

Hybrid multi-task deep learning framework for joint forecasting of solar and wind power

Assemgul Tynykulova¹, Raushan Moldasheva², Elmira Eldarova³, Yerlan Izbassarov², Shynar Kodanova⁴, Lunara Diyarova⁵, Saya Baigubenova⁵, Zhanar Azhibekova⁶

¹Higher School of Information Technology and Engineering, Astana International University, Astana, Kazakhstan

²Department of Software Engineering, Faculty of Physics, Mathematics and Information Technology, Kh. Dosmukhamedov Atyrau University, Atyrau, Kazakhstan

³Department of Computer Science, Faculty of Physics, Mathematics and Information Technology, Kh. Dosmukhamedov Atyrau University, Atyrau, Kazakhstan

⁴Department of Production Automation and Information Technologies, Faculty of Information Technology, Atyrau University of Oil and Gas named after S. Utebayev, Atyrau, Kazakhstan

⁵Institute of Digital Economy and Sustainable Development, Non-Profit Joint-Stock Company «Zhangir Khan West Kazakhstan Agrarian-Technical University», Uralsk, Kazakhstan

⁶Department of Information and Communication Technologies, Non-profit Joint Stock Company S.Asfendiyarov Kazakh National Medical University, Almaty, Kazakhstan

Article Info

Article history:

Received Jan 28, 2026

Revised Aug 27, 2026

Accepted Sep 4, 2026

Keywords:

Extreme events

Hybrid deep learning models

Machine learning

Multi-task learning

Renewable energy

ABSTRACT

This study presents the adaptive three-expert ensemble (A3E), a reproducible framework for joint one-hour-ahead solar and wind power forecasting. A3E combines temporal, physically informed, and high-generation Extra Trees experts through a causal local-error gate and is evaluated under a strictly chronological, leakage-free protocol. After timestamp reconciliation and causal feature construction, the primary source yielded 8,700 model-ready observations: 5,775 for training, 1,463 for validation, 719 for internal testing, and 743 for a same-source December holdout. The primary scale-neutral comparison used Macro train-SD normalized root mean square error (NRMSE) on 672 common forecast origins. A3E achieved 0.1978, compared with 0.2314 for the official photovoltaic (PV)-Client 2026 comparator and 0.2340 for TimeMixer. However, validation-optimized static weights achieved 0.2017 on the complete 719-origin internal test, compared with 0.2025 for adaptive A3E; the difference was not statistically significant under dependence-aware analysis. On the independent CISO and ERCO systems, TimeMixer outperformed A3E, demonstrating source-dependent model rankings. Event evaluation used 98 CISO and 320 ERCO high-generation events. The results support the value of heterogeneous expert integration, but not a universal accuracy advantage of adaptive weighting.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Raushan Moldasheva

Department of Software Engineering, Faculty of Physics, Mathematics and Information Technology

Kh. Dosmukhamedov Atyrau University

Atyrau, Kazakhstan

Email: raushan85_07@mail.ru

1. INTRODUCTION

The increasing penetration of photovoltaic (PV) and wind generation is central to power-sector decarbonisation, but it also amplifies short-term operational uncertainty. Solar production is governed by irradiance, cloud cover, temperature, and the diurnal cycle, whereas wind production depends on wind speed,

air density, and rapidly changing local flow conditions. Accurate short-term forecasts are therefore needed for balancing, reserve scheduling, storage dispatch, congestion management, and secure system operation [1]–[3].

Data-driven forecasting has progressed from linear and seasonal models to tree ensembles, recurrent and convolutional networks, Transformer-based architectures, and physics-informed hybrids [4]–[7]. Recent studies have also emphasised uncertainty quantification, extreme-event analysis, cross-domain transfer, and reproducible evaluation [8]–[14]. Nevertheless, model performance remains strongly dependent on data volume, feature availability, target scale, and the evaluation protocol.

The principal method families have distinct limitations. Linear and seasonal models are interpretable and computationally efficient, but they cannot fully represent nonlinear weather-generation relationships. Tree ensembles are effective for heterogeneous tabular inputs, yet a single fixed model cannot test whether different physical and temporal representations are locally preferable. Recurrent, convolutional, and transformer-based models can learn sequential structure, but their data requirements and source sensitivity may limit their advantage on short, heterogeneous energy records. Physics-informed and hybrid models incorporate domain knowledge, but commonly rely on fixed fusion rules and often treat solar and wind forecasting separately [15]–[17]. Finally, random train-test splits, test-derived scaling, and unreported filtering rules can produce optimistic or irreproducible estimates.

This work addresses a narrower, falsifiable question: does causal local-error weighting of three specialised experts improve joint one-hour-ahead solar and wind forecasts relative to strong single models and simpler ensemble rules? Adaptive three-expert ensemble (A3E) is a sequential tree ensemble rather than an end-to-end neural network. Its temporal, physical, and high-generation experts share an Extra Trees backbone but differ in feature contracts, hyperparameters [18]–[21], and sample weighting. A separate gate estimates each expert's local normalised error from information available at the forecast origin and converts these estimates into target-specific weights [8], [9], [22].

The study is organised around three research questions:

- Is A3E competitive with reproducible baseline models and official modern forecasting architectures under an identical one-hour protocol?
- Does the adaptive gate outperform equal weights, validation-optimised static weights, inverse-error weighting, stacking, and a non-deployable oracle control across repeated initialisations and operating regimes?
- How do model rankings change under external-domain replication, strict zero-shot transfer, high-generation events, and longer forecast horizons?

The contributions are: i) a causal, fully specified heterogeneous-expert forecasting framework; ii) an equal-target, train-only normalised evaluation metric; iii) repeated ablations and dependence-aware statistical comparisons; iv) matched experiments with official TimeXer, TimeMixer, iTransformer, SOFTS, TimeKAN, and the renewable-energy-specific PV-Client 2026 architecture; and v) independent in-domain and zero-shot evaluation on CISO and ERCO. Importantly, negative findings are retained: adaptive weighting is not claimed to be universally superior to strong static aggregation. The literature review therefore serves two roles. The renewable-energy studies in [10], [12], [13] establish the application context, including recent PV, wind, hybrid, and zero-shot forecasting methods. The official general time-series architectures evaluated directly in this work are cited in [14], [23]–[25]. Their codebases were executed without substituting the defining attention, multiscale mixing, inverted-token, series-core fusion, or frequency-decomposition mechanisms. The remainder of the paper is organised as follows. Section 2 describes the data, leakage-free preprocessing, A3E architecture, metrics, comparators, and statistical protocol. Section 3 reports the primary benchmark, aggregation controls, event analysis, external transfer, and multihorizon results. Section 4 discusses the scientific and operational implications and limitations. Section 5 concludes the study.

2. METHOD

The primary source was the publicly available wind and solar power generation dataset [26]. It contains hourly solar and wind generation values and meteorological variables including wind speed, air density, temperature, humidity, and solar-radiation measurements. A renewed audit of the Mendeley record, DataCite/OAI metadata, and all eight source workbooks found no authoritative specification of the target units, exact site, time zone, or installed solar and wind capacities. Accordingly, the primary targets are reported in source units and are not capacity-normalised. Independent external evaluation used hourly balancing-authority data from the U.S. Energy Information Administration (EIA-930) for CISO and ERCO and 2024 nameplate capacities from EIA-860.

The source records were ordered by timestamp and audited before modelling. The published solar file contained 8,753 rows and the wind file 8,752 rows. One conflicting duplicated solar timestamp was resolved deterministically against the unique wind index, leaving 8,752 aligned hourly observations. At forecast origin t , the observed information vector is (1):

Hybrid multi-task deep learning framework for joint forecasting of solar and ... (Assemgul Tynykulova)

$$\mathbf{x}_t = [x_{1,t}, x_{2,t}, \dots, x_{p,t}]^T \quad (1)$$

The timestamp was not used as an unrestricted numerical predictor. Hour, day, month, and seasonal position were encoded cyclically. For a periodic variable a_t with period P (2):

$$\sin_p(a_t) = \sin\left(\frac{2\pi a_t}{P}\right), \quad \cos_p(a_t) = \cos\left(\frac{2\pi a_t}{P}\right) \quad (2)$$

Physically motivated features included meteorological interactions available at the origin. Wind-energy potential was represented through the cubic wind-speed relationship (3):

$$\phi_t^{\text{wind}} = \rho_t v_t^3 \quad (3)$$

Where ρ_t is air density and v_t is wind speed. A radiation-contrast feature was used for the solar representation (4):

$$\phi_t^{\text{solar}} = R_t^{\text{atm}} - R_t^{\text{surface}} \quad (4)$$

only causal lag, rolling, and ramp features were constructed. For any measured or derived signal z_t (5)-(7):

$$z_{t,\ell}^{\text{lag}} = z_{t-\ell}, \quad \ell \in \mathcal{L} \quad (5)$$

$$z_{t,W}^{\text{roll}} = \frac{1}{W} \sum_{u=t-W+1}^t z_u \quad (6)$$

$$z_{t,\ell}^{\text{ramp}} = z_t - z_{t-\ell} \quad (7)$$

The primary task was direct, one-hour-ahead, two-output regression. The target vector was (8):

$$\mathbf{y}_{t+1} = [y_{s,t+1}, y_{w,t+1}]^T \quad (8)$$

Rows without a complete 48-hour causal history or an observed future target were removed. The resulting target matrix was (9):

$$\mathbf{Y} = [y_{t_1+1}, \dots, y_{t_N+1}]^T \in \mathbb{R}^{N \times 2} \quad (9)$$

Service fields, raw strings, metadata, target columns, and any unavailable future variables were excluded from the modelling matrix (10):

$$\mathbf{X} = [x_{t_1}, \dots, x_{t_N}]^T \in \mathbb{R}^{N \times p} \quad (10)$$

The primary-source event threshold for target j was calculated only from the final training split (11):

$$q_{j,0.95}^{\text{train}} = Q_{0.95}(\{y_{j,t} : t \in \mathcal{D}_{\text{train}}\}) \quad (11)$$

The joint primary-source event label was positive when either target exceeded its training-only threshold (12):

$$E_t = \mathbb{I}(y_{s,t} \geq q_{s,0.95}^{\text{train}} \vee y_{w,t} \geq q_{w,0.95}^{\text{train}}) \quad (12)$$

For CISO and ERCO, official capacities allowed a physically interpretable event definition at a fixed capacity-factor threshold of 0.60 (13):

$$E_t^{\text{cap}} = \mathbb{I}\left[\max\left(\frac{y_{s,t}}{c_s}, \frac{y_{w,t}}{c_w}\right) \geq 0.60\right] \quad (13)$$

Chronological partitioning was performed without random mixing. After the duplicate resolution, 48 initial history rows, one final shifted-target row, and three boundary-crossing rows were removed, yielding 8,700 model-ready observations (14):

$$\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}} \cup \mathcal{D}_{\text{test}} \cup \mathcal{D}_{\text{Dec}} \quad (14)$$

The validation interval was further divided into four non-overlapping chronological roles: 359 meta-training rows, 359 probability-calibration rows, 359 threshold-selection rows, and 383 interval-calibration rows; three boundary rows were purged. This separation prevents the gate, probability calibrator, event threshold, and interval procedure from sharing future validation information (15):

$$\mathcal{D}_{\text{val}} = \mathcal{D}_{\text{meta}} \cup \mathcal{D}_{\text{prob}} \cup \mathcal{D}_{\text{thr}} \cup \mathcal{D}_{\text{int}} \cup \mathcal{D}_{\text{purged}} \quad (15)$$

when scaling was required, the transformation parameters were fitted exclusively on training data and then frozen (16):

$$\tilde{x}_{m,t} = \frac{x_{m,t} - \mu_{m,\text{train}}}{\sigma_{m,\text{train}}} \quad (16)$$

The same causal feature contracts and train-only preprocessing rules were applied throughout model selection. The December segment is a later same-source holdout, not an independent external domain. CISO and ERCO are evaluated separately under official common-origin in-domain replication and strict zero-shot protocols. Evaluation metrics. Target-specific errors are reported separately, and every aggregate metric has a unique definition. For target j with N evaluated origins (17)-(21):

$$\text{RMSE}_j = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_{ij} - \hat{y}_{ij})^2} \quad (17)$$

$$\text{RMSE}_{\text{pool}} = \sqrt{\frac{1}{2N} \sum_{i=1}^N \sum_{j \in \{s,w\}} (y_{ij} - \hat{y}_{ij})^2} \quad (18)$$

$$\text{RMSE}_{\text{macro}} = \frac{\text{RMSE}_s + \text{RMSE}_w}{2} \quad (19)$$

$$\text{MacroNRMSE}_{\text{train-SD}} = \frac{1}{2} \left(\frac{\text{RMSE}_s}{s_{s,\text{train}}} + \frac{\text{RMSE}_w}{s_{w,\text{train}}} \right) \quad (20)$$

$$\text{MAE}_j = \frac{1}{N} \sum_{i=1}^N |y_{ij} - \hat{y}_{ij}| \quad (21)$$

Macro train-SD normalized root mean square error (NRMSE) is the primary ranking metric because it assigns equal weight to the two targets after train-only scaling. Capacity-normalised errors are reported only for CISO and ERCO, where official nameplate capacities are available. The primary-source capacity-normalised metric is intentionally unavailable rather than estimated from observed maxima. A3E comprises temporal, physical, and high-generation Extra Trees experts. The temporal expert used 28 trees, maximum depth 14, minimum leaf size 2, and max_features=0.8; the physical expert used 24 trees, depth 14, minimum leaf size 2, and max_features=0.65; the high-generation expert used 24 trees, depth 12, minimum leaf size 3, max_features=1.0, and event-weighted training observations. The realised normalised squared loss of expert k is (22):

$$e_{k,j,t} = \left(\frac{\hat{y}_{k,j,t+1} - y_{j,t+1}}{s_{j,\text{train}}} \right)^2 \quad (22)$$

A causal reliability history of length W uses only errors already observable at origin t (23):

$$r_{k,j,t} = \frac{1}{|W_t|} \sum_{u \in W_t, u < t} e_{k,j,u} \quad (23)$$

A target-specific Extra Trees gate predicts each expert's local error from the reliability history, calendar context, available generation features, and, in the full configuration, the calibrated event probability (24):

$$\hat{e}_{k,j,t} = g_j(z_t)_k \quad (24)$$

The predicted errors are converted to non-negative, unit-sum weights with temperature τ (25):

$$w_{k,j,t} = \frac{\exp(-\hat{e}_{k,j,t}/\tau)}{\sum_{\ell=1}^K \exp(-\hat{e}_{\ell,j,t}/\tau)} \quad (25)$$

The final target-specific forecast is (26):

$$\hat{y}_{j,t+1} = \sum_{k=1}^K w_{k,j,t} \hat{y}_{k,j,t+1} \quad (26)$$

Experts are trained on the main training period. The local-error gate is fitted sequentially on the meta-training interval. The event classifier is a HistGradientBoosting model with 80 iterations, learning rate 0.06, maximum depth 4, minimum leaf size 20, and L2 regularisation 10-3. Its probabilities are calibrated by Platt scaling on a disjoint interval, and the operating threshold is selected on the subsequent threshold-selection interval. Four gate configurations were compared on validation data; the selected configuration used gate depth 3, reliability window 72, and temperature 1.0.

Comparators and statistical analysis. Baselines included persistence, 24-hour seasonal persistence, Ridge, random forest, Extra Trees, extreme gradient boosting (XGBoost), and tabular neural models. Official TimeXer, TimeMixer, iTransformer, SOFTS, TimeKAN, and PV-Client code was executed with pinned versions [19], [23]. The official PV-Client interface is single-target; therefore, two independent unmodified instances were trained for solar and wind, and their target-specific errors were combined only at the evaluation stage. Model selection used training and validation data only. A3E ablations used ten fixed seeds; official deep comparators used five. Pairwise inference was performed on identical origins using 24-hour moving-block bootstrap intervals, a Diebold–Mariano-type comparison with heteroscedasticity-and-autocorrelation-consistent variance and maximum lag 24, and Holm correction for multiple testing. The previous per-timestamp Wilcoxon analysis was not used. Table 1 summarizes the notation used throughout the mathematical formulation of the proposed framework. It defines the indices, target and forecast variables, error and reliability terms, expert weights, and event-related parameters used in the subsequent equations. These definitions provide a consistent notation and facilitate the interpretation of the proposed model.

Table 1. Symbols used in the mathematical formulation

Symbol	Description
t	Forecast origin
h	Forecast horizon
$j \in \{s, w\}$	Target index: solar or wind
k	Expert index
y_j	Observed target value
$\hat{y}_{j,t+1}$	
$\hat{y}_{k,j,t+1}$	Forecast from expert k
$s_{j,\text{train}}$	Training-set standard deviation of target j
$e_{k,j,t}$	Realized normalized squared error
$r_{k,j,t}$	Causal reliability history
$\hat{e}_{k,j,t}$	Predicted local expert error
$w_{k,j,t}$	Normalized expert weight
τ	SoftMax temperature
E_t	Binary high-generation event label
C_j	Official installed capacity
$q_{w,0.95}^{\text{train}}$	Training-only 95th-percentile event threshold

The notation in Table 1 distinguishes the two forecasting targets, solar and wind, while maintaining separate indices for the individual experts. The expert-specific error and reliability terms are used to estimate local forecasting performance and subsequently determine the normalized expert weights. The event-related variables define high-generation conditions using training-only thresholds for the primary dataset and official installed capacities for the external datasets. Together, these symbols provide the mathematical basis for the causal weighting and final forecast aggregation procedures described in the following equations.

3. RESULTS

The primary one-hour benchmark was evaluated on 672 forecast origins shared by A3E and all sequence-based comparators. The mean Macro train-SD NRMSE across the available fixed seeds was used as the primary scale-neutral evaluation metric. As presented in Table 2, lower Macro train-SD NRMSE values indicate better forecasting accuracy.

Table 2. Primary one-hour common-origin benchmark

Model	Publication/status	Seeds	Origins	Macro train-SD NRMSE
A3E	Proposed heterogeneous ensemble	10	672	0.197813
PV-Client	Official renewable-energy model, 2026	5	672	0.231383±0.005743
TimeMixer	Official model, 2024	5	672	0.233988
SOFTS	Official model, 2024	5	672	0.248499
iTransformer	Official model, 2024	5	672	0.248666
TimeXer	Official model, 2024	5	672	0.258287
TimeKAN	Official model, 2025	5	672	0.345878±0.039449

A3E obtained the lowest primary Macro train-SD NRMSE (0.197813). The renewable-energy-specific PV-Client 2026 comparator achieved 0.231383±0.005743, and TimeMixer achieved 0.233988. The paired loss difference A3E minus PV-Client was −0.022453 for solar and −0.013301 for the equal-target macro loss; both remained significant after Holm correction. The wind difference was smaller (−0.004149) and was not statistically significant (adjusted p=0.235768). These findings support A3E on this source but do not imply universal superiority. To isolate the value of adaptive aggregation, all 13 A3E ablations were independently retrained for ten common seeds. Table 3 presents the controls most directly relevant to the aggregation claim.

Table 3. Aggregation controls on the complete 719-origin internal test

Aggregation mechanism	Macro train-SD NRMSE, mean±SD	Macro MAE, mean±SD	Deployable
Oracle best-expert one-hot	0.159159±0.001152	111.157979±1.111278	No
Validation-optimized static weights	0.201720±0.001491	156.414367±1.985133	Yes
Full adaptive A3E	0.202478±0.001335	159.531851±1.572841	Yes
Equal weights	0.202516±0.001337	159.582094±1.571763	Yes
Stacking without reliability	0.218661±0.003364	203.162210±4.996776	Yes

Static weights produced the lowest mean error among deployable aggregation rules, but the paired macro difference relative to adaptive A3E was 0.000255 with a 95% 24-hour block interval of −0.000379 to 0.000833 and Holm-adjusted p=0.550131. A3E was marginally better than equal weights (loss difference −0.000014; 95% interval −0.000021 to −0.000007), but the absolute effect was practically negligible. A3E substantially outperformed stacking without reliability. The non-deployable oracle demonstrates remaining headroom if local expert competence can be predicted more accurately (Figure 1).

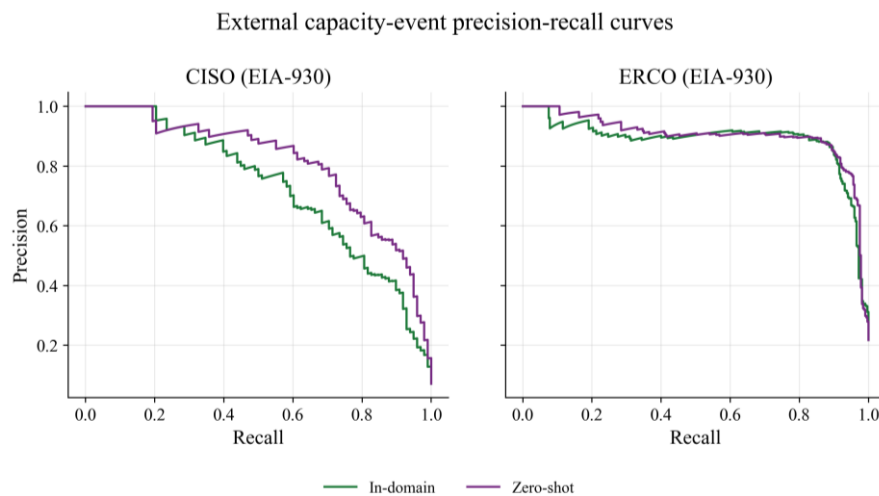


Figure 1. Precision–recall curves for capacity-factor events on CISO and ERCO

Event detection was evaluated independently from regression ranking. On the 719-origin internal test, A3E achieved precision 0.7397, recall 0.9643, F1 0.8372, and precision–recall area under the curve (PR-AUC) 0.9506. This was not the highest recall: Ridge reached 1.0000 at much lower precision (0.2617), while random forest and residual XGBoost achieved higher F1 values. Therefore, A3E is interpreted as one high-sensitivity operating point rather than the universally best detector. No application-specific false-negative versus false-positive cost function was available. Table 4 summarises dependence-aware paired comparisons. Differences are defined as A3E normalised squared loss minus comparator loss; negative values favour A3E. The old Wilcoxon table was removed.

Table 4. Dependence-aware paired comparisons on common forecast origins

Comparison	Δ loss	95% block CI	Holm p
PV-Client/solar	−0.022453	See supplement	0.000937
PV-Client/wind	−0.004149	See supplement	0.235768
PV-Client/macro	−0.013301	See supplement	0.000029
Equal/macro	−0.000014	[−0.000021, −0.000007]	0.001389
Static/macro	+0.000255	[−0.000379, +0.000833]	0.550131
Stacking/macro	−0.006790	[−0.009422, −0.004560]	<0.000001
TimeKAN/macro	−0.080865	[−0.094496, −0.066443]	<0.000001

The dependence-aware results distinguish statistical detectability from practical importance. In particular, the adaptive-versus-equal difference is extremely small despite its adjusted significance, whereas the adaptive-versus-static interval contains zero. This supports a heterogeneous-expert contribution but not a general superiority claims for the local gate (Table 5).

Table 5. Event-detection results at fixed, validation-selected operating points

Method/setting	Precision	Recall	F1	PR-AUC
A3E—primary (56/719)	0.7397	0.9643	0.8372	0.9506
Random forest-primary	0.9412	0.8571	0.8972	0.9509
Residual XGBoost-primary	0.8909	0.8750	0.8829	0.9585
Ridge—primary	0.2617	1.0000	0.4148	0.8682
A3E—CISO (98/1375)	0.8148	0.4490	0.5789	0.7277
A3E—ERCO (320/1470)	0.9010	0.8250	0.8613	0.8863

The December holdout contains one positive event among 743 observations and is therefore retained only as a descriptive same-source case. Quantitative external event conclusions use CISO and ERCO, which contain 98 and 320 positive events, respectively. Table 6 summarizes the external-domain evaluation populations and forecasting performance for the CISO and ERCO systems. It reports the number of forecast origins and high-generation events together with the Macro train-SD NRMSE obtained by TimeMixer, target-domain A3E, and strict zero-shot A3E.

Table 6. External-domain regression and event-evaluation populations

Source	Origins	Events	TimeMixer	A3E domain	A3E zero
CISO	1321	98/1375	0.1565	0.1794	0.1918
ERCO	1423	320/1470	0.1897	0.2011	0.2405

The external results do not establish universal generalisation. TimeMixer outperformed A3E on both independent systems. A3E zero-shot performance was weaker than target-domain replication, particularly on ERCO, confirming a substantial distribution shift. Strict zero-shot runs used no external fitting, scaler estimation, validation, threshold selection, or calibration; model hashes before and after transfer were identical. In-domain replication and zero-shot transfer are therefore reported as distinct protocols. The EIA-860 nameplate capacities were 23,085.9 MW solar and 6,334.7 MW wind for CISO, and 22,179.5 MW solar and 38,317.7 MW wind for ERCO. Figure 2 reconciles the denominators used by the primary benchmark, the practical multihorizon diagnostic, official external common-origin comparisons, external event analyses, and the December holdout. These populations must not be compared as if they were identical.

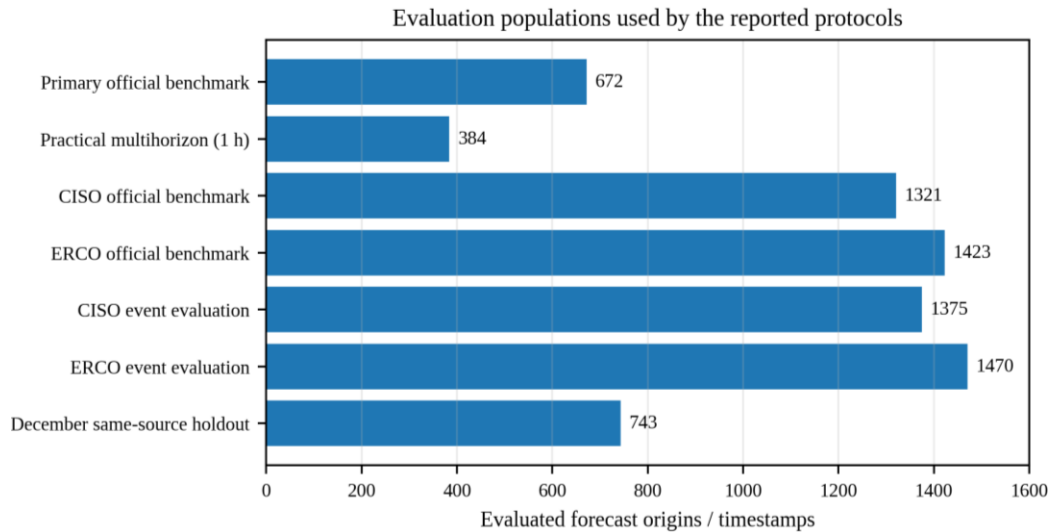


Figure 2. Evaluation populations used by the reported protocols

Figure 3 combines two complementary experimental comparisons conducted using a unified scale-invariant metric, Macro train-SD NRMSE. Figure 3(a) presents the primary one-hour benchmark based on the initial data source, covering 672 common forecast origins shared by A3E and all comparative sequence-based models. A3E achieved the lowest Macro train-SD NRMSE value of 0.1978. The values for the other models were 0.2314 for the official specialized model PV-Client 2026, 0.2340 for TimeMixer, 0.2485 for SOFTS, 0.2487 for iTransformer, 0.2583 for TimeXer, and 0.3459 for TimeKAN. Thus, on the primary source, A3E demonstrates the lowest error among the models evaluated, given identical forecast origins and normalization rules. Figure 3(b) illustrates the cross-domain evaluation on the independent CISO and ERCO power systems, clearly distinguishing between in-domain training and strict zero-shot transfer. TimeMixer achieved a Macro train-SD NRMSE of 0.1565 on CISO and 0.1897 on ERCO. For A3E, the corresponding values were 0.1794 and 0.2011 under in-domain training, whereas under strict zero-shot transfer, the error increased to 0.1918 on CISO and 0.2405 on ERCO. These results indicate that the relative performance of the models depends on the data source and transfer protocol: while A3E exhibits the lowest error on the primary benchmark, TimeMixer outperforms A3E on both external systems. Therefore, the external results are interpreted as evidence of source-dependent model behavior rather than as confirmation of the universal superiority of A3E.

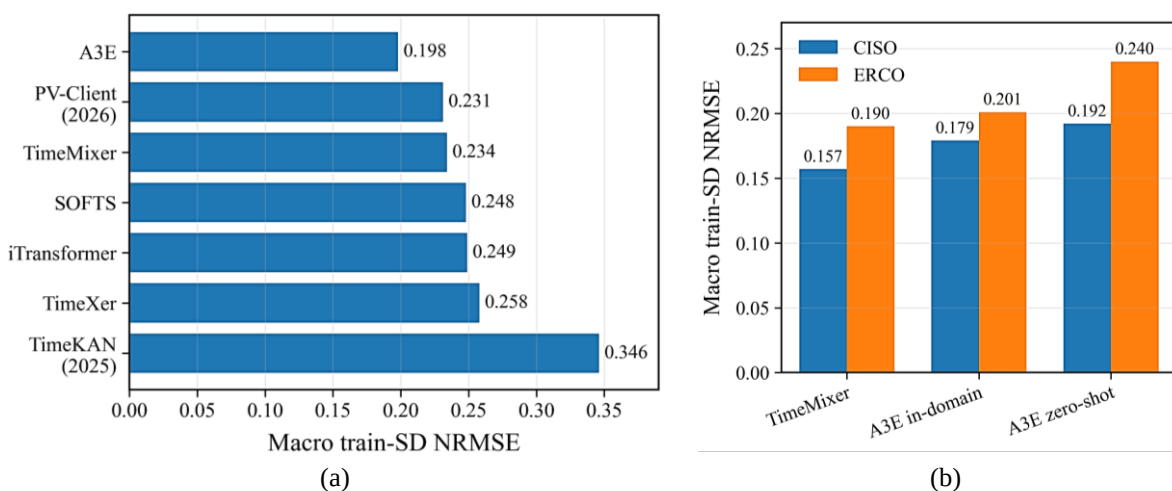


Figure 3. Comparison of forecasting performance: (a) primary-source benchmark based on 672 common forecast origins and (b) external in-domain and zero-shot evaluation on the CISO and ERCO systems

The approximately 0.20 and 0.23 one-hour values in earlier drafts referred to different evaluation populations. The primary common-origin result is 0.197813 on 672 origins. The practical one-hour point in the horizon-balanced experiment is 0.229560 on 384 origins shared by the 1-, 24-, 168-, and 336-hour tasks. The latter is a diagnostic multihorizon protocol and is not the primary one-hour benchmark. Figure 4 shows all retrained A3E ablations. The displayed intervals are seed-level uncertainty summaries; the formal paired comparisons are reported in Table 4.

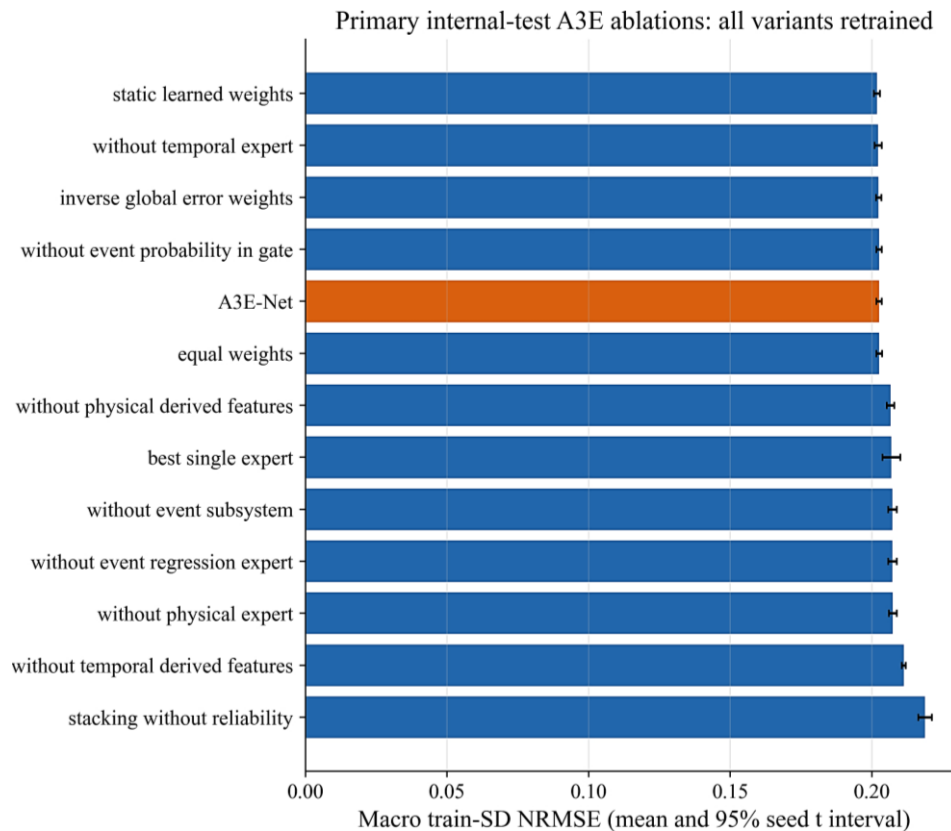


Figure 4. Primary internal-test A3E ablations; all variants independently retrained

Several aggregation rules cluster near Macro train-SD NRMSE 0.20. Removing temporal features or using stacking without reliability causes larger degradation, whereas removing event probability from the gate has negligible effect. Thus, the empirical contribution is better characterised as integrated expert diversity and a leakage-controlled evaluation framework than as a demonstrated universal benefit of adaptive weighting. Figure 5 presents the first seven consecutive days of the internal test, selected without cherry-picking and averaged over ten A3E seeds. The visual diagnostic is supplementary to, not a replacement for, the numerical and interval estimates.

Multihorizon evaluation used direct models for 1-, 24-, 168-, and 336-hour forecast horizons. A practical history-only protocol was distinguished from a non-deployable weather-oracle protocol using realized target-time meteorological information. All horizons were evaluated on the same 384 forecast origins to ensure a consistent comparison. Figure 6 illustrates the forecasting performance of the practical and weather-oracle protocols across these four horizons, allowing the effect of forecast horizon and the availability of future meteorological information to be directly compared.

Practical error increased from 0.229560 at one hour to approximately 0.73 at 24 hours and 0.84 at seven days, then decreased slightly at 14 days. The non-monotonic 14-day point is not interpreted as improved long-horizon predictability because the 336-hour task has only about 1.14 effectively non-overlapping intervals within the common-origin window. Weather-oracle results are diagnostic upper bounds and are not reported as deployable performance. Figure 7 reports 95% 24-hour moving-block intervals for CISO and ERCO event metrics. The larger positive-event support on ERCO yields more stable estimates than CISO, but neither result defines an application-optimal operating point without an explicit cost model.

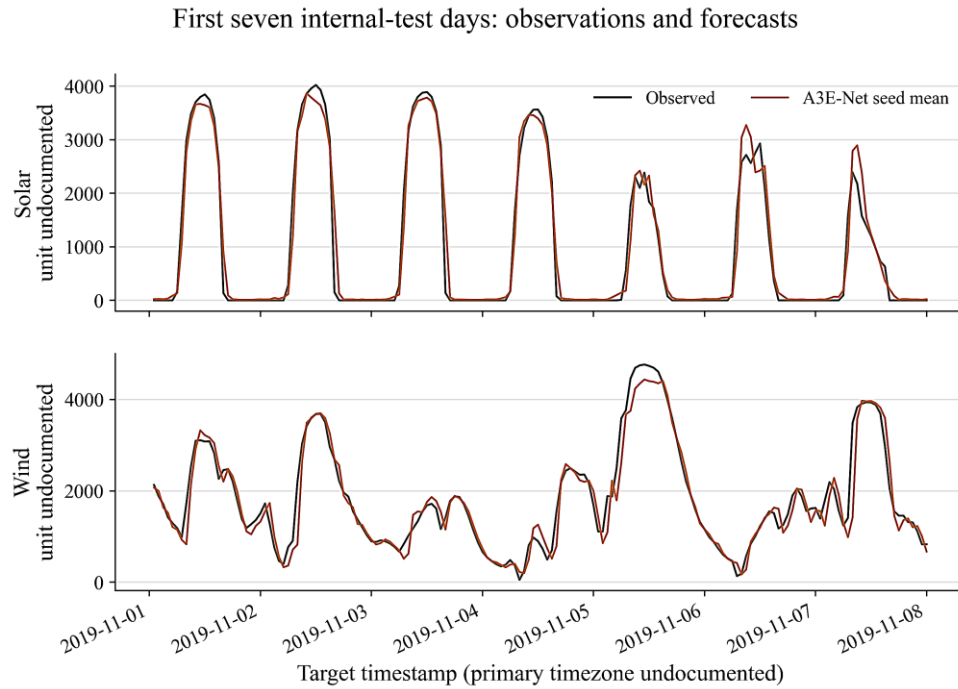


Figure 5. First seven internal-test days: observations and A3E seed-mean forecasts

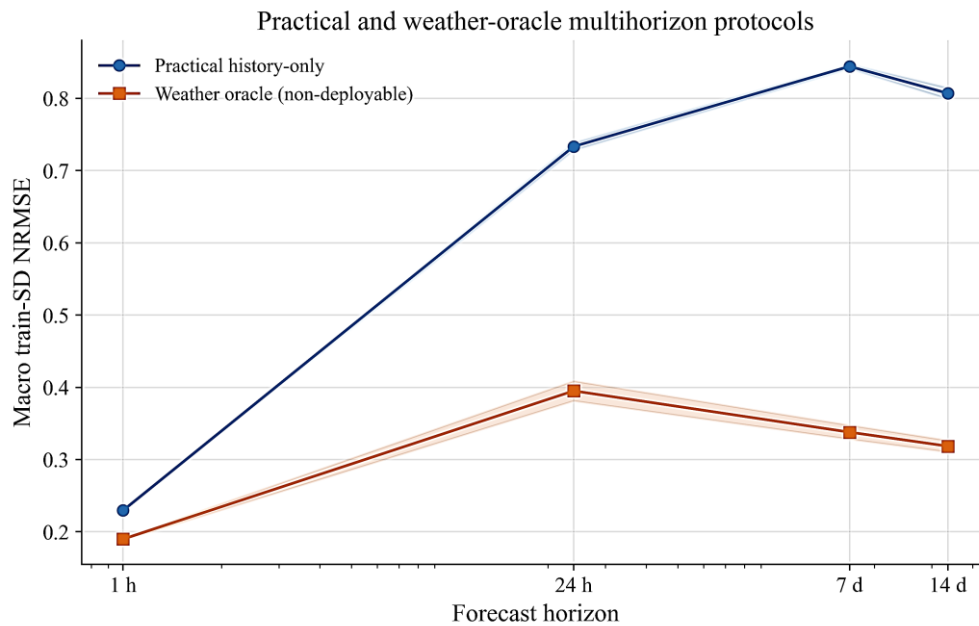


Figure 6. Practical and non-deployable weather-oracle multihorizon protocols

Across the regression, event-detection, and transfer experiments, model performance varied depending on the data source and evaluation setting. The observed rankings were also influenced by the selected operating point and evaluation population. Therefore, the findings support protocol-specific conclusions rather than a universal ranking of the evaluated models.

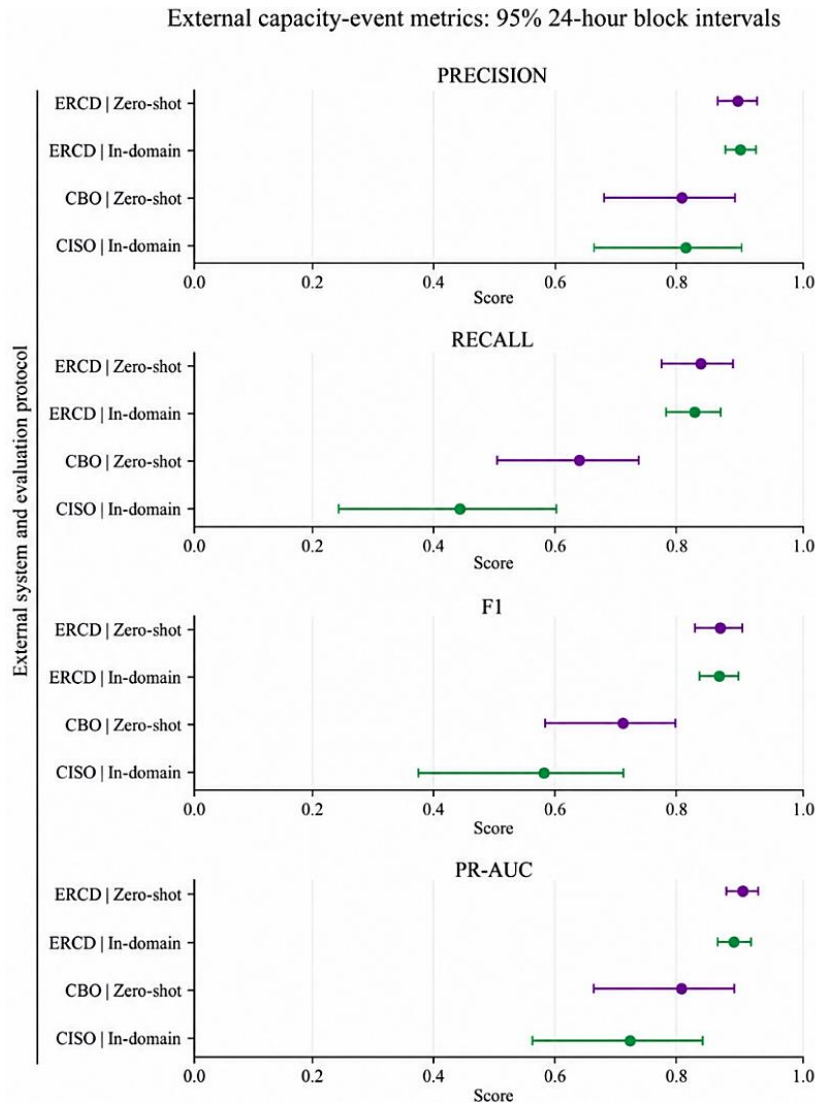


Figure 7. External capacity-event metrics with 95% 24-hour block intervals

4. DISCUSSION

The revised experiments materially narrow the interpretation of A3E. On the primary common-origin benchmark, A3E outperformed the official 2024 architectures, TimeKAN 2025, and the renewable-energy-specific PV-Client 2026 comparator. This result is meaningful because all methods used the same one-hour target, causal 48-hour history, train-only selection, and common forecast origins. However, the external CISO and ERCO comparisons reverse the ranking: TimeMixer achieves lower error than A3E on both systems. The evidence therefore supports source-dependent competitiveness rather than universal state-of-the-art performance. The aggregation analysis also changes the novelty claim. Validation-optimised static weights achieved a slightly lower mean error than the adaptive gate, and the dependence-aware interval for their difference included zero. Equal weighting was almost indistinguishable in practical magnitude. By contrast, stacking without reliability was clearly worse, and the non-deployable oracle was substantially better. These findings indicate that expert diversity is useful and that local competence contains exploitable information, but the present gate does not extract enough of that information to establish a general adaptive advantage.

The event module should be interpreted as an auxiliary risk-screening component. A3E has high internal recall, but other models achieve higher recall, F1, precision, or PR-AUC at different thresholds. External evaluation on CISO and ERCO provides adequate positive support for interval estimation, unlike the single-event December holdout. Nevertheless, an operationally optimal threshold cannot be specified without application-specific costs for missed events and false alarms. The explicit separation of zero-shot transfer and in-domain replication is equally important. Zero-shot A3E uses a common dimensionless causal representation

and no target-domain fitting. Its deterioration, especially on ERCO, quantifies distribution shift rather than adaptation failure. The stronger in-domain result demonstrates learnability within the target system but must not be treated as unseen-domain generalisation.

Several limitations remain. The primary source lacks authoritative units, location, time zone, and installed capacities; consequently, physical-unit and capacity-normalised claims are not made for that dataset. The primary test spans a limited historical period. External validation covers two U.S. balancing authorities and does not establish global transportability. Long-horizon practical forecasts do not use operational numerical-weather-prediction inputs, and the 336-hour common-origin support is small. Finally, the public repository URL must be inserted before resubmission so that the audited local reproducibility package can be independently accessed. Future work should focus on improving the causal prediction of local expert competence rather than merely increasing gate complexity. Evaluation should be extended to multi-year regional datasets, operational weather forecasts, additional distribution shifts, and application-defined event costs. Such studies can determine whether the oracle gap observed here can be converted into reliable operational gains.

5. CONCLUSION

In this study, an A3E model was developed and systematically evaluated for one-hour-ahead joint solar and wind power forecasting under a strictly time-based protocol with leakage control. On a core benchmark dataset of 672 common forecast sources, A3E achieved the lowest NRMSE macro train-SD of 0.1978, outperforming the renewable energy-specific PV-Client 2026 model (0.2314), TimeMixer (0.2340), SOFTS (0.2485), iTransformer (0.2487), TimeXer (0.2583), and TimeKAN (0.3459). The difference in macrometric loss between A3E and PV-Client was -0.0133 and remained statistically significant after Holm's correction ($p=0.000029$), demonstrating the competitive performance of the heterogeneous expert model on the main dataset.

Aggregation experiments allowed for a more detailed evaluation of the adaptive mechanism. On a full internal test with 719 sources, the adaptive A3E achieved a Macro train-SD NRMSE of 0.2025, compared to 0.2017 for validation-optimized static weighting, 0.2025 for equal weighting, and 0.2187 for reliability-neutral stacking. The difference between adaptive and static weighting was not statistically significant (95% confidence interval: -0.000379 to 0.000833; adjusted $p=0.550131$), indicating that the primary empirical benefit arises from heterogeneous expert integration rather than a universal advantage of adaptive weighting. For in-system high-generation event detection, A3E achieved precision of 0.7397, recall of 0.9643, F1-score of 0.8372, and PR-AUC of 0.9506.

External evaluation further demonstrated that model performance is source-dependent. On CISO, A3E achieved a Macro train-SD NRMSE score of 0.1794 when evaluated in the target domain and 0.1918 when strictly transferring without pretraining, while TimeMixer achieved 0.1565. On ERCO, the corresponding A3E errors were 0.2011 and 0.2405, compared to 0.1897 for TimeMixer. For external event detection, A3E achieved F1 scores of 0.5789 on CISO and 0.8613 on ERCO. These results indicate that A3E is competitive on the primary source but do not support claims of overall superiority in independent power systems. However, the study has several limitations. The primary dataset does not contain reliable information on target units, geographic location, time zone, or installed generating capacity, which hinders the interpretation of the primary results given physical units and capacity. External validation was limited to two US balancing authorities, and practical experiments over a long horizon did not include operational numerical weather forecasting data. Furthermore, limited effective support over a 336-hour horizon limits the ability to draw firm conclusions about very long-term forecasting. Therefore, future research should evaluate A3E on multi-year and multi-regional datasets, integrate operational weather forecasts, explore more efficient causal assessment of local expert competence, and evaluate event detection thresholds using application-specific costs for missed events and false positives. Overall, the results confirm that A3E is a reproducible and leakage-controlled heterogeneous forecasting system, while providing a transparent assessment of both its strengths and current limitations.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Assemgul Tynykulova	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Raushan Moldasheva		✓				✓		✓	✓	✓	✓	✓		
Elmira Eldarova	✓		✓	✓			✓			✓	✓		✓	✓
Yerlan Izbassarov		✓				✓		✓	✓	✓	✓	✓		
Shynar Kodanova					✓		✓			✓		✓		✓
Lunara Diyarova		✓				✓		✓	✓	✓	✓	✓		
Saya Baigubenova					✓		✓			✓		✓		✓
Zhanar Azhibekova					✓		✓			✓		✓		✓

C : **C**onceptualization
M : **M**ethodology
So : **S**oftware
Va : **V**alidation
Fo : **F**ormal analysis

I : **I**nterpretation
R : **R**esources
D : **D**ata Curation
O : **O**riginal Draft
E : **E**diting

Vi : **V**isualization
Su : **S**upervision
P : **P**roject administration
Fu : **F**unding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] E. A. Manzano, R. E. Nogales, and A. Rios, "A systematic review of wind energy forecasting models based on deep neural networks," *Wind*, vol. 5, no. 4, p. 29, Nov. 2025, doi: 10.3390/wind5040029.

[2] J. Selvaraj, L. Muthuramalingam, V. Karthikeyan, A. Karthick, and V. Sathiyaseelan, "Optimizing wind energy integration: a review of forecasting techniques and emerging trends," *Archives of Computational Methods in Engineering*, vol. 33, no. 3, pp. 4261–4286, Apr. 2026, doi: 10.1007/s11831-025-10442-1.

[3] Z. Hou, Y. Zhang, Q. Liu, and X. Ye, "A hybrid machine learning forecasting model for photovoltaic power," *Energy Reports*, vol. 11, pp. 5125–5138, Jun. 2024, doi: 10.1016/j.egyr.2024.04.065.

[4] R. Rimbawati, H. Ambarita, T. B. Sitorus, and M. Irwanto, "Artificial neural network (ANN) backpropagation for forecasting 100% renewable energy in north sumatera," *Environmental Research, Engineering and Management*, vol. 81, no. 1, pp. 87–101, Mar. 2025, doi: 10.5755/j01.ere.m.81.1.37767.

[5] R. Raman, S. Gunasekar, D. Kaliyaperumal, and P. Nedungadi, "Navigating the nexus of artificial intelligence and renewable energy for the advancement of sustainable development goals," *Sustainability*, vol. 16, no. 21, p. 9144, Oct. 2024, doi: 10.3390/su16219144.

[6] H. Zhou, Q. Ai, and R. Li, "Short-term multi-energy load forecasting method based on transformer spatio-temporal graph neural network," *Energies*, vol. 18, no. 17, p. 4466, Aug. 2025, doi: 10.3390/en18174466.

[7] D. Szpilkó, X. Fernando, E. Nica, K. Budna, A. Rzepka, and G. Lăzăroiu, "Energy in smart cities: technological trends and prospects," *Energies*, vol. 17, no. 24, p. 6439, Dec. 2024, doi: 10.3390/en17246439.

[8] A. Verdone, M. Panella, E. De Santis, and A. Rizzi, "A review of solar and wind energy forecasting: from single-site to multi-site paradigm," *Applied Energy*, vol. 392, p. 126016, Aug. 2025, doi: 10.1016/j.apenergy.2025.126016.

[9] S. Shringi, L. M. Saini, and S. K. Aggarwal, "A review of data-driven deep learning models for solar and wind energy forecasting," *Renewable Energy Focus*, vol. 55, p. 100739, Dec. 2025, doi: 10.1016/j.ref.2025.100739.

[10] X. Xu, Q. Cao, R. Deng, Z. Guo, Y. Chen, and J. Yan, "A cross-dataset benchmark for neural network-based wind power forecasting," *Renewable Energy*, vol. 254, p. 123463, Dec. 2025, doi: 10.1016/j.renene.2025.123463.

[11] M. K. G. Deshmukh, M. Sameeroddin, D. Abdul, and M. A. Sattar, "Renewable energy in the 21st century: a review," *Materials Today: Proceedings*, vol. 80, pp. 1756–1759, 2023, doi: 10.1016/j.matpr.2021.05.501.

[12] B. Xiong, L. Lou, X. Meng, X. Wang, H. Ma, and Z. Wang, "Short-term wind power forecasting based on attention mechanism and deep learning," *Electric Power Systems Research*, vol. 206, p. 107776, May 2022, doi: 10.1016/j.epsr.2022.107776.

[13] K. Kumar, P. Prabhakar, A. Verma, S. Saroha, and K. Singh, "Advancements in wind power forecasting: a comprehensive review of artificial intelligence-based approaches," *Multimedia Tools and Applications*, vol. 84, no. 10, pp. 8331–8360, May 2024, doi: 10.1007/s11042-024-18916-3.

[14] H. Liu and Z. Zhang, "Development and trending of deep learning methods for wind power predictions," *Artificial Intelligence Review*, vol. 57, no. 5, p. 112, Apr. 2024, doi: 10.1007/s10462-024-10728-z.

[15] E. M. Al-Ali et al., "Solar energy production forecasting based on a hybrid cnn-lstm-transformer model," *Mathematics*, vol. 11, no. 3, p. 676, Jan. 2023, doi: 10.3390/math11030676.

[16] Y. Liu, Y. Wang, Q. Wang, K. Zhang, W. Qiang, and Q. H. Wen, "Recent advances in data-driven prediction for wind power," *Frontiers in Energy Research*, vol. 11, Jun. 2023, doi: 10.3389/fenrg.2023.1204343.

[17] Y. Yang, H. Lou, J. Wu, S. Zhang, and S. Gao, "A survey on wind power forecasting with machine learning approaches," *Neural Computing and Applications*, vol. 36, no. 21, pp. 12753–12773, Jul. 2024, doi: 10.1007/s00521-024-09923-4.

[18] Y. Song, J. Zhou, Y. Peng, and J. Yan, "Power generation forecasting based on hybrid deep learning models for a multi-energy power generation system," *Engineering Applications of Artificial Intelligence*, vol. 157, p. 111380, Oct. 2025, doi: 10.1016/j.engappai.2025.111380.

- [19] F. Li, H. Wang, D. Wang, D. Liu, and K. Sun, "A review of wind power prediction methods based on multi-time scales," *Energies*, vol. 18, no. 7, p. 1713, Mar. 2025, doi: 10.3390/en18071713.
- [20] Q. Xia and W. Kang, "MSTFNet: multi-scale temporal fusion network with frequency-enhanced attention for financial time series forecasting," *Mathematics*, vol. 14, no. 8, p. 1391, Apr. 2026, doi: 10.3390/math14081391.
- [21] R. Han, H. Jiang, M. Wei, and R. Guo, "Research on joint forecasting technology of cold, heat, and electricity loads based on multi-task learning," *Electronics*, vol. 13, no. 22, p. 4396, Nov. 2024, doi: 10.3390/electronics13224396.
- [22] A. Ilinca, "The future of renewable energy: 2nd edition," *Energies*, vol. 19, no. 12, p. 2837, Jun. 2026, doi: 10.3390/en19122837.
- [23] Y. Mei, J. Che, Q. Sun, and W. Dong, "An overview of short-term wind power forecasting: multi-scale decomposition and multi-model deep learning fusion," *Energy Strategy Reviews*, vol. 63, p. 102021, Jan. 2026, doi: 10.1016/j.esr.2025.102021.
- [24] M. Sambane, B. Mendu, and B. B. Monchusi, "Advanced neural network and hybrid models for wind power forecasting: a comprehensive global review," *Future Energy*, vol. 3, no. 4, pp. 67–79, Nov. 2024, doi: 10.55670/fpl.fuen.3.4.5.
- [25] P. K. Singh, A. Prakash, A. Saraswat, Y. Gupta, and J. N. Sabhahit, "A new hybrid deep-learning model with hyperparameter optimization for short-term wind power generation forecasting," *Unconventional Resources*, vol. 12, p. 100365, Jul. 2026, doi: 10.1016/j.unres.2026.100365.
- [26] Y. Liu, "Wind and solar power generation dataset," *Mendeley Data*, 2024, doi: 10.17632/gxc6j5btrx.1.

BIOGRAPHIES OF AUTHORS



Assemgul Tynykulova    has teaching experience at leading universities in Kazakhstan. Her main research interests include optimization methods, decision-making support, and expert systems. Over the past five years, she has published more than ten scientific articles, including four in journals indexed in the Scopus database. She currently holds a Ph.D. degree and works as an Acting Associate Professor at the Higher School of Information Technology and Engineering of Astana International University, Astana. Astana International University, Ph.D., Senior Lecturer at the Higher School of Information Technology and Engineering, Astana, Kazakhstan. She can be contacted at email: asem_110981@mail.ru.



Raushan Moldasheva    graduated from Atyrau State University named after H. Dosmukhamedov in 2006 with a degree in Physics and Mathematics. In 2013, she received a master's degree in Natural Science. In 2023, she graduated from the doctoral program of the Kazakh agrotechnical research university named after S. Seifullin with a degree in 8D06101–"Big Data Analytics". In 2024, she defended his doctoral dissertation, and from 2025 to the present—Ph.D. doctor of philosophy of Atyrau State University named after H. Dosmukhamedov in the specialty 8D06101 - "Big Data Analytics". She is the author of more than 40 scientific papers. Her research interests include big data processing and analysis techniques, machine learning and deep learning algorithms, intelligent data analysis (data mining), artificial intelligence systems, statistical modeling, and probability analysis. She can be contacted at email: raushan85_07@mail.ru.



Elmira Eldarova    graduated in 2002 from the Atyrau State University named after H. Dosmukhamedov with a degree in Computer Science. In 2013, she obtained a Master's degree in Computer Science. In 2021, she completed her doctoral studies at the L. N. Gumilyov Eurasian National University in the specialty 6D070400—Computer Engineering and Software. From 2025 to the present, she has been a doctor of philosophy (Ph.D.) in the specialty 6D070400 — Computer Engineering and Software at the Atyrau University named after H. Dosmukhamedov. She is the author of more than 20 scientific publications. Her research interests include digital image processing, digital image quality enhancement, machine learning, and neural networks. She can be contacted at email: doctorphd_eldarova@mail.ru.



Yerlan Izbassarov    is master of technical sciences, senior lecturer at the Department of Software Engineering at Dosmukhamedov Atyrau University. In 2003, he graduated with distinction from Korkyt Ata Kyzylorda University with a degree in Automated Information Processing and Control Systems, and in 2010 he obtained a master's degree in Information Systems. He takes an active part in fundamental and applied research. He is the author of 1 patent, 5 author's certificate and more than 20 scientific works. He can be contacted at email: erlan_1081@mail.ru.



Shynar Kodanova    in 2011 received a candidate of Technical Sciences degree, 05.13.18 - "Mathematical modeling, numerical methods and a set of programs" (2011). She completed an internship at BV Engineering in Automation and Artificial Intelligence. Currently, she works at the Atyrau University of Oil and Gas named after Safi Utebaev as the Dean of the Faculty of Information Technologies. Her research interests include mathematical modeling and control of complex objects, formalization and development of methods for solving decision-making problems in a fuzzy environment, as well as the creation of intelligent decision-making systems and artificial intelligence. She has more than 30 years of scientific and pedagogical experience and more than 25 scientific papers, including 6 article based on Scopus, 2 textbooks with ISBN, and two monographs. She can be contacted at email: S.Kodanova@aogu.edu.kz.



Lunara Diyarova    is a master's degree in Technical Sciences, specializing in Information Systems. She is a senior lecturer at the Institute of Digital Economy and Sustainable Development. She has over 33 scientific and educational publications, including more than 10 textbooks and teaching aids. Performer of more than 2 scientific and innovative international projects of Erasmus+. Her research areas are information technology, machine learning, and robotics. She can be contacted at email: lunara.diyarova@mail.ru.



Saya Baigubenova    is a master's degree in Technical Sciences, specializing in Information Systems. She is a senior lecturer at the Institute of Digital Economy and Sustainable Development. Her research interests focus on modern methods of teaching information technologies, including digital and intelligent educational tools. She can be contacted at email: saya_bsk@mail.ru.



Zhanar Azhibekova    graduated from Al-Farabi Kazakh National University with major Physics and has obtained qualification teacher of physics in 1993. From 1993 till 1995 she worked as an assistant of Department of Physics at Kazakh Autotransport Institute. In 1995 she was a senior laboratory assistant of Department of Physics at Asfendiyarov State Medical University. In 1995 she became a teacher and from 2000 has started to work as a senior teacher of Department of Medical Biophysics, Informatics, and Mathematics and from 2003 she became an associate professor of the Department. In 09.10.2003 on Dissertation Council of Abai Kazakh National Pedagogical University has defended master's thesis on topic methodic of studying ways of information representation in informatics course in medical high schools and got candidate master's degree of pedagogical sciences. As a result of her research, three educational guides, 60 scientific articles, and five electronic textbooks were published. Since 2012, she has been working as the head of the Department of Information and Communication Technologies at Asfendiyarov Kazakh National Medical University. She can be contacted at email: Azhibekova.z@kaznmu.kz.