

ARTICLE

Game-Guarded Transformer Forecasting of Hydropower Output

Stoyan Popov^{1,2}, Silvia Baeva^{1,2} and Nikolay Hinov^{2,3,*}

¹Department of Mathematical Modelling and Numerical Methods, Faculty of Applied Mathematics and Informatics, Technical University of Sofia, Sofia, Bulgaria

²National Centre of Excellence “Mechatronics and Clean Technologies”, Sofia, Bulgaria

³Department of Computer Systems, Faculty of Computer Systems and Technologies, Technical University of Sofia, Sofia, Bulgaria

*Corresponding Author: Nikolay Hinov. Email: hinov@tu-sofia.bg

Received: 25 July 2026; Accepted: 04 September 2026; Published: 21 October 2026

ABSTRACT: Short-horizon hydropower forecasting is difficult to validate because slowly changing operation can make persistence appear deceptively strong, while random splitting and overlapping windows can inflate the reported performance of learned models. This study develops and audits a leakage-safe, fifteen-minute-ahead active-power forecasting pipeline using supervisory-control records obtained from one Bulgarian hydropower plant over the hydrological year 01 October 2020–30 September 2021. The archive contains 361 available daily files, 508,234 unique timestamps and four unavailable days. After minute normalization and schema harmonization, complete-window screening requires all 30 one-minute history records and the target record (31 records in total) to be observed within one continuity segment; 8298 hourly forecast origins remain. Six experts—persistence, ridge regression, histogram gradient boosting, a gated recurrent unit, a long short-term memory network and a compact self-attention Transformer—share a persistence-residual formulation based on a thirty-minute multivariate history. A finite zero-sum game minimizes the worst calendar-day expected randomized-expert loss on May 2021. Entropic multiplicative-weights dynamics generate a candidate strategy. An exact linear program certifies the minimax solution. An in-block development screening guard determines deployment. Models are fitted on October–March, neural stopping uses April, game tuning uses May, conformal calibration uses June, and July–September is a held-out evaluation block. On 1999 test origins, persistence attains 1.8087 kW mean absolute error and 6.8910 kW root-mean-square error. The Transformer has a slightly lower root-mean-square error of 6.8587 kW but a higher mean absolute error of 2.1434 kW. Its squared-error Diebold–Mariano comparison is not significant after Holm adjustment ($p = 0.7012$). The 5000-step evolutionary candidate assigns 98.36% to persistence but is rejected because both its overall and worst-day tuning errors exceed persistence. The exact minimax strategy and final guarded forecast are pure persistence. Seven-day moving-block bootstrap intervals, multiplicity-adjusted paired tests and calendar-day maximum-residual conformal intervals quantify uncertainty. The negative selection result is operationally informative: under the observed low-ramp test regime, the Transformer shows limited complementary behavior in the largest-ramp strata and a slightly lower RMSE without a statistically supported overall gain, whereas the guarded game prevents an unsupported mean-error claim. The evaluated scope is one hourly phase (HH:30 origin to HH:45 target), one fixed neural seed and one plant-year; the claims are limited accordingly. A separate one-day minute-rolling diagnostic confirms within-hour phase heterogeneity and does not support an attention-specific advantage; it is not used to replace the annual result.

KEYWORDS: Hydropower forecasting; active power; transformer; self-attention; evolutionary game; minimax aggregation; conformal prediction; uncertainty quantification; chronological validation

1 Introduction

Hydropower is renewable generation whose short-term output is determined jointly by water availability, hydraulic state, turbine-generator dynamics and supervisory control. Unlike wind and solar generation, the electrical response does not follow an exogenous weather input alone. Gate or nozzle commands, unit availability, rotational speed, plant-level set points and dispatch decisions can alter active power over minutes even when slowly varying environmental conditions appear unchanged. Mechanistic turbine-governor and micro-hydropower models provide an essential physical description [1,2], but their calibration and online state estimation can be demanding when only routine supervisory-control logs are publicly available.

Accurate forecasts fifteen minutes ahead can assist reserve scheduling, operator awareness, ramp detection and coordination with other renewable sources. At this horizon, however, the present active power is a formidable baseline. If most test observations lie in quasi-stationary regimes, the persistence forecast $\hat{P}_{t+15} = P_t$ can achieve a small average error without learning any plant dynamics. A flexible model may improve a small number of large transitions and therefore lower root-mean-square error (RMSE), yet introduce small corrections during many stable periods and consequently worsen mean absolute error (MAE). Reporting only one metric, or omitting persistence, can therefore reverse the practical interpretation.

Published hydropower forecasting research includes artificial neural networks, machine-learning comparisons, neuro-fuzzy optimization, deep ensembles and hybrid recurrent-attention systems [3–7]. The related rainfall–runoff literature provides strong evidence for long short-term memory (LSTM) networks in representing delayed hydrological responses [8–10]. Transformer models offer direct, content-dependent interaction between observations in a history window [11–15]. Patch-based, linear and variable-oriented designs extend the recent time-series comparison [16–18]. Nevertheless, architecture complexity alone is not a guarantee of better out-of-time performance. The evidence that simple linear forecasting models can remain highly competitive [17] is particularly relevant when the plant-specific sample is small and the operating distribution changes across seasons.

Three validation problems motivate this study. First, random train/test splitting permits adjacent and overlapping temporal windows to enter both partitions, creating leakage and optimistic estimates. Second, average-loss training may hide calendar days on which a model fails badly. Third, a point estimate of an error difference does not establish that the difference is stable under serial dependence. Forecast combination [19], weighted majority [20], replicator dynamics [21], multiplicative-weights extensions [22], electricity-demand expert aggregation [23], deep-model uncertainty estimation [24], dependence-aware testing [25–27] and conformal calibration [28,29] provide complementary tools for addressing these problems.

The objective is not to force a deep model to win. It is to determine, under an auditable year-long protocol, whether a compact Transformer contributes useful forecast information and whether a game-theoretic selector should deploy it. The complete workflow is summarized in Fig. 1. The experiment uses one fixed neural seed and one hourly phase; it is not an exhaustive architecture search.

The main contributions are as follows:

1. A transparent audit of a one-hydrological-year acquisition period with 361 available daily files obtained from the portal, including schema changes, duplicate timestamps, incomplete rows, minute collisions, four missing dates and 1395 continuity segments.
2. A common persistence-residual formulation for linear, tree, recurrent and self-attention experts. Each learned branch predicts only the fifteen-minute active-power change, while train-only clipping bounds extrapolation.

3. An evolutionary zero-sum game whose adverse player represents the 31 calendar days of the weight-tuning month. The finite 5000-step multiplicative-weights candidate is compared with an exact linear-programming minimax certificate.
4. An in-block development screening guard that rejects a candidate unless both its aggregate and worst-calendar-day tuning MAE are no greater than those of persistence. Rejection is a valid outcome. The rule is not an independent safety guarantee.
5. A reserved three-month evaluation block with classical and deep baselines, ablations, seven-day moving-block bootstrap intervals, corrected Diebold–Mariano tests with multiplicity control, and calendar-day maximum-residual conformal intervals.

The study addresses three research questions.

RQ1: Does a compact Transformer improve fifteen-minute-ahead active-power accuracy over persistence, recurrent networks and conventional learners?

RQ2: Does a calendar-day minimax game support deployment of a mixture rather than persistence?

RQ3: Are any observed differences robust to temporal dependence and uncertainty analysis?

The results answer RQ1 in a metric-dependent way, RQ2 negatively for this year, and RQ3 by showing that the small squared-error improvement is not statistically decisive after dependence and multiplicity are considered.

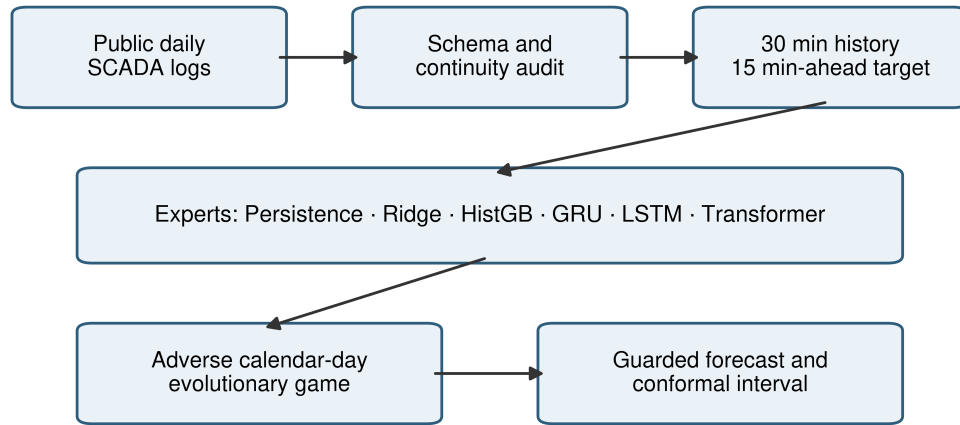


Figure 1: Audited forecasting and guarded-deployment workflow. Supervisory control and data acquisition (SCADA) files obtained from the HPP flash portal are harmonized and screened into continuous thirty-minute histories with fifteen-minute targets. The expert pool contains persistence, ridge regression, histogram gradient boosting (HistGB), a gated recurrent unit (GRU), a long short-term memory network (LSTM) and a transformer. An evolutionary calendar-day game, exact minimax certificate and development-only guard determine the deployed forecast before separate conformal calibration and held-out evaluation.

2 Related Work and Research Gap

2.1 Hydropower Generation and Hydrological Sequence Forecasting

Hydropower models range from first-principles turbine, penstock and governor equations to fully data-driven predictors. Kishor et al. [1] reviewed dynamic plant and control representations, and Márquez et al. [2] demonstrated detailed modeling and control for a micro-hydropower system. Such mechanistic descriptions are valuable for simulation and controller design, but their parameters and internal states are not necessarily identifiable from a public operational log. A data-driven forecaster can instead learn the mapping from recent

measured conditions to future electrical output, provided its validation does not conflate temporal proximity with generalization.

At plant level, Barzola-Monteses et al. [3] applied artificial neural networks to hydropower production. Sattar Hanoon et al. [4] compared machine-learning algorithms for Three Gorges generation. Dehghani et al. [5] combined grey-wolf optimization with adaptive neuro-fuzzy inference. Tebong et al. [6] proposed a two-level deep ensemble. Zhang et al. [7] recently combined LSTM and Transformer components for medium- and long-term hydropower generation. These studies establish the relevance of nonlinear learning, but direct numerical comparison is inappropriate when horizons, aggregation intervals, plants, covariates and train/test periods differ. The present study therefore uses their methodological lessons while performing all accuracy comparisons on one common trace.

Rainfall-runoff and streamflow forecasting offer a larger body of sequence-model evidence. Kratzert et al. [8] showed that LSTM networks can represent catchment memory. Gauch et al. [9] trained a single recurrent model across multiple timescales. Hunt et al. [10] examined operational streamflow improvements. Klotz et al. [24] addressed uncertainty estimation in deep rainfall-runoff models. These results motivate recurrent comparators and careful uncertainty analysis. They do not remove the need for plant-level validation, because active power is mediated by dispatch and electromechanical control rather than by hydrological response alone.

2.2 Transformers, Recurrent Models and Strong Simple Baselines

Self-attention maps a sequence to content-dependent weighted combinations of all its positions [11]. It shortens the information path between two historical observations relative to a recurrent state update and can, in principle, focus on a gate movement, unit transition or power ramp wherever it occurs in the thirty-minute context. Temporal Fusion Transformers combine recurrent processing, gating and interpretable attention [12]. Informer [13], Autoformer [14], FEDformer [15], PatchTST [16] and iTransformer [18] introduce sparse attention, decomposition, frequency-domain processing, patching or variable-wise attention for longer forecasting problems.

Those large architectures are not transferred wholesale here. The available fitting block contains 4274 forecast windows, so an overparameterized network would be poorly justified. Instead, one single-head, single-block Transformer with 2641 trainable parameters is used as a controlled test of self-attention. A gated recurrent unit (GRU) and an LSTM network of comparable scale separate the benefit of attention from the benefit of nonlinear sequence modeling. Ridge regression and histogram gradient boosting test whether a fixed lag representation is sufficient. Persistence remains the reference against which all learned corrections must earn their inclusion.

This conservative design follows an important empirical lesson: simple time-series baselines can outperform elaborate Transformers when the data, horizon or inductive bias favor local continuity [17]. The appropriate scientific question is consequently not “Can a Transformer be inserted?” but “Under which metric and regime does attention add information beyond the current value, and should that information be deployed?”

2.3 Forecast Aggregation as a Finite Zero-Sum Game

Bates and Granger [19] established the value of combining imperfect forecasts. Weighted-majority and multiplicative-weights procedures formalize repeated allocation among experts [20,22], while replicator dynamics describe how population shares evolve according to relative fitness [21]. Expert aggregation has also been used in electricity-demand forecasting [23]. Most conventional combinations minimize average

validation loss. That objective may choose a model whose gains are concentrated in easy periods while tolerating a severe failure on one operating day.

The formulation in this paper treats the forecaster and calendar day as opposing players. The forecaster chooses an expert distribution. The adverse player chooses a May calendar day. The payoff is daily MAE. The resulting game minimizes the worst calendar-day expected loss under randomized expert selection and, by convexity, an upper bound on the worst-day MAE of the deterministic forecast mixture. Entropic multiplicative-weights updates provide a discrete-time evolutionary learning mechanism, while an exact linear program distinguishes an ergodically averaged numerical candidate from the actual minimax solution. Finally, the deterministic weighted forecast is assessed by a separate guard because the game payoff is built from randomized-expert losses and convexity provides an upper bound, not equality, for the absolute error of a deterministic mixture.

2.4 Dependence-Aware Comparison and Predictive Uncertainty

Forecast errors sampled in time are serially dependent. Diebold and Mariano [25] proposed testing equal predictive accuracy from a loss-difference sequence. Harvey et al. [26] provided a small-sample correction. A heteroskedasticity-and-autocorrelation-consistent (HAC) long-run variance is needed here because operating conditions persist within days. Moving-block bootstrap resampling [27] supplies complementary confidence intervals without treating all 1999 origins as independent.

Split conformal prediction constructs intervals from held-out residual quantiles [28]. Dynamic time-series adaptations recognize that temporal dependence and distribution shift complicate the exchangeability argument [29]. This study therefore uses June calendar-day maximum residuals as conservative non-conformity scores and reports the resulting July–September coverage diagnostically. It does not claim distribution-free time-series coverage under seasonal nonstationarity.

Table 1 locates the proposed pipeline relative to representative alternatives and identifies the gap addressed by their combination.

Table 1: Methodological comparison and research gap.

Approach	Temporal Representation	Protection in a Small Plant-Specific Sample	Explicit Regime Robustness	Statistical and Interval Evidence
Static neural or neuro-fuzzy hydropower models [3,5]	Fixed feature vector	Architecture-specific regularization	Usually average-loss fitting	Often point metrics
Recurrent hydrological models [8–10]	Compressed hidden state	Gating and early stopping	Learned implicitly	Study-dependent

(Continued)

Table 1 (continued)

Approach	Temporal Representation	Protection in a Small Plant-Specific Sample	Explicit Regime Robustness	Statistical and Interval Evidence
Transformer models [12–16,18]; linear baseline [17]	Attention, decomposition, patching, variable-wise attention or linear lags	Commonly require substantial data and tuning	Learned implicitly	Model-dependent
Recent close-domain plant studies [30,31]	Temporal Fusion Transformer, seasonal autoregression, autoregressive neural network or LSTM	Plant-specific rolling or pseudo-forecast evaluation	No minimax calendar-day selector	Probabilistic scores or prediction-interval coverage, but not the present matched 15-min task
Conventional forecast combinations [19,23]	Expert outputs	Diversification	Usually average validation loss	Intervals not intrinsic
Proposed guarded game	Residual persistence, linear, tree, GRU, LSTM and compact Transformer experts	Train-only preprocessing, compact networks, early stopping and fallback vertex	Minimax loss over calendar days plus an explicit guard	Moving-block bootstrap, corrected Diebold–Mariano tests, multiplicity control and calendar-day conformal diagnostics

Two recent close-domain studies sharpen the comparison. Francisco et al. applied a Temporal Fusion Transformer to probabilistic hourly potential-power forecasting for a Portuguese run-of-the-river plant, with climatology and persistence as explicit benchmarks. At their shortest evaluated lead of 1 h, persistence markedly outperformed the Transformer. Uncertainty is assessed using quantile loss, continuous ranked probability score, quantile–quantile reliability and relative resolution [30]. Kåge et al. compared 24-h-ahead hourly forecasts from a seasonal autoregressive integrated moving-average model, an autoregressive neural network and an LSTM at a manually scheduled Swedish hydropower station. Their point errors are similar, whereas 95% interval coverage is 0.93, 0.27 and 0.26, respectively [31]. The latter study included neither persistence nor a Transformer.

Those experiments are not numerically comparable with the present 15-min-ahead SCADA-only task because plants, targets, covariates, sampling frequencies and horizons differ. They nevertheless reinforce the need for a strong persistence benchmark, direct recurrent-vs.-attention comparison on the same trace and

explicit uncertainty diagnostics. The present study adds a compact standalone Transformer, calendar-day minimax game aggregation, an exact saddle certificate, a development-block guard, dependence-aware intervals and a one-hydrological-year acquisition period with four explicitly missing daily files. It is not claimed to be an exhaustive benchmark of every recent large time-series architecture; evaluating Temporal Fusion Transformer, DLinear, PatchTST and iTransformer under matched tuning budgets remains future work.

3 Data Source, Audit and Forecasting Problem

3.1 Public Source and Hydrological-Year Study Period

Daily text logs were obtained from the HPP Flash supervisory-control portal during the acquisition period [32]. The study interval, 1 October 2020 through 30 September 2021, follows the hydrological year used for this analysis. The portal supplied 361 of 365 daily files. Files for 17, 18, 19 and 20 July 2021 are unavailable and are not synthesized. The source timestamps have no time-zone or daylight-saving offset, so they are treated as time-zone-naïve local clock values. This is a limitation rather than silently assigning a zone.

3.2 Schema Harmonization and Timestamp Integrity

The portal layout changes during the year. Early records have the timestamp, two plant water channels and two 31-field hydro-generator blocks, yielding a common 70-field core. Later records append load and meteorological channels, producing a dominant 104-field layout. Only the first 70 fields common to both layouts are retained. This decision avoids a structural pattern in which weather variables are absent throughout the early season. Three data rows contain fewer than 70 fields—one each on 20 November 2020, 6 April 2021 and 8 June 2021—and are rejected. Occasional longer rows are restricted to the common core.

Parsing accepts 508,502 timestamped rows before exact-time de-duplication. After sorting, 268 duplicate timestamps are removed by retaining the first occurrence, leaving 508,234 unique timestamps between 00:00:03 on 1 October 2020 and 23:59:03 on 30 September 2021. At the source-audit level, flooring all unique timestamps to the minute produces 1156 same-minute collisions and 507,078 occupied cells, or 96.476% of the 525,600 theoretical minutes. The modeling stream first removes 293 rows whose total active power is missing; its remaining 507,941 exact-time rows produce 1145 same-minute collisions and 506,796 target-valid occupied cells (96.422%) when the first record in each minute is retained. These occupancy percentages are audit measures, not counts of accepted multivariate forecast windows.

A new continuity segment begins at the first observation or when the difference between successive exact timestamps is less than 30 s or greater than 90 s. The lower threshold separates multiple observations too close to represent successive nominal minutes. The upper threshold marks missing nominal-minute observations. This rule produces 1395 segments. The largest interruption is 405,697 s, extending from 16 July 2021 at 18:23:03 to 21 July 2021 at 11:04:40. It contains the four entirely unavailable dates, 17–20 July, together with partial-day loss on the two boundary dates. No interpolation is used in the final annual experiment, and a forecast sample is accepted only if every history minute and its target belong to the same segment.

Fig. 2 displays the annual active-power trace and the number of target-valid occupied minute cells per day. The seasonal rise into April–May and the decline toward September are visible, as is the four-day July gap. These changes make a chronological test substantially harder and more informative than a random split.

Table 2 provides the corresponding numerical audit.

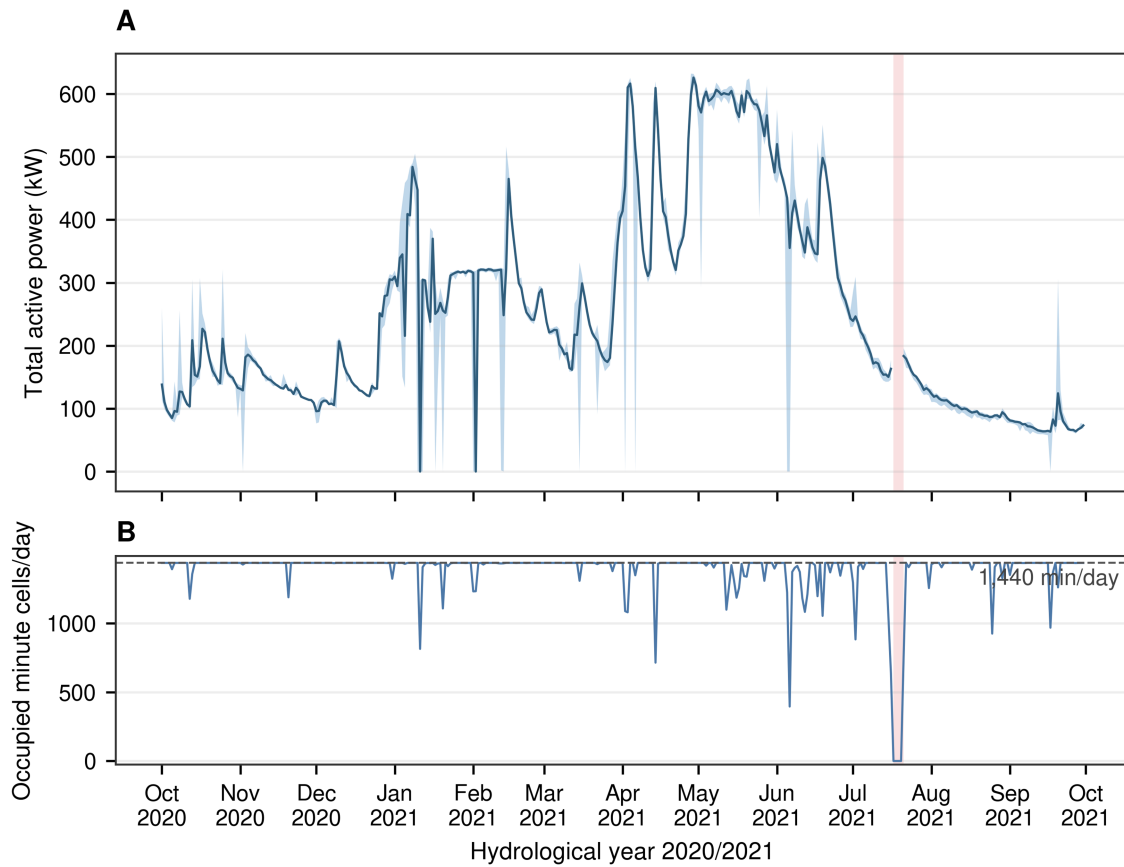


Figure 2: Hydrological-year source audit. Panel A shows the daily median and 10th–90th percentile band of total active power computed as the sum of the two hydro-generator active-power channels. Panel B shows the target-valid occupied minute cells after removal of missing-target rows and same-minute collisions. The dashed reference is 1440 min/day. The four entirely missing dates, 17–20 July, appear as a light-red gap; the associated interruption also includes partial-day loss on 16 and 21 July. Source: HPP flash portal [32], authors’ processing.

Table 2: Reproducible data-audit summary. Source: HPP flash portal [32]; authors’ processing.

Audit Item	Result
Requested calendar days	365
Available daily files	361
Missing dates	17–20 July 2021
Accepted rows before exact-time de-duplication	508,502
Exact duplicate timestamps removed	268
Unique exact timestamps	508,234
Same-minute collisions in source-level minute audit	1156
Source-level occupied minute cells	507,078
Theoretical minute cells	525,600
Source-level minute-cell coverage	96.476%
Rows removed for missing total active power	293
Target-valid exact-time rows before minute de-duplication	507,941

(Continued)

Table 2 (continued)

Audit Item	Result
Same-minute collisions in modeling stream	1145
Target-valid occupied minute cells/coverage	506,796/96.422%
Rows shorter than the common 70-field core	3
Continuity segments	1395
Largest observed gap	405,697 s
Complete, same-segment forecasting windows	8298

3.3 Target, Plant Data Dictionary, Inputs and Sample Construction

Let $P_t^{(1)}$ and $P_t^{(2)}$ denote the active-power channels of hydro-generators HG1 and HG2 at the forecast origin. The plant-level instantaneous active power is

$$P_t = P_t^{(1)} + P_t^{(2)} \quad (1)$$

The supervised target is $y_t = P_{t+15}$ in kW. It is active power, not electrical energy. Energy over an interval would require time integration and would be expressed in kWh or MWh; no such integration is performed. Each sample contains the thirty one-minute vectors ending at origin t , and origins are placed every 60 min. On the regular grid the first origin is 00:30 with a 00:45 target. The input tensor for sample i is

$$\mathbf{X}_i = [\mathbf{x}_{t_i-29}, \dots, \mathbf{x}_{t_i}] \in \mathbb{R}^{30 \times 20}, \quad y_i = P_{t_i+15} \quad (2)$$

Accordingly, every retained origin is at HH:30 and predicts HH:45; no other within-hour phase is evaluated.

All 30 history rows must be complete for the twenty inputs, the target must be observed, and all 31 required records must have one identical continuity-segment identifier. These rules exclude windows that touch a missing minute, a discontinuity or the four-day portal outage. They yield 8298 valid hourly-origin windows.

Table 3 lists the predictors. The two trigonometric variables use minute-of-day m_t and are computed as $\sin(2\pi m_t/1440)$ and $\cos(2\pi m_t/1440)$. No future schedule, weather observation or target-derived regime label is included.

Table 3: Compact public-source dictionary for the 20 predictors. Indices are zero-based positions in the common 70-field schema. Current portal metadata are not an operator-certified historical dictionary. Source: HPP flash daily-file header and public portal metadata [32]; authors' cross-walk.

Channel(s); Zero-Based Index	Interpretation	Unit/Provenance	Model Use and QA
level, setlvl; 6, 7	Level-related measurement and set-point-related channel	Current plot module: mm; historical TXT header silent	Inputs; no conversion; common scale/datum must be operator-confirmed

(Continued)

Table 3 (continued)

Channel(s); Zero-Based Index	Interpretation	Unit/Provenance	Model Use and QA
HG1/HG2 P; 8, 39	Active electrical power	kW; historical header and portal	Inputs and plant-level target; energy would require time integration
HG1/HG2 Q_e , PF; 9–10, 40–41	Reactive power and power factor	kVAr and dimensionless	Inputs; check $ PF \leq 1$ and $ PF \approx$ $ P /\sqrt{(P^2+Q_e^2)}$ only under compatible signs/units
HG1/HG2 open, rpm, press; 20–22, 51–53	Opening-related signal, rotational speed, pressure-related signal	Current portal: %, rpm, m; partial historical provenance	Inputs; historical scale, sensor location and valid range require confirmation
HG1/HG2 temp1–temp2; 24–25, 55–56	Upper-bearing and winding-1 temperatures	Current generator pages: °C; historical header silent	Inputs; apply only operator-approved validity/sentinel limits
temp3–temp5; 26–28, 57–59	Other winding and lower-bearing temperatures	Current generator pages: °C	Not used; retained in dictionary for explicit feature selection
tod_sin, tod_cos	Cyclic minute-of-day encoding	Dimensionless; author-engineered	Inputs; one consistent local-clock convention

The historical daily-file header explicitly documents active power P in kW, reactive electrical power Q_e in kVAr, voltage in V, current in A, and rotational speed in rpm, but it does not provide units for level, setlvl, open, press, or temp1–temp5. Channel semantics are therefore cross-walked among the common 70-field positions, the current HTML labels, and the JavaScript field indices of the public portal [32]. The current portal identifies open as opening (%), press as “Pressure, m”, and temp1–temp5 as upper-bearing, three winding, and lower-bearing temperatures (°C). The current level-plot module labels level and setlvl in millimetres. Because these web assets may post-date the 2020–2021 files, they are treated as secondary public provenance rather than an operator-certified historical schema. No unit conversion is applied to the historically unverified channels; their source-native numeric values are standardized using statistics fitted only on the model-fitting period.

Throughout the electrical feature definitions, Q_e denotes reactive electrical power in kVAr. If a future physics-informed extension introduces water discharge, Q_w shall denote volumetric water flow in m^3/s ; it must not be confused with Q_e . The hydraulic relation is $P_h = \eta \rho g Q_w H$ and is not evaluated here because operator-certified flow and head metadata are unavailable.

Across 508,234 exact-deduplicated timestamps, the observed source-native ranges are 0–99 for open, 0–85.3/85.9 for press, and 0–1340.0/1363.7 for HG1/HG2 rpm. Repeated extremes in level, setlvl, P , PF , and temperature channels are compatible with, but do not prove, sentinel or register-encoding effects. No undocumented unwrap, offset, or unit conversion is applied in the primary analysis. Diagnostic thresholds

are used only for data-quality flags and *post hoc* sensitivity analysis. Fitting-period standardization makes the inputs numerically dimensionless for optimization, but it does not validate their physical units.

3.4 Chronological Partitions and Information Barriers

The partitions are defined by target timestamp (Table 4 and Fig. 3). October–March is the only fitting interval. April is used for ridge-penalty selection and neural early stopping. May supplies the calendar-day payoff matrix and the guard calculation. June supplies conformal nonconformity scores. July–September is a held-out evaluation block: its observations are not used to fit an expert, estimate a game weight, apply the guard, or calibrate an interval width.

Table 4: Chronological partitions, sample counts and active-power regimes. Source: authors' processing of HPP flash data [32].

Role	Target Dates	n	Mean Target (kW)	Mean $ P_{t+15} - P_t $ (kW)
Model fitting	1 Oct 2020–31 Mar 2021	4274	212.671	4.500
Internal early stopping	1–30 Apr 2021	685	442.708	4.578
Game and guard tuning	1–31 May 2021	694	574.210	7.312
Conformal calibration	1–30 Jun 2021	646	384.413	6.711
Held-out evaluation	1 Jul–30 Sep 2021	1999	111.849	1.809

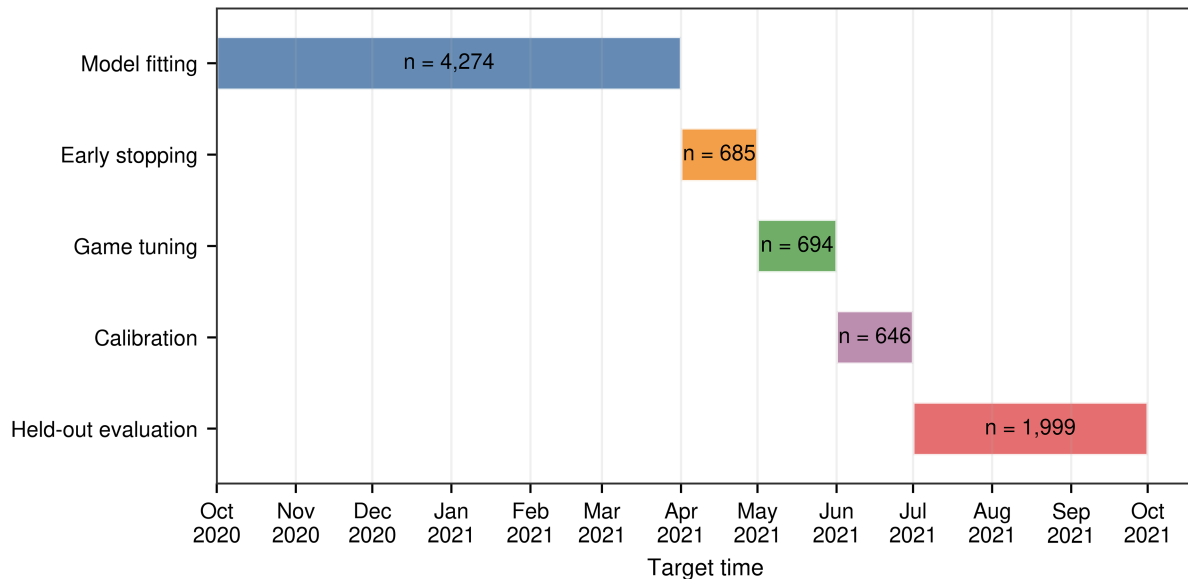


Figure 3: Chronological experimental partitions and information barriers. From left to right, the blocks are model fitting (October–March), neural early stopping (April), game and guard tuning (May), conformal calibration (June), and held-out evaluation (July–September). Sample counts refer to valid forecast origins classified by target timestamp. No random split or backward information flow is permitted. Source: authors' processing of HPP flash data [32].

The mean target falls from 574.210 kW in May to 111.849 kW in testing, and the mean absolute fifteen-minute change falls from 7.312 to 1.809 kW. This is a substantial operating-distribution shift. It also explains why persistence is particularly difficult to beat in the final quarter. The protocol evaluates that shift rather than eliminating it through shuffled sampling.

4 Proposed Forecasting and Game-Theoretic Method

4.1 Persistence-Residual Learning

The persistence expert is

$$\hat{P}_{t+15|t}^{(0)} = P_t \quad (3)$$

Every learned expert predicts the change relative to persistence,

$$r_{t+15} = P_{t+15} - P_t \quad (4)$$

If expert m outputs $\hat{r}_{t+15}^{(m)}$, its active-power forecast is

$$\hat{P}_{t+15|t}^{(m)} = P_t + \text{clip}\left(\hat{r}_{t+15}^{(m)}, c_{0.005}, c_{0.995}\right) \quad (5)$$

where $c_{0.005} = -52.3985$ kW and $c_{0.995} = 72.9970$ kW are the 0.5th and 99.5th percentiles of fitting-block residuals. Preprocessing differs only where required by the estimator. The GRU, LSTM and Transformer receive channel-wise standardized sequence inputs, with means and scales fitted to all historical tokens in October–March. Their residual targets are standardized with fitting-block moments and optimized by mean squared error. Ridge receives a separately standardized fixed-lag vector, whereas histogram gradient boosting receives the corresponding unscaled fixed-lag vector. All transformations are then frozen. Eq. (5) supplies the persistence skip connection and a train-derived extrapolation cap, not a physical safety guarantee. The 0.5% quantile choice is a disclosed regularization decision whose effect on rare ramps is considered as a limitation. Mean squared error is retained as the common differentiable neural training loss, whereas deployment uses MAE for its direct kW interpretation and reduced domination by rare ramps. This deliberate objective mismatch helps explain why lower RMSE need not imply lower MAE; a matched MAE/Huber-loss retraining ablation is not available in the fixed experiment.

4.2 Ridge and Histogram Gradient-Boosting Experts

The fixed-vector experts use lags 0, 1, 4, 9, 14 and 29 min for all twenty channels. Five summaries of the thirty-minute total-power history—mean, standard deviation, minimum, maximum and end-minus-start change—are appended. Ridge regression [33] estimates a linear residual map with L_2 penalty. Candidate penalties $\{0.01, 0.1, 1, 10, 100\}$ are fitted on October–March and compared by April MAE, $\alpha = 100$ is selected.

Histogram gradient boosting [34] fits nonlinear partitions to the same fixed representation. It uses squared-error loss, 250 iterations, learning rate 0.05, at most 15 leaf nodes and L_2 regularization 1.0. Its internal early stopping is disabled so that the number of iterations remains fixed. Both fixed-vector models apply the train-only residual limits in Eq. (5).

4.3 Recurrent Experts

The GRU [35] and LSTM [36] experts each contain one recurrent layer with 16 hidden units, followed by an 8-unit rectified-linear readout and one linear residual output. Their 1921 and 2513 trainable parameters, respectively, are of the same order as the Transformer. Both are trained on the standardized residual with mean squared error, Adam at learning rate 0.0015, batch size 128, no dropout and no shuffling. The maximum is 12 epochs, and training stops after three successive epochs without an April standardized-residual validation-loss improvement of at least 10^{-5} . The best validation-loss weights are restored. Under the fixed seed, the GRU completes seven epochs and the LSTM five.

Recurrent models compress the ordered history into a hidden state. Their gates can preserve or discard earlier information, which is appropriate for short plant dynamics, but all distant interactions pass through the recurrent chain. This provides a meaningful architectural comparator for the direct pairwise interactions of self-attention.

4.4 Compact Self-Attention Transformer

For each standardized input vector $\tilde{x}_j$, a learned linear projection to dimension $d = 16$ is added to a deterministic sinusoidal positional vector p_j :

$$h_j^{(0)} = W_e \tilde{x}_j + b_e + p_j \quad (6)$$

The single attention head forms query, key and value matrices

$$Q = HW_Q, \quad K = HW_K, \quad V = HW_V \quad (7)$$

and computes

$$\text{Attn}(H) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) VW_O \quad (8)$$

All thirty positions precede the forecast origin, so every interaction is historical and a causal mask within the input window is unnecessary. Residual connections and layer normalization surround attention and a position-wise feed-forward network of width 32:

$$U = \text{LayerNorm}(H + \text{Attn}(H)) \quad (9)$$

$$H' = \text{LayerNorm}[U + W_2 \text{GELU}(W_1 U + b_1) + b_2] \quad (10)$$

The last historical token is sent through an 8-unit rectified-linear layer and a scalar output. The model has 2641 trainable parameters. Optimization, standardized-residual mean squared error and the absence of dropout match the recurrent networks. It completes six epochs under early stopping. The theoretical reason to include attention is its ability to compare a recent state directly with any preceding minute and condition those comparisons on content. The practical reason to keep it compact is that the fitting sample is only 4274 windows. Its value is tested through out-of-time errors, not inferred from attention alone. An architecture-matched no-attention retraining control is unavailable for the annual experiment. A separate one-day minute-rolling control is reported in [Sections 5.7](#) and [6.10](#). It does not identify an annual causal attention effect.

4.5 Calendar-Day Zero-Sum Game

Let $M = 6$ experts and let May contain $B = 31$ available calendar days. Day b has a set $\mathcal{I}_b$ of 17–24 valid hourly origins. The payoff of expert m on day b is its daily MAE,

$$L_{bm} = \frac{1}{|\mathcal{I}_b|} \sum_{i \in \mathcal{I}_b} |y_i - \hat{y}_i^{(m)}| \quad (11)$$

Let $w \in \Delta_M$ denote the forecaster's distribution over experts and $q \in \Delta_B$ the adverse distribution over May calendar days. The minimax value of this finite zero-sum game is

$$V^* = \min_{w \in \Delta_M} \max_{q \in \Delta_B} q^T Lw = \min_{w \in \Delta_M} \max_b (Lw)_b \quad (12)$$

Starting from uniform distributions, the two players are updated simultaneously using exponentiated-gradient, or entropic multiplicative-weights, steps [20–22]:

$$w_m^{(k+1)} = \frac{w_m^{(k)} \exp \left\{ -\eta \left[L^T q^{(k)} \right]_m \right\}}{\sum_j w_j^{(k)} \exp \left\{ -\eta \left[L^T q^{(k)} \right]_j \right\}} \quad (13)$$

$$q_b^{(k+1)} = \frac{q_b^{(k)} \exp \left\{ +\eta \left[L w^{(k)} \right]_b \right\}}{\sum_c q_c^{(k)} \exp \left\{ +\eta \left[L w^{(k)} \right]_c \right\}} \quad (14)$$

The learning rate is $\eta = 0.03 \text{ kW}^{-1}$ and $K = 5000$ iterations are run; the inverse-power unit makes the exponential arguments dimensionless. The candidate is the ergodic average $\bar{w} = K^{-1} \sum_{k=1}^K w^{(k)}$, which reduces cycling and follows multiplicative-weights convergence theory [20,22]. The corresponding deterministic forecast is

$$\hat{y}_i^{(G)} = \sum_{m=1}^M \bar{w}_m \hat{y}_i^{(m)} \quad (15)$$

The distinction between Eqs. (12) and (15) matters. Eq. (12) is the expected loss of randomly choosing an expert. The relevant deterministic-mixture bound can be stated explicitly.

Proposition 1: For any day b and any expert weights $w \in \Delta_M$, the deterministic mixture satisfies $MAE_b \left(\sum_m w_m \hat{y}_i^{(m)} \right) \leq (Lw)_b$ and therefore $\max_b MAE_b \left(\sum_m w_m \hat{y}_i^{(m)} \right) \leq \max_b (Lw)_b$.

Proof: At every origin, the triangle inequality gives $\left| y_i - \sum_m w_m \hat{y}_i^{(m)} \right| = \left| \sum_m w_m (y_i - \hat{y}_i^{(m)}) \right| \leq \sum_m w_m |y_i - \hat{y}_i^{(m)}|$. Averaging over the origins of day b gives the first inequality, taking the maximum over days gives the second. Equality is not guaranteed, so the actual deterministic forecast is evaluated separately by the development screening guard. $\square$

4.6 Exact Minimax Certificate

Finite multiplicative-weights iterations need not return an exact equilibrium. The forecaster's exact minimax problem is also solved as a linear program:

$$\min_{w, v} v \text{ subject to } Lw \leq v \mathbf{1}, \mathbf{1}^T w = 1, w \geq 0 \quad (16)$$

Its solution supplies a numerical certificate $V^* = v$ and allows the finite-candidate upper gap $\max_b (L\bar{w})_b - V^*$ to be reported. A simple pure-strategy saddle certificate is also available: if some day b^* and expert m^* satisfy $L_{b^* m^*} = \max_b L_{b m^*}$ and $L_{b^* m^*} = \min_m L_{b^* m}$, then $L_{b^* m^*}$ simultaneously lower-bounds every forecaster strategy and upper-bounds expert m^* , so (b^*, m^*) is a saddle point. This comparison prevents a nearly converged evolutionary strategy from being described as the exact game solution. The program is solved with SciPy linprog using the HiGHS solver and 1×10^{-9} primal and dual feasibility tolerances; acceptance requires reported optimality and residuals within those tolerances.

4.7 In-Block Development Screening Guard

The candidate is accepted only if both

$$\text{MAE}_{\text{May}}(G) \leq \text{MAE}_{\text{May}}(0) \quad (17)$$

and

$$\max_b \text{MAE}_b(G) \leq \max_b \text{MAE}_b(0) \quad (18)$$

where superscript 0 denotes persistence. If either condition fails, the deployed vector is the persistence vertex $(1, 0, 0, 0, 0, 0)$. This guard cannot guarantee future superiority; it only blocks a candidate that is already inferior on its designated May selection period. Its computation uses May observations only, June is used only for interval calibration and July–September only for held-out evaluation. The equal-weight ensemble reported later is an analytical ablation and is not eligible for deployment under this selection pipeline. Because the same May block constructs the game weights and evaluates the guard, this is an in-block development screening rule rather than an independent validation test. It provides no formal out-of-sample safety guarantee.

4.8 Calendar-Day Maximum-Residual Conformal Intervals

For each of the 30 June calendar days, a maximum absolute residual score is formed,

$$S_d = \max_{i \in \mathcal{I}_d} |y_i - \hat{y}_i| \quad (19)$$

At nominal coverage $1 - \alpha = 0.90$, the finite-sample rank is

$$k = \lceil (n_{\text{cal}} + 1) (1 - \alpha) \rceil = \lceil 31 \times 0.90 \rceil = 28 \quad (20)$$

and $\hat{q}$ is the 28th ordered daily score. The interval at each test origin is

$$C_i = [\hat{y}_i - \hat{q}, \hat{y}_i + \hat{q}] \quad (21)$$

Pointwise coverage and simultaneous calendar-day coverage are both reported, i.e., the fraction of test days on which every available origin is covered. The interval is intentionally conservative because a single large residual controls each day's score. Seasonal distribution shift violates the simple exchangeability argument, so the result is an empirical calibration diagnostic rather than an unconditional coverage guarantee. Intervals are reported on the unbounded model-output scale; no zero floor or plant-capacity clipping is applied because operator-certified physical bounds are unavailable. Sharpness is summarized relative to mean test power and the observed test range.

5 Experimental Protocol and Statistical Analysis

5.1 Fixed Implementation Settings

[Table 5](#) records the principal settings. They are intentionally compact to match the sample size. Python processing uses version 3.12.13 with NumPy 2.3.5, pandas 2.2.3, SciPy 1.17.0 and scikit-learn 1.8.0. Neural models use Node.js 24.14.0 and TensorFlow.js 4.22.0 on the CPU backend. The common seed is 20260714. Deterministic seed control improves repeatability, although CPU kernels and library changes can still

produce small platform differences. These are fixed configurations rather than a compute-matched model-family search: ridge alone receives an explicit penalty grid, and seed variability and architecture uncertainty are not estimated.

Table 5: Fixed model, game and reproducibility settings; the configurations do not constitute a compute-matched model-family search.

Component	Fixed Setting	Data Used for Choice
History/horizon/origin stride	30 min/15 min/60 min	Defined before fitting
Residual limits	$[-52.3985, 72.9970]$ kW	October–March only
Ridge	$\alpha \in \{0.01, 0.1, 1, 10, 100\}$; selected 100	Fit October–March; choose by April MAE
Histogram gradient boosting	250 iterations; learning rate 0.05; 15 leaves; $L_2 = 1$	Fixed; fit October–March
GRU	16 units; 8-unit readout; 1921 parameters	April stopping; 7 epochs completed
LSTM	16 units; 8-unit readout; 2513 parameters	April stopping; 5 epochs completed
Transformer	$d = 16$; one head; one block; feed-forward 32; readout 8; 2641 parameters	April stopping; 6 epochs completed
Neural optimization	Standardized-residual MSE; Adam 0.0015; batch 128; maximum 12 epochs; patience 3; no dropout; no shuffle	October–March fit, April stopping
Evolutionary game	31 May days; $\eta = 0.03 \text{ kW}^{-1}$; 5000 iterations	May only
Exact game	Linear program in Eq. (16)	May only
Conformal interval	$\alpha = 0.10$; daily maximum score; rank 28/30	June only
Reproducibility	seed 20260714; CPU neural backend	Fixed

5.2 Point Metrics

For test errors $e_i = \hat{y}_i - y_i$, the primary selection metric is MAE because it has a direct kW interpretation and limits domination by rare large ramps. RMSE is reported to emphasize large deviations:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |e_i|, \quad \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n e_i^2} \quad (22)$$

Symmetric MAPE (sMAPE), weighted absolute percentage error (WAPE), coefficient of determination and bias are

$$\text{sMAPE} = \frac{100}{n} \sum_i \frac{2|e_i|}{|y_i| + |\hat{y}_i|}, \quad \text{WAPE} = 100 \frac{\sum_i |e_i|}{\sum_i |y_i|} \quad (23)$$

$$R^2 = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \bar{y})^2}, \quad \text{Bias} = \frac{1}{n} \sum_i e_i \quad (24)$$

Terms with a zero sMAPE denominator are assigned zero. Conventional MAPE is unstable near zero, so it is calculated only where $|y_i| \geq 1$ kW; 1995 of 1999 test origins satisfy this guard. Percentage metrics are secondary because the target can approach zero. This thresholded MAPE uses $n_g = 1995$ eligible test origins and is distinct from the game screening guard.

5.3 Moving-Block Bootstrap

Uncertainty intervals use 2000 moving-block bootstrap replicates over daily summaries reindexed to the complete 92-day July–September calendar. The four missing July dates contribute zero observations and are retained as calendar positions rather than being compressed out of time. Each replicate concatenates sampled seven-calendar-day blocks until 92 positions are reached. Within-day forecast origins remain together, preserving local dependence. Percentile 95% intervals are reported for MAE, RMSE, sMAPE and R^2 . Paired intervals reuse identical sampled calendar blocks for the compared forecasts and summarize model-minus-persistence differences.

5.4 Corrected Diebold–Mariano Tests and Multiplicity

The loss difference for model a vs. persistence 0 is $d_i = e_{i,a}^2 - e_{i,0}^2$. The Diebold–Mariano statistic is the sample mean divided by its HAC standard error. A Bartlett/Newey–West form with lag 23 origin indices represents an approximately 24-h dependence window under the nominal hourly schedule. Where origins are missing, it is an index lag rather than an exact elapsed-time lag [37]. Because forecast origins are 60 min apart and the horizon is 15 min, adjacent forecast targets do not overlap and the Harvey–Leybourne–Newbold horizon parameter is $h = 1$. Its finite-sample multiplier is $\sqrt{(n-1)/n} = 0.999750$ for $n = 1999$. Two-sided p -values use a Student reference distribution with 1998 degrees of freedom [25,26]. The analysis uses a significance level of $\alpha = 0.05$.

Holm’s procedure [38] controls family-wise error across all eight squared-error comparisons with persistence reported in Section 6: five learned experts and three aggregation variants. This complete family is used because the study is not preregistered and the resulting tests are retrospective rather than confirmatory. A negative statistic favors the model in squared-error loss. It does not imply lower MAE.

5.5 Ablations and Interpretation Rules

Three aggregation variants clarify the contribution of weighting and of the Transformer:

1. the uniform deterministic mean of all six experts;
2. a five-expert evolutionary game excluding the Transformer;
3. the full six-expert game.

The equal-weight result is purely analytical and does not participate in the guard. The no-Transformer game uses the same May loss construction and 5000 multiplicative-weights updates. The exact linear program is recomputed for its reduced pool. None of these strategy calculations uses July–September observations.

5.6 Locked-Prediction Robustness Analyses

Additional robustness analyses use only the locked May, June and July–September targets and forecasts. They do not retrain, retune or reselect any expert from held-out outcomes. For the evolutionary game, $\eta \in$

$\{0.01, 0.02, 0.03, 0.05, 0.075, 0.10\}$ kW⁻¹ and $K \in \{1000, 2500, 5000, 7500, 10,000\}$ are crossed. Daily May MAE differences are also computed for the five-expert game without the Transformer.

Two secondary selection diagnostics are reported. First, the game is estimated on 1–21 May and the deterministic guard is evaluated on the later 22–31 May block, with engineering tolerances $\tau \in \{0, 0.005, 0.01, 0.05, 0.10\}$ kW applied to both guard inequalities. Second, a simplex-constrained linear program minimizes origin-level May MAE for a deterministic convex ensemble. Its May worst-day loss and locked June/test behavior are then evaluated without changing the primary selection rule. These retrospective analyses are sensitivity checks, not pre-specified deployment guarantees.

Held-out diagnostics stratify calendar days by the median daily mean actual power (95.924702 kW) and origins by realized ramp magnitude. The output-mode threshold is test-derived and descriptive. Ramp-stratum Transformer-minus-persistence MAE intervals use 5000 paired bootstrap replications clustered by calendar day. Finally, the already-fitted forecasts are summarized across May, June and July–September to show temporal variation; May and June retain their development/calibration roles and are not relabeled as independent test sets.

5.7 Supplementary One-Day Rolling-Origin and No-Attention Controls

The archived 7 April 2021 SCADA file is verified against the annual source manifest by filename, byte count and SHA-256 checksum. A model-free phase audit uses every admissible minute with the same 30-min history eligibility and 15-min horizon. The separate model control contains 1396 chronologically ordered windows partitioned into 698 fitting, 139 early-stopping, 139 game-tuning, 139 conformal-calibration and 221 held-out test windows, with 15-origin purges between blocks and no shuffling.

The feed-forward-only control removes self-attention while retaining the 16-dimensional token projection, sinusoidal positional vector, 32-unit position-wise feed-forward residual, layer normalization, last-token 8-unit readout and persistence-residual target. It has 1585 trainable parameters and uses Adam at learning rate 0.0015, batch size 32, fixed seed 20260714, train-only residual clipping and early stopping. Paired differences use 5000 moving-block-bootstrap replications with 15-min blocks. This is a one-day sensitivity check, not a replacement for annual rolling-origin retraining.

6 Results

6.1 Operating Distribution and Persistence Difficulty

The seasonal regime changes strongly across the five partitions (Table 4). May has the highest mean active power (574.210 kW) and mean fifteen-minute absolute change (7.312 kW). Testing has the lowest mean active power (111.849 kW) and change (1.809 kW). Thus, the test is not an identically distributed extension of the tuning months. A method selected for average performance in the high-output spring regime may add unnecessary corrections in the low-ramp late-summer regime.

This distributional fact is central to interpretation. Persistence error equals the realized absolute change $|P_{t+15} - P_t|$. Its 1.8087 kW test MAE is therefore also a direct description of typical fifteen-minute movement in July–September. Any learned residual must improve enough transition cases to compensate for the errors it introduces across the many near-flat cases.

With the fitted experts held fixed, persistence/Transformer MAE is 7.3121/7.7002 kW in May, 6.7108/6.8034 kW in June and 1.8087/2.1434 kW in July–September; the corresponding RMSE values are 32.9847/32.9299, 19.4079/19.3726 and 6.8910/6.8587 kW. Thus, the Transformer's small RMSE advantage is directionally repeated across the three locked blocks, whereas persistence retains lower MAE. This is a temporal stress diagnostic, not multi-year or multiple independent holdout validation.

6.2 Test Accuracy of Individual Experts

Table 6 and Fig. 4 present the held-out evaluation results. Persistence attains the lowest MAE (1.8087 kW), sMAPE (1.9550%), WAPE (1.6170%) and thresholded MAPE (1.5783%). The Transformer is the strongest learned individual model, as its MAE is 2.1434 kW and RMSE 6.8587 kW. Relative to persistence, Transformer MAE is 0.3348 kW (18.5%) higher, while RMSE is 0.0323 kW (0.47%) lower. This split is consistent with limited complementary behavior in some large-error cases but worse corrections across typical stable observations; it does not identify attention as the causal mechanism.

Table 6: Held-out July–September metrics ($n = 1999$). Best value among the six individual experts is italicized for each error metric. Source: authors' calculations from HPP flash data [32].

Forecast	MAE (kW)	RMSE (kW)	Thresholded MAPE ($ y \geq 1$ kW; %)	sMAPE (%)	WAPE (%)	R ²	Bias (kW)
Persistence	1.8087	6.8910	1.5783	1.9550	1.6170	0.97910	−0.1505
Ridge	5.1984	8.4092	4.8581	5.1533	4.6476	0.96888	3.4258
Histogram gradient boosting	5.2890	8.6986	5.7271	6.3232	4.7287	0.96670	−3.2965
GRU	6.3187	8.7321	6.7137	6.9523	5.6493	0.96645	5.2549
LSTM	3.2899	7.1978	3.3986	3.8426	2.9414	0.97720	1.3733
Transformer	2.1434	6.8587	1.9921	2.5357	1.9163	0.97930	0.4554

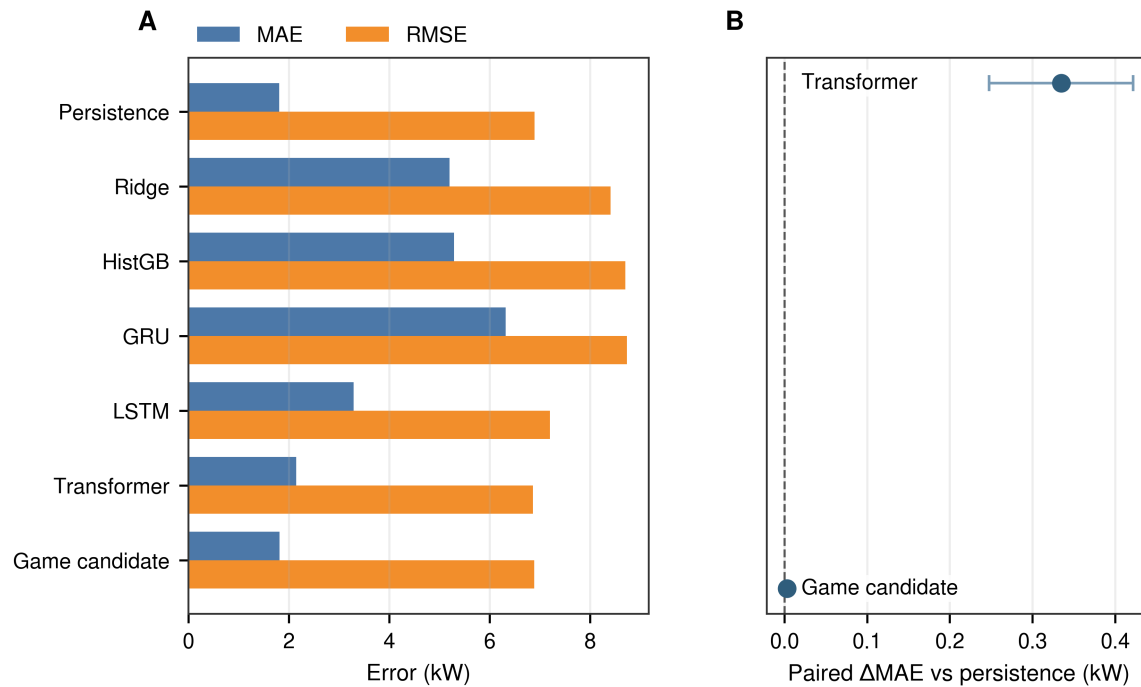


Figure 4: Held-out accuracy and paired uncertainty. Panel A compares mean absolute error (MAE) and root-mean-square error (RMSE) for the six individual experts and the full evolutionary-game candidate. Panel B shows point estimates and 95% seven-calendar-day paired moving-block-bootstrap intervals for the Transformer-minus-persistence and game-candidate-minus-persistence Δ MAE; positive values favor persistence. Source: authors' calculations from HPP flash data [32].

The LSTM is the second-best learned model by MAE (3.2899 kW), whereas ridge, histogram gradient boosting and GRU have MAEs between 5.1984 and 6.3187 kW. All models retain high R^2 because the total-power level varies far more than the short-horizon residual. Consequently, R^2 alone would obscure operational differences.

6.3 Sampling Uncertainty and Predictive-Accuracy Tests

The seven-day moving-block intervals are wide relative to the 0.0323 kW Transformer RMSE difference. Persistence has a 95% MAE interval of [1.1957, 2.0750] kW and an RMSE interval of [4.0979, 7.4385] kW. The Transformer intervals are [1.5702, 2.3836] kW and [4.1372, 7.4012] kW, respectively. Individual intervals overlap because both forecasts experience the same difficult episodes, which is why the paired differences are more informative.

The paired Transformer-minus-persistence MAE difference is positive with 95% interval [0.2477, 0.4217] kW. The paired RMSE difference interval [-0.0599, 0.0445] kW includes zero. The squared-error Diebold–Mariano–Harvey–Leybourne–Newbold statistic is -0.9337 with unadjusted $p = 0.3506$, so the null of equal predictive accuracy is not rejected. The full game candidate has a tiny RMSE advantage and an unadjusted $p = 0.0347$, but Holm adjustment across all eight reported comparisons gives $p = 0.1389$. Neither the Transformer nor the game candidate is significant at 0.05 after multiplicity control. Tables 7 and 8 report every member of the eight-comparison family.

Table 7: Paired effect estimates relative to persistence on the held-out evaluation block. Confidence intervals use 2000 seven-calendar-day paired moving-block percentile-bootstrap replications. Source: authors' calculations from HPP flash data [32].

Model Minus Persistence	Point Δ MAE (kW)	Paired 95% CI	Point Δ RMSE (kW)	Paired 95% CI
Ridge	3.3897	[2.6345, 4.4985]	1.5182	[1.1893, 2.8038]
Histogram gradient boosting	3.4803	[2.6809, 4.3229]	1.8076	[1.4704, 3.4431]
GRU	4.5101	[3.8415, 5.2363]	1.8411	[1.6501, 2.9229]
LSTM	1.4812	[1.0297, 1.8871]	0.3068	[0.2109, 0.7640]
Transformer	0.3348	[0.2477, 0.4217]	-0.0323	[-0.0599, 0.0445]
Equal-weight mean	0.8401	[0.7092, 1.0737]	-0.1448	[-0.2621, 0.3466]
Game without Transformer	0.0028	[0.0020, 0.0038]	-0.0023	[-0.0032, 0.0000]
Full game candidate	0.0033	[0.0022, 0.0047]	-0.0027	[-0.0038, -0.0003]

Table 8: Dependence-aware squared-error comparisons with persistence. Δ MSE is the model-minus-persistence mean squared-loss difference in kW². DM–HLN denotes the Diebold–Mariano statistic with Harvey–Leybourne–Newbold correction; Holm p is adjusted across all eight comparisons. Source: authors' calculations from HPP flash data [32].

Model Minus Persistence	Δ MSE (kW ²)	DM–HLN Statistic	Raw p	Holm p
Ridge	23.2287	5.3191	$p < 0.001$	$p < 0.001$
Histogram gradient boosting	28.1790	3.5652	$p < 0.001$	0.002232
GRU	28.7643	10.6671	$p < 0.001$	$p < 0.001$
LSTM	4.3222	2.6572	0.007943	0.03972
Transformer	-0.4444	-0.9337	0.3506	0.7012
Equal-weight mean	-1.9751	-0.7513	0.4525	0.7012
Game without Transformer	-0.0323	-1.9575	0.05043	0.1513
Full game candidate	-0.0378	-2.1129	0.03473	0.1389

The p -value and effect-size conclusions are intentionally separated. For the game candidate, the bootstrap identifies a very small but consistent MAE deterioration and a very small RMSE improvement. Even if one considered the unadjusted squared-error p -value, a 0.0027 kW RMSE reduction would require an operational cost analysis before being described as useful. The selection guard appropriately uses development MAE and worst-day MAE rather than test significance.

Positive DM–HLN statistics indicate larger squared error than persistence. Thus, the significant differences for ridge, histogram boosting, GRU and LSTM indicate deterioration, not improvement. Sensitivity calculations using HAC lags of 0, 6, 12, 23, 47 and 71 hourly origins leave the substantive result unchanged, as the Transformer's Holm-adjusted p ranges from 0.5883 to 0.7012 and the full game candidate's from 0.0624 to 0.1389. Ridge, histogram boosting and GRU remain significant throughout, LSTM crosses 0.05 only at lag 71. These lag checks are retrospective robustness analyses, not additional selection criteria.

6.4 Transition Episode

Fig. 5 shows the first four test days, selected by a fixed calendar rule rather than by searching for a favorable event. Persistence and the full game candidate almost overlap because the candidate assigns 98.36% weight to persistence. The Transformer deviates more during both ramps and stable periods. Some deviations reduce the largest squared errors, while many small deviations increase absolute error. The plot visually supports the metric split in Table 6 but is illustrative; all inference uses the complete test.

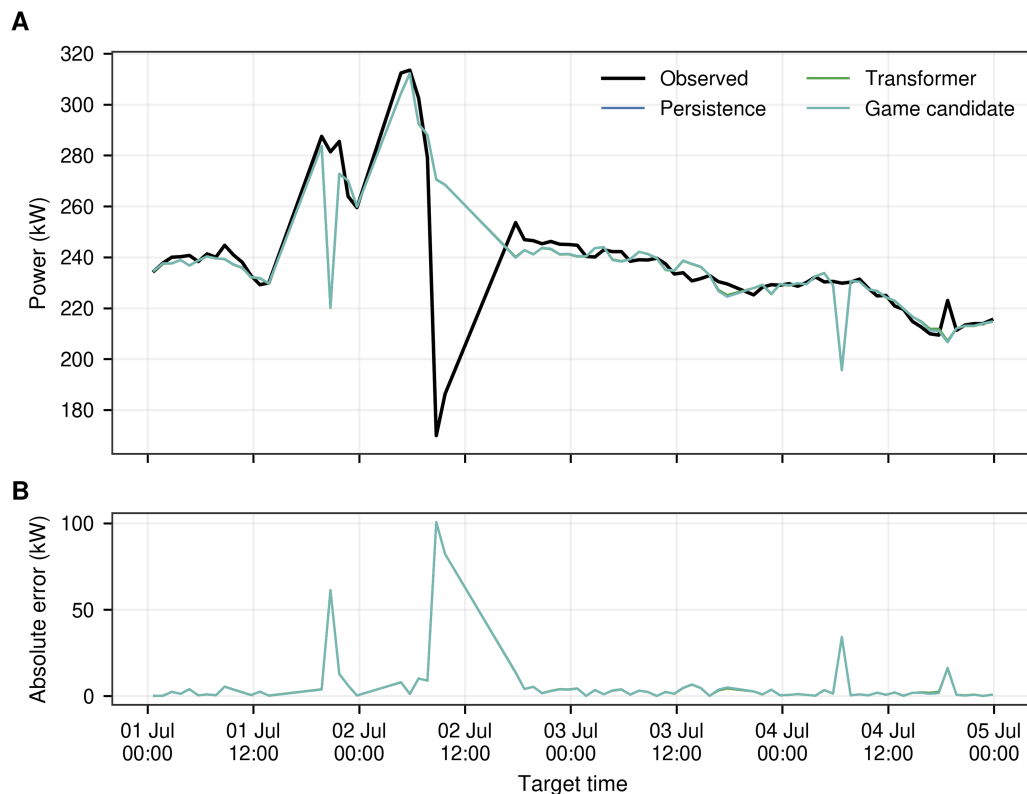


Figure 5: Illustrative fixed held-out episode, 1–4 July 2021. Panel A compares observed active power with persistence, the compact transformer and the 5000-step ergodically averaged evolutionary-game candidate. Panel B shows absolute forecast errors. The calendar episode is descriptive; selection and inference use their designated full periods. Source: authors' calculations from HPP flash data [32].

6.5 Evolutionary Game, Exact Solution and Guard Decision

The finite evolutionary candidate places 0.983649 on persistence, 0.011450 on the Transformer, 0.003200 on LSTM and less than 0.001 on each remaining expert (Table 9). Its worst expected May daily loss under the randomized-expert payoff is 32.250981 kW. The exact linear program selects pure persistence and certifies a game value of 32.237504 kW. The multiplicative-weights upper gap is therefore 0.013477 kW after 5000 iterations. Fig. 6 displays the final ergodic weights and the convergence of the upper objective toward the exact value. The adverse multiplicative-weights population becomes effectively concentrated on 2 May, the difficult day that controls the maximum.

The deterministic candidate has May MAE 7.312412 vs. 7.312105 kW for persistence. Its worst May calendar-day MAE is 32.242955 vs. 32.237504 kW. Both inequalities in the zero-margin guard fail, although only by 0.000307 and 0.005451 kW. These margins are below the 0.1 kW decimal precision printed for each unit's active power, so they are a deterministic acceptance decision rather than evidence of operationally material harm. The candidate is consequently rejected before testing. The selected weights are exactly (1, 0, 0, 0, 0, 0) and the guarded test forecast is identical to persistence.

The pure saddle condition from Section 4.6 is directly verified. On 2 May, daily MAE is 32.237504 kW for persistence, 36.305288 for ridge, 40.198447 for histogram boosting, 35.115500 for GRU, 33.055350 for LSTM and 32.534518 for the Transformer. This is persistence's worst May day and persistence is the row minimum on that day. Consequently, pure persistence against 2 May is a saddle point, and its 32.237504 kW loss is both a lower and an upper bound on the game value.

Table 9: Calendar-day game and development-screening-guard results on May 2021. Source: authors' calculations from HPP flash data [32].

Quantity	Persistence	Ridge	HistGB	GRU	LSTM	Transformer	Value/Decision
5000-step candidate weight	0.983649	0.000762	0.000327	0.000611	0.003200	0.011450	Maximum expected daily MAE: 32.250981 kW
Exact linear-program weight	1.000000	0	0	0	0	0	Game value 32.237504 kW
Selected guarded weight	1.000000	0	0	0	0	0	Candidate rejected
Overall deterministic May MAE	7.312105	—	—	—	—	—	Candidate: 7.312412 kW
Worst deterministic May-day MAE	32.237504	—	—	—	—	—	Candidate: 32.242955 kW

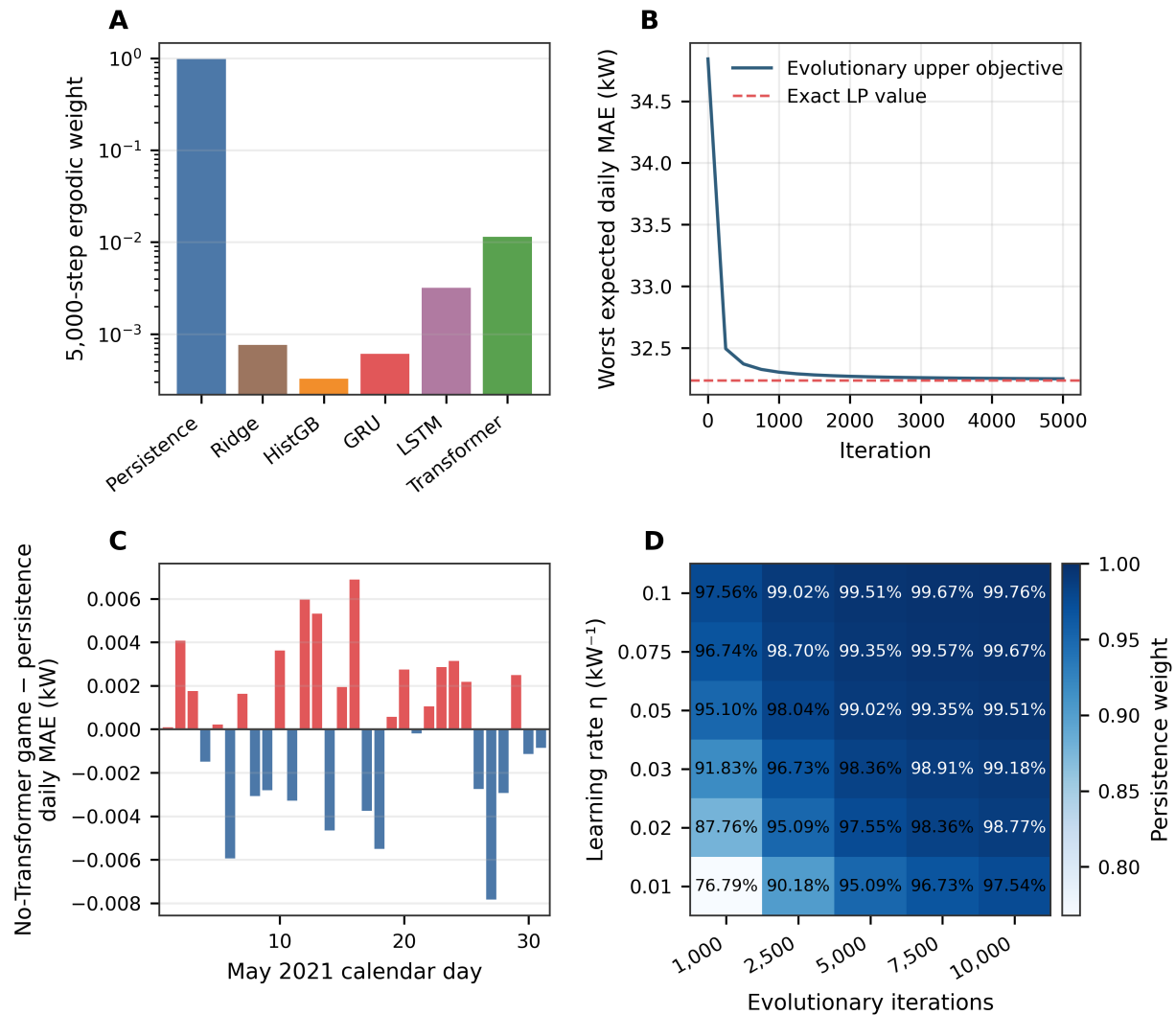


Figure 6: May game robustness from locked forecasts. Panel A shows baseline 5000-step expert weights; Panel B compares the evolutionary upper objective with the exact minimax value; Panel C reports daily MAE of the five-expert game without the transformer minus persistence; Panel D shows persistence weight across 30 η -iteration combinations. Positive Panel C values favor persistence. Source: authors' calculations from locked May predictions.

Across all 30 η -iteration combinations, the ergodic persistence weight ranges from 0.767865 to 0.997559 and the upper saddle gap from 0.001971 to 0.199665 kW. Every finite candidate fails the zero-margin screening guard, while the exact linear program remains pure persistence. The numerical path is therefore hyperparameter-sensitive, but the certified game solution and primary decision are not.

For the 5000-step game without the Transformer, daily May MAE is below persistence on 14 days and above it on 17; the median difference is +0.000218 kW and the largest deterioration is +0.006880 kW on 16 May. On the controlling 2 May day, persistence is 32.237504 vs. 32.241574 kW for the reduced-pool candidate. Daily evidence therefore supports, rather than conceals, its guard rejection.

In the chronological May sensitivity, the game fitted on 1–21 May again has an exact pure-persistence solution. On the independent 22–31 May guard block, its finite candidate improves overall MAE (7.025060 vs. 7.026523 kW) but worsens the worst-day MAE (30.118637 vs. 30.104175 kW). It fails $\tau = 0, 0.005$ and 0.01 kW but passes $\tau = 0.05$ and 0.10 kW; on the original full-May in-block calculation it passes at $\tau \geq 0.01$ kW.

Because τ was examined after review, these values quantify sensitivity and do not replace the zero-margin primary rule.

The result illustrates the benefit of coupling evolutionary dynamics with an exact certificate. A report based only on the 98.36%/1.15% mixture could imply that a nontrivial equilibrium had been found. The linear program shows instead that the equilibrium is the persistence vertex. The residual non-persistence weights are numerical mass remaining in the finite-run ergodic approximation. The guard reaches the same deployment decision from the actual deterministic forecasts.

6.6 Aggregation Ablations

Table 10 separates the effects of simple averaging, attention and game weighting. The May-MAE-optimal convex ensemble has the lowest test RMSE (6.7356 kW) and highest R^2 (0.98004) among the reported variants. The equal-weight mean has an RMSE of 6.7462 kW and R^2 of 0.97997, but its MAE is 2.6487—0.8401 kW above persistence. It is not an eligible deployment strategy and its test behavior did not influence the guard.

Table 10: Aggregation and validation-optimized convex-ensemble ablations on the held-out evaluation block. Source: authors' calculations from locked predictions.

Variant	MAE (kW)	RMSE (kW)	sMAPE (%)	R^2	Bias (kW)	Status
Persistence/final guarded strategy	1.8087	6.8910	1.9550	0.97910	−0.1505	Deployed vertex selected in May
Equal-weight mean of six experts	2.6487	6.7462	2.9162	0.97997	1.1771	Analytical ablation; never eligible for deployment
Five-expert game without Transformer	1.8114	6.8887	2.1576	0.97912	−0.1374	Reduced-pool guard rejects; exact LP is persistence
Six-expert evolutionary candidate	1.8120	6.8883	2.1582	0.97912	−0.1337	Rejected on May before test
May-MAE-optimal convex ensemble	2.2737	6.7356	2.5686	0.98004	0.4267	May average improves; worst-day guard rejects

Removing the Transformer from the game produces a candidate with weights 99.3472% persistence, 0.0952% ridge, 0.0400% histogram boosting, 0.0795% GRU and 0.4381% LSTM. Its exact minimax solution remains pure persistence. On May its overall MAE is 7.312014 kW, marginally below persistence's 7.312105 kW, but its worst-day MAE is 32.241574 vs. 32.237504 kW. The reduced-pool guard therefore rejects it. On testing, its MAE is 1.8114 and RMSE 6.8887 kW. The full game adds a 1.1450% Transformer share and shifts those values only to 1.8120 and 6.8883 kW. Thus, attention contributes a minute tail-error reduction within the finite candidate, but not enough to improve the primary MAE or change the exact/guarded selection.

The May-MAE-optimal convex ensemble assigns weights 0.668795 to persistence, 0.210709 to ridge, 0.077093 to HistGB and 0.043403 to LSTM, with zero GRU and Transformer weight. It lowers May average MAE from 7.312105 to 7.242327 kW, but raises the worst-day MAE from 32.237504 to 32.945520 kW and therefore fails the screening guard. Its held-out MAE/RMSE is 2.2737/6.7356 kW: lower RMSE than equal weighting, but higher MAE than persistence. Average-loss optimization does not overturn the primary MAE-and-worst-day decision.

6.7 Conformal Interval Diagnostics

The June calendar-day maximum score for the final guarded strategy is $\hat{q} = 151.3000$ kW, giving constant total interval width 302.6000 kW. The Transformer width is 302.4654 kW, and the game-candidate width is 302.6049 kW. All three cover 100% of the 1999 test points and 100% of the 88 available test calendar days (Table 11 and Fig. 7), above the nominal 90% level.

Table 11: June-calibrated 90% calendar-day maximum-residual intervals on July–September. Source: authors' calculations from HPP flash data [32].

Forecast	Calibration Days	Rank	$\hat{q}$ (kW)	Mean Width (kW)	Point Coverage	Simultaneous Day Coverage	Test Days
Persistence/ guarded	30	28	151.3000	302.6000	100%	100%	88
Transformer	30	28	151.2327	302.4654	100%	100%	88
Full game candidate	30	28	151.3025	302.6049	100%	100%	88

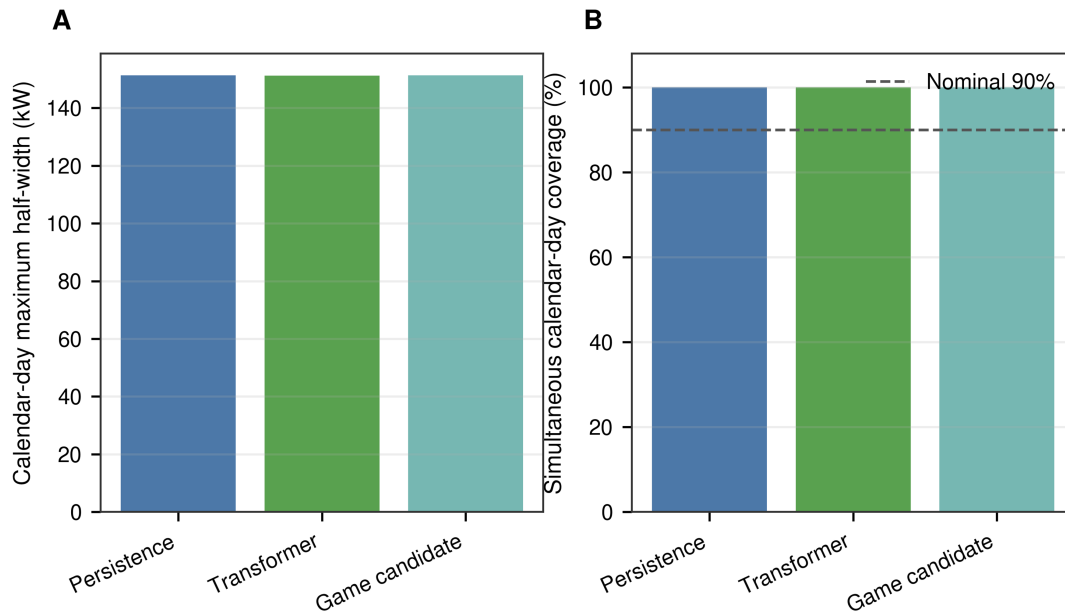


Figure 7: Calendar-day maximum-residual conformal diagnostics. Panel A compares June-calibrated interval half-widths; Panel B compares simultaneous calendar-day coverage over the held-out block with the nominal 90% level. The 100% observed coverage is accompanied by very wide intervals. Source: authors' calculations from HPP flash data [32].

The 302.6000 kW persistence interval width equals 270.54% of mean held-out power and 96.52% of the observed held-out power range. No physical clipping is applied, so lower limits may be negative. The 100% point and day coverage therefore reflects conservative calibration rather than useful sharpness. These intervals are not operationally informative at the present scale. Future work should evaluate locally scaled or adaptive time-series conformal methods with multiple calibration seasons and operator-certified bounds.

6.8 Error by Realized Ramp Magnitude

As summarized in Table 12 and illustrated in Fig. 8, held-out origins are grouped *post hoc* by realized absolute fifteen-minute power change, which is unavailable at forecast time. Persistence/Transformer MAE is 0.2742/0.7327 kW for ≤ 1 kW changes, 2.1512/2.2323 for 1–5 kW, 9.1143/9.0427 for 5–20 kW and 48.3857/47.6853 for >20 kW. Paired calendar-day-cluster bootstrap intervals for Transformer-minus-persistence Δ MAE are [0.4199, 0.4965], [0.0065, 0.1565], [−0.1896, 0.0434] and [−1.0836, −0.3111] kW, respectively. The >20 kW estimate is based on only 28 origins across 18 days; it is exploratory evidence of a different error pattern, not a general tail-superiority claim.

Table 12: Descriptive held-out MAE by realized fifteen-minute active-power change. Source: authors' calculations from HPP flash data [32].

Realized Change	<i>n</i>	Share	Persistence MAE (kW)	Transformer MAE (kW)	Game Candidate MAE (kW)
≤ 1 kW	1444	72.24%	0.2742	0.7327	0.2785
(1, 5] kW	422	21.11%	2.1512	2.2323	2.1541
(5, 20] kW	105	5.25%	9.1143	9.0427	9.1125
> 20 kW	28	1.40%	48.3857	47.6853	48.3627

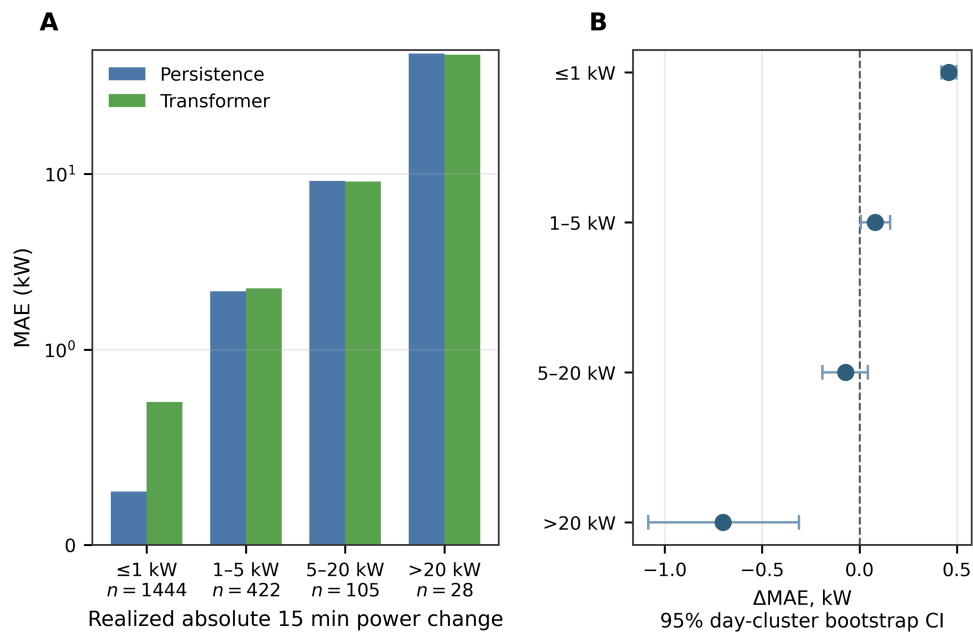


Figure 8: Held-out ramp-stratified diagnostics. Panel A compares persistence and transformer MAE on a logarithmic scale; Panel B shows Transformer-minus-persistence Δ MAE with 95% paired day-cluster bootstrap intervals (5000 replications). Positive differences favor persistence. Groups use realized targets and are retrospective, not forecast-time inputs. Source: authors' calculations from locked predictions.

6.9 Post Hoc Performance by Daily Output Mode

As summarized in Table 13, held-out calendar days are split descriptively at the median daily mean actual power (95.924702 kW), yielding 44 low-output days (1017 origins; mean 77.305 kW) and 44 high-output days (982 origins; mean 147.624 kW). Persistence has the lowest MAE in both modes (1.5347 and 2.0924 kW). The Transformer has higher MAE (1.9847 and 2.3078 kW) but slightly lower RMSE (6.8417 vs. 6.8844 kW; 6.8762 vs. 6.8979 kW). Because the threshold uses held-out targets, this is a *post hoc* diagnostic, not a forecast-time gate or selection rule.

Table 13: Held-out accuracy by *post hoc* daily output mode. The split threshold is the median held-out calendar-day mean actual power. Source: authors' calculations from locked predictions.

Forecast	Low-Output MAE (kW)	Low-Output RMSE (kW)	High-Output MAE (kW)	High-Output RMSE (kW)
Persistence	1.5347	6.8844	2.0924	6.8979
Ridge	3.6175	7.6867	6.8356	9.0972
HistGB	6.5088	9.3167	4.0256	8.0082
GRU	7.0662	9.1502	5.5446	8.2769
LSTM	3.7207	7.4110	2.8438	6.9701
Transformer	1.9847	6.8417	2.3078	6.8762

6.10 Supplementary One-Day Rolling-Origin and No-Attention Results

The byte-verified 7 April file contains 1440 valid 104-field records from 00:00:03 to 23:59:04, with no duplicate timestamps or missing minute buckets and 59–61 s spacing. All 12 local-load channels are missing throughout; all other fields and the total-active-power target are complete. The historical layout lacks the two flow fields present in the current 106-field live schema. A generic conspicuous-value screen flags no rows, but this does not exclude undocumented operator-specific codes.

Across all 1396 admissible one-minute origins, persistence MAE/RMSE is 4.8898/11.8245 kW. The 24 HH:30 origins yield a persistence MAE/RMSE of 3.6083/4.4597 kW, the HH:30 MAE ranks seventh-lowest among the 60 phase-specific MAEs, which range from 3.1826 to 11.8522 kW. Thus, HH:30 is relatively easy on this date and within-hour phase materially affects the reported error.

As shown in Table 14, the FFN-only control on the 221-origin held-out block reduces MAE by 4.3632 kW relative to the Transformer (95% moving-block-bootstrap interval $[-7.4641, -1.6992]$ kW), whereas its 0.1033 kW MAE difference from persistence has an interval crossing zero $[-1.0430, 0.7455]$ kW). The full ensemble attains the lowest MAE, 2.8173 kW. This one-date ranking differs from the annual HH:30 result and does not support an attention-specific advantage. It demonstrates why the annual claim remains schedule-, date- and configuration-specific.

6.11 Post Hoc Encoding-Sentinel Sensitivity

Some source-native channels contain isolated values that resemble encoding sentinels, but no authoritative plant dictionary defines valid ranges. The primary analysis therefore retains every finite numeric value and does not tune a cleaning rule on test errors. As a transparent *post hoc* sensitivity only, an origin is flagged when any value in its thirty-minute input satisfies at least one conspicuous rule: absolute unit active power above 1000, absolute power factor above 10, absolute level or setlvl above 10, or absolute selected temp above 1000, all in the source-native numeric scales. Three of 1999 test origins are flagged. After their exclusion ($n = 1996$), persistence MAE/RMSE is 1.8006/6.8871 kW, Transformer MAE/RMSE is 2.1361/6.8558 kW, and the full candidate is 1.8039/6.8844 kW. The ordering and substantive conclusion are unchanged: persistence retains the lowest MAE, while the Transformer retains a slightly lower RMSE. Because the rule is

retrospective and metadata are incomplete, these filtered values are not substituted for the primary results. This test-only exclusion does not evaluate whether analogous undocumented codes affected model fitting or development-block selection.

Table 14: Supplementary one-day rolling-origin accuracy on 7 April 2021 (221 held-out consecutive one-minute origins). Source: authors' calculations from the archived HPP flash daily file [32].

Forecast	n	MAE (kW)	RMSE (kW)
Persistence	221	4.5109	5.3157
Ridge	221	3.1733	4.0806
HistGB	221	22.8248	26.0165
GRU	221	4.6050	5.8526
LSTM	221	5.9942	7.2007
Transformer	221	8.9774	10.5750
Game without Transformer	221	2.8241	3.4521
Full evolutionary ensemble	221	2.8173	3.4557
FFN-only no-attention	221	4.6142	5.8146

7 Discussion

7.1 Answers to the Research Questions

For RQ1, the compact Transformer is the best learned expert and has the lowest RMSE among individual experts, but it does not beat persistence on primary MAE, percentage errors or the multiplicity-adjusted squared-error test. Its forecasts have a different error pattern: they are worse in the two common low-ramp groups, indistinguishable within uncertainty for 5–20 kW changes and better in the sparse >20 kW group. This does not isolate attention as the causal mechanism, and the evidence is insufficient for stand-alone deployment under an MAE-centered objective.

For RQ2, the exact game and the zero-margin screening rule select pure persistence. The finite candidate is close to, but not equal to, the minimax solution. An independent late-May sensitivity still fails on worst-day MAE, while tolerances of 0.05 kW or more would accept it; those tolerances were not pre-specified. The result is therefore a conservative development-selection decision, not a formal safety guarantee.

For RQ3, dependence-aware uncertainty weakens any claim based on the 0.0323 kW Transformer RMSE difference. Its paired RMSE interval crosses zero and its Holm-adjusted squared-error p -value is 0.7012. The conformal intervals cover every observed test day but span 270.54% of mean held-out power, so they are not sharp. The defensible conclusion is limited to the evaluated plant, block, horizon, phase and MAE-centered rule.

7.2 Why Persistence Dominates this Test

The data partition intentionally exposes a seasonal shift. Spring tuning contains high output and larger mean changes. Late-summer testing contains low output and 72.24% of origins move no more than 1 kW in fifteen minutes. A residual learner has asymmetric opportunity under this composition. During a stable observation, the optimal residual is near zero and any learned bias adds error. During a rare ramp, a learned correction may remove tens of kW. MAE weights every kW linearly, so the many small deteriorations outweigh modest gains on 133 origins above 5 kW. RMSE squares the errors and therefore reflects those tail gains more strongly.

The behavior is not a reason to replace chronological testing with stratified random sampling. The seasonal composition is part of the operational forecasting problem. Instead, it suggests that a future selector may need observable, forecast-time regime information, such as a validated controller-state transition indicator, to decide when a tail-sensitive expert should be activated. The realized ramp classes in [Section 6.8](#) cannot serve this role because they use future active power.

7.3 Theoretical and Practical Role of the Transformer

Within a thirty-minute window, self-attention can relate the latest electrical state directly to any preceding token [11]. In this fixed configuration, the Transformer yields a slightly lower RMSE and lower MAE only in the rare >20 kW realized-ramp group, while persistence dominates common stable conditions. These observations establish forecast diversity for the fitted model, not a causal benefit of attention.

However, the record does not establish that attention itself causes the improvement. Only one compact architecture, one seed and one plant-year are evaluated. The Transformer, GRU and LSTM differ in internal dynamics even though training settings and parameter scale are controlled. A causal architectural claim would require repeated seeds, hyperparameter budgets matched by compute, multi-year data and preferably multiple plants. The present contribution is narrower, as it supplies an explicit attention implementation, scale-matched local comparators under unequal but disclosed tuning budgets, and a selection rule that can reject it. For the annual experiment, a no-attention/feed-forward control, repeated seeds and compute-matched tuning remain unresolved. The supplementary one-day control yields lower MAE for the FFN-only model than for the Transformer, but this single-date result cannot establish an annual causal attention effect.

7.4 Value of Exact Certification and Guarding

The entropic multiplicative-weights scheme instantiates the evolutionary-game layer as a discrete-time exponential-weights analogue of two-population replicator learning and provides an interpretable trajectory of expert shares. Yet an iteration count is not a proof of equilibrium. Here the ergodically averaged candidate retains 1.64% total mass outside persistence and its worst expected daily loss is only 0.013477 kW above the exact value. Without an exact check, this might be called a mixed minimax solution. Linear programming and the pure-saddle argument show that the true solution is pure. The example demonstrates why computational game studies should report a saddle residual or exact certificate whenever the finite game is small enough.

The deterministic guard addresses another subtlety. The game matrix contains the daily losses of randomized expert selection, whereas deployment averages forecast values. Convexity links the two, but their losses are not identical. Testing the actual candidate on the same authorized May block avoids relying on the bound. Requiring both overall and worst-day non-inferiority also prevents an average gain from masking a severe day. These controls are inexpensive and applicable to other renewable-energy expert pools. The post-review 1–21/22–31 May split confirms that an independent guard can change the overall-loss comparison while leaving the worst-day failure; engineering tolerance must be chosen prospectively.

7.5 Comparison with Prior Work

Prior hydropower studies demonstrate that nonlinear and hybrid models can be effective [3–7], and hydrological sequence studies show the expressive value of recurrent networks [8–10]. The present result does not contradict them because it concerns a different target, minute-scale plant active power, a different sample size and a strict seasonal holdout. Transformer architectures [12–16,18] and the simple linear baseline of Zeng et al. [17] also show that rankings vary with representation and horizon. In particular, the competitiveness of the linear model reported by Zeng et al. [17] supports the broader principle that strong simple baselines are indispensable.

Broader renewable-energy benchmarks reinforce the reviewer's point that one architecture per family is not exhaustive. Rathnayake et al. compared 24 short-term wind-power models [39], while a cascaded ANFIS hydropower study compared several AI/ML alternatives [40]. Their plants, inputs, sampling and horizons are not numerically comparable with this SCADA-only fifteen-minute task, but they motivate future compute-matched multi-model and multi-seed benchmarking.

The game component extends ordinary forecast combination [19,23] by optimizing a worst-calendar-day payoff rather than an average alone, and it connects multiplicative weighting [20,22] with evolutionary dynamics [21]. Its empirical contribution is not a headline accuracy gain but an auditable rejection. The uncertainty layer follows the same conservative stance: moving blocks [27] and corrected DM tests [25,26] account for dependence, while conformal methods [28,29] are reported with an explicit warning about temporal nonexchangeability.

7.6 Operational Implications

For the evaluated Bulgarian plant, July–September 2021 block, fifteen-minute horizon, HH:30→HH:45 hourly phase and MAE-centered decision rule, persistence is the selected forecast. This result must not be generalized to other plants, phases, horizons or cost functions. A cost-weighted game could rationally accept higher MAE to reduce costly ramps, but it would require a prospectively specified utility and new validation.

The current pipeline remains valuable even when it selects persistence. It provides automated file auditing, continuity-safe sample construction, a consistent expert interface, exact and approximate game solvers, an auditable screening guard, and uncertainty diagnostics. As more years become available, the same workflow can test whether the Transformer share increases in years with more frequent transitions or whether regime-conditioned experts become supportable. The released package reproduces analyses from locked predictions but not end-to-end model training from the restricted SCADA archive.

8 Limitations and Threats to Validity

The findings should be interpreted within the following boundaries.

1. One plant and one acquisition year. The interval spans 1 October 2020–30 September 2021, but it is still one plant-year and four daily files are absent. Interannual hydrology, maintenance and dispatch variability are not represented. External validity requires additional years and sites.
2. Four complete dates and boundary-day coverage are unavailable. The largest 405,697-s interruption includes the four entirely unavailable dates, 17–20 July, and partial-day loss on 16 and 21 July. Windows cannot cross the interruption, but performance throughout the missing episode is unknown.
3. Source metadata and dimensional provenance. Table 3 provides the most complete channel dictionary recoverable from the public TXT header, HTML pages, and JavaScript field mappings. The wider portal identifies open as opening (%), press as a pressure-related signal expressed in metres, and temp1–temp5 as upper-bearing, three winding, and lower-bearing temperatures (°C). For level and setlvl, the current plotting module states millimetres, whereas the historical daily-file header gives no unit and the observed codes do not establish the historical scale or datum unambiguously. Because the current web assets may post-date the study period, Table 3 is a public-source reconstruction rather than an operator-certified historical dictionary. Operator confirmation of storage units, scale/offset, sensor locations, reference datums, sign conventions, valid ranges, sentinel codes, calibration records, time-zone rules, and schema-effective dates remains necessary before physics-informed dimensional constraints are imposed.
4. The files contain local-looking clock fields without a zone or offset. The 25 October 2020 file has 1500 records, including 120 records labelled in local hour 03 and 37 repeated exact timestamp labels; the

- 28 March 2021 file has 1380 records and no local hour 03. This pattern is consistent with daylight-saving clock changes, but that interpretation cannot be proved from the source metadata. Retaining the first exact timestamp and then the first record per minute deterministically collapses part of the repeated autumn hour, so any distinct second-occurrence state at the same local label is unavailable to the model.
5. Schema and sentinel uncertainty. Restricting the analysis to the common 70-field core avoids relying on appended channels that are absent during part of the year, and incomplete rows are rejected. Nevertheless, undocumented numeric sentinel values may remain because no authoritative plant schema defines their valid ranges. Aggressive manual filtering would risk deleting legitimate operating states. The *post hoc* three-origin exclusion in [Section 6.11](#) leaves the ranking unchanged but cannot replace a plant-approved data dictionary. The *post hoc* exclusion assesses only flagged evaluation inputs; it does not quantify possible fitting-period contamination, which would require operator-certified validity rules and complete retraining. A full-timeline cleaning rule and confirmed-channel retraining experiment cannot be defined responsibly without an operator-certified schema and therefore remains unresolved.
 6. Unobserved future dispatch. The inputs include recent measured control and electrical channels but no future unit schedule, dispatch instruction or planned set-point trajectory. If an operator changes the requested output after the forecast origin, the resulting target is not identifiable from the supplied history alone. The study therefore evaluates statistical predictability from SCADA history, not perfect anticipation of exogenous decisions.
 7. Feature-pool choice. Twenty channels were selected from the common 70-field core to cover output, water/control, dynamics, condition and clock information. This choice prevents seasonal schema leakage but was not subjected to a nested feature-selection ablation. Some omitted common-core fields could contain useful state information.
 8. Hourly origins. The annual experiment retains one-minute histories but issues forecasts only at HH:30 for HH:45. The supplementary 7 April analysis evaluates all one-minute origins and confirms phase heterogeneity, but it is one day with within-day fitting and testing. The annual conclusion therefore remains restricted to HH:30; a full plant-year rolling retraining is still required.
 9. Single forecast horizon. Fifteen minutes was fixed. Persistence strength, model ranking and game weights can differ at 5, 30 or 60 min.
 10. Single neural seed and limited tuning. One deterministic seed and one compact configuration per neural family are evaluated. Parameter counts and core optimizer settings are comparable, but ridge alone receives an explicit penalty grid and the neural/tree search budgets are not compute-matched. Training variance and architecture uncertainty are therefore unquantified. No-attention and repeated-seed controls are not estimated for the annual experiment. The supplementary one-day FFN-only control outperforms the one-day Transformer but not persistence within paired uncertainty; it cannot establish an annual causal attention effect.
 11. Residual clipping. Every learned residual is capped at fitting-block 0.5% and 99.5% quantiles. This common extrapolation regularizer limits unstable corrections but may suppress a legitimate future ramp outside the fitting distribution. The quantile level was fixed rather than optimized through a nested sensitivity analysis. Unclipped and alternative-quantile forecasts cannot be recovered from the locked outputs and would require retraining; tail conclusions are correspondingly exploratory.
 12. Seasonal nonstationarity. Calibration in June and testing in July–September are chronologically correct but not exchangeable. The 100% conformal coverage is descriptive and conservative, not a formal guarantee under drift.
 13. Metric and guard dependence. MAE favors persistence, whereas RMSE is lowest for the May-MAE-optimal convex ensemble among the reported aggregation ablations and, among individual experts, for the Transformer. Deployment depends on the loss function. The zero-margin guard is deliberately strict

and its two numerical failures are below the source's printed power precision; an operational deployment would need an engineering tolerance and asymmetric cost function fixed before evaluation. Moreover, the guard is evaluated on the same May block used to construct the game weights and is therefore not an independent validation or formal safety guarantee. Post-review sensitivity shows acceptance at $\tau \geq 0.01$ kW in-block but only at $\tau \geq 0.05$ kW on the independent late-May guard; these retrospective thresholds are not deployment recommendations.

14. *Post hoc* strata. Realized-ramp groups and the 95.924702 kW daily-output threshold depend on held-out targets and cannot be used online. Although the >20 kW day-cluster interval favors the Transformer, it contains only 28 origins across 18 days. Neither analysis establishes a pre-specified regime gate.
15. Power rather than energy. The target is instantaneous active power in kW. The article does not estimate produced energy, revenue or reserve cost; those require integration and economic assumptions.

These limitations narrow the claims but do not invalidate the internal comparison. Every expert receives identical accepted windows, the partitions are chronological, and the July–September block does not enter fitting, game weighting, guard calculation or interval calibration. The rejection result is reported rather than optimized away.

9 Conclusions

This study presents a detailed, auditable hydrological-year evaluation of fifteen-minute-ahead hydropower active-power forecasting. The archive audit yields 508,234 unique timestamps and 8298 continuity-safe hourly-origin windows from 361 daily files. A common persistence-residual formulation supports ridge, histogram gradient boosting, GRU, LSTM and compact Transformer experts. Calendar-day entropic multiplicative-weights dynamics, an exact zero-sum linear program and a two-part development screening guard provide an explicit bridge from evolutionary game theory to operational model selection.

On the 1999-origin July–September test, persistence achieves the lowest MAE of 1.8087 kW. The Transformer lowers RMSE slightly to 6.8587 kW but raises MAE to 2.1434 kW; its squared-error difference is not significant. The evolutionary candidate allocates 98.36% to persistence yet is marginally worse on both May guard criteria. The exact minimax strategy and final guarded strategy are pure persistence. Equal averaging produces a lower RMSE but a substantially higher MAE and is only an ablation. Block-bootstrap and corrected-test results do not support a general superiority claim, while calendar-day conformal intervals are fully covering but very wide. A May-MAE-optimal convex ensemble lowers test RMSE to 6.7356 kW but raises MAE to 2.2737 kW and fails the May worst-day guard, so it does not change the decision.

The principal conclusion is methodological: complex forecasts and game aggregation are useful when the protocol permits them to lose. Here the Transformer exhibits a different, exploratory large-ramp error pattern and slightly lower RMSE without a statistically supported overall gain; exact certification and development screening decline deployment. Multi-year, multi-plant, multi-phase, repeated-seed and compute-matched experiments are required before broader claims. A supplementary one-day rolling-origin experiment covers all within-hour phases and produces a different ranking, reinforcing rather than removing the annual schedule limitation.

Acknowledgement: None.

Funding Statement: This work was supported by the European Regional Development Fund under the “Research Innovation and Digitization for Smart Transformation” program 2021–2027 under Project BG16RFPR002-1.014-0006 “National Centre of Excellence Mechatronics and Clean Technologies”, and the APC was funded by Project BG16RFPR002-1.014-0006.

Author Contributions: Stoyan Popov, Silvia Baeva and Nikolay Hinov were involved in the full process of producing this paper, including conceptualization, methodology, modeling, validation, visualization, and preparing the manuscript. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The raw annual SCADA files were obtained from the third-party HPP Flash portal [32] during the acquisition period, and current access to historical files is controlled by the portal operator. The 7 April 2021 audit used an archived source copy whose filename, byte count and SHA-256 checksum match the annual source manifest; the raw file is not redistributed. The complete annual raw archive, processed multivariate table and trained weights cannot be distributed. Supplementary Material S1 contains source-manifest metadata, partition identifiers, locked May, June and July–September targets and forecasts, derived one-day rolling-origin outputs, FFN-only control code and results, final and post-review analysis outputs, software specifications and validation reports. It supports verification from locked/derived outputs but not end-to-end annual reconstruction or retraining.

Ethics Approval: This study used operational machine data and did not involve identifiable human data or animals.

Conflicts of Interest: Given his role as Guest Editor of this Special Issue, Nikolay Hinov had no involvement in the peer review of this article and had no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to another journal editor. The authors declare no other conflicts of interest.

Supplementary Materials: Supplementary Material S1 contains fixed annual partition identifiers and locked predictions; performance, uncertainty, game-certificate, sensitivity and data-quality outputs; source-manifest metadata; the derived one-day rolling-origin audit and FFN-only control; software-environment and validation reports; analysis and figure-generation code; and publication figures. Raw SCADA files, the complete processed annual modeling table, access credentials and trained model weights are not included. The supplementary material is available online at <https://www.techscience.com/doi/10.32604/cmcs.2026.089814/sl>.

Abbreviations

Abbreviation	Definition
SCADA	Supervisory control and data acquisition
HistGB	Histogram gradient boosting
GRU	Gated recurrent unit
LSTM	Long short-term memory
FFN	Feed-forward network
MAE	Mean absolute error
MSE	Mean squared error
RMSE	Root-mean-square error
MAPE	Mean absolute percentage error
sMAPE	Symmetric mean absolute percentage error
WAPE	Weighted absolute percentage error
DM–HLN	Diebold–Mariano test with Harvey–Leybourne–Newbold correction
HAC	Heteroskedasticity-and-autocorrelation consistent
LP	Linear programming
PF	Power factor

References

1. Kishor N, Saini RP, Singh SP. A review on hydropower plant models and control. *Renew Sustain Energy Rev*. 2007;11(5):776–96. doi:10.1016/j.rser.2005.06.003.
2. Márquez JL, Molina MG, Pacas JM. Dynamic modeling, simulation and control design of an advanced micro-hydro power plant for distributed generation applications. *Int J Hydrogen Energy*. 2010;35(11):5772–7. doi:10.1016/j.ijhydene.2010.02.100.

3. Barzola-Monteses J, Gómez-Romero J, Espinoza-Andaluz M, Fajardo W. Hydropower production prediction using artificial neural networks: an Ecuadorian application case. *Neural Comput Appl.* 2022;34(16):13253–66. doi:10.1007/s00521-021-06746-5.
4. Sattar Hanoon M, Najah Ahmed A, Razzaq A, Oudah AY, Alkhayyat A, Feng Huang Y, et al. Prediction of hydropower generation via machine learning algorithms at three Gorges Dam, China. *Ain Shams Eng J.* 2023;14(4):101919. doi:10.1016/j.asej.2022.101919.
5. Dehghani M, Riahi-Madvar H, Hooshyaripor F, Mosavi A, Shamshirband S, Zavadskas EK, et al. Prediction of hydropower generation using grey wolf optimization adaptive neuro-fuzzy inference system. *Energies.* 2019;12(2):289. doi:10.3390/en12020289.
6. Tebong NK, Simo T, Takougang AN. Two-level deep learning ensemble model for forecasting hydroelectricity production. *Energy Rep.* 2023;10(5):2793–803. doi:10.1016/j.egy.2023.09.091.
7. Zhang G, Li H, Wang L, Wang W, Guo J, Qin H, et al. Research on medium- and long-term hydropower generation forecasting method based on LSTM and transformer. *Energies.* 2024;17(22):5707. doi:10.3390/en17225707.
8. Kratzert F, Klotz D, Brenner C, Schulz K, Herrnegger M. Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrol Earth Syst Sci.* 2018;22(11):6005–22. doi:10.5194/hess-22-6005-2018.
9. Gauch M, Kratzert F, Klotz D, Nearing G, Lin J, Hochreiter S. Rainfall–runoff prediction at multiple timescales with a single long short-term memory network. *Hydrol Earth Syst Sci.* 2021;25(4):2045–62. doi:10.5194/hess-25-2045-2021.
10. Hunt KMR, Matthews GR, Pappenberger F, Prudhomme C. Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States. *Hydrol Earth Syst Sci.* 2022;26(21):5449–72. doi:10.5194/hess-26-5449-2022.
11. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*; Dec 4–9; Long Beach, CA, USA.
12. Lim B, Arik SÖ, Loeff N, Pfister T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *Int J Forecast.* 2021;37(4):1748–64. doi:10.1016/j.ijforecast.2021.03.012.
13. Zhou H, Zhang S, Peng J, Zhang S, Li J, Xiong H, et al. Informer: beyond efficient transformer for long sequence time-series forecasting. *Proc AAAI Conf Artif Intell.* 2021;35(12):11106–15. doi:10.1609/aaai.v35i12.17325.
14. Wu H, Xu J, Wang J, Long M. Autoformer: decomposition transformers with auto-correlation for long-term series forecasting. *Adv Neural Inf Process Syst.* 2021;34:22419–30.
15. Zhou T, Ma Z, Wang X, Sun L, Jin R, Wen Q. Fedformer: frequency enhanced decomposed transformer for long-term series forecasting. In: *AI for time series*. Boca Raton, FL, USA: CRC Press; 2026. p. 10–34. doi:10.1201/9781003612742-2.
16. Nie Y, Nguyen NH, Sinthong P, Kalagnanam J. A time series is worth 64 words: long-term forecasting with transformers. In: *Proceedings of the Eleventh International Conference on Learning Representations*; 2023 May 1–5; Kigali, Rwanda.
17. Zeng A, Chen M, Zhang L, Xu Q. Are transformers effective for time series forecasting? *Proc AAAI Conf Artif Intell.* 2023;37(9):11121–8. doi:10.1609/aaai.v37i9.26317.
18. Liu Y, Hu T, Zhang H, Wu H, Wang S, Ma L, et al. iTransformer: inverted transformers are effective for time series forecasting. In: *Proceedings of the Twelfth International Conference on Learning Representations*; 2024 May 7–11; Vienna, Austria.
19. Bates JM, Granger CWJ. The combination of forecasts. *J Oper Res Soc.* 1969;20(4):451–68. doi:10.2307/3008764.
20. Littlestone N, Warmuth MK. The weighted majority algorithm. *Inf Comput.* 1994;108(2):212–61. doi:10.1006/inco.1994.1009.
21. Taylor PD, Jonker LB. Evolutionary stable strategies and game dynamics. *Math Biosci.* 1978;40(1–2):145–56. doi:10.1016/0025-5564(78)90077-9.
22. Arora S, Hazan E, Kale S. The multiplicative weights update method: a meta-algorithm and applications. *Theory Comput.* 2012;8(6):121–64. doi:10.4086/toc.2012.v008a006.
23. Devaine M, Gaillard P, Goude Y, Stoltz G. Forecasting electricity consumption by aggregating specialized experts. *Mach Learn.* 2013;90(2):231–60. doi:10.1007/s10994-012-5314-7.

24. Klotz D, Kratzert F, Gauch M, Keefe Sampson A, Brandstetter J, Klambauer G, et al. Uncertainty estimation with deep learning for rainfall–runoff modeling. *Hydrol Earth Syst Sci.* 2022;26(6):1673–93. doi:10.5194/hess-26-1673-2022.
25. Diebold FX, Mariano RS. Comparing predictive accuracy. *J Bus Econ Stat Am Stat Assoc.* 1995;13(3):253–63. doi:10.1080/07350015.1995.10524599.
26. Harvey D, Leybourne S, Newbold P. Testing the equality of prediction mean squared errors. *Int J Forecast.* 1997;13(2):281–91. doi:10.1016/S0169-2070(96)00719-4.
27. Kunsch HR. The jackknife and the bootstrap for general stationary observations. *Ann Statist.* 1989;17(3):1217–41. doi:10.1214/aos/1176347265.
28. Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L. Distribution-free predictive inference for regression. *J Am Stat Assoc.* 2018;113(523):1094–111. doi:10.1080/01621459.2017.1307116.
29. Xu C, Xie Y. Conformal prediction interval for dynamic time-series. In: *Proceedings of the 38th International Conference on Neural Information Processing Systems*; 2024 Dec 10–15; Vancouver, BC, Canada.
30. Francisco R, Matos JP, Marinheiro R, Lopes N, Portela MM, Barros P. Application of temporal fusion transformers to run-of-the-river hydropower scheduling. *Hydrology.* 2025;12(4):81. doi:10.3390/hydrology12040081.
31. Kåge L, Milic V, Andersson M, Wallén M, Trépout S. Hourly hydropower production forecasting with machine learning: a case study in linköping, Sweden. In: *Proceedings of the 10th World Congress on New Technologies*; 2024 Aug 25–27; Barcelona, Spain. doi:10.11159/icert24.102.
32. Solar Dispatcher. HPP flash supervisory-control portal, hydrogenerator metadata pages and historical daily operational logs [Internet]. [cited 2026 Aug 14]. Available from: https://www.solar-dispatcher.com/scada/HPP_Flash/.
33. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics.* 1970;12(1):55–67. doi:10.1080/00401706.1970.10488634.
34. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Statist.* 2001;29(5):1189–232. doi:10.1214/aos/1013203451.
35. Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; 2014 Oct 25–29; Doha, Qatar.
36. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput.* 1997;9(8):1735–80. doi:10.1162/neco.1997.9.8.1735.
37. Newey WK, West KD. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica.* 1987;55(3):703. doi:10.2307/1913610.
38. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat.* 1979;6(2):65–70.
39. Rathnayake N, Jayasinghe J, Semasinghe R, Rathnayake U. Predicting short-term wind power generation at musalpetti wind farm: model development and analysis. *Comput Model Eng Sci.* 2025;143(2):2287–305. doi:10.32604/cmesci.2025.064464.
40. Rathnayake N, Rathnayake U, Dang TL, Hoshino Y. A cascaded adaptive network-based fuzzy inference system for hydropower forecasting. *Sensors.* 2022;22(8):2905. doi:10.3390/s22082905.