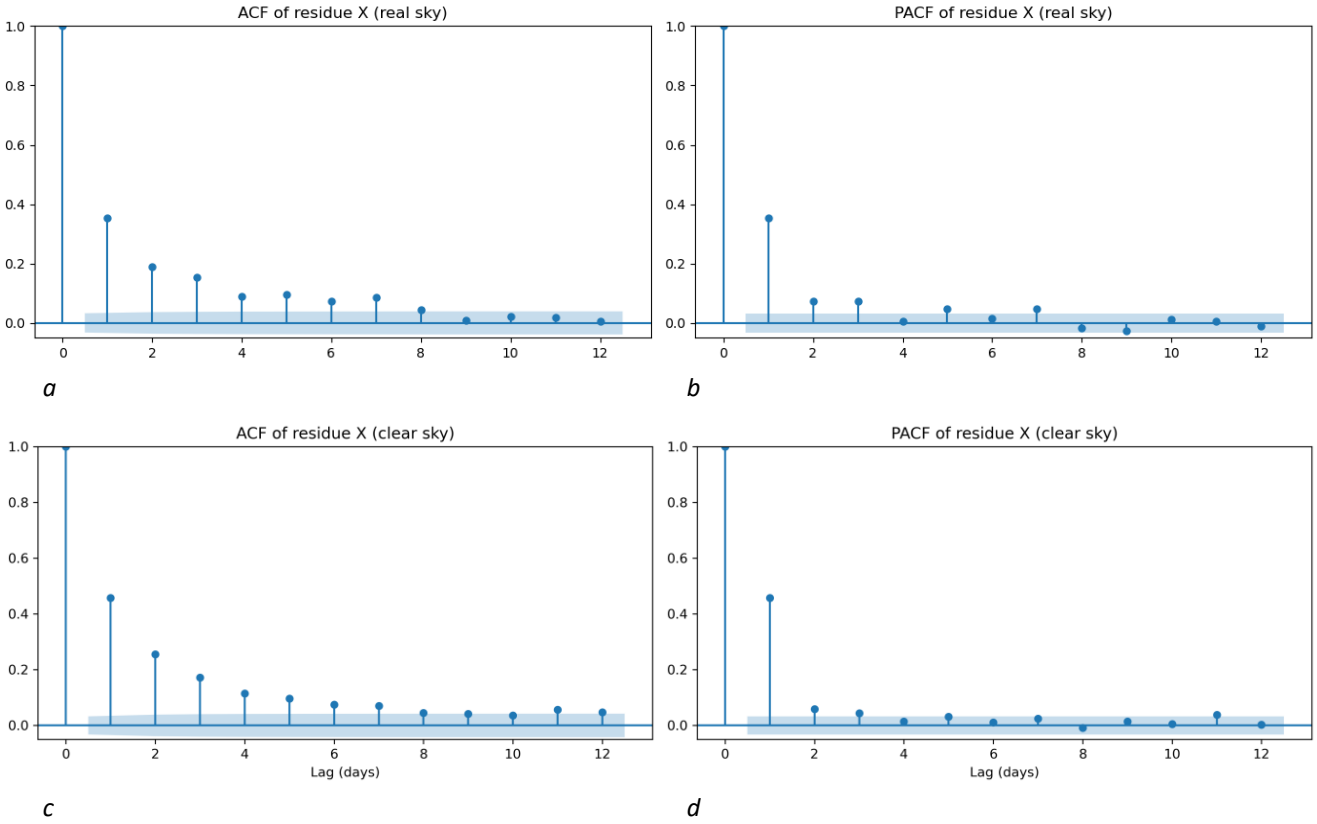


Fig. 4. Autocorrelation (ACF) (a, c) and partial autocorrelation (PACF) (b, d) coefficients of residuals X for actual meteorological data (a, b) and the clear-sky model (c, d). The blue shading indicates the insignificance zone

The results show rapid decay in both functions: the ACF bc comes statistically insignificant by lag 9, while the PAC does so by lag 3 or 4. For actual meteorological data, th first lag value is PACF(1) = 0.354; for clear-sky condition PACF(1) = 0.458. The PACF(1)/PACF(2) ratios are 4.72 an 7.75, respectively. This suggests that a first-order mod may be appropriate for both cases. Based on the correlo-gram behavior, the most likely candidates for describing *X* dynamics are AR(1), AR(2), or ARMA(1,1) models.

ARMA Model Selection and Verification. In the ARMA(*p,q*) model, the time series X_t is represented as a linear combination of its previous values (autoregression) and previous errors (moving average). The equation for X_t is:

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \phi_i X_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} , \quad (4)$$

where ϕ_i are AR parameters, θ_j are MA parameters, ε_t represents white noise (error) at time *t* and *c* is a constant.

The parameters ϕ_i quantify the direct influence of X_{t-i} on X_t, accounting for the influence of intermediate lags. These are calculated via the Yule–Walker equations using ACF coefficients [5, 6]. The θ_j parameters describe the inertia of atmospheric processes; a large θ suggests that a past "error" (e.g., a random cloud cover change) continues to influence current radiation. This typically appears as a short burst in the first 1–3 lags of the ACF. Generally, if the ACF cuts off while the PACF decays, the MA component dominates; conversely, if the PACF cuts off while the ACF decays, the AR component dominates.

The identification of an adequate model within the ARMA class – specifically, the selection of *p* and *q* values – can be achieved through two primary approaches:

- *Estimation of ϕ_i and θ_j based on ACF/PACF correlograms:* To find a specific parameter ϕ_k at lag *k*, the aforementioned system of *k* Yule–Walker linear equations must be solved, using ρ_k as coefficients. Calculating the θ_j parameters is more complex, requiring the solution of a system where ρ_k is expressed as a nonlinear function of the variable set $\{\theta_j\}$. This is typically performed using successive iterations within the method of moments or the recursive innovations algorithm [5].
- *Estimation of ϕ_i, and θ_j from the empirical time series X_t:* This involves fitting the parameters to the residual series using the maximum likelihood estimation (MLE) method [6].

To ensure identification robustness, both approaches were employed. For the second approach, a Grid Search was used to iterate through parameter sets. The optimal model was selected based on the Akaike Information Criterion (AIC) [19] minimality condition: a lower AIC indicates a superior model that describes the data with a minimum number of parameters.

After evaluating three ARMA models relevant for describing Kyiv's insolation, the ϕ_i and θ_j coefficients were determined for both actual weather and clear-sky scenarios (Table). Since the moving average contribution is significant and exhibits a damping character (θ_j < 0), the ARMA (1,1) model was selected for further analysis.

Table. Correlation coefficients of AR(1), AR(2), and ARMA(1,1) models

Model	Coefficient	Actual weather:	Clear sky
AR(1)	ϕ_1	0.354	0.457
AR(2)	ϕ_1	0.327	0.430
	ϕ_2	0.075	0.059
ARMA(1,1)	θ_1	-0.308	-0.139

After applying the ARMA model to X_t, the resulting residuals and their correlogram are shown in Fig. 5 (for actual meteorological conditions). At this stage, it is essential to verify that these residuals approximate ideal white noise. Achieving white noise in the residuals ensures that the model does not underestimate extreme periods (overcast vs. clear). This is critical for autonomous systems, where errors in distribution "tails" can lead to power supply failure.

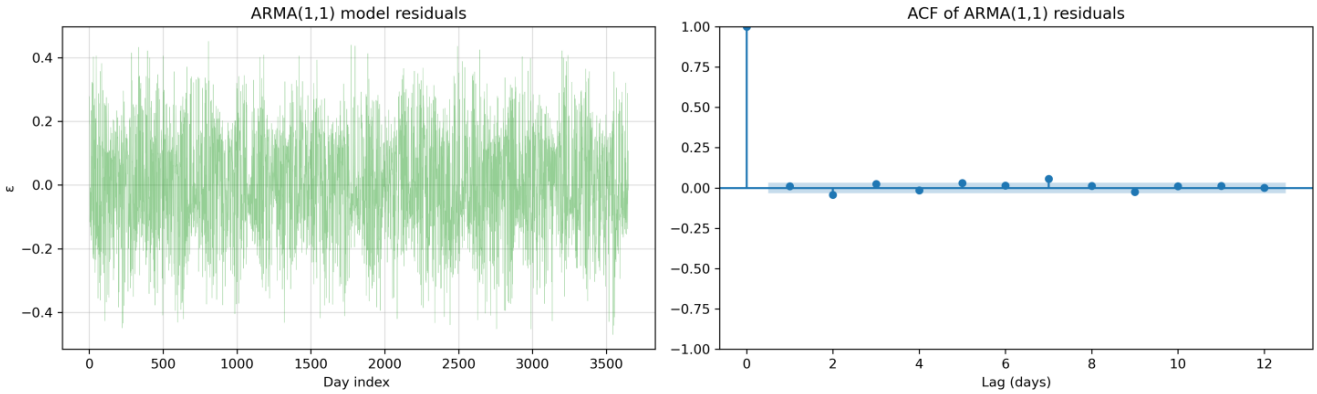


Fig. 5. Results for ARMA(1,1) residuals: (a) ARMA residual time series; (b) ACF correlogram

The ACF plot (Fig. 5b) visually approximates white noise, as nearly all values fall within the confidence interval (the blue band), except for a minor spike at lag 7. The Ljung–Box test yields p-values of 0.001 for lag 10, 0.022 for lag 20, and 0.063 for lag 30. This indicates that some weak autocorrelation persists at short to medium intervals (up to 20 days).

Specifically, the spike at lag 7 suggests a residual weekly cyclicity that may influence short-term indicators. Even after applying the ARMA(1,1) model, complex meteorological processes—such as a series of fronts or prolonged blocking anticyclones—may leave "micro-dependencies" that a linear model cannot entirely eliminate. Although these minor deviations formally reduce the Ljung–Box p-value at lag 10, they do not significantly compromise modeling accuracy. Crucially, the overall ACF amplitudes at the initial lags are reduced by an order of magnitude compared to the original ACF of the *X* series.

Furthermore, given the large sample size (exceeding 3,650 observations), the Ljung–Box test becomes hyper-sensitive, detecting minor deviations from ideal white noise that have negligible impact on forecasting accuracy. Therefore, the ARMA (1,1) model is deemed adequate for generating synthetic series and for subsequent use in photovoltaic system modeling.

Discussion and Conclusions. This study presented a comprehensive statistical analysis of long-term daily solar irradiation series for Kyiv. Using global horizontal irradiation (GHI) data for both actual meteorological and clear-sky conditions, the clearness index K_t was derived and decomposed into seasonal and stochastic components.

The results demonstrate that the frequency distribution of K_t exhibits a distinct bimodal structure, with pronounced peaks corresponding to "overcast" and "clear-sky" states. The "saddle" zone between these peaks indicates that intermediate states of variable cloudiness are less frequent and inherently less stable. This statistical inhomogeneity is further confirmed by the cumulative distribution function (CDF), where the median resides within the "saddle" zone. Consequently, a key conclusion of this study is that the average clearness index K_t for the Kyiv region represents a 'physical artifact'—a state that occurs with the lowest statistical probability. This bimodal distribution challenge implies that the atmosphere predominantly occupies either 'overcast' or 'clear-sky' regimes. Therefore, PV system design based solely on long-term monthly averages leads to significant systematic errors, potentially overestimating generation or failing to account for the critical reliability margins required during prolonged overcast periods.

The primary focus of this investigation was the stochastic residuals *X* of the K_t series, which are essential for energy production forecasting and PV storage capacity sizing. The residual time series was confirmed to be stationary via the ADF test. While these residuals retain a bimodal frequency distribution, their maxima on the PDF(X) curve are more closely spaced and centered around the median (zero). The "inflection" observed near the median on the CDF(X) plot is a direct consequence of this bimodality. The sign change of the second derivative and the presence of two distinct modes provide robust evidence against using a Gaussian normal distribution for residual modeling. The proposed stochastic model, incorporating empirical PDF and CDF characteristics, provides a solid framework for generating ensembles of synthetic time series necessary for evaluating PV system efficiency and reliability indicators.

To elucidate the internal structure and dependencies within the residual series, the ACF and PACF were computed for both actual meteorological and clear-sky scenarios. Although the Ljung-Box test indicated the presence of autocorrelation, the analysis revealed that it decays rapidly over time. For both datasets, the PACF(1) coefficients were significant, while subsequent PACF(2) values were substantially lower, indicating a short-term memory process.

Parameter estimation for the ARMA(p,q) model family demonstrated that the ARMA(1,1) model is the most adequate in terms of both accuracy and parsimony, as confirmed by the AIC criterion. Applying this model effectively transformed the complex bimodal structure of the *X* deviations into a stationary process. The resulting ARMA residuals closely approximate white noise, despite the Ljung-Box test detecting a weak residual autocorrelation. The minor spike at lag 7 likely reflects a fundamental synoptic periodicity—the typical cycle of air mass transformation in temperate continental climates. While the ARMA(1,1) model effectively eliminates the primary stochastic inertia, the presence of this 7-day 'micro-dependency' suggests that future research could benefit from integrating seasonal components, such as SARIMA models, to further refine the method for synthetic series generation.

The identified bimodal features and the determined ARMA coefficients provide a robust stochastic 'weather generator' specifically tailored for the Kyiv region. These results can be directly integrated into energy modeling software for the optimization of PV-microgrids, enabling more accurate LCOE assessments and ensuring the stability of decentralized power supply systems under volatile meteorological conditions.

REFERENCES

1. Ding Y. Data Science for Wind Energy. Boca Raton, FL: CRC Press, Taylor & Francis Group, 2020. 387 pp. DOI: 10.1080/00401706.2020.1744901.
2. Zobaa A.F., Bihl T.J. (Eds). Big Data Analytics in Future Power Systems. Boca Raton: CRC Press, 2020. 188 pp.
3. Unni A., Channi H.K. Data Analytics for Performance Optimization in Renewable Energy. In: Optimization in Sustainable Energy. Eds: Chatterjee P., Khosla A., Kumar A., Demir G. Wiley, 2026. Ch. 9. https://doi.org/10.1002/9781394242139.ch9.
4. Kuznetsov M. P. Features of combined energy systems with renewable energy sources: monograph. — Kyiv: IRE, 2022. 142 p.
5. Brockwell P.J., Davis R.A. Introduction to Time Series and Forecasting. Third Edition. Springer, 2016. 425 pp.
6. Box G.E.P., Jenkins G.M., Reinsel G.C., Ljung G.M. Time Series Analysis: Forecasting and Control. 5th Edition.

Hoboken, New Jersey: John Wiley and Sons Inc., 2015. 712 pp. https://doi.org/10.1111/jtsa.12194.

7. Said S.E., Dickey D.A. Testing for unit roots in autoregressive-moving average models of unknown order. Biometrika, — 1984. Vol. 71. Issue 3. P. 599–607.
8. Larrañeta M., Fernandez-Peruchena C., Silva-Pérez M.A., Lillo-bravo I., Grantham A., Boland J. Generation of synthetic solar datasets for risk analysis. Solar Energy, – 2019, Vol. 187, P.212-225. https://doi.org/10.1016/j.solener.2019.05.042.
9. Gaevskii O., Gaevska H. A synthetic insolation series in sizing calculations of PV plants. Vidnovlyuvana Energetica. –2025, No3, https://doi.org/10.36296/1819-8058.2025.3(82).97-106
10. Gaevskii O., Gaevska H. Two clustering approaches for generating synthetic insolation series. XXVII Conference "Renewable energy and energy efficiency in the XXI century". Kyiv, — 2026. May 20–22.
11. Atmosphere Data Store. [Электронный ресурс]. URL: https://ads.atmosphere.copernicus.eu/.
12. Peixeiro M. Time Series Forecasting in Python. O'Reilly, 2022. 456 pp. ISBN 9781617299889.
13. Time-series analysis and forecasting with Python. Tiger Data. URL: https://www.tigerdata.com/learn/time-series-analysis-and-forecasting-with-python.
14. Khatib T., Ibrahim I.A., Mohamed A. A review on sizing methodologies of photovoltaic array and storage battery in a standalone photovoltaic system. Energy Conversion and Management, — 2016. Vol. 120. P. 430–448. https://doi.org/10.1016/j.enconman.2016.05.011.
15. Cristea M., Cristea C., Tîrnovan R.A., Șerban F.M. Levelized Cost of Energy (LCOE) of Different Photovoltaic Technologies. Appl. Sci., — 2025. Vol. 15. Issue 12. 6710. https://doi.org/10.3390/app15126710.
16. Pillai D.S., Bayindir A.B., Thiruchutan A., Lopez Garcia J. et al. A comprehensive review of CAPEX-driven LCOE optimization strategies for utility-scale PV systems. Solar Energy, — 2026. Vol. 306. 114296. https://doi.org/10.1016/j.solener.2025.114296.
17. Chen Y.C. A Tutorial on Kernel Density Estimation and Recent Advances. Biostatistics & Epidemiology, — 2017. Vol. 1. Issue 1. DOI: 10.1080/24709360.2017.1396742.
18. U. Amato, A. Andretta, B. Bartoli et al. Markov processes and Fourier Analysis as a tool to describe and simulate daily solar irradiance. Solar Energy, 1986, Vol. 37, N 3, 179-194. https://doi.org/10.1016/0038-092X(86)90075-7
19. Konishi S., Kitagawa G. Information Criteria and Statistical Modeling. Springer Nature, 2008. 274 pp.