energies

MDPI

Article

Cross-Renewable Generalization of PEM Electrolyzer Hydrogen-Flow Forecasting: Multi-Horizon Evaluation and Physics Anchoring

Xianhuan Wu [1], Cenglin Yao [2,*] and Lesheng Liu [3]

[1] School of Intelligent Transportation, Wuhan Business University, Wuhan 430056, China
[2] School of Artificial Intelligence and Big Data, Wuhan Business University, Wuhan 430056, China
[3] Industry-Education Integration Center, Wuhan Business University, Wuhan 430056, China
* Correspondence: lhua79966@gmail.com

Abstract

Forecasting renewable-powered proton-exchange-membrane (PEM) electrolysis across unfamiliar power trajectories requires evaluation at the renewable-source level rather than within familiar profile families. We evaluated experimentally recorded hydrogen flow from one nominal 1.25 MW PEM-electrolyzer platform under four renewable-source operating domains—Simulated Wind, Simulated Wave, Historical Solar Photovoltaic, and Simulated Tidal—using leave-one-renewable-source-out validation, a 60 s process history, and direct 1, 5, 10, and 30 s horizons. Ridge, HistGradientBoosting, and LightGBM transferred usefully at short horizons but degraded as the horizon increased; Wave had negative R^2 for all three models at 30 s. Source balancing and worst-source selection produced only local gains, and model-disagreement error awareness weakened with horizon. For one prespecified five-epoch Conv1d masked-reconstruction design and linear probe, self-supervised embeddings underperformed handcrafted descriptors in source-macro forecasting, without establishing a general limitation of self-supervision. A current-command-based operational gray-box anchor (F1) outperformed its residual-augmented version (G1) at all four horizons; G1 improved on the prespecified Ridge comparator but was less accurate than F1. Equal-weight source-macro errors describe these four held-out operating domains, not a population of systems or sites. The findings do not establish transfer across manufacturers, stacks, capacities, or sites, and no electrochemical parameters were uniquely identified.

Keywords: PEM water electrolysis; green hydrogen; renewable energy integration; cross-domain generalization; multi-horizon forecasting; physics-informed machine learning; gray-box modeling; self-supervised learning; selective prediction

Check for updates

Academic Editor: Felix Barreras

Received: 29 September 2026
Revised: 23 October 2026
Accepted: 27 October 2026
Published: 30 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Renewable-powered water electrolysis is expected to operate under increasingly variable electrical conditions as wind and solar generation expand. Proton-exchange-membrane (PEM) electrolyzers are attractive in this setting because of their fast load-following capability, but dynamic operation couples electrical input with thermal, pressure, gas-production, and efficiency responses over different time scales. Recent reviews emphasize that dynamic PEM operation under fluctuating power must be evaluated beyond steady-state behavior [1,2]. Experimental work under fluctuating load likewise reports pronounced transient responses when operating trajectories change rapidly or repeatedly [3].

This complexity has motivated a growing use of statistical, machine-learning, and hybrid physical–data models for PEM electrolysis. Recent work has shown that response-surface and shallow neural-network models can predict hydrogen-production variables within compiled PEM datasets [4]. Graph-temporal networks can emulate dynamic multiphysics fields generated by high-fidelity simulations [5]. Physics-informed neural networks can model transient temperature behavior by embedding governing structure into learning [6]. These advances demonstrate substantial modeling capacity, but predictive accuracy within a given dataset does not by itself establish that the learned mapping will remain valid when the renewable trajectory family changes.

The distinction is important because time-series domain shift is not equivalent to ordinary random-sample variation. Forecasting distributions can change together with temporal dependence, lag structure, and the conditional relationship between history and future response. Time-series transfer and domain-generalization studies have consequently shown that methods designed for static covariate shift can become unstable when domains differ in their temporal patterns, and that transferable forecasting requires explicit attention to domain-dependent conditional structure [7,8]. For renewable-powered electrolysis, the renewable source itself can therefore act as a domain: wind, wave, solar-photovoltaic, and tidal trajectories can expose the same electrolyzer to different spectra, ramp patterns, occupancy ranges, and persistence structures.

This creates a specific scientific ambiguity. A model trained on several renewable profiles may learn process relationships that are genuinely transferable across source domains, or it may exploit correlations that are predictive only for the trajectory families represented during training. Standard random splits, and even within-source chronological splits, cannot fully distinguish these possibilities because closely related trajectories remain on both sides of the evaluation. A stricter test is to withhold an entire renewable-source domain and ask whether the model can forecast the same PEM output without seeing that source during fitting, preprocessing, model selection, representation learning, or physical-parameter estimation.

The present study addresses this ambiguity with public megawatt-scale PEM-electrolysis data spanning Simulated Wind, Simulated Wave, Historical Solar PV, and Simulated Tidal operation. Cross-renewable transfer is evaluated by withholding one complete renewable-source domain at a time and retaining a common 60 s forecasting history across direct 1, 5, 10, and 30 s horizons. Historical Wind is used separately as a within-wind provenance control rather than as a fifth primary domain.

We first examined whether source-balanced fitting or worst-source selection changed the failure pattern. We then evaluated label-free error ranking, a source-held-out self-supervised temporal representation, and a low-complexity physical anchor. This design separates a source-held-out test of forecasting transfer from within-domain fitting. Its principal contribution is the comparison of forecasting approaches under the same source-level exclusion rule, with physical anchoring assessed alongside data-driven alternatives rather than presented as a new universal forecasting architecture. Figure 1 summarizes the study design.

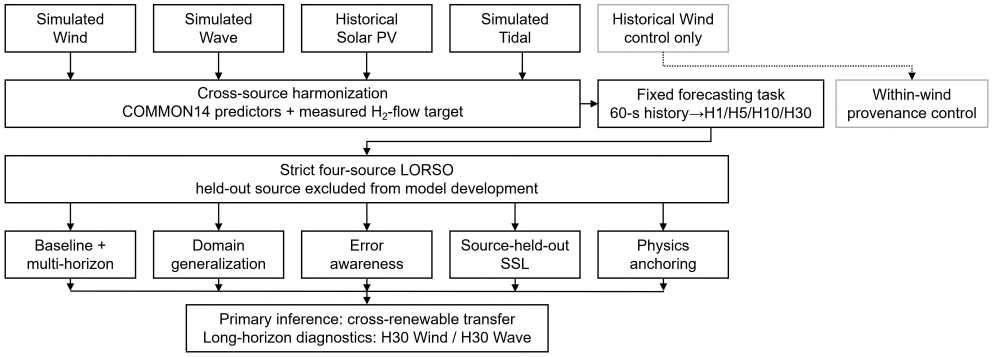


Figure 1. Cross-renewable-source forecasting and analytical framework. Four primary renewable-source domains were harmonized to a common PEM process schema and evaluated under strict four-source leave-one-renewable-source-out (LORSO) validation. Historical Wind was retained only as a separate within-wind provenance control. The same 60 s history and H1/H5/H10/H30 forecast registry supported parallel analyses of baseline transfer, domain-generalization controls, label-free error awareness, source-held-out self-supervised learning, and physics anchoring.

2. Materials and Methods

2.1. Renewable-Source Domains and Study Scope

Four publicly released renewable-powered PEM-electrolysis datasets were treated as the primary source domains (Table 1): Simulated Wind, Simulated Wave, Historical Solar PV, and Simulated Marine Hydrokinetic Tidal. Simulated Wind and Simulated Wave correspond to National Laboratory of the Rockies (NLR) Data Catalog releases 305 [9] and 306 [10], respectively. Historical Solar PV and Simulated Marine Hydrokinetic Tidal correspond to releases 318 [11] and 317 [12], respectively. Across this reference-data program, the releases provide one-second process measurements under distinct renewable forcing patterns [9,12]. The catalog documentation identifies the Nel Hydrogen MC250 megawatt-scale platform with a single 1.25 MW stack for this reference program [9,12]. The study therefore evaluates transfer across renewable-source operating domains on this platform; it does not test cross-manufacturer, cross-stack, or cross-capacity generalization.

The term "simulated" in the Wind, Wave, and Tidal releases refers to the renewable forcing trajectories. In particular, the wave profile was model-derived, whereas the hydrogen-flow target was measured during operation of the physical electrolyzer [10].

A fifth release, NLR Historical Wind (Data Catalog 316 [13]), was retained only as a within-wind provenance and trajectory control. It was never treated as a fifth primary domain and did not participate in the four-source leave-one-renewable-source-out (LORSO) tuning or subsequent multi-horizon, domain-generalization, representation-learning, and gray-box evaluations. This separation was fixed before the multi-horizon and representation-learning experiments to avoid expanding the validation design after observing model performance. A public PT-MELT model artifact associated with the Wind and Wave releases was also available [14] and was treated only as an external reproducibility comparator, as described in Section 2.6.

Table 1. Renewable-source domains, canonical records, and forecast-anchor counts.

Domain	Catalog/DOI	Canonical Records	Paired Anchors	Role
Simulated Wind	305/10.7799/3007835	61,285	42,213	Primary LORSO
Simulated Wave	306/10.7799/3007836	12,006	11,294	Primary LORSO
Historical Solar PV	318/10.7799/3372313	110,302	109,750	Primary LORSO
Simulated Tidal	317/10.7799/3364185	47,502	47,260	Primary LORSO
Historical Wind	316/10.7799/3024096	28,800	28,559 at H1	Control only

2.2. Canonical Harmonization and COMMON14 Process Representation

All the source releases were harmonized into a common one-second schema before cross-source modeling. The forecasting target, q, was stored canonically as `h2_flow_kg_h` and mapped directly from the recorded hydrogen-flow field `IVAL_f_FM011_Flow` (reported as kg/h where units were explicit in the source tables). The combined source tables contain this flow signal separately from `H2E_f_Elec_CalcProdRate`, the calculated hydrogen-production-rate channel; the target was not reconstructed from the current command or from `H2E_f_Elec_CalcProdRate`. NLR documentation for the Wind and Wave releases describes hydrogen mass production as a measured flow signal obtained with an Emerson Coriolis flow meter [9,10]. The same measured hydrogen-flow field was present in the Solar PV and Tidal source tables used in this study [11,12]. The COMMON14 process representation comprised 1 current-command channel and 13 pressure, resistivity, and temperature channels shared by the cross-source schema.

COMMON14 channels:

`current_cmd_pct`, `H2E_f_PT264_Pressure`, `H2E_f_PT307_Pressure`, `H2E_f_PT312_Pressure`, `H2E_f_PT604_Pressure`, `H2E_f_RS209_Resistivity`, `H2E_f_RS507_Resistivity`, `H2E_f_TE218_Temp`, `H2E_f_TE219_Temp`, `H2E_f_TE338_Temp`, `H2E_f_TE601_Temp`, `H2E_f_TT645_Temp`, `H2E_f_TT646_Temp`, and `H2E_f_TT647_Temp`.

For the tabular models, each channel was summarized over the 60 s history using six descriptors fixed before model comparison: the last value, mean, population standard deviation, minimum, maximum, and the 59 s change between the first and last history samples. COMMON14 therefore yielded $14 \times 6 = 84$ predictors. A diagnostic extension appended `renewable_input_kW`, producing 90 predictors, but renewable input was not part of the primary multi-horizon representation and was excluded entirely from the primary gray-box models.

2.3. Forecast Anchors, History Window, and Prediction Horizons

The forecasting task was strictly forward looking. For an eligible anchor at time t, the model received only the preceding 60 one-second observations, including the observation at t, and predicted future hydrogen flow at one of four direct horizons $h \in \{1, 5, 10, 30\}$ s. No predictor from $t + h$ or from the intervening future interval was available to the model. The paired-anchor registry retained only anchors for which the complete 60 s history and all four future targets were valid, yielding 210,517 common anchors: 42,213 Wind, 11,294 Wave, 109,750 Solar, and 47,260 Tidal.

$$\begin{aligned} X(t) &= [x(t-59), \ldots, x(t)], \\ y(h,t) &= q(t+h), \\ h &\in \{1, 5, 10, 30\}\ \text{s}. \end{aligned} \tag{1}$$

Because adjacent forecasting windows overlap, these 210,517 anchors provide paired temporal evaluations across the four held-out source domains rather than independent replications of source transfer.

The use of a single paired registry ensured that comparisons across horizons, representations, and later gray-box variants were not confounded by different eligible sample sets. Direct non-recursive forecasting was used throughout: each horizon had its own target and, where applicable, its own supervised model, rather than obtaining H5–H30 recursively from H1 predictions.

2.4. Four-Source LORSO Validation and Historical Wind Control

The primary validation unit was the renewable-source domain. In each outer fold, one complete source domain was held out and every data-dependent operation was restricted to the other three domains. Thus, the four outer folds held out Wind, Wave, Solar, and Tidal in turn. Random sample-level splitting was not used because adjacent one-second records and overlapping histories belong to the same renewable trajectory and would otherwise permit source-trajectory leakage.

For model families requiring selection, the three outer-training domains were further divided by leave-one-training-source-out inner validation. Hyperparameters or strategy choices were selected only from these inner source-held-out evaluations. The outer held-out domain was not used for preprocessing, feature scaling, hyperparameter selection, self-supervised pretraining, physical-parameter estimation, or residual fitting. Historical Wind was evaluated separately with the Wind-fold training configuration held fixed; it did not enter selection and was used only to assess whether the difficult Wave transfer could be dismissed as a generic consequence of historical-data provenance.

2.5. Evaluation Metrics and Source-Macro Aggregation

Mean absolute error (MAE) was the primary prediction metric used because the central question concerned transfer performance across heterogeneous source domains rather than pooled record-level fit. RMSE and R^2 were retained as secondary diagnostics. Metrics were first computed within each held-out source and were then averaged with equal source weight. This source-macro aggregation prevents the much larger Solar domain from dominating the cross-source conclusion.

$$\begin{aligned} \mathrm{MAE}(d) &= \mathrm{mean}\{|y_i - \hat{y}_i| : i \in d\}, \\ \text{source-macro MAE} &= \frac{1}{4}[\mathrm{MAE}(\mathrm{Wind}) + \mathrm{MAE}(\mathrm{Wave}) \\ &\quad + \mathrm{MAE}(\mathrm{Solar}) + \mathrm{MAE}(\mathrm{Tidal})]. \end{aligned} \tag{2}$$

The same equal-weight principle was used for source-macro RMSE and R^2: each quantity was computed separately in the held-out Wind, Wave, Solar, and Tidal domains and then averaged arithmetically across the four domains. The source-macro R^2 was therefore not recomputed from pooled predictions. Equal source weighting was also used for the domain-generalization selection criteria and the error-awareness summaries. H30 Wind and H30 Wave were retained as long-horizon failure diagnostics after the multi-horizon analysis established that transfer was especially difficult in these domains.

MAE is expressed in kg h^{-1}. Source-macro averaging gives each of the four observed domains one equal weight; it neither creates additional independent sources nor estimates population-level uncertainty. Differences between domain-level averages are descriptive and are not claims of statistical significance or generalization to an unseen population of renewable systems.

2.6. Data-Driven Baselines and Fixed History Descriptors

The baseline model family comprised the archived training-source constant-mean predictor reported as "mean" in the baseline outputs and Ridge regression [15]. HistGradientBoosting (HGB), based on gradient-boosted regression trees, provided one nonlinear ensemble [16]. LightGBM provided a second boosting implementation [17]. The primary feature set was the 84-dimensional COMMON14 descriptor vector. For each channel c, the history operator returned the final value, mean, population standard deviation, minimum, maximum, and 59 s change:

$$\phi(c,t) = [\text{last, mean, population SD, minimum,} \\ \text{maximum, } x(c,t) - x(c,t-59)]. \tag{3}$$

The initial baseline analysis used a 1 s next-step task to establish whether transferable signal was present and to examine whether directly appending `renewable_input_kW` provided consistent incremental value. Ridge, HGB, and LightGBM were trained within each outer LORSO fold; candidate settings for the learned models were selected by inner leave-one-training-source-out performance. All fitted preprocessing and model selection were restricted to the outer-training domains. The renewable-input representation remained diagnostic because its incremental value varied by model and held-out source.

The public PT-MELT 60 s attention-LSTM artifact [14] was investigated as an external baseline, but strict reproduction was not forced. The public artifact itself was available, whereas the exact public raw-input provenance required to reconstruct all 15 official channels was incomplete across the required Wind and Wave archives. Missing official inputs were neither substituted nor imputed. PT-MELT is therefore reported as a reproducibility limitation rather than as a negative model result.

2.7. Multi-Horizon Direct Forecasting

The multi-horizon analysis extended the fixed 60 s history task from H1 to H1, H5, H10, and H30 while preserving the same source memberships, outer LORSO folds, and 210,517 paired anchors. Forecasts were direct and non-recursive. COMMON14 remained the primary representation, and COMMON14 plus `renewable_input_kW` remained diagnostic. Each model–horizon combination was fit and evaluated on the same held-out-source registry, allowing horizon effects to be separated from sample-composition changes.

For learned models with tunable settings, inner source-held-out selection was performed within the three outer-training domains. The primary comparison used equal-weight source-macro MAE; per-source MAE, RMSE, and R^2 were retained to expose source-specific failure modes. No horizon was dropped or redefined after observing the multi-horizon results.

2.8. Domain-Generalization Strategies

The domain-generalization analysis tested whether two simple forms of control could reduce the cross-source failures without changing the feature set or model families. A prespecified 2 × 2 factorial crossed the training loss weighting with the inner model-selection criterion. D0_ERM_MEAN used ordinary empirical-risk minimization and selected by mean inner-source MAE. D1_ERM_WORST retained ordinary fitting but selected by the worst inner-source MAE. D2_SOURCEBAL_MEAN used source-balanced sample weights and selected by mean inner-source MAE. D3_SOURCEBAL_WORST combined source-balanced fitting with worst-source inner selection.

For source-balanced fitting, each training domain received the same total weight regardless of its number of anchors. If N is the total number of rows across K training sources and n_d is the number of rows from source d, each sample from d received

$$w(i) = \frac{N}{K\, n(d)}. \tag{4}$$

The COMMON14 feature set was unchanged and source identity was not added as a predictor. D0 served as the empirical-risk reference. D1–D3 were evaluated as prespecified diagnostic alternatives; outer-test results were not used to select a different deployment strategy for each model–horizon combination. The analysis therefore tested whether source-size imbalance or average-risk inner selection could account for the established failure pattern.

These interventions operate through sample weights and the inner-source selection criterion; unlike broader time-series transfer methods concerned with changing temporal conditional relationships [7,8], they do not directly enforce representation invariance across sources.

2.9. Label-Free Error Awareness and Selective Prediction

The error-awareness analysis did not attempt to construct calibrated prediction intervals or a formal binary out-of-distribution classifier. Instead, it evaluated whether label-free scores available at prediction time could rank cases by expected error and thereby support selective prediction, following the broader reject-option and distribution-shift motivation in the literature [18,19]. Two predictor states were fixed before evaluation: D0_ERM_MEAN as the canonical reference and D3_SOURCEBAL_WORST as a prespecified combined-DG state. No D1/D2/D3 strategy was selected separately by horizon after test performance was observed.

The first score, U1_MODEL_DISAGREEMENT, was the population standard deviation of the three prespecified learned-model predictions (Ridge, HGB, and LightGBM) for the same case:

$$U1(i) = \mathrm{SD}\left\{\hat{y}_{\mathrm{Ridge}}(i), \hat{y}_{\mathrm{HGB}}(i), \hat{y}_{\mathrm{LightGBM}}(i)\right\}. \quad (5)$$

The second score, U2_FEATURE_NOVELTY, measured distance from the outer-training source distributions in COMMON14 descriptor space. For each training source s, the outer-training data defined a feature-wise center μ_s and scale σ_s. The standardized distance and raw novelty score were

$$d_s(x) = \left[J^{-1}\sum_j \left(\frac{x_j - \mu_{sj}}{\sigma_{sj}}\right)^2\right]^{1/2}, \qquad U2_{\mathrm{raw}}(x) = \min_s d_s(x).$$

A training-only reference distribution was constructed by comparing each training source with the other training sources; the held-out raw distance was then mapped to its percentile in that reference distribution. Neither held-out labels nor the held-out feature distribution were used to fit the score. Numerical safeguards for near-zero feature scales and the executable implementation are provided in the reproducibility package.

For selective prediction, cases were ordered from lower to higher score and the lowest-score fraction was retained at coverages $c = 0.50, 0.55, \ldots, 1.00$. Let $R(c)$ denote the MAE among retained cases and $R(1)$ the MAE of the full held-out set. The reported risk ratio was $RR(c) = R(c)/R(1)$, with RR50 and RR80 evaluated at 50% and 80% coverage. Normalized AURC, defined here as the trapezoidal average of RR(c) over 50–100% coverage, was $(1/0.5)\int_{0.5}^{1} RR(c)\,dc$; values below 1 indicate lower average selective risk than the full-set reference. Spearman correlation between each score and absolute forecast error was also reported. These are descriptive measures of label-free error awareness and do not provide guaranteed coverage, calibrated uncertainty, or a formal OOD classifier.

2.10. Source-Held-Out Self-Supervised Representation Learning

The self-supervised analysis tested whether a raw temporal representation learned without hydrogen-flow labels could replace the handcrafted descriptors, motivated by prior self-supervised multivariate time-series representation learning [20,21]. For every paired anchor, the exact 60 × 14 COMMON14 history was retained at 1 Hz. Within each outer LORSO fold, the held-out source was excluded completely from self-supervised pretraining, including as unlabeled data. Per-channel normalization statistics were estimated from the three training sources only.

The pretext task masked exactly 12 of the 60 time steps (20%) and masked all 14 normalized process channels at each selected step. A binary mask indicator was appended as a fifteenth input channel. The encoder comprised Conv1d layers 15→32 (kernel 5), 32→64

(kernel 5), and 64→64 (kernel 3), each followed by GELU, with global mean pooling to a 64-dimensional embedding. A 64→32→14 decoder reconstructed the process channels, and the loss was the mean squared reconstruction error evaluated only at masked positions:

$$L_{\mathrm{SSL}} = \mathrm{mean}\{[x(j,c) - \hat{x}(j,c)]^2 \text{ over masked channel–time positions}\}. \quad (6)$$

Each outer-fold encoder was trained for five epochs with batch size 512 using AdamW (learning rate: 10^{-3}; weight decay: 10^{-4}), without dropout, early stopping, architecture selection, or test-driven retuning. The random seed was fixed at 14 September 2026.

Three fixed downstream probes were evaluated with StandardScaler followed by Ridge(alpha = 1.0), with no inner cross-validation or hyperparameter tuning. S0_DESC_LINEAR used the 84 handcrafted COMMON14 descriptors; S1_SSL_LINEAR used the 64-dimensional self-supervised embedding and was the primary SSL comparison; S2_DESC_PLUS_SSL_LINEAR concatenated both representations into 148 dimensions as a diagnostic incremental-representation test. The same four-source LORSO and the same paired anchors were used for all probes.

2.11. Physics-Anchored Gray-Box Forecasting

The gray-box analysis tested a deliberately low-complexity physical inductive bias rather than a full electrochemical polarization model, consistent with the broader physics-informed machine-learning principle of constraining learned mappings with verified physical structure [22]. The available cross-source schema did not contain the independently verified stack quantities required to identify a complete mechanistic PEM model. The physical branch was therefore limited to a current-command-to-hydrogen proportionality consistent with Faraday-law electrolysis [23], using an effective nonnegative scale coefficient. Because the canonical signal is a current command rather than an independently verified realized stack current, the fitted coefficients are not interpreted as Faradaic efficiency, cell count, or a uniquely identifiable electrochemical constant. Throughout this paper, "physics-anchored" denotes a reduced-order operational command-to-flow inductive bias; F0 and F1 are not fully identified electrochemical or multiphysics models.

The unit-bearing `current_cmd_A` field was admitted only after a prespecified provenance gate established, separately in all four primary sources, that it was an essentially affine rendering of the already available COMMON14 `current_cmd_pct` signal ($R^2 \geq 0.999$ and normalized RMSE ≤ 0.01). It therefore did not introduce a new independent sensor channel.

2.11.1. Static Physical Anchor (F0)

The static anchor predicted future hydrogen flow from the last observed current command I(t):

$$\hat{q}_{F0}(t+h) = \kappa_0 I(t), \qquad \kappa_0 \geq 0. \quad (7)$$

Within each outer fold, κ_0 was fitted by through-origin least squares using H1 labels from the three outer-training sources only. The same κ_0 was then reused at H1, H5, H10, and H30. F0 was a physics-only ablation and used no `renewable_input_kW`, target history, or future command.

2.11.2. First-Order Dynamic Physical Anchor (F1)

F1 introduced a single effective response constant τ. For a candidate τ, the observed 60 s current history was filtered with a first-order state, using $a = \exp(-1/\tau)$:

$$z(j) = a\, z(j-1) + (1-a) I(j), \qquad a = \exp(-1/\tau). \quad (8)$$

Future current was not assumed known. Instead, the last observed command was held constant, giving the zero-order-hold forecast state

$$z(t+h \mid t) = I(t) + [z(t) - I(t)] \exp(-h/\tau). \tag{9}$$

and the dynamic physical forecast

$$\hat{q}_{F1}(t+h) = \kappa_1 z(t+h \mid t), \qquad \kappa_1 \geq 0. \tag{10}$$

The τ grid was fixed in advance at {1, 5, 10, 30, 60} s. Within each outer fold, τ was selected by inner leave-one-training-source-out H1 source-macro MAE using the physics-only model; ties were resolved in favor of the smaller τ. After selection, κ_1 was refitted on all three outer-training sources using H1 labels only. The selected τ and κ_1 were horizon-invariant within the outer fold and were reused for H1, H5, H10, and H30. τ is therefore an effective first-order response constant under this forecasting representation, not a uniquely identified electrochemical time constant.

The future current trajectory is not observed at forecast origin. Accordingly, zero-order hold is an explicit prediction-time assumption that carries the most recent current command forward; it is not intended to reproduce actual future dispatch schedules. An oracle forecast supplied with future commands could separate some forecast-forcing error from response-model error, but such an experiment was not included in the frozen protocol and is not claimed here.

2.11.3. Residual Gray-Box Model (G1)

The primary residual gray box added a horizon-specific linear residual correction to the fitted F1 physical branch:

$$\hat{q}_{G1}(t+h) = \hat{q}_{F1}(t+h) + \hat{r}_h[\phi(t)]. \tag{11}$$

The residual feature vector $\phi(\mathrm{t})$ contained the same six descriptors from the 13 non-current COMMON14 channels, giving 78 predictors. Both `current_cmd_pct` and `current_cmd_A` were excluded from the residual branch so that the direct current-to-production relationship remained confined to the physical anchor. `renewable_input_kW`, hydrogen-flow target history, calculated hydrogen-production-rate channels, Historical Wind, and all future values were prohibited.

For each outer fold and horizon, the residual target was $\mathrm{q(t + h)} - \hat{q}\mathrm{F1(t + h)}$. StandardScaler was fitted on outer-training rows only, followed by Ridge(alpha = 1.0, fit_intercept = True). The residual penalty $\alpha = 1.0$ was fixed, not optimized by inner cross-validation; this experiment cannot isolate how an alternative penalty would affect transfer. No residual-model family, alpha, weighting scheme, or feature subset was selected from test outcomes. Primary predictions were not clipped after fitting; any negative-prediction frequency was retained as a diagnostic rather than repaired post hoc.

The primary gray-box comparison was specified in advance as G1 versus the Ridge COMMON14 comparator from the multi-horizon analysis at the same horizon. G1 was also reported descriptively against the lowest-MAE data-driven model from the preceding multi-horizon analysis at each horizon. F1 versus F0 assessed the value of first-order dynamics, and G1 versus F1 assessed the incremental value of the process-state residual. No formal significance claim was prespecified for these point comparisons.

G1 is physics-anchored but not strictly physics-constrained: the nonnegative F1 component can be offset by an unconstrained residual and the combined prediction may be negative. Such outputs violate the admissible sign of the modeled hydrogen-production mass flow. We retain the original negative-output rates and unmodified MAE values as

failure diagnostics; the reported G1 must not be construed as a physically guaranteed or deployment-ready controller.

2.12. Reproducibility, Software, and Leakage Control

The analytical modules (Table 2) were specified before execution. Negative outcomes were retained without outcome-driven expansion of parameter grids, source domains, random seeds, representations, or model families. This applies to the domain-generalization alternatives, error-awareness scores, self-supervised architecture and training constants, and the gray-box τ grid and residual model.

Leakage control was source-based rather than record-based. In every outer fold, the held-out renewable-source domain was excluded from model fitting, hyperparameter or strategy selection, fitted preprocessing, self-supervised pretraining, physical-parameter estimation, and residual fitting. For self-supervised learning, the exclusion also applied to unlabeled held-out data; for the physics-anchored analysis, future current and renewable input were unavailable by construction. Prediction-level outputs were retained so that MAE, RMSE, and R^2 could be recomputed independently.

The software versions, configuration files, source manifests, fold-level audit tables, execution records, and file-integrity hashes are supplied in the Supplementary Materials. The self-supervised run used Python 3.12.10, NumPy 2.3.5, pandas 2.2.3, scikit-learn 1.8.0, and PyTorch 2.13.0 + cu130 on an NVIDIA GeForce RTX 5060 Laptop GPU. The gray-box execution metadata recorded Python 3.13.14, NumPy 2.5.3, and pandas 3.0.5.

Generative-AI disclosure: OpenAI ChatGPT (GPT-5.6 Sol) was used as an auxiliary tool for drafting and checking Python code and Windows PowerShell 5.1 scripts, manuscript organization, English-language editing, and figure-formatting assistance. All analytical scripts were executed against the documented public datasets, and numerical results, tables, and figures were checked against executed outputs. The tool was not used to create missing experimental observations, alter measured data, or fabricate model results. The authors retain responsibility for the analysis, references, interpretation, and final manuscript.

Table 2. Analytical program used to test cross-renewable-source transfer.

Analysis	Question	Design/Representation	Role in Analysis
Baseline	Does transferable next-step structure exist?	COMMON14 descriptors; Ridge/HGB/ LightGBM	Cross-source baseline
Wind control	Is the Wave failure explained by historical provenance?	Historical Wind control with fixed Wind-fold training	Provenance-shift diagnostic
Multi-horizon	How does transfer change with forecast horizon?	Direct H1/H5/H10/H30 forecasting	Multi-horizon reference
DG	Can simple DG controls remove the failure?	ERM/worst-source selection × source balancing	Domain-shift diagnostic
Error awareness	Can label-free scores rank forecast risk?	Model disagreement (U1); feature novelty (U2)	Selective-risk diagnostic
SSL	Does raw temporal SSL improve transfer?	Source-held-out masked Conv1d representation	Representation test
Gray box	Does physical anchoring improve transfer?	F0/F1 physical anchors; G1 residual gray box	Physics-anchored test

3. Results

3.1. Baseline Cross-Renewable Transfer and Historical Wind Control

The initial LORSO screening showed transferable predictive information in the common PEM process history, while also revealing marked dependence on the held-out renewable-source domain. With COMMON14, HGB gave the lowest source-macro MAE (1.310),

followed by LightGBM (1.341) and Ridge (1.431); the archived mean baseline had a source-macro MAE of 6.008. Ridge had the highest source-macro R^2 among the learned COMMON14 models (0.850). Model ranking changed across held-out sources, with Wave the most difficult domain.

Directly appending `renewable_input_kW` did not create a source-invariant improvement. Relative to COMMON14, source-macro MAE decreased by 4.38% for HGB and 4.28% for LightGBM but increased by 3.05% for Ridge. The direction also changed by held-out source: for example, the added renewable trajectory improved HGB and LightGBM on Wave but degraded Ridge on the same domain, while the nonlinear models were slightly worse on Simulated Wind. The additional trajectory signal was therefore informative in some transfers without constituting a generally portable predictor.

The Historical Wind control showed that difficult transfer could not be attributed to historical provenance alone. With the Wind-fold model choices held fixed, COMMON14 MAE on Historical Wind was 1.001 for Ridge, 1.403 for HGB, and 1.515 for LightGBM, compared with 1.037, 1.695, and 1.726, respectively, for the Simulated Wind evaluation under the same fixed training regime. Historical Wind was therefore not harder for any learned model in this control. Because the simulated and historical wind trajectories were not matched interventions, this descriptive contrast does not identify a causal provenance effect or exclude one. It shows that the historical label was not uniformly associated with larger errors in this control, but it cannot determine whether differences in provenance, trajectory structure, or both contributed to the Wave results. Strict reproduction of the public PT-MELT model was not scored as a performance result because exact public provenance for all required official input channels was incomplete.

3.2. Multi-Horizon Cross-Renewable Transfer

The initial 1 s screening and the H1 result in the multi-horizon analysis used different eligibility registries. The multi-horizon registry retained only anchors with a complete 60 s history and valid targets at H1, H5, H10, and H30, whereas the initial screening was a separate 1 s next-step probe; their H1 point estimates are therefore not expected to be numerically identical. Within the common 210,517-anchor multi-horizon registry, H1 and H5 were similar in average difficulty, degradation was evident at H10, and it became pronounced at H30. Relative to paired H1, H30 source-macro MAE increased by approximately 95.6% for HGB, 100.3% for LightGBM, and 73.8% for Ridge (Table 3 and Figure 2).

Table 3. Primary COMMON14 source-macro MAE (kg h^{-1}) across forecast horizons.

Horizon	HGB	LightGBM	Ridge	Lowest MAE
H1	1.292	1.312	1.439	HGB
H5	1.355	1.335	1.344	LightGBM
H10	1.517	1.567	1.566	HGB
H30	2.527	2.628	2.502	Ridge

The long-horizon degradation was not evenly distributed across sources. At H30, Wave had negative R^2 for all three primary models (HGB −0.331, LightGBM −0.337, Ridge −0.625). Unseen Wind also deteriorated materially, although its H30 R^2 remained positive (approximately 0.172, 0.108, and 0.412 for HGB, LightGBM, and Ridge). Thus, Wave was the most persistent challenge overall, but the long-horizon problem was not confined to Wave. The renewable-input diagnostic remained mixed across horizons: it improved several H5/H10 cells but worsened H30 HGB and LightGBM, reinforcing that the extra source trajectory did not provide a stable cross-domain remedy.

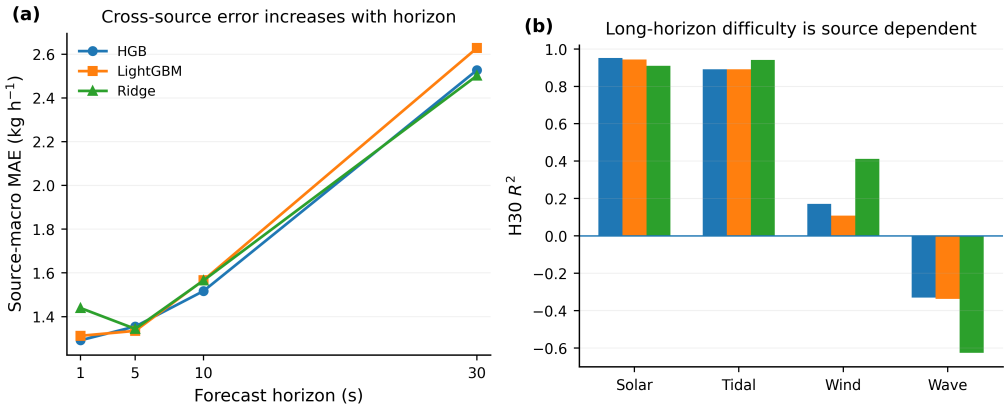


Figure 2. Multi-horizon cross-renewable transfer. (**a**) Source-macro MAE for Ridge, HGB, and LightGBM using COMMON14 under four-source LORSO. (**b**) Per-source H30 R^2 shows that long-horizon difficulty was concentrated in held-out Wind and especially Wave rather than being uniform across sources.

3.3. Domain-Generalization Controls

The domain-generalization factorial tested whether source-size imbalance or mean-risk model selection could account for the established failures. The clearest descriptive gains occurred at H5. D3 reduced HGB source-macro MAE from 1.355 to 1.294 (4.49%) and LightGBM from 1.335 to 1.266 (5.16%), while D1 reduced Ridge from 1.344 to 1.323 (1.55%). Figure 3 summarizes the percentage change in source-macro MAE for D1–D3 relative to D0 across all model–horizon combinations, including these H5 gains, without treating the lowest observed outer-test value as a strategy-selection rule.

The pattern did not persist across models and horizons. At H10, D2 improved HGB and D3 improved LightGBM, whereas D0 remained lowest for Ridge. At H30, D0 remained lowest for HGB and Ridge; LightGBM showed only a very small source-macro reduction with D2, accompanied by a worse worst-source MAE. Across the 12 model-by-horizon cells, D1 reduced source-macro MAE in 3 cells, D2 in 4, and D3 in 4. These results do not support source balancing or worst-source selection as a universal remedy for the long-horizon transfer problem.

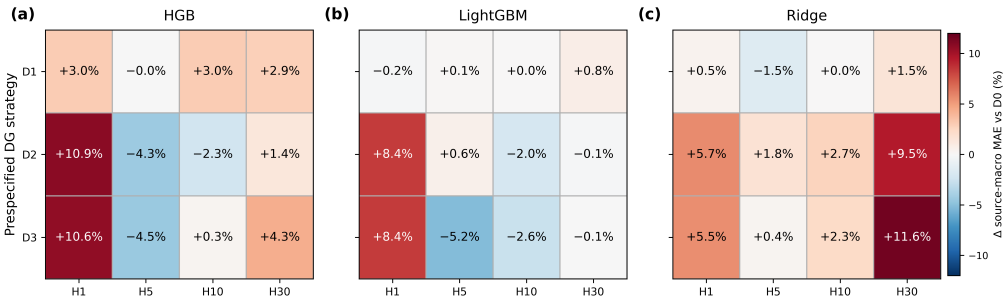


Figure 3. Domain-generalization effects relative to D0. Each panel shows the percentage change in source-macro MAE for the three prespecified DG alternatives relative to D0 across horizons. Negative values indicate improvement. Gains were concentrated at H5/H10 and did not provide a consistent H30 remedy.

3.4. Label-Free Error Awareness and Selective Prediction

Model disagreement (U1) was the stronger of the two prespecified error-awareness scores. Across 96 outer-source/model/horizon cells, U1 had a positive association with absolute error in 83 cells, normalized AURC below 1 in 87, and an 80% coverage risk ratio below 1 in 88. The corresponding counts for feature novelty (U2) were 59, 64, and 71. In source-macro paired comparisons, U1 outperformed U2 in 22 of 24 comparisons for

Spearman ρ, normalized AURC, and the 80% risk ratio, and in 21 of 24 comparisons for the 50% risk ratio. Figure 4 summarizes the horizon-dependent source-macro Spearman association, normalized AURC, and representative U1 risk–coverage curves at H1 and H30.

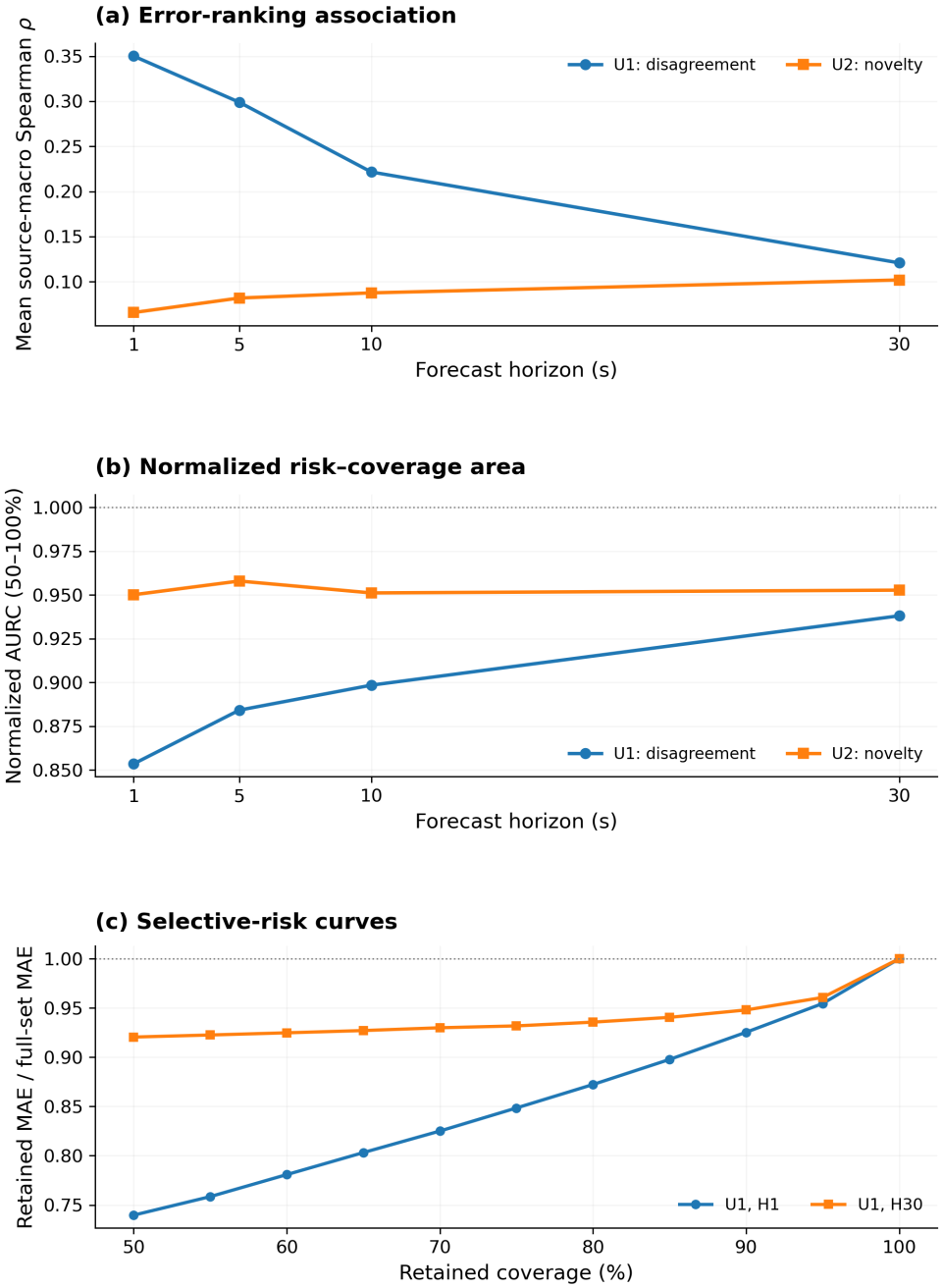


Figure 4. Label-free error awareness and selective prediction. (**a**) Mean source-macro Spearman association between uncertainty score and absolute error across predictor states and learned models. (**b**) Corresponding normalized AURC over 50–100% coverage. (**c**) Mean U1 risk–coverage curves at H1 and H30. Model disagreement was more informative than feature novelty overall, but its ranking value weakened with horizon.

Error awareness nevertheless weakened as the forecast was pushed farther into the future. The source-macro U1 Spearman correlation fell from approximately 0.350 at H1 to 0.299 at H5, 0.222 at H10, and 0.121 at H30. Over the same horizons, normalized AURC increased from approximately 0.854 to 0.884, 0.898, and 0.938, indicating progressively less effective selective ranking. On H30 Wave, U1 could even become negatively correlated

with error for HGB and LightGBM. U2 was weaker overall but supplied complementary information in some of these difficult H30 Wave cases. These results support label-free error awareness and selective deferral, not calibrated uncertainty or formal OOD classification.

3.5. Self-Supervised Representation Transfer

The source-held-out SSL objective optimized cleanly: masked-reconstruction MSE decreased monotonically in all four outer folds, by approximately 94.8–96.2% from epoch 1 to epoch 5. S0 in this analysis was a fixed linear probe (StandardScaler + Ridge, $\alpha = 1.0$) applied to the 84 handcrafted descriptors, not the separately selected Ridge pipeline used in the multi-horizon baseline; small numerical differences between their MAEs are therefore expected. Forecast transfer moved in the opposite direction from reconstruction loss: S1 produced source-macro MAE values of 3.817, 4.075, 4.314, and 4.721 at H1, H5, H10, and H30, compared with 1.429, 1.316, 1.558, and 2.540 for S0. The relative deterioration ranged from 85.9% to 209.7% (Figure 5).

Concatenating the SSL embedding with the descriptors also failed to recover an incremental benefit. S2 source-macro MAE was 2.191, 2.105, 2.303, and 4.296 across H1, H5, H10, and H30, all above the corresponding S0 values. At the per-source level, S1 improved on S0 in only 1 of 16 source-by-horizon comparisons: held-out Wave at H30 decreased from 4.649 to 4.407 (−5.22%). That isolated exception did not alter the overall result. The single Wave-H30 improvement does not identify why the embedding helped that one held-out condition: reconstruction may preserve a useful aspect of the local history, but no repeat-seed, matched-trajectory, or feature-attribution analysis was performed to distinguish this possibility from sampling-specific variation. Under the fixed encoder, objective, and probe, successful temporal reconstruction was therefore not evidence of a more transferable forecast representation.

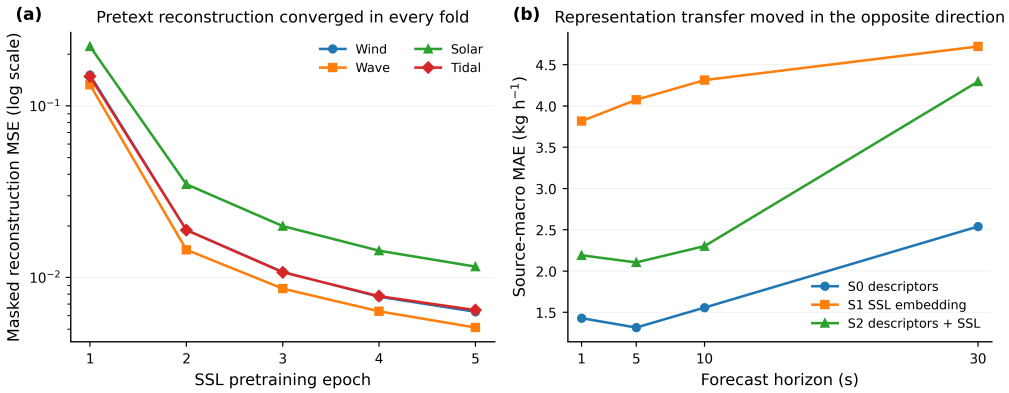


Figure 5. Self-supervised pretraining versus forecast transfer. (**a**) Masked-reconstruction loss decreased sharply in every source-held-out SSL fold. (**b**) Despite pretext convergence, the SSL embedding (S1) and descriptor-plus-SSL representation (S2) had higher source-macro forecast MAE than the handcrafted descriptor probe (S0) at every horizon.

3.6. Physics-Anchored Forecasting and Ablations

G1 reduced source-macro MAE relative to the Ridge comparator from the multi-horizon analysis at every horizon: 18.46% at H1, 15.00% at H5, 13.58% at H10, and 11.53% at H30. It was also descriptively below the lowest observed data-driven MAE at each horizon by 9.15%, 14.42%, 10.83%, and 11.53%, respectively. These comparisons (Table 4) were specified before the gray-box run and are reported as descriptive point estimates rather than formal significance tests. Figure 6 shows the comparator and ablation results.

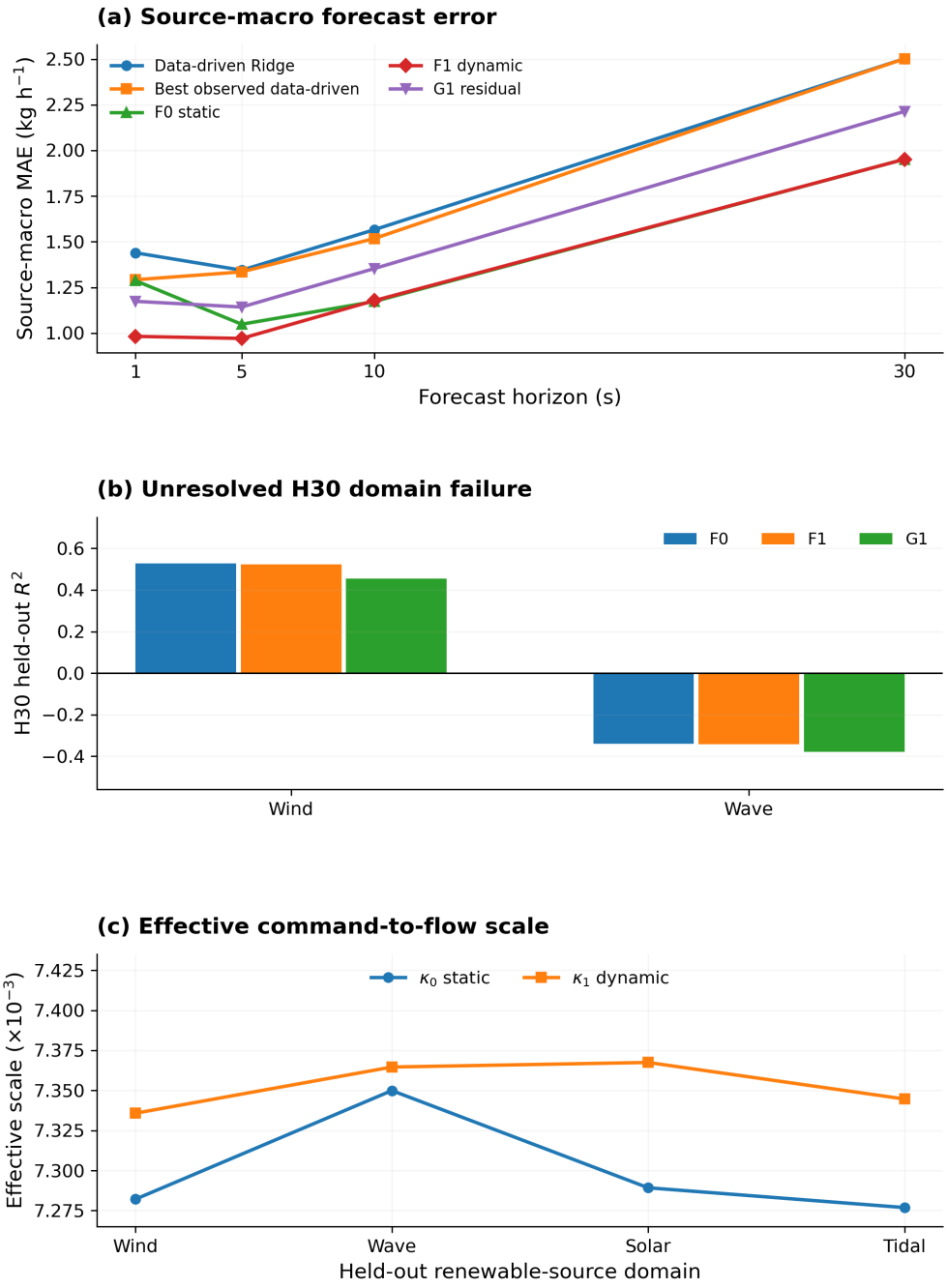


Figure 6. Physics anchoring and residual gray-box behavior. (**a**) Source-macro MAE for data-driven references and the F0/F1/G1 physics-anchored variants. (**b**) H30 R^2 for held-out Wind and Wave shows that Wave remained unresolved even after physical anchoring. (**c**) Effective current-command-to-flow scale coefficients were stable across outer holdouts; these fitted quantities are operational gray-box parameters rather than identified electrochemical constants.

The held-out-source results also show marked heterogeneity. At H30, the F1 MAE ranges from 0.597 to 3.479 kg h^{-1}, versus 0.835 to 3.691 kg h^{-1} for G1 (Table 5). These are observed across-source ranges, not confidence intervals. At H1 and H5, G1 improved on F1 for held-out Wind and Wave even though the source-macro mean favored F1 at both horizons.

F1 achieved lower source-macro MAE than G1 at every horizon. Relative to F0, first-order current-command dynamics reduced MAE at H1 (0.982 vs. 1.287) and H5 (0.971 vs. 1.048), while F0 and F1 were nearly indistinguishable at H10 and H30. Adding the 78-feature Ridge residual then increased MAE relative to F1 by 19.53%, 17.72%, 14.90%,

and 13.43% from H1 through H30. The main transfer gain therefore came from the low-dimensional current-command-to-hydrogen physical branch rather than the higher-dimensional residual correction.

Table 4. Physics-anchored forecasting and prespecified ablations (source-macro MAE) (kg h^{-1}).

Horizon	Ridge Reference	Lowest Data-Driven MAE	F0 Static	F1 Dynamic	G1 Residual Gray Box
H1	1.439	1.292	1.287	0.982	1.174
H5	1.344	1.335	1.048	0.971	1.142
H10	1.566	1.517	1.173	1.178	1.353
H30	2.502	2.502	1.952	1.952	2.214

Table 5. H30 MAE by held-out renewable-source domain for the physics-anchored models (kg h^{-1}). Source-level values summarize dependent prediction windows.

Held-Out Source	F1 Dynamic	G1 Residual Gray Box
Wind	2.826	3.142
Wave	3.479	3.691
Solar PV	0.597	0.835
Tidal	0.904	1.186

The effective current-command-to-flow scale coefficients were stable across outer folds (cross-fold coefficient of variation ≈0.47% for the static anchor and ≈0.21% for the dynamic anchor). Inner selection chose an effective response constant of 10 s for Wind, Solar, and Tidal holdouts and 5 s for the Wave holdout. These fitted quantities are operational gray-box parameters under the present representation and are not interpreted as identified Faradaic efficiencies or unique electrochemical time constants.

3.7. H30 Wind and Wave Diagnostics

Average improvement from the physics anchor did not eliminate the prespecified H30 failure modes. For held-out Wave at H30, R^2 remained negative for F0 (-0.339), F1 (-0.342), and G1 (-0.380), with MAE values of 3.476, 3.479, and 3.691. Held-out Wind remained difficult but retained positive H30 R^2: 0.527 for F0, 0.523 for F1, and 0.454 for G1. The residual layer therefore worsened both Wind and Wave relative to the simpler physical anchors.

G1 produced a small number of negative hydrogen-flow predictions because the unconstrained residual could pull the physical branch below zero; the source-macro mean negative-prediction fraction remained below 0.65% at every horizon, and no post hoc clipping was applied. The result is therefore appropriately described as physics anchored rather than strictly physically constrained. Long-horizon Wind and Wave difficulties persisted despite source weighting, error-awareness diagnostics, self-supervised representation learning, and residual augmentation.

4. Discussion

4.1. Cross-Source Transfer and Domain Dependence

The LORSO results show that recent PEM process history contains information that transfers across renewable-source domains. All three learned COMMON14 models substantially outperformed the archived Mean baseline in the initial screening, and Ridge retained a source-macro R^2 of 0.850. The Historical Wind control also transferred well when evaluated with the same fixed Wind-fold training regime. These observations support a transferable component in the measured process state, but they do not imply a single invariant mapping across renewable sources.

Domain dependence remained visible in model ranking, the difficulty of Wave, and the sign of the incremental effect of `renewable_input_kW`. Historical Wind supplies an unmatched descriptive contrast, not a causal intervention. Its lower observed errors relative to Simulated Wind under the fixed Wind-fold setting show that historical provenance was not uniformly associated with greater error in this comparison; they neither eliminate a provenance effect nor establish why Wave was difficult. The trajectories were not matched for their forcing or state distributions. The empirical conclusion is narrower: cross-renewable transfer is feasible, but its accuracy depends materially on the held-out source domain.

4.2. Effect of Forecast Horizon

The multi-horizon analysis revealed a limitation that was much less visible in the next-step task. Average error changed little from H1 to H5, increased at H10, and rose sharply at H30. Wave reached negative R^2 for all three data-driven models at H30, and unseen Wind also deteriorated substantially. The same Wave difficulty remained after physics anchoring. Short-horizon transfer should therefore not be used as a surrogate for history-only forecasting tens of seconds beyond the observed window.

Longer-horizon error cannot be attributed solely to PEM internal dynamics because future current commands were deliberately unavailable. As the horizon increases, a history-only forecast depends both on how predictable the future electrical forcing is from the preceding 60 s and on how the PEM system responds to that forcing. Ramp rate, intermittency, transport delay, and related effects may contribute, but none was isolated causally here. The LORSO results therefore establish source-dependent long-horizon predictability under the specified information set, not a uniquely identified source-specific process mechanism.

For the F1 anchor, zero-order hold carries the last observed command $I(t)$ forward to $t + h$. Any intervening change in the actual dispatch therefore introduces a forcing mismatch that can become consequential at H30. The Wave input trajectory was model-derived, but the hydrogen-flow response was measured on the physical electrolyzer [10]. The observed negative Wave R^2 may reflect future-dispatch unpredictability, characteristics of the imposed trajectory, and inadequacies in the response approximation. The present history-only comparison cannot quantify these contributions separately or establish that simulation artifacts caused the failure. We did not calculate source-matched ramp-rate distributions, forcing spectra, command-occupancy distributions, or operating-state occupancy contrasts. Without such diagnostics, a mechanistic explanation for the Wave H30 failure would be premature.

4.3. Effect of Added Modeling Complexity

None of the more flexible learning interventions removed the cross-source problem consistently. The model-dependent effects of adding `renewable_input_kW` likewise argue against treating that signal as automatically source invariant: it supplies a potentially informative description of the forcing, but a fitted relationship to future hydrogen flow can depend on the held-out trajectory and model class. The observed Wave benefit for nonlinear baselines, alongside Ridge deterioration, cannot by itself reveal the responsible interaction or guarantee transfer to another source. Source balancing and worst-source selection helped selected H5 and H10 cases but did not dominate across models or horizons. The self-supervised encoder reduced masked-reconstruction loss by more than 94% in every outer fold, yet its embedding was substantially less transferable for hydrogen-flow forecasting than the handcrafted descriptors. The 78-feature residual Ridge likewise degraded the simpler dynamic physical anchor at every horizon.

These results are specific to the tested designs; they are not evidence that domain generalization, self-supervision, or hybrid models fail in general. They show why the large number of one-second anchors should not be confused with abundant evidence about domain invariance. Transportability was evaluated across only four primary renewable-source domains, so a flexible representation could fit extensive within-source regularity while receiving relatively little independent information about which relationships survive a domain change.

The G1 residual used training-source standardization and a fixed Ridge penalty ($\alpha = 1.0$), without an inner search for that penalty. Consequently, its higher held-out error cannot be assigned uniquely to residual overfitting, an unsuitable penalty, or incompatibility between auxiliary descriptors and the current-command anchor. Pressure, temperature, and resistivity descriptors may have source-dependent links to future flow, but no attribution or alternative-penalty experiment was conducted. Additional regularization or feature reduction would require separate source-held-out validation. Likewise, the SSL conclusion concerns only the five-epoch masked-reconstruction Conv1d model and its fixed linear probe, not self-supervised learning generally. The isolated S1 advantage over S0 at Wave H30 (4.407 versus 4.649 kg h^{-1}) does not reveal a unique mechanism and was not replicated across the other 15 source–horizon cells.

4.4. Role of the Physics-Based Anchor

First-order dynamics improved on F0 at H1 and H5, but not at H10 (F0 1.173 versus F1 1.178 kg h^{-1}), and the two models rounded to the same H30 source-macro MAE (1.952 kg h^{-1}). In the zero-order-hold expression, the transient term $[z(t) - I(t)]\exp(-h/\tau)$ decays as the horizon increases, so the F1 forecast state approaches the last observed command. This reduces the structural distinction from F0 at longer horizons; with separately fitted scale factors and unseen future commands, it does not guarantee numerical identity or causal identification of the response dynamics. F1 achieved lower source-macro MAE than G1 at every horizon even though G1 remained below the prespecified data-driven comparators. The residual model therefore did not create the principal transfer improvement. Instead, the strongest performance among the tested representations came from the low-dimensional dynamic current-command-to-hydrogen branch. The fitted current-command-to-flow coefficients also varied little across outer folds, although that stability should be interpreted only within this platform and forecasting representation.

The effective coefficients and response constants are not identified Faradaic efficiencies, cell counts, or unique electrochemical parameters. The available cross-source schema lacks the measurements required for that interpretation. The result is better viewed as evidence that a modest physical inductive bias can be useful when the same relationship must survive a domain change. More elaborate residual learning may still help with richer physical measurements or more independent training domains, but that was not observed here.

The small cross-fold variations in effective κ_0 and κ_1, together with the selected values of τ, reflect source-fold stability on the same equipment under the fixed 60 s command representation, rather than parameter stability across membrane technologies, aging states, manufacturers, or operating scales. Although pressure and temperature channels were available for descriptor models, the F1 branch did not explicitly represent thermal accumulation, pressure dynamics, or degradation. These unmodeled effects may alter the effective response under other operating conditions.

4.5. Selective Error Awareness

Model disagreement (U1) provided useful error ranking in most source/model/horizon cells, but its performance weakened with horizon: the source-macro Spearman

association with absolute error fell from approximately 0.350 at H1 to 0.121 at H30. Feature novelty (U2) was weaker on average and only locally complementary. Error awareness was therefore easiest where the underlying forecasts were also easier to transfer.

The practical interpretation is selective rather than probabilistic. U1 and U2 can be used to defer higher-score cases or trigger additional review, but they are not calibrated uncertainty estimates and do not constitute formal OOD detection. No controller, alarm policy, or closed-loop fallback combining these scores with the physical anchor was evaluated, so operational responses remain outside the evidence reported here.

Selective retention trades automated coverage for potentially lower error among accepted cases. At 80% coverage, for example, one in five forecasts would be deferred by definition; such deferral may incur operator workload, delayed decisions, or fallback-model cost, none of which was quantified. The H30 Wave reversals in U1–error association further caution against applying one threshold indiscriminately across domains. An operational workflow would need independently validated thresholds, explicit costs for accepting versus deferring a forecast, and a safe supervisory action. The reported retrospective risk–coverage curves do not establish a net operational benefit, calibrated uncertainty, or a safety guarantee.

4.6. Limitations and Scope of Generalization

The main limitation is the small number of independent transfer domains. The four primary datasets describe renewable-driven operation of the same NLR megawatt-scale PEM platform, so the study does not establish cross-manufacturer, cross-stack, cross-capacity, or cross-site generalization. Because the inferential unit for transportability is the renewable-source domain rather than the individual time anchor, four outer domains cannot support precise population-level uncertainty statements about a broader universe of renewable operating environments. Equal-weight source-macro summaries prevent the much larger Solar record count from dominating the reported averages, but they do not create additional independent domains.

The source set also mixes simulated Wind, Wave, and Tidal trajectories with historical Solar PV, while Historical Wind is a separate provenance control. This design evaluates deployment-style transfer but does not isolate renewable-source physics from provenance or operating-profile differences. Matched simulation/field trajectories would be needed for that purpose. No matched ramp-rate, frequency-domain, current-distribution, or operating-occupancy analysis was available within the reported frozen evaluation, so the stronger Wave H30 error remains an observed forecast result rather than a mechanistically decomposed failure. Strict reproduction of the public PT-MELT model was also blocked because exact public provenance for all required official input channels was incomplete; missing channels were not substituted or imputed, so this is a reproducibility boundary rather than a negative PT-MELT performance result.

The target is hydrogen-flow forecasting from a fixed 60 s history. Because future current commands were deliberately unavailable, long-horizon forecast error combines uncertainty in future electrical forcing with PEM response and model error; the design therefore evaluates history-only forecastability rather than system dynamics conditional on a known future command trajectory. The analysis does not address degradation, remaining life, faults, safety events, economics, or closed-loop control. The physics-anchored model also lacks independently verified stack voltage, active area, cell count, and other measurements required for a fully identified electrochemical formulation. Because the additive G1 residual has no nonnegativity constraint, it occasionally returned physically inadmissible negative hydrogen-flow predictions. These predictions were retained in the reported evaluation.

Practical use would require a separately validated, constraint-preserving formulation, along with testing on independent PEM systems and additional operating conditions.

Changing the renewable-source domain exposed persistent forecast errors, especially at H30. The physical anchor offered a more consistent cross-source reference than the more flexible alternatives tested here, within the limits of a single electrolyzer platform.

5. Conclusions

This study evaluated forecasting across four held-out renewable operating domains imposed on one nominal 1.25 MW PEM-electrolyzer platform. The equal-weight four-source summaries are descriptive and do not support inference to a population of manufacturers, stacks, capacities, or sites. History-only accuracy deteriorated by H30, especially for Wave; no matched forcing/occupancy diagnostic isolated the mechanism. Source balancing and worst-source selection did not consistently resolve this difficulty. The negative self-supervised finding applies specifically to the tested five-epoch masked-reconstruction Conv1d encoder and Ridge linear probes, rather than to SSL methods more broadly. Model disagreement provided retrospective, horizon-dependent error ranking, not a validated deferral policy.

The reduced-order first-order current-command anchor F1 was more accurate than the residual gray box G1 at all four horizons, although F1 provided no material improvement over the static F0 anchor at H10 or H30. G1 improved on the prespecified data-driven Ridge comparator but introduced larger error than F1 and occasional negative predictions. These models are operational gray-box forecasting representations, not physically identified electrochemical models: realized stack current and parameters such as Faradaic efficiency were not independently identified. Validation across independent sites and hardware, additional command-trajectory diagnostics, and deployment-specific cost–risk assessment remain future work rather than demonstrated results.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/en19215161/s1, Supplementary Information Final, including the extended source-domain, domain-generalization, error-awareness, self-supervised-learning, and gray-box audit tables and figures; and Reproducibility Package V2 Final, containing analysis scripts, configuration files, audit records, machine-readable summary outputs, figure data, and SHA256 manifests.

Author Contributions: Conceptualization, X.W. and C.Y.; methodology, X.W. and C.Y.; software, C.Y.; validation, X.W. and C.Y.; formal analysis, X.W. and C.Y.; data curation, X.W. and C.Y.; visualization, X.W. and C.Y.; writing—original draft preparation, X.W. and C.Y.; writing—review and editing, X.W., C.Y., and L.L.; supervision, L.L.; project administration, X.W. and C.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study analyzed publicly released engineering data from a PEM electrolyzer and involved no human participants or animals.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Simulated Wind and Simulated Wave source datasets are publicly available through the National Laboratory of the Rockies Data Catalog [9,10]. The Historical Solar PV and Simulated Tidal datasets are available through the same catalog [11,12]. The Historical Wind control dataset is available separately in the NLR Data Catalog [13]. Analysis code, configuration files, audit tables, derived summary outputs, figure data, and file-integrity manifests are provided in the accompanying Reproducibility Package. The public PT-MELT artifact is available through the repository described in [14].

Acknowledgments: The authors used OpenAI ChatGPT (GPT-5.6 Sol) as an auxiliary tool for code drafting and checking, manuscript organization, English-language editing, and figure-formatting assistance. All tool-assisted material was reviewed and verified by the authors, who take full responsibility for the content of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sayed-Ahmed, H.; Toldy, Á.I.; Santasalo-Aarnio, A. Dynamic operation of proton exchange membrane electrolyzers—Critical review. *Renew. Sustain. Energy Rev.* **2024**, *189*, 113883. [CrossRef]
2. Endrődi, B.; Trapp, C.A.; Szén, I.; Bakos, I.; Lukovics, M.; Janáky, C. Challenges and opportunities of the dynamic operation of PEM water electrolyzers. *Energies* **2025**, *18*, 2154. [CrossRef]
3. Zeng, Q.; Yang, X.; Bu, E.; Liu, Z.; Ji, F.; Yuan, F.; Jin, Y.; Gao, X.; Li, C. Comprehensive experimental investigation of dynamic response to fluctuating load of proton exchange membrane water electrolyzer. *Int. J. Hydrogen Energy* **2025**, *103*, 310–326. [CrossRef]
4. Tawalbeh, M.; Sabri, M.; Kazim, H.; Al-Othman, A.; Alshraideh, H.; Almomani, F. Statistical and machine learning methods for modeling the next-generation PEM water electrolysis systems: Response surface methodology and artificial neural networks. *Energy Convers. Manag. X* **2026**, *29*, 101450. [CrossRef]
5. Long, J.; Cheng, W.; Lai, J.; Yu, H.; Yang, S. Graph-temporal neural network for accelerated, accurate prediction of dynamic multiphase fields in PEM water electrolyzers. *Int. J. Hydrogen Energy* **2026**, *216*, 153847. [CrossRef]
6. Zerrougui, I.; Li, Z.; Hissel, D. Physics-Informed Neural Network for modeling and predicting temperature fluctuations in proton exchange membrane electrolysis. *Energy AI* **2025**, *20*, 100474. [CrossRef]
7. Deng, S.; Sprangers, O.; Li, M.; Schelter, S.; de Rijke, M. Domain Generalization in Time Series Forecasting. *ACM Trans. Knowl. Discov. From Data* **2024**, *18*, 1–24. [CrossRef]
8. Li, Z.; Cai, R.; Fu, T.Z.J.; Hao, Z.; Zhang, K. Transferable Time-Series Forecasting Under Causal Conditional Shift. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 1932–1949. [CrossRef] [PubMed]
9. Abel, R.; Schwarz, M.; Leighton, D.; Starke, G.; Bay, C.; Wimer, N. Public Reference Data for Megawatt-Scale Hydrogen Electrolysis—Simulated Wind. In *NLR Data Catalog*; National Laboratory of the Rockies: Golden, CO, USA, 2025; updated 2026. [CrossRef]
10. Abel, R.; Schwarz, M.; Leighton, D.; Gevorgian, V.; Wimer, N. Public Reference Data for Megawatt-Scale Hydrogen Electrolysis—Simulated Wave. In *NLR Data Catalog*; National Laboratory of the Rockies: Golden, CO, USA, 2025; updated 2026. [CrossRef]
11. Abel, R.; Schwarz, M.; Leighton, D.; Nagasawa, K. Public Reference Data for Megawatt-Scale Hydrogen Electrolysis—NLR Historical Solar PV. In *NLR Data Catalog*; National Laboratory of the Rockies: Golden, CO, USA, 2026. [CrossRef]
12. Abel, R.; Schwarz, M.; Leighton, D.; Ross, H.; Zalkind, D.; Wiley, W. Public Reference Data for Megawatt-Scale Hydrogen Electrolysis—Simulated Marine Hydrokinetic Tidal Turbine. In *NLR Data Catalog*; National Laboratory of the Rockies: Golden, CO, USA, 2026. [CrossRef]
13. Abel, R.; Schwarz, M.; Leighton, D.; Nagasawa, K.; Wimer, N. Public Reference Data for Megawatt-Scale Hydrogen Electrolysis—NLR Historical Wind. In *NLR Data Catalog*; National Laboratory of the Rockies: Golden, CO, USA, 2026. [CrossRef]
14. National Laboratory of the Rockies. PT-MELT hydrogen electrolysis model repository. In *Hugging Face*; National Laboratory of the Rockies: Golden, CO, USA, 2025. Available online: https://huggingface.co/NatLabRockies/ptmelt-hydrogen-electrolysis (accessed on 29 September 2026).
15. Hoerl, A.E.; Kennard, R.W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
16. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
17. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30* , 3146–3154.
18. Geifman, Y.; El-Yaniv, R. Selective classification for deep neural networks. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4878–4887.
19. Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.V.; Lakshminarayanan, B.; Snoek, J. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 13991–14002.
20. Zerveas, G.; Jayaraman, S.; Patel, D.; Bhamidipaty, A.; Eickhoff, C. A Transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*; ACM: New York, NY, USA, 2021; pp. 2114–2124. [CrossRef]
21. Yue, Z.; Wang, Y.; Duan, J.; Yang, T.; Huang, C.; Tong, Y.; Xu, B. TS2Vec: Towards universal representation of time series. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 8980–8987. [CrossRef]

22. Karniadakis, G.E.; Kevrekidis, I.G.; Lu, L.; Perdikaris, P.; Wang, S.; Yang, L. Physics-informed machine learning. *Nat. Rev. Phys.* **2021**, *3*, 422–440. [CrossRef]
23. Carmo, M.; Fritz, D.L.; Mergel, J.; Stolten, D. A comprehensive review on PEM water electrolysis. *Int. J. Hydrogen Energy* **2013**, *38*, 4901–4934. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.