ISSN: 2458-8989

Natural and Engineering Sciences

NESciences, 2026, 11 (3): 845-866 doi: 10.28978/nesciences.263066

Biochemical and Electrochemical Degradation Mechanisms in Hydrogen Fuel Cells: Implications for Material Stability and Sustainable Energy Systems

Ge Lixia[1*], Mohammad Nizamuddin Inamdar[2], Aiman Alodainia[3]

[1]Faculty of Engineering, Lincoln University College, Petaling Jaya, Malaysia.
Email: xia02240108@163.com (lixia.phdscholar@lincoln.edu.my)

[2]Faculty of Engineering, Lincoln University College, Petaling Jaya, Malaysia.
Email: nizamuddin@lincoln.edu.my

[3]Faculty of Engineering, Lincoln University College, Petaling Jaya, Malaysia.
Email: aiman@lincoln.edu.my

Abstract

This study audits the public SANformer forecasting repository and evaluates its supplied checkpoints alongside three independently computed baselines on the same chronologically held-out rows of two version-identified processed PEMFC series (FC1: 1,155 hourly observations; FC2: 1,021). Six source and checkpoint combinations are identified by SHA-256 hashes, and 2,607 SANformer target forecasts are matched by target row to baseline forecasts. At one, three, and five hours, SANformer RMSE is 0.000676, 0.001210, and 0.001947 V for FC1, and 0.001304, 0.002405, and 0.004385 V for FC2. SANformer improves on last-value and linear-trend baselines in all six comparisons, whereas lagged Ridge has lower RMSE in five of six; SANformer has lower RMSE than Ridge only for FC2 at three hours. These checkpoint test results are distinct from the repository's archived prediction-segment outputs. Checkpoint training provenance is not established by file hashes or test alignment. Neither series has an observed end-of-life label, so no remaining-useful-life accuracy is reported. The degradation pathways behind these records are also examined. Radical attack on the perfluorosulfonic acid membrane, platinum dissolution, carbon-support corrosion, hardware corrosion, and fuel impurities are linked to activation, ohmic, and mass-transport losses. Stack voltage falls by 4.03% in FC1 and 5.78% in FC2, and the cell-voltage spread widens from 8.1 to 11.7 mV and from 11.4 to 17.1 mV, indicating gradual, uneven ageing that voltage data alone cannot assign to one mechanism. Implications for platinum use, fluorinated materials, and service life are discussed.

Keywords: PEMFC; total stack voltage; short-horizon forecasting; chronological evaluation; prediction archives; platinum dissolution; carbon-support corrosion; perfluorosulfonic acid membrane.

Available Online: 10/09/2026

1. Introduction

Hydrogen fuel cells are electrochemical devices which produce electricity in transportation and stationary power applications. They perform poorly in operation, however. Voltage can be impacted due to degradation of the catalyst, carbon corrosion, damage to the membrane, changes in gas transport, flooding, drying, contamination, and repeated operating transients. Other loadings such as dynamic

Corresponding Author: Ge Lixia, E-mail: lixia.phdscholar@lincoln.edu.my

loadings, start/stop cycles, humidity and temperature changes further increase the challenge for durability [1,2,3,4,5]. A stack could lose sufficient voltage or power before failing to a terminal such that its efficiency or operating margins are impacted. For the purpose of this paper, the biochemical (biological and bio-based) route(s) for producing hydrogen include dark fermentation and biogas reforming, and the impurities that may be introduced through these routes include contaminants in the fuel. The PEMFC does not contain any biological processes operating in the cell, and these impurities operate in the cell in the same chemical and electrochemical pathways as the other degradation modes discussed here (Section 4.2.5) such as catalyst poisoning and ionomer contamination.

Near-term forecasting and remaining useful life (RUL) estimation are two aspects of prognostics and health management. Voltage forecasting can predict the next operating interval. Another event necessary for RUL is an end-of-life (EOL) event defined, and forecast uncertainty assessed. Thus, accurate short-horizon voltage prediction does not lead to the accurate estimation of RUL. There are no documented run-to-failure events in public PEMFC datasets [6,7].

Recent studies have shifted from a single signal to multiple signals and hybrid models, which combine multiple signals combined together. Thus, the explicit state structure of particle filters and variants of Kalman filters is also important for their intuitive physical interpretation [6,7]. The use of LSTM, GRU, convolutional networks, attention models, transfer learning, Transformers, decomposition methods, and hybrid stochastic-neural methods is becoming more popular [8,9,10,11]. These models can represent nonlinear temporal relationships which are not easily represented analytically. There are challenges, however, in the validation of prediction flexibility for the benefits. Early and late ageing states can be mixed if the data is subjected to random cross validation. Smooth data can conceal very rapid recovery behavior. But sometimes models can seem complicated because they have highly correlated electrical channels but are never tested for generalization ability. However, information from the test period can also enter training, if no care is taken with data boundaries.

The public SANformer repository features two multivariate PEMFC ageing series and pre-processing and forecasting code, as well as checkpoints of the model and prediction files of the stored, preprocessed series. This study audits these artifacts, and assesses the provided checkpoints and three independent voltage-only baselines, using a set of matched chronological held-out test targets in processed files identified as versions. Prediction-segment scores from previous years are not included in the calculation. In Section 1.3, the research questions and conditional scope of RUL are given. The source repository can be found at https://github.com/yangyijunSdu/SANformer.

This is an analysis of the origin and analysis of past forecasts. The target is the total stack voltage and the inputs are electrical and operating measurements. The prediction files are checked separately and are not included in the repository's chronological test split. The remaining-useful-life rule with a threshold is mentioned as a potential future use, since there is no confirmed end-of-life event in these data.

1.1 Degradation signals and prognostic observability

PEMFC degradation occurs within the membrane, catalyst layer, gas-diffusion layer and at the interfaces of these components. This is influenced by its operating conditions. Damage to the membranes may be caused by chemical, mechanical and thermal action. Low hydration increases ionic resistance and high hydration decreases reactant transport. The catalyst layer can be influenced by platinum dissolution, particle growth, as well as support corrosion [1,2,3,4,5]. Such mechanisms are not regularly recorded as part of the stack. A voltage change may be due to irreversible ageing, reversible recovery from ageing, or a temporary change in load, temperature, pressure, humidity, or supply of reactants and is easily measurable. Section 4.2 sets out these chemical and electrochemical pathways in more detail and relates each one to the channels recorded in FC1 and FC2.

This is why health modelling has moved towards multivariate approaches. As work on dynamic health indicators has shown, voltage or power alone is not always sufficient when the operating point changes [8]. Relative power loss, polarization resistance, normalized voltage, and model-derived ageing states have been proposed to distinguish operational effects from gradual performance loss. Wang et al. have developed a nonlinear degradation model and particle filter for online RUL estimation under variable loading [6]. To reduce computational cost without sacrificing degradation-tracking accuracy, Ao et al. proposed a frequency-domain Kalman filtering method [7]. In such model-based approaches, an interpretable state equation is attractive, but its assumptions become critical when degradation rates change or when reversible recovery events are large.

1.2 Data-driven and hybrid prognostics

Neural sequence models consider a different part of the problem. The LSTM networks were adopted early due to their ability to keep the long term memory without being affected by the gradient instability of simple recurrent networks. In more recent studies, decomposition prior to LSTM, deep belief

networks, bidirectional recurrent units, WaveNet, convolutional feature extraction and attention [9,12,13,14] are used. Ma et al. presented a transfer-learning method by fully leveraging inter-cell differences to enhance degradation prediction in various fuel cells [9]. Xie et al. talked about uncertainty-aware deep learning and optimized the deep belief network structures [15,16]. The application of transfer learning to PEMFCs is especially important since experimental information on ageing is expensive, and a model created for one stack might require modification for use with another. There is no assurance of transferability, however. Differences in the actual design of the stack, operating procedure and sensor calibration can result in domain shift for the same nominal task.

The use of structured hybridization has become a popular approach in the literature. Tian et al. combined voltage-recovery modelling with the Bayesian optimisation of a multi-kernel relevance vector machine [17]. Hu et al. combined a Wiener process, state estimation and Bayesian GRU to take into consideration temporal uncertainty, individual uncertainty, and measurement uncertainty [14]. To enhance voltage and RUL prediction, Izadi et al. employed a WaveNet-GRU model and transfer learning techniques [13]. Kebede et al. fused features obtained by the variational autoencoder with a bidirectional LSTM with attention and combined the knee-point detection with transfer learning [18]. This important assumption of these studies is that the degradation is not a stationary series. It encompasses unit-to-unit variation, effects of operating conditions, trends, and recovery effects. Either explicitly divide these into separate components or learn them robustly from the data.

Relationships between far away observations can be represented by attention-based sequence models. Transformer variants have been employed in PEMFC studies to include voltage recovery [11], screen input variables [19] and model longer sequences or health states [20,21,22]. An attention signal can indicate dependencies between electrical and thermal signals, pressure and flow signals and so on, but it does not create a degradation mechanism. The weights can also be adjusted according to correlated measurements and operating protocols. A closely related sequence model to the degradation-trajectory forecasting is the diffusion-based sequence model [23].

The variations in operating conditions are particularly relevant for fuel-cell vehicles. According to the literature on automotive degradation, dynamic loads, idling, cold starts, thermal or humidity excursions and contamination are stressors that trigger the various failure pathways [3]. Experimental tests conducted over a long period of time have demonstrated dependence of the degradation rate of the stack on the current degradation level and the driving cycle [24,25]. The observations made here drive the importance of including humidity, flow, temperature, pressure, current density and current in a forecasting model. They also decrease the validity of analysis when data are randomly divided. If so, then there is no difference in the data of the same ageing trajectory between the training and test sets, and a model can take advantage of local continuity rather than learn a transferable degradation relationship.

1.3 Research gap, originality, and research questions

It is also mentioned that there is a general lack of gap between degradation forecasting and RUL estimation in the literature. Typically, in the field of RUL studies, a health indicator is defined and an EOL boundary is defined and then a model is propagated until the EOL boundary is reached [6,8,14,17]. Uncertainty-aware studies do not report a crossing time, but rather an interval [14,15,26,27] or a distribution of crossing times. This is better since the length of the horizon makes the point forecasts unstable. Public datasets, however, typically stop when a failure threshold that has a physical meaning has not yet been reached. If the last sample is an EOL event, then it is valid to subtract each sample's time from the previous sample's time to assign RUL labels. Using an arbitrary end of testing as a failure will throw off labels and both training and testing.

3 gaps in the literature are identified. The first is the mismatch between the rich multivariate ageing signals and the sparse explicit ageing labels. The second is a validation gap: whereas high one-step accuracy is often observed, the decrease in performance at longer horizons is given less attention. The third is a gap in reproducibility. Published methods might include architectures and pre-processing, but code paths to generate the evaluation samples and scale the variables are not always obvious. Here, the third gap is addressed by using a public repository that shares the artifacts from the preprocessing as well as the prediction. It also gives the chance to directly study the first two gaps, so long as the analysis does not exceed what is possible in the data. These gaps align with those found in recent work on uncertainty-aware and health monitoring tools [14,21,28].

The study is based on three questions. RQ1 investigates the change in archived forecast error at different forecast horizons (1h, 3h, and 5h) in FC1 and FC2. RQ2 is whether the published data and prediction files are able to back up the reported evaluation design. The comparison of the checkpoints supplied to the process with last-value, linear-trend, and lagged Ridge baselines on the same chronological held-

out target rows is requested by RQ3. Only extension is considered as remaining useful life, no RUL accuracy is reported.

In parallel, the main chemical and electrochemical degradation routes in PEMFCs and their impact on material degradation and the loss of voltage are explored (Section 4.2 and 4.3), along with what these two trajectories can and cannot reveal about PEMFCs.

There is a gap between laboratory forecasting and operational decision support. Numerous studies evaluate a model using one degradation path or a few stacks typically with a controlled protocol. These datasets can be used for controlled model development but a deployed model is subject to effects from sensor drift, stack-to-stack manufacturing variation, maintenance actions, ambient changes, and changes in the control policy. Predicators that are successful with one protocol might not be successful with another. Temporal modeling can be evaluated in a lab setting with laboratory data. They are not sufficient to provide much information on their own regarding performance in a larger stack population or failure risk of a fleet. The two series in the present collection are to be understood within that scope. Long-duration vehicular and dynamic-load tests [3,20,21] suffer from the same problems.

Another deficiency is the lack of documentation about EOL in datasets. Even though experiments may be terminated due to logistical, funding, safety or scheduling considerations, the last recorded observation is often considered the termination of the useful life. If there is no pre-established failure criterion, and it is documented, then the last date is not the end of the observation. This separation alters the learning issue. With a true run-to-failure data set, supervised RUL labels can be obtained. A censored ageing dataset can be used to perform survival or threshold-extrapolation analysis (at best) and degradation forecasting. If such designs can be handled as interchangeable, the impression is created that there is a full RUL benchmark based on a set of labels that have no engineering significance. This boundary is defined by the present study, rather than added as a restriction at the end of the evaluation of the model. Defensible RUL estimation is thus a key element of an explicit threshold definition [6,14,18].

Last but not least, in this context, explanation should be understood as predictive attribution and not as mechanistic evidence. The measured channels on which the model relies can be identified using permutation importance, grouped ablation, attention summaries, and SHAP values. They cannot demonstrate that altering a variable would alter the degradation in the same direction because operating controls are interdependent and may be affected by potential latent system condition(s). An extensive explainability analysis should, consequently, be based on physically meaningful feature groups and temporally blocked perturbations. The incremental predictive value of grouped ablation doesn't suggest a causal sequence. This gives a more defensible connection between the use of machine learning for interpretation and fuel-cell engineering. Recently, feature screening or attribution has been used to perform this task for PEMFC research [19,23].

2. Materials and methods

2.1 Data source and study design

This study uses the public SANformer PEMFC ageing repository. The supplied processed CSV snapshot contains 1,155 FC1 rows and 1,021 FC2 rows. Their respective SHA-256 hashes are db1685ec6b2b7b8623bae972117f145c1ce830afb55373fd8021adb2f56eafcf and 94ed5e097de20ba87d976a25fa731989dab327129281cc6fd54c1207e563e49e. The model, training code, checkpoints, and archived neural-network predictions originate from that repository. Independent baseline predictions and an execution of the supplied SANformer checkpoints on the same held-out test target rows were evaluated in this study. The evaluation package supplies six checkpoint SHA-256 hashes, target-row predictions, and summary metrics. Matching hashes establish file identity, but cannot independently establish when or on which rows the checkpoints were trained.

The processed data contain 25 columns: time and 24 measured or derived channels. Electrical variables comprise individual cell voltages (U1–U5), total stack voltage (Utot), current density (J), and current (I). Thermal variables include hydrogen, air, and coolant inlet and outlet temperatures. Pressure variables include air and hydrogen inlet and outlet pressures. Flow variables include hydrogen inlet and outlet flow, air inlet and outlet flow, and water flow. Air relative humidity is also recorded at the fuel-cell side. The forecasting target is Utot. The repository implementation uses the other 23 non-time channels as predictors and includes preceding Utot values in the input sequence. Table 1 defines the variable groups. The variable groups used in the analysis are defined in Table 1.

Table 1. Variable definitions and analytical roles.

Variable(s)	Measurement / description	Analytical role
Time	Hourly operating-time index after preprocessing	Index / sequencing variable
U1-U5	Five cell-voltage channels	Electrical predictors; cell-level performance history
Utot	Total stack voltage	Primary dependent forecasting target; historical values also enter the input sequence
J	Current density	Electrical operating predictor
I	Stack current	Electrical operating predictor
TinH2, ToutH2	Hydrogen inlet and outlet temperatures	Thermal operating predictors
TinAIR, ToutAIR	Air inlet and outlet temperatures	Thermal operating predictors
TinWAT, ToutWAT	Coolant-water inlet and outlet temperatures	Thermal operating predictors
PinAIR, PoutAIR	Air inlet and outlet pressures	Pressure operating predictors
PinH2, PoutH2	Hydrogen inlet and outlet pressures	Pressure operating predictors
DinH2, DoutH2	Hydrogen flow channels	Reactant-flow predictors
DinAIR, DoutAIR	Air flow channels	Reactant-flow predictors
DWAT	Coolant-water flow	Thermal-management predictor
HrAIRFC	Air relative humidity at fuel-cell side	Humidity predictor

2.2 Data preprocessing

The repository preprocessing script converts time to numeric values, removes invalid times, sorts records, and retains the last duplicate. It interpolates numeric channels to an hourly grid and applies LOWESS smoothing. The FC1 CSV in the supplied source snapshot has 1,155 records from 0 to 1154 h, consistent with the grid in the script. FC2 has 1,021 records from 0 to 1020 h. All newly computed baseline results and the dataset plot use the 1,155-row FC1 file identified by its hash in Section 2.1.

Each numeric signal is then smoothed individually using LOWESS. The smoothing fraction is 20 divided by the series length for FC1 and 10 divided by the series length for FC2, which corresponds to relatively local smoothing windows. Smoothing can improve the estimation of slow degradation trends, but it can also dampen genuine rapid events. Smoothed signals are therefore suitable for trend forecasting, whereas high-frequency residuals should be retained or analysed separately when the aim is to detect fault onset. In the present article, smoothed changes are not used to attribute the voltage decline to any particular degradation mechanism. The repository preprocessing steps are described in Section 2.2.

2.3 Temporal partitioning and scaling

The source code defines chronological training, validation, and test portions of 50:10:40. In the version-identified FC1 file (n = 1,155), these contain 577, 115, and 463 rows, with test targets beginning at row 692. FC2 (n = 1,021) contains 510, 102, and 409 rows, with test targets beginning at row 612. The repository test loader uses the preceding input window for context. Feature and target MinMax scalers are fitted on training rows, whereas the repository's time scaler is fitted on all timestamps. A separate prediction loader starts at row 578 for FC1 and row 509 for FC2, both before the corresponding test boundaries. The repository saves prediction-loader outputs, rather than test-loader outputs, in its predictions.csv files. Thus its archived scores are not an independently verified held-out evaluation.

Input windows contain 24 hourly steps (24 hours) for FC1 and 12 hourly steps (12 hours) for FC2. Forecast lengths are one, three, or five hourly steps. Comparisons of archived forecast errors are made within each series.

2.4 Forecasting architecture

The archived repository implements a spatio-temporal forecasting model with 24 input channels, model dimension 128, two attention heads, two encoder layers, feed-forward dimension 256, dropout 0.11, and moving-average decomposition kernel 5. Its code combines embedding, decomposition, attention, recurrent factor learning, and projection to future total stack voltage. These are properties of the published implementation, not a newly trained architecture in this study.

Electrical histories include cell voltages and total stack voltage. Temperature, pressure, flow, current, and humidity channels describe operating conditions. These inputs support voltage prediction; they do not establish that reversible operational changes have been separated from irreversible degradation.

2.5 Evaluation metrics

The errors made in forecast are presented in terms of root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE) and coefficient of determination (R2). Their definitions are given below. The predicted values and the residuals are looked at the row level in the archived prediction files.

The numerical model scores are taken from the repository's archived metrics and predictions CSV files. The model, training code and saved outputs were provided by the repository. This study independently verified the file counts, the time grids for processing, the boundaries for the predictions in the loaders and the patterns of the forecast errors, and graphed the saved row level predictions and residuals. It was not retrained and the archived scores were not retested; the supplied checkpoints were tested separately (Sections 2.8 and 3.5). The predictions in FC1 archive do not necessarily constitute a single, fully reconciled execution of the FC1, but only the predictions in the currently processed FC1.

$RMSE = sqrt[(1/n)\ \Sigma\ (y_i - \hat{y}_i)^2]$

$MAE = (1/n)\ \Sigma\ |y_i - \hat{y}_i|$

$MAPE = (100/n)\ \Sigma\ |(y_i - \hat{y}_i) / y_i|$

$R^2 = 1 - \Sigma(y_i - \hat{y}_i)^2 / \Sigma(y_i - \bar{y})^2$

2.6 Potential threshold-based RUL extension

An independently determined end-of-life (EOL) threshold and an independently confirmed endpoint would be needed for a future RUL analysis. Given a fixed reference voltage V_ref and a fractional threshold θ, a predicted voltage path could be a voltage path that predicts a conditional first crossing for the smallest future time step k for which $\hat{V}_(t+k) \leq \theta V_ref$. This manuscript doesn't choose θ, measure crossing times, or assess the accuracy of RULs. The performance of RULs can only be obtained by short-horizon voltage forecasts.

A crossing time for a forecast horizon may not be reported if the forecast does not cross a defined threshold during the horizon. This framework would facilitate maintenance decisions if the follow-up (after) prediction, uncertainty calibration, and independently observed EOL events were available.

$SOH_t = V_t / V_ref$

$RUL_t = inf\{k > 0 : \hat{V}_(t+k) \leq \theta\ V_ref\}$

2.7 Robustness, reliability, and reproducibility checks

Horizon sensitivity is used to measure robustness. The larger the forecast lead time, the greater the local inaccuracy of the model should be. Therefore, the RMSE, MAE, MAPE and R2 are calculated separately for each FC1 and FC2. This offers a more robust in-house stress test than just looking at one best-case metric. The 6 aggregate metric files are available, and there are no replicate runs or bootstrap distributions available, so a statistical significance test is not applied. Pseudo-replication would occur if the p-values were computed from individual runs which have been archived.

The physical sensors are not interchangeable items on a psychometric scale and therefore Cronbach's alpha is not applicable. The numbers presented in this section refer to lengths of files, times on the regular hour, missing values in published processed CSV files, split boundaries, and the origin of the saved prediction files. These checks are not causing a degradation signal. Table 2 contains the data and output checks.

Table 2. Data-quality, temporal-validity, and reproducibility diagnostics.

Diagnostic	Implementation / status	Interpretation
Temporal ordering	Chronological 50:10:40 split	Defined in code, but archived CSVs are exported from a separate prediction segment.

Diagnostic	Implementation / status	Interpretation
Sampling grid	Hourly interpolation	Regular sequence spacing; introduces temporal dependence and interpolation assumptions.
Duplicate time handling	Keep final duplicate before interpolation	Documented in preprocessing code.
Feature scaling	MinMax fit on training features	Appropriate; test feature extrema do not set scaling parameters.
Target scaling	MinMax fit on training Utot	Appropriate; future target values do not set target scaling.
Time scaling	Scaler fit on full timestamp vector	Minor reproducibility caveat; strict rerun should fit on training timestamps.
Psychometric reliability	Cronbach alpha not used	Not applicable to physical sensor channels.
Correlation matrix	Not numerically reported here	Full-series execution required; raw trend correlation would be confounded by autocorrelation.
Predictive validity	RMSE, MAE, MAPE, R2 across 1/3/5 steps	Describes archived prediction-segment outputs; held-out test validity is unverified.
RUL validity	No observed EOL label	Threshold-based RUL remains conditional until an external EOL rule is supplied.

2.8 Independent chronological test baselines

Three baselines were evaluated on the hashed processed FC1 and FC2 files: persistence of the last observed total voltage, linear extrapolation from the first and last values in the input window, and a lagged Ridge regression. FC1 uses 24 prior hourly voltage values, and FC2 uses 12. Each model directly predicts the next one, three, or five values. Training targets end before the 50% boundary; validation targets end before the 60% boundary; test targets begin at 60%. Ridge inputs are standardized using training windows only. Its regularization parameter was selected by validation mean squared error from {0.001, 0.01, 0.1, 1, 10, 100}. Test origins are spaced by the forecast horizon so exported target rows do not repeat. RMSE, MAE, MAPE, and R^2 are calculated after flattening each horizon's test predictions. These newly computed baselines use neither the SANformer checkpoints nor its archived prediction-segment scores. The supplied SANformer checkpoints are subsequently evaluated on the same origin and target indices as these baselines; model metrics use flattened target-level errors. Matching target indices establishes a direct test-set comparison of these forecasts, subject to the checkpoint-training provenance limitation described in Section 4.4.

3. Results

3.1 Dataset profile

The supplied, version-identified processed files contain 1,155 FC1 hourly observations from 0 to 1154 h and 1,021 FC2 observations from 0 to 1020 h. Together they contain 2,176 time points, analyzed as separate sequences. Neither CSV has missing values. Their first total-stack-voltage values are 3.34784 V and 3.33711 V. Figure 1 plots these files and their chronological split boundaries. Table 3 reports the file profiles. New baseline calculations use these same hashed files.

Table 3. Verified dataset profile and forecasting configuration.

Series	Processed N	Time span	Grid	Initial Utot	Input context	Forecast horizons
FC1	1,155	0–1154 h	1 h	3.34784	24 h	1, 3, 5 h

Series	Processed N	Time span	Grid	Initial Utot	Input context	Forecast horizons
FC2	1,021	0-1020 h	1 h	3.33711	12 h	1, 3, 5 h

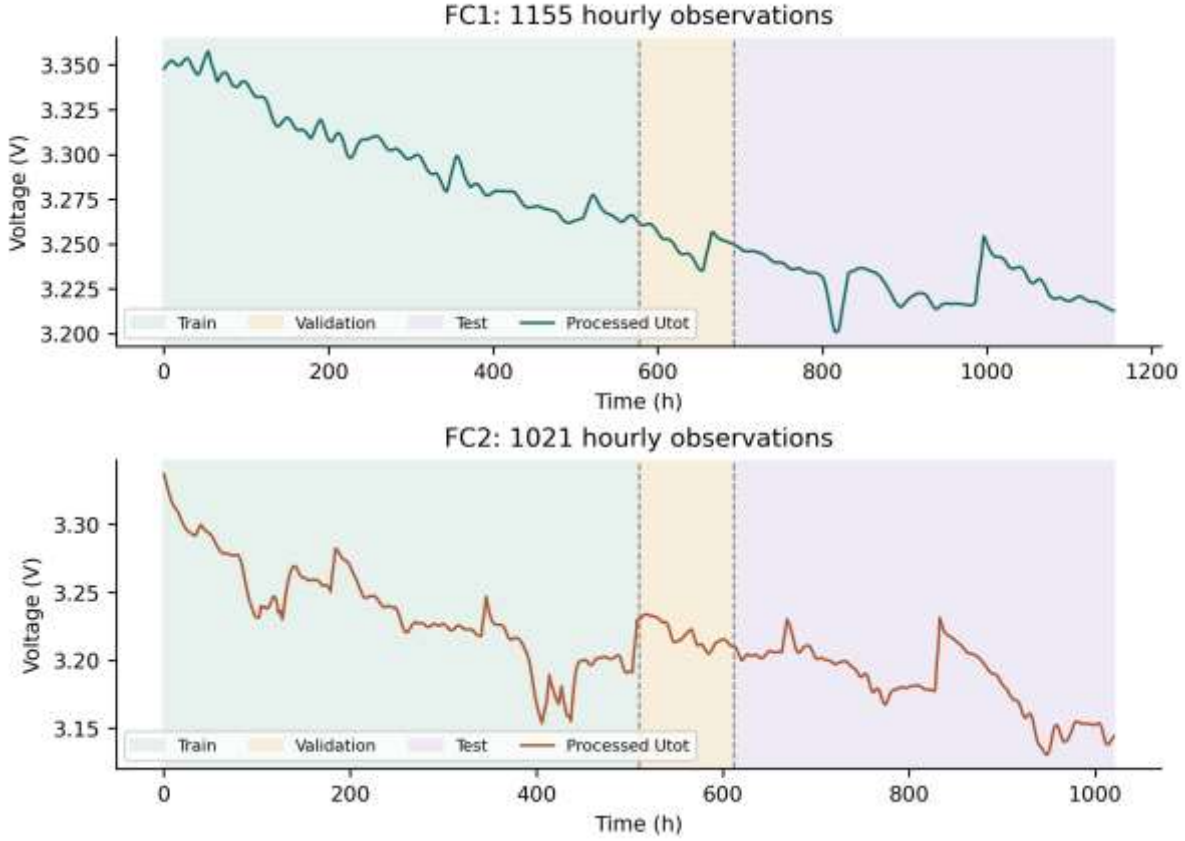

Figure 1. Processed total stack voltage for the supplied 1,155-row FC1 and 1,021-row FC2 snapshot. Shading marks chronological training, validation, and test portions calculated from the published split code.

3.2 Archived multi-horizon prediction performance

The archived single prediction CSVs have 577 rows of FC1 and 512 rows of FC2. Their saved RMSE values are 0.000618 and 0.001196 V, respectively, and corresponding R^2 values are 0.998045 and 0.997971. In FC1, 577 is the length of the source (1,155)-the published prediction (578). This fixes the reported row-count mismatch issue in this snapshot, but doesn't imply full training and checkpoint provenance. The prediction starts both before the respective test boundaries, meaning that the saved metrics can not be called held-out test accuracy. Figures 2 and 3 show the archived one-step values and residuals. Archived error increases with forecast length. FC1 RMSE is 0.001109 V at three steps and 0.001711 V at five steps. FC2 RMSE rises from 0.001196 V at one step to 0.002474 V at three steps and 0.004830 V at five steps. These values are descriptive results for the saved prediction-segment outputs. Table 4 contains all archived metrics.

To examine rapid changes in the saved one-step outputs, the absolute change between successive observed voltage rows was calculated within each series. Rows in the upper decile of this change were compared with the remaining rows. FC1 contained 58 rapid-change rows: RMSE was 0.001551 V, compared with 0.000395 V for other rows. FC2 contained 52 rapid-change rows: RMSE was 0.003042 V, compared with 0.000739 V for other rows. The grouping and errors were calculated from the archived prediction CSVs. Because the files contain no event annotations and are not verified held-out test outputs, these results describe forecast sensitivity to rapid observed changes, not performance during confirmed recovery events.

Table 4. Archived forecasting metrics for total stack voltage (Utot)

Series	Horizon	Archived output rows	RMSE	MAE	MAPE (%)	R2
FC1	1	577	0.000618	0.000296	0.009143	0.998045
FC1	3	576	0.001109	0.000520	0.016102	0.993690
FC1	5	575	0.001711	0.000846	0.026185	0.984963
FC2	1	512	0.001196	0.000659	0.020663	0.997971
FC2	3	510	0.002474	0.001525	0.047794	0.991237
FC2	5	510	0.004830	0.002557	0.080165	0.966609

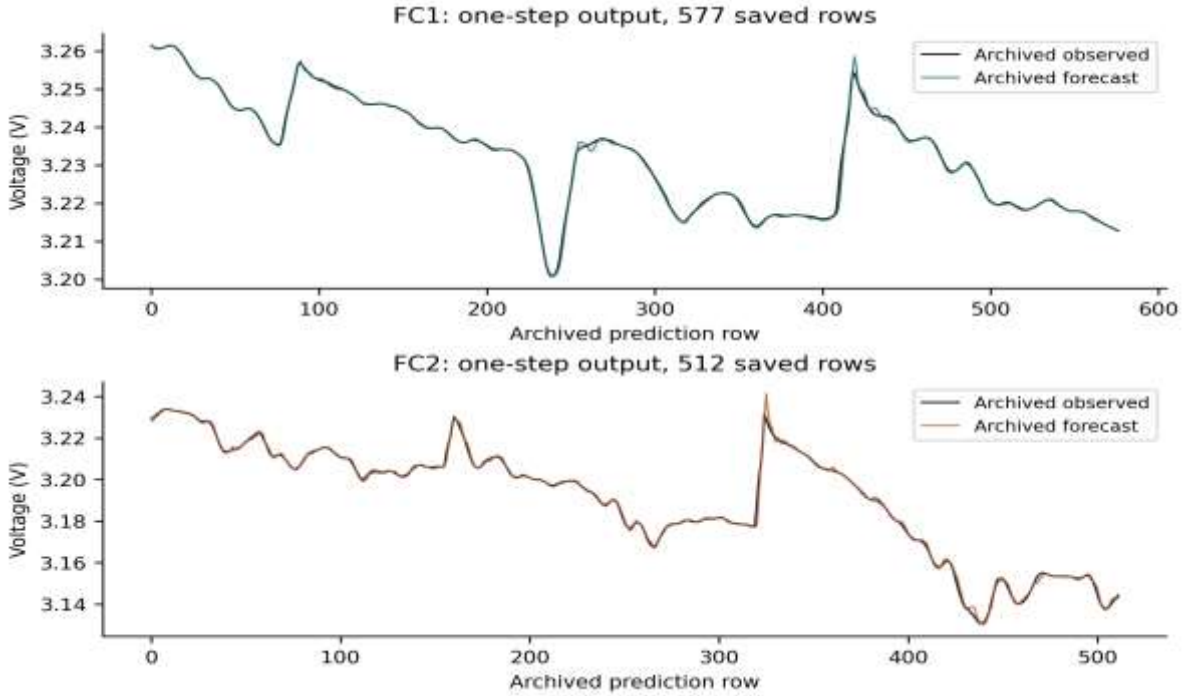

Figure 2. Observed and forecast total stack voltage in the repository's one-step prediction CSVs. The horizontal axis gives saved prediction-row order, not a verified held-out test index.

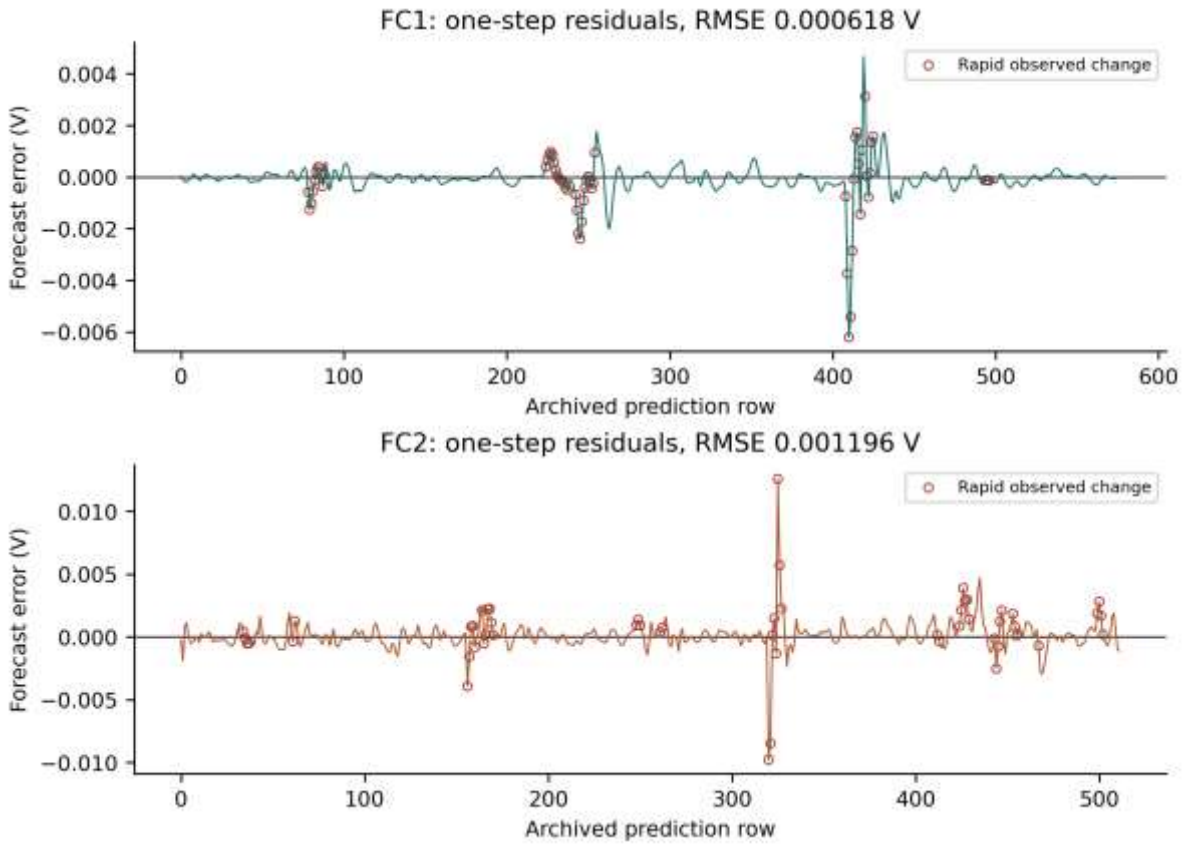

Figure 3. One-step archived residuals (forecast minus observed voltage). Markers identify rows in the top decile of the absolute change in observed voltage relative to the preceding saved row; these are rapid changes, not verified recovery events. The archived files are not verified held-out test results.

3.3 Robustness and horizon sensitivity

From one to five steps, archived FC1 RMSE rises by 176.9% and archived FC2 RMSE by 303.8%. These within-archive changes are descriptive because the saved files come from the prediction segment. Figure 4 displays the archived horizon-wise values. The independent test baselines are reported separately in Section 3.4; no physical cause is assigned to differences between the series.

RQ1 is answered descriptively: archived prediction error rises as the horizon increases. This pattern does not establish held-out generalization, resistance to recovery events, or a causal account of fuel-cell ageing.

RQ2 identifies an evaluation distinction: the repository saved forecast files originate from a prediction segment that starts before its test boundary, while this study's newly calculated baselines use only the test segment of the hashed processed files. The FC1 archived row count agrees with the 1,155-row snapshot, but the original checkpoint provenance has not been independently verified. Table 5 reports archived horizon sensitivity, not RUL accuracy.

Table 5. Horizon sensitivity from one-step to five-step forecasting.

Series	RMSE change	MAE change	MAPE change	R2 change	Interpretation
FC1	+176.9%	+185.8%	+186.4%	-1.31 percentage points	Error rises materially, but R2 remains above 0.98.
FC2	+303.8%	+288.0%	+288.0%	-3.14 percentage points	Stronger horizon sensitivity; long-horizon caution is required.

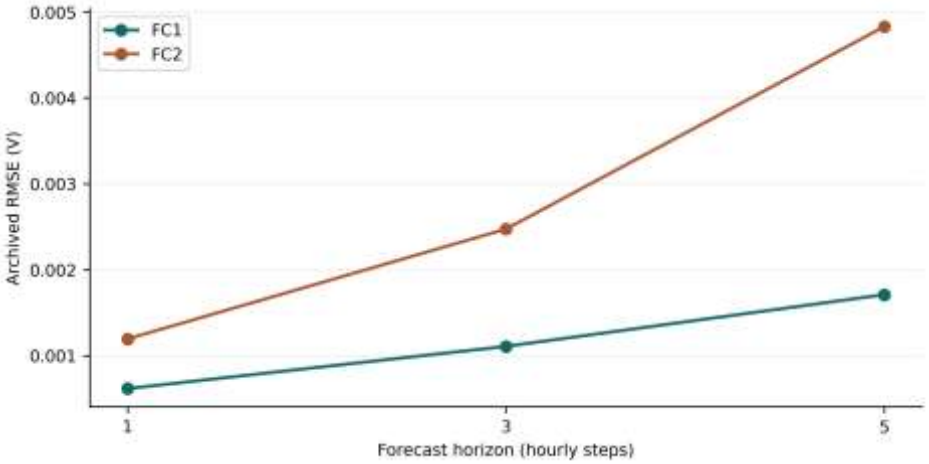

Figure 4. RMSE recalculated from the archived prediction CSVs across one-, three-, and five-step outputs. Each horizon's saved file contains overlapping output rows and is not an independent held-out test set.

3.4 Independent held-out baseline results

To address RQ3, Table 6 reports the newly calculated chronological test baselines; Figure 5 shows one-hour Ridge trajectories and Figure 6 shows baseline RMSE across horizons. Each SANformer checkpoint was then evaluated on the corresponding baseline origin and target rows. Table 7 and Figures 7–8 present this matched evaluation. The archived repository prediction-segment metrics are separate and are not used in these rankings.

Table 6. Newly computed chronological held-out test RMSE (V) for three baseline models. Each row uses the hashed processed file and nonoverlapping target windows; Ridge alpha was selected on the validation portion.

Series	Horizon (h)	Test origins	Test values	Last value RMSE	Linear trend RMSE	Ridge lag RMSE
FC1	1	463	463	0.000883	0.000935	0.000281
FC1	3	154	462	0.001838	0.002008	0.000931
FC1	5	92	460	0.002882	0.003283	0.001481
FC2	1	409	409	0.001724	0.001831	0.001085
FC2	3	136	408	0.003266	0.003822	0.002772
FC2	5	81	405	0.004607	0.005259	0.003500

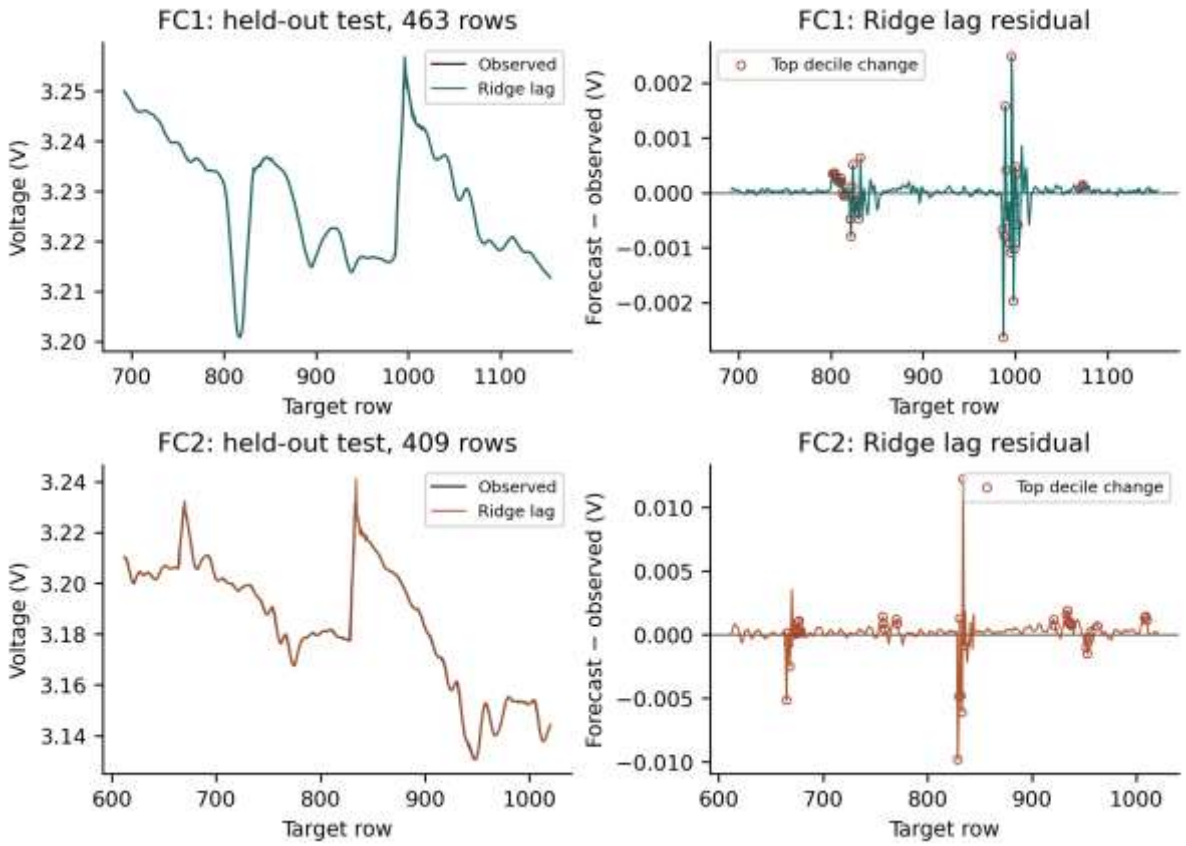

Figure 5. Independent one-hour held-out test forecasts from the lagged Ridge baseline for FC1 and FC2, with time-ordered residuals. Markers indicate the upper decile of absolute changes in observed voltage between adjacent test rows, not confirmed physical recovery events.

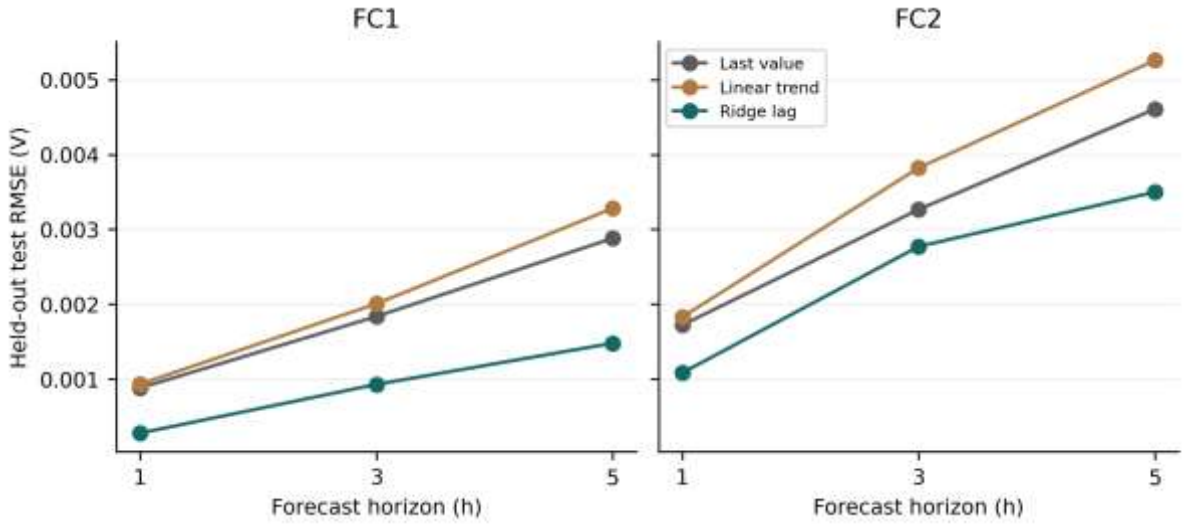

Figure 6. Newly computed held-out test RMSE of the last-value, linear-trend, and lagged Ridge baselines at one-, three-, and five-hour horizons. The matched SANformer results appear in Table 7 and Figures 7–8; archived prediction-segment scores are separate.

3.5 Matched SANformer checkpoint evaluation

The six supplied checkpoints were run against the version-identified processed series. For each series and forecast horizon, SANformer and all three baselines share the same forecast origins and test target rows (FC1: 463, 462, and 460 target values at one, three, and five hours; FC2: 409, 408, and 405). There are 2,607 matched SANformer target forecasts in total. Source and checkpoint SHA-256 values, row identifiers, observed values, forecasts, and residuals accompany the evaluation. Recalculation from row-level predictions reproduces Table 7 to numerical rounding. File and row alignment validate the reported comparison, while checkpoint training history remains an independent limitation.

Table 7. Matched chronological held-out test evaluation of supplied SANformer checkpoints. RMSE and MAE are in volts; MAPE is in percent. Ridge RMSE is included as the strongest baseline in five of six cases. Each model is evaluated on identical target rows.

Series	Horizon (h)	Origins	Targets	SANformer RMSE (V)	Ridge RMSE (V)	SANformer MAE (V)	SANformer R^2
FC1	1	463	463	0.000676	0.000281	0.000323	0.9966
FC1	3	154	462	0.001210	0.000931	0.000574	0.9890
FC1	5	92	460	0.001947	0.001481	0.000930	0.9715
FC2	1	409	409	0.001304	0.001085	0.000708	0.9972
FC2	3	136	408	0.002405	0.002772	0.001438	0.9905
FC2	5	81	405	0.004385	0.003500	0.002523	0.9680

SANformer has lower RMSE than last-value and linear-trend forecasts in all six series-horizon comparisons. Ridge has lower RMSE than SANformer at every FC1 horizon and FC2 one and five hours; SANformer has lower RMSE than Ridge at FC2 three hours (0.002405 versus 0.002772 V). The accompanying matched_sanformer_test_metrics.csv reports SANformer MAPE from 0.0100% to 0.0793%; Table 7 reports R^2 from 0.9680 to 0.9972. These results support a model-specific forecast comparison on these processed series, rather than a general advantage for either architecture.

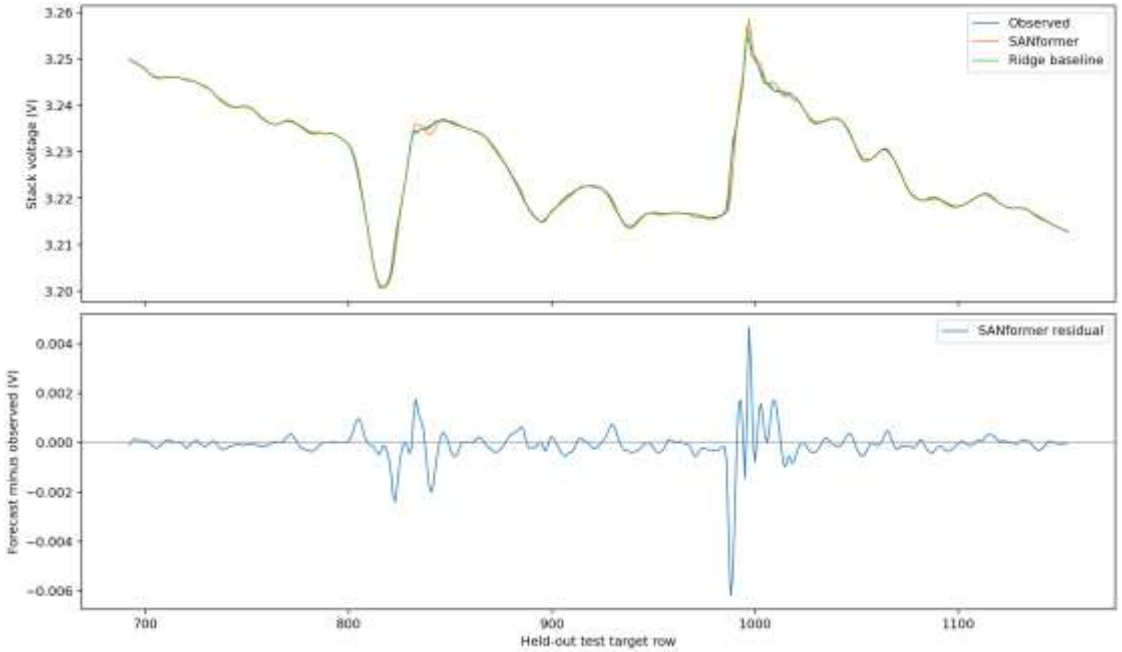

Figure 7. One-hour matched held-out test observations, supplied SANformer checkpoint forecasts, lagged Ridge forecasts, and SANformer residuals for FC1. The residual is forecast minus observed voltage; the same test target rows are used for both methods.

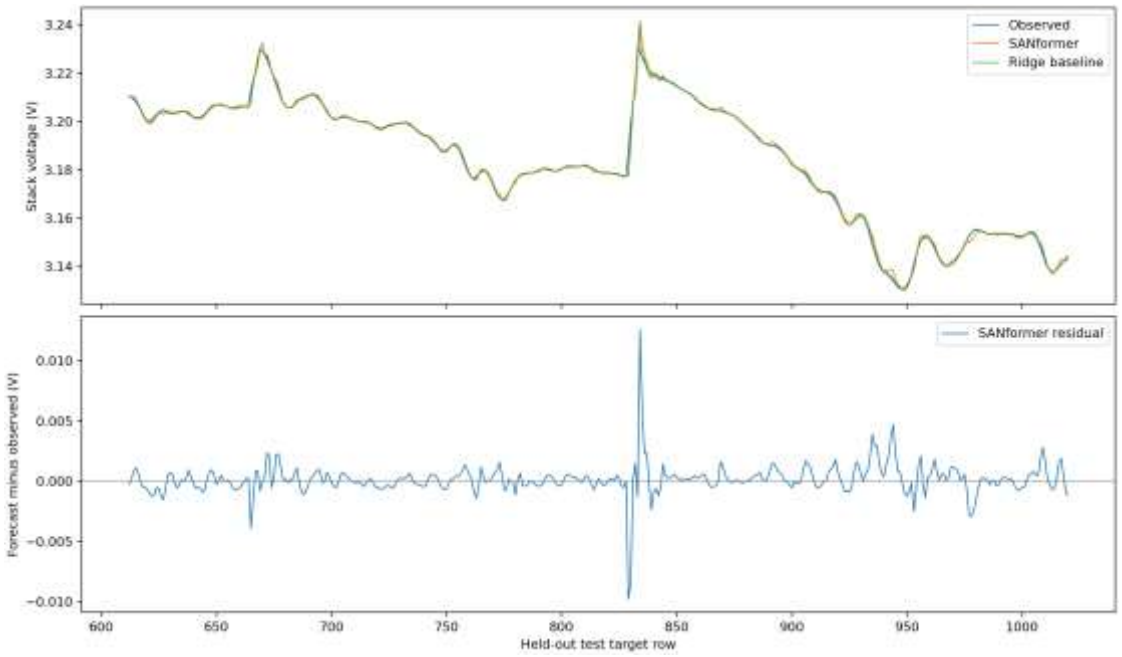

Figure 8. One-hour matched held-out test observations, supplied SANformer checkpoint forecasts, lagged Ridge forecasts, and SANformer residuals for FC2. The residual is forecast minus observed voltage; the same test target rows are used for both methods.

For a sensitivity check on held-out one-step Ridge forecasts, rapid changes were defined as the upper decile of the absolute observed voltage difference from the preceding test row. FC1 has 47 such rows; RMSE is 0.000771 V in these rows versus 0.000144 V elsewhere. FC2 has 41 rapid-change rows; RMSE is 0.003097 V versus 0.000490 V elsewhere. The test files do not identify physical recovery events, so this comparison does not measure recovery-event accuracy.

4. Discussion

4.1 Interpretation of the forecasting evidence

The R^2 of the model is not necessarily the most informative result. It's more instructive to consider how performance varies with horizon. The accuracy is very good in both series with increasing errors at 3 and 5 hours. Recursive and multistep prognostics should obey this pattern, as the further away from the last observation, the higher the uncertainty. However, the size of the increase is important when deploying in FC2. It is important to realize that the error observed might be sufficient for a controller who only needs one hour-ahead support, but might be much more difficult for a maintenance planner who needs to obtain a long-range RUL estimate. Uses should not be compared.

The length of the sequence, the starting point, operating conditions, and the fraction of training are reported as sensitive parameters for PEMFC forecasting studies published [6,9,10,11]. Based on transfer-learning results, the accuracy of one stack may not be transferred to another [9,13,18]. The studies of forecast error also reveal the rationale for the diminishing informativeness of a point forecast at longer horizons [14,15,26,27]. Although these RMSE and MAPE values are high with long horizons, they increase as horizon length increases in the archived SANformer outputs. If you just reported one-step accuracy, you would miss out on that pattern.

The FC2 archived outputs are subject to careful interpretation. They have a 5-step RMSE that is over 4 times as large as the one-step, but the R^2 is still high. This is possible because of differences in the operating variability or in the length of files in the input window, but the existing files cannot separate either explanation. These forecasts are not sufficient to determine a physical cause. This would necessitate operating events with annotations or an identification design where the operating events are separated to show the relevant stressors.

4.2 Relationship to degradation mechanisms and operating conditions

The operating state is described by cell voltages, current, current density, temperature, pressure, flow and humidity. Such variables could be used for predicting near-term stack voltage, however, the present analyses do not account for voltage responses due to reversible performances or irreversible component ageing [1,2,3,4,5]. Historical cell voltages are also strongly correlated to the overall required target stack voltage, and do not in and of themselves provide a mechanism.

The purpose of the pre-processing influences the forecasting task. LOWESS smoothing is suitable for highlighting the gradual changes in voltage and may help smooth out short voltage fluctuations. Hourly interpolation smooths the time grid, while LOWESS smoothing emphasizes gradual voltage changes and may help to smooth out short voltage fluctuations. It is not possible to prove accuracy on raw, real time, sensor readings. For online use, a separate evaluation on the raw or lightly filtered data is required. STVRs may be due to reversible recovery, not to component health recovery [11].

A repository is used to define chronological training, validation and testing splits. Its archived prediction CSVs are exported from a separate segment which is placed before the test boundary in both series and its time scaler is fitted on all timestamps. The scores are thus archived for that prediction segment. The newly evaluated checkpoints are evaluated with the specified test target rows, but file checks do not ensure that those rows were not included in the checkpoint training and selection.

4.2.1 Voltage-loss components as a link between mechanisms and measurements

Degradation becomes visible in a stack record through the voltage balance of each cell. Under load, the cell voltage can be written as

$$V_{cell} = E_{rev} - \eta_{act} - j \cdot R_{\Omega} - \eta_{mt}$$

where E_{rev} is the reversible potential at the prevailing temperature, pressure, and humidity; η_{act} is the activation overpotential, which in a hydrogen–air cell is dominated by the oxygen reduction reaction at the cathode; j is current density; R_{Ω} is the area-specific ohmic resistance of the membrane, ionomer, and contact interfaces; and η_{mt} is the loss caused by restricted transport of oxygen and water. Each pathway described below raises one or more of these terms, and Utot records their sum over five cells [2,29]. In both processed files, current density stays between 0.699 and 0.706 A cm^{-2}, so the decline in Utot cannot be explained by a shift in the load point. Table 8 summarizes the pathways, the loss components they mainly affect, and the recorded channels that bear on them.

4.2.2 Chemical and mechanical degradation of the membrane

Chemical attack on the perfluorosulfonic acid (PFSA) membrane begins with hydrogen peroxide. Peroxide forms when oxygen is reduced through a two-electron route, $O_2 + 2H^+ + 2e^- \rightarrow H_2O_2$, either at the cathode or at the anode, where oxygen that has crossed the membrane meets platinum held near the hydrogen potential. Trace Fe^{2+} or Cu^{2+}, released for example from corroding hardware, converts peroxide into hydroxyl radicals through the Fenton reaction, $Fe^{2+} + H_2O_2 \rightarrow Fe^{3+} + {\bullet}OH + OH^-$, and further reaction with peroxide gives hydroperoxyl radicals (•OOH) [1,4,29]. These radicals attack carboxylic acid end groups that remain on the polymer backbone from manufacture. Each attack removes a CF_2 unit, releases hydrogen fluoride, and leaves a new acid end group, so the chain is shortened step by step. Side chains are also cut at the ether and C–S bonds, which removes sulfonic acid sites [1,4].

The material consequences reinforce one another. The membrane thins, its proton conductivity and mechanical strength fall, and pinholes may form. A thinner or perforated membrane lets more hydrogen and oxygen cross, which lowers the open-circuit voltage, adds a mixed-potential loss at the cathode, produces more peroxide, and creates local hot spots where the gases combine directly. Fluoride in the product water is therefore a common indicator of chemical membrane ageing [1,29]. Mechanical damage acts alongside this chemistry. The membrane swells when wet and shrinks when dry, and repeated humidity or temperature cycling produces fatigue cracks, particularly where the polymer is already weakened [4]. In voltage terms, loss of sulfonic acid sites and drying raise R_{Ω}, whereas thinning and pinholes act mainly through crossover; the second route tends to end in abrupt failure rather than gradual decline. Low humidity, elevated cell temperature, and long periods at or near open circuit accelerate both routes. Cerium or manganese additives that scavenge radicals are the most widely used mitigation [1,4].

In FC1 and FC2, processed air-side relative humidity stays close to 50% and 52%, with standard deviations of 0.11 and 0.17 percentage points, and coolant outlet temperature remains near 55.4 °C. At the resolution of the processed files, humidity and thermal cycling are therefore small in both records. This does not exclude membrane ageing, because peroxide formation continues at constant load, but it makes hydration cycling an unlikely driver of the trends observed here. Fluoride release and gas crossover were not recorded, so chemical membrane ageing cannot be quantified from these files.

4.2.3 Dissolution and coarsening of the platinum catalyst

Platinum in the cathode is thermodynamically unstable at the potentials of a working cell. It dissolves directly, $Pt \rightarrow Pt^{2+} + 2e^-$ ($E^\circ \approx 1.19$ V), and also through an oxide route in which $Pt + H_2O \rightarrow PtO + 2H^+ + 2e^-$ is followed by chemical dissolution of the oxide, $PtO + 2H^+ \rightarrow Pt^{2+} + H_2O$ [5,29]. Dissolution is fastest when the electrode potential moves repeatedly between oxide formation and oxide reduction, as happens during load changes, idling, and transitions to open circuit. Small particles dissolve first because of their higher surface energy. The dissolved Pt^{2+} either redeposits on larger particles (Ostwald

ripening) or diffuses into the membrane, where crossover hydrogen reduces it to a band of isolated platinum. Particles can also migrate and coalesce on the carbon support. Where the catalyst is a platinum alloy, the less noble metal, such as Co or Ni, leaches out, and the released cations contaminate the ionomer [5].

Each of these processes lowers the electrochemically active surface area (ECSA). For a Tafel slope b, the additional activation loss caused by a fall in ECSA is approximately $\Delta\eta_{act} \approx b \cdot \log_{10}(ECSA_0/ECSA)$. With b ≈ 70 mV per decade, halving the active area adds about 21 mV per cell. For comparison, the mean cell voltage falls by about 27 mV in FC1 and 34 mV in FC2 between the first and last 24 h of the processed records. A moderate ECSA loss could therefore account for a large part of the observed decline, although the files contain no cyclic voltammetry or polarization data from which this share could be estimated. A platinum band in the membrane removes metal from the catalyst layer and, depending on where it forms, can either promote or suppress radical formation, which couples catalyst ageing to membrane ageing [5,29].

The two stacks differ in load profile. The processed files share channel names and test durations with the FCLAB stacks released for the IEEE PHM 2014 Data Challenge, in which FC1 was operated under a stationary regime and FC2 with high-frequency current ripples superimposed on the nominal current [30]. Hourly interpolation and LOWESS smoothing remove the ripple from the processed J and I channels, so its effect cannot be read directly. Repeated changes in electrode potential are the condition under which platinum dissolution and oxide cycling are expected to be strongest [3,5,25], but whether a high-frequency ripple of this kind is large enough to matter cannot be judged from the processed files. The larger voltage loss of FC2, together with its larger forecast errors and stronger horizon sensitivity (Tables 5–7), is consistent with a more demanding load profile, but two stacks are not enough to establish the link.

4.2.4 Corrosion of the carbon support

The carbon that supports the platinum is itself oxidized in the cathode, $C + 2H_2O \rightarrow CO_2 + 4H^+ + 4e^-$ ($E^\circ \approx 0.21$ V). Despite its low equilibrium potential, the reaction is slow at normal operating potentials, but platinum catalyzes it and its rate rises steeply above about 1.0 V [29]. Such potentials arise mainly during start-up and shut-down. When a hydrogen–air front moves through the anode, the part of the cell that still contains air drives a reverse current, and the opposite cathode region can reach about 1.4–1.5 V [31]. Fuel starvation has a comparable effect at the anode, where cell reversal forces carbon oxidation to supply the current.

Carbon loss thins and compacts the cathode catalyst layer, reduces its porosity, detaches and agglomerates platinum particles, and leaves oxygen-containing surface groups that make the layer more hydrophilic. The result is mainly a mass-transport loss. Oxygen access becomes poorer and liquid water is retained more readily, so η_{mt} grows, especially at high current density, while the platinum detached from the support adds to the activation loss [2,5,29]. Graphitized carbon, oxide-based supports, and control strategies that limit cathode potential during start–stop events are the main counter-measures [3]. The processed FC1 and FC2 files do not identify start-up, shut-down, or fuel starvation events, so this pathway cannot be assessed from them.

4.2.5 Gas-diffusion media, bipolar plates, and impurities

Outside the catalyst layer, the gas-diffusion layer and microporous layer gradually lose part of their PTFE coating and their carbon fibers oxidize, so the hydrophobicity that keeps pores open for gas declines over time. Compression and creep can also reduce porosity and raise the contact resistance between layers. These changes favor cathode flooding and increase both η_{mt} and R_Ω [2,3]. Metallic bipolar plates corrode in the warm, acidic, and humid cell environment. Iron, chromium, and nickel ions released from stainless steel enter the ionomer, where multivalent cations bind to sulfonic acid sites more strongly than protons and lower conductivity; iron also feeds the Fenton chemistry described in Section 4.2.2. At the same time, the passive oxide film on the plate surface raises interfacial contact resistance [3,29].

Impurities in the fuel and air act on the catalyst and ionomer directly. Carbon monoxide adsorbs strongly on anode platinum and blocks hydrogen oxidation even at sub-ppm concentrations, although much of this loss can be recovered by oxidizing the adsorbed CO. Hydrogen sulfide forms strongly bound Pt–S species and is largely irreversible at normal anode potentials. Ammonia is converted to NH_4^+ in the ionomer and raises ohmic and cathode losses, while sulfur dioxide and nitrogen oxides in the air poison the cathode [3,29]. These impurities are directly relevant to hydrogen produced from biological and bio-based feedstocks. Hydrogen obtained by dark fermentation or by reforming biogas can carry sulfur and nitrogen compounds, carbon monoxide, and organic species that originate in the biochemical conversion of the feedstock, and it has to be purified before use. ISO 14687 limits hydrogen supplied to PEMFC road vehicles to 0.2 μmol mol^{-1} of carbon monoxide, 0.004 μmol mol^{-1} of total sulfur

compounds, and 0.1 μmol mol^{-1} of ammonia [32]. Gas composition was not recorded for FC1 or FC2, so contamination cannot be examined in these records.

4.2.6 Reversible and irreversible losses in the FC1 and FC2 records

Part of the voltage loss in an operating stack is recoverable. Platinum oxide coverage builds up during continuous operation and is reduced when the cathode potential falls; water that accumulates in the electrodes and diffusion media can be removed by changes in flow or temperature; and some adsorbed contaminants are oxidized away. These reversible components lie on top of the irreversible losses described above, and a stack voltage record contains both [11,29].

The processed records show a long-term voltage decline with superimposed upward excursions, consistent with combined irreversible and reversible losses; these components cannot be separated conclusively from voltage data alone. Linear fits over the full records give a decline in Utot of 117 μV h^{-1} for FC1 and 114 μV h^{-1} for FC2, equivalent to about 23 μV h^{-1} per cell and to 3.5% and 3.4% of the initial stack voltage per 1,000 h. FC2 lost about 41 mV within its first 24 h, whereas FC1 changed little over the same period, which contributes to the larger first-to-last decline of FC2 (5.78% against 4.03%) even though the long-term slopes are similar. Upward steps are superimposed on the trend. The smoothed Utot rises by more than 1 mV from one hour to the next 71 times in FC1 and 109 times in FC2, and the largest hourly rises are 5.6 and 15.0 mV. These rises cluster in short episodes, for example around 822–830 h and 987–995 h in FC1 and 829–833 h in FC2. This is the pattern expected from the recovery of reversible losses after a change in operation, rather than from any restoration of degraded material.

The cell records show that the loss is not uniform. Between the first and last 24 h, the five cell voltages of FC1 fall by 3.50% to 4.73% and those of FC2 by 4.58% to 5.88%. The mean difference between the highest and lowest cell voltage widens from 8.1 mV over the first 100 h to 11.7 mV over the last 100 h in FC1, and from 11.4 to 17.1 mV in FC2. A pathway acting evenly on all cells would shift the cell voltages together. A widening spread instead points to local differences in hydration, temperature, reactant distribution, or catalyst loss, which are typical of stack operation [3,21]. The cell with the largest loss differs between the stacks (U3 in FC1 and U5 in FC2), and the records contain no information on cell position or local conditions that would explain this.

These observations limit what the forecasting results can say about mechanisms. The models in Section 3 learn the combined trajectory of reversible and irreversible losses on smoothed signals, so high forecast accuracy does not show that any particular pathway has been captured. Separating the pathways would require diagnostic measurements alongside the ageing record: impedance spectra to follow ohmic and charge-transfer resistance, polarization curves and cyclic voltammetry for ECSA, fluoride emission in the product water, and post-test microscopy of the membrane electrode assembly. With such data, mechanism-based degradation terms can be combined with data-driven forecasting, as in recent hybrid lifetime models [33], and the loss assignments in Table 8 can be tested rather than inferred.

Table 8. Principal chemical and electrochemical degradation pathways in PEMFCs, their effects on materials and voltage, and the related evidence in the FC1 and FC2 records.

Pathway	Principal reactions or processes	Main stressors	Effect on materials and voltage	Evidence available in FC1 and FC2
Membrane chemical degradation	H_2O_2 formation; Fenton reaction with Fe^{2+} or Cu^{2+}; •OH and •OOH attack on PFSA end groups and side chains; HF release	Low humidity; high temperature; open-circuit and idle periods; metal ions	Thinning, pinholes, loss of sulfonic acid sites; higher R_Ω and gas crossover; lower open-circuit voltage	Not measured (no fluoride or crossover data); little hourly humidity or thermal cycling
Membrane mechanical and thermal degradation	Swelling and shrinkage under humidity and temperature cycling; creep under compression	Humidity cycling; freeze–thaw; uneven compression	Cracks and pinholes; crossover; abrupt failure	HrAIRFC nearly constant (SD 0.11–0.17 percentage points)

Pathway	Principal reactions or processes	Main stressors	Effect on materials and voltage	Evidence available in FC1 and FC2
Platinum dissolution and coarsening	$Pt \rightarrow Pt^{2+} + 2e^{-}$; PtO formation and dissolution; Ostwald ripening; particle migration and coalescence; Pt band in membrane; alloy leaching	Potential cycling; load changes; high cathode potential	ECSA loss; higher η_{act} (about 21 mV per cell for halved ECSA at 70 mV per decade)	Inferred only from Utot and U1–U5 decline; FC2 current ripple removed by preprocessing
Carbon-support corrosion	$C + 2H_2O \rightarrow CO_2 + 4H^{+} + 4e^{-}$; reverse-current mechanism; cell reversal under fuel starvation	Start-up and shut-down; potentials above about 1.0 V; fuel starvation	Catalyst-layer thinning and collapse; Pt detachment; hydrophilic surface; higher η_{mt} and η_{act}	No start–stop or starvation events identified
Gas-diffusion media ageing	PTFE loss; carbon-fiber oxidation; compression creep	High current; liquid water; temperature	Flooding; higher η_{mt} and contact resistance	Flow and pressure channels recorded; no direct water-content data
Bipolar-plate corrosion	Release of Fe, Cr, and Ni ions; passive-film growth	Warm, acidic, humid environment; high potential	Cation contamination of ionomer; Fenton source; higher contact resistance	Not measured
Fuel and air impurities	CO and H_2S adsorption on Pt; NH_4^{+} exchange in ionomer; SO_2 and NO_x at cathode	Impure hydrogen, including hydrogen from biological or bio-based routes; polluted air	Blocked catalyst sites; higher η_{act} and R_{Ω}; partly reversible	Gas composition not recorded
Reversible losses	Platinum oxide coverage; water accumulation; adsorbed species	Continuous operation at constant load	Recoverable voltage loss	Hourly rises above 1 mV: 71 in FC1 and 109 in FC2

4.3 Implications for RUL, maintenance, and clean-energy systems

The end of an observation record is not an observed failure event. Accordingly, the last row in either series cannot serve as an RUL label. A maintenance-based EOL rule would need a threshold tied to required stack voltage, power, or a manufacturer's health limit. That threshold should be specified before evaluating forecasts, followed by sensitivity and uncertainty analyses. This is consistent with prognostic methods that model health-state crossing [6,14]. No such endpoint is evaluated here.

For maintenance decisions, calibrated prediction intervals would be more informative than point forecasts alone. The archived files used here do not supply validated forecast distributions or interval-coverage results. Such results would require a specified uncertainty method and evaluation over independent forecast origins. Accordingly, this study makes no probabilistic RUL or interval-calibration claim [3,21].

Model complexity should be evaluated against a defined forecasting objective. Table 7 presents a matched chronological test comparison of supplied SANformer checkpoints with independent baselines. SANformer improves on persistence and linear trend in every series-horizon case, yet Ridge

obtains lower RMSE in five of six. The neural model uses multivariate inputs whereas Ridge uses lagged voltage only; consequently these results do not isolate the effect of architecture from differences in available inputs [10,11,19,21].

The row-matched comparison provides direct evidence for these supplied models on the two identified processed files. Its rankings are conditional on their preprocessing, fixed checkpoints, and test partitions. The separate archived prediction-segment scores cannot substitute for these held-out test results. Checkpoint training and selection provenance has not been independently documented, so the held-out interpretation remains conditional on the checkpoints having been finalized without access to the evaluated targets.

The durability and environmental impact of a stack are also dependent on the pathways in section 4.2. Limited durability is primarily due to irreversible components. Catalyst degradation due to platinum dissolution and carbon corrosion continues to degrade the activity of the catalyst, while severe crossover can cause abrupt catalyst degradation due to chemical and mechanical damage to the membrane. The two records analyzed here span only 1154 h and 1020 h, which is much less than what is expected in vehicle and stationary applications, and the rates observed here cannot be extrapolated; for example, membrane thinning can occur at a higher rate later in life.

The stability of materials is related to sustainability. The dissolution loss and carrying of the metal to the membrane increases the amount of platinum needed to produce the same amount of energy delivered and decreases the amount to be recovered at EoL; platinum is rare and energy-consuming to produce. Chemical degradation will lead to the release of fluoride during operation of PFSA membranes and ionomers, and will result in the leaving of fluorinated material to be processed at disposal, which is increasingly being discussed with the regulatory authorities. More costs are incurred as a result of the coating and replacement costs associated with corrosion of the bipolar metal plates. This is because extending the stack life will ensure a reduction of material usage and cost. The main levers are derived from the pathways themselves; radical scavengers in the membrane; more stable catalyst supports; control of start–stop and idle potentials; careful humidity-thermal management; and fuel purity, which is of particular importance if the hydrogen is generated via biological or bio-based methods [3,4,5,32]. Voltage forecasting is only indirectly involved in this situation. A short-horizon forecast can highlight deviations from the expected course and, therefore, enable operating decisions to be made which prevent harmful conditions, but it does not provide any information on which path is being followed. Such control is not tested with this study and the sustainability benefits outlined above would need to be shown in operating systems [3,21].

4.4 Generalization, uncertainty, and limitations

The next extension is called cross stack validation. As it is currently implemented, FC1 and FC2 are assessed independently. A better generalization test would have trained on FC1 and tested on FC2, both with the same feature scaling as given during training, and then do the opposite. Studies using transfer learning demonstrate the feasibility and difficulty of such domain shifts [9,13,18]. The model would also be evaluated using a cross-stack approach to determine if it learned degradation structure or stack specific signal geometry. The same temporal and EOL rules can be used to compare two different adaptation methods if the transfer performance is dramatically reduced.

The matched evaluations are made only on the two processed, smoothed stack series, and not on raw sensor signals or other stacks. The multivariate inputs to SANformer are different from the Ridge baseline's historical voltage, so there are not only differences attributable to model architecture. The provided checkpoints files and hashes do not stand alone as evidence that training data is valid, that checkpoints are correctly selected or that the absence of test-target exposure. There is no re-training uncertainty or a cross-stack transfer estimate reported. There is no confirmed end-of-life event for either series and so no evaluation of RUL accuracy.

There are increasing voltage errors in both archived and matched test forecasts, and there are clear data segments and provenance. Providing checkpoint selection or being able to retrain from the source data would be beneficial in establishing the checkpoint selection and measuring variation across runs.

There would need to be calibrated uncertainty and low point error for maintenance decisions. A useful prediction interval should have its coverage close to an independently specified EOL boundary. Incomplete intervals without observations may be misinterpreted when making decisions on replacement. Empirical coverage, interval width and threshold-crossing coverage over rolling forecast origins could be evaluated in future work. Multiple training runs also would be useful for the purpose of quantifying neural-model performance variation.

Greater forecast accuracy doesn't equate to reduced emission or resource use. They might aid in the scheduling of maintenance and in preventing early stack replacement, but the impact needs to be

investigated in a working system. Besides predictive accuracy, future lifecycle studies will involve measuring replacement timing, downtime, hydrogen use and material turnover [3,21].

5. Conclusions

This study audited archived SANformer outputs and evaluated the supplied six SANformer checkpoints and three independent baselines on matched chronological test target rows of two version-identified processed PEMFC series. FC1 contains 1,155 rows and FC2 contains 1,021. The checkpoint test results are kept separate from the repository's earlier archived prediction-segment scores.

SANformer improves on persistence and linear trend in all six matched comparisons, while lagged Ridge achieves lower RMSE in five of six. Only for FC2 at three hours does SANformer outperform Ridge by RMSE. This conditional result gives a concrete comparison on the supplied processed files without establishing a general neural-model advantage or independently verifying the checkpoints' training history.

Degradation mechanisms are examined, establishing a physical connection between the voltage records and their causes. Each of these is a cause of specific activation, ohmic or mass-transport losses, and is a potential cause of radical attack on the membrane, dissolution and coarsening of platinum, corrosion of the carbon support, ageing of diffusion media and bipolar plates, and impurities in the fuel or air. The total stack voltage across all the cells decreases by approximately 3.4-3.5% per 1,000 h in both stacks, FC2 has an additional early voltage loss, and the difference in the voltage of the cells increases over time. These trends are just as expected if the ageing is gradual, non-uniform, and if there is some reversible loss, these losses would be measured as impedances, as active areas, and as effluents. The service life of a stack and the observed loss rates are not extrapolable for such a long service life, because the irreversible pathways determine this. In doing so, they would have a radical scavenger, a stable catalyst support, a start–stop mechanism and idle operation, and a high-purity hydrogen that is compliant with ISO 14687, among other attributes, which would extend the life of the stack and reduce the quantity of platinum and fluorinated materials consumed per unit of energy produced, essential for the sustainability of the hydrogen fuel-cell system.

The repository does not offer or maintain an observed end-of-life label. This study does not provide any RUL estimate or RUL accuracy. A threshold crossing rule for maintenance decisions would demand a valid EOL criterion, longer calibrated forecasts and independent endpoint validation.

Supplementary materials

Figures 1–4 use the identified processed files and archived prediction CSVs as indicated in their captions. Figures 5–6 and Table 6 report independent test baselines. Figures 7–8 and Table 7 report matched test outputs of the supplied SANformer checkpoints. The accompanying evidence package contains source and checkpoint hashes, row-level SANformer and baseline predictions, and summary metrics.

Data availability statement

The public SANformer repository is available at https://github.com/yangyijunSdu/SANformer. The processed FC1 and FC2 files and supplied checkpoints used for matched evaluation were accessed on 24 September 2026. The source-file SHA-256 hashes are reported in Section 2.1 and individual checkpoint hashes accompany the analysis package. Independent baseline code, summary results, and row-level predictions, together with matched SANformer test forecasts, figures, and metrics, accompany the evaluation. An earlier repository inspection was made on 31 August 2026.

Ethics statement

This secondary analysis uses public engineering sensor data and does not involve human participants, animals, personal information, or clinical records. An institutional ethics approval number is therefore not asserted.

AI use statement

AI tools assisted with language editing, analysis-code preparation, and the construction of data-based plots. Calculations reported in this manuscript were checked against the identified CSV files.

Funding

This research received no external funding.

Conflict of interest
The authors declare no competing interests.

References
1. Okonkwo PC, Ben Belgacem I, Emori W, Uzoma PC. Nafion degradation mechanisms in proton exchange membrane fuel cell (PEMFC) system: A review. International Journal of Hydrogen Energy. 2021. https://doi.org/10.1016/j.ijhydene.2021.06.032
2. Dafalla AM, Habte BT, Guo J, Jin F. Membrane Electrode Assembly Degradation Modeling of Proton Exchange Membrane Fuel Cells: A Review. Energies. 2022;15(23):9247. https://doi.org/10.3390/en15239247
3. Liu Z, Cai S, Tu Z, Chan SH. Recent development in degradation mechanisms of proton exchange membrane fuel cells for vehicle applications: problems, progress, and perspectives. Energy Storage and Saving. 2024. https://doi.org/10.1016/j.enss.2024.02.005
4. Madhav D, Wang J, Keloth R, Mus J, Buysschaert F, Vandeginste V. A Review of Proton Exchange Membrane Degradation Pathways, Mechanisms, and Mitigation Strategies in a Fuel Cell. Energies. 2024;17(5):998. https://doi.org/10.3390/en17050998
5. Okonkwo PC. Proton Exchange Membrane Fuel Cell Catalyst Layer Degradation Mechanisms: A Succinct Review. Catalysts. 2025;15(1):97. https://doi.org/10.3390/catal15010097
6. Wang P, Liu H, Hou M, Zheng L, Yang Y, Geng J, Song W, Shao Z. Estimating the Remaining Useful Life of Proton Exchange Membrane Fuel Cells under Variable Loading Conditions Online. Processes. 2021;9(8):1459. https://doi.org/10.3390/pr9081459
7. Ao Y, Laghrouche S, Depernet D, Chen K. Proton Exchange Membrane Fuel Cell Prognosis Based on Frequency-Domain Kalman Filter. IEEE Transactions on Transportation Electrification. 2021. https://doi.org/10.1109/TTE.2021.3077506
8. Hua Z, Zheng Z, Pahon E, Pera MC, Gao F. Multi-Timescale Lifespan Prediction for PEMFC Systems Under Dynamic Operating Conditions. IEEE Transactions on Transportation Electrification. 2022. https://doi.org/10.1109/TTE.2021.3103921
9. Ma J, Liu X, Zou X, Yue M, Shang P, Kang L, Jemei S, Lu C, Ding Y, Zerhouni N, Cheng Y. Degradation prognosis for proton exchange membrane fuel cell based on hybrid transfer learning and intercell differences. ISA Transactions. 2021. https://doi.org/10.1016/j.isatra.2020.06.005
10. Min H, Zhao H, Zeng B. A Data-Driven Prediction Method for Proton Exchange Membrane Fuel Cell Degradation. Energies. 2024;17(4):968. https://doi.org/10.3390/en17040968
11. Meng X, Mei J, Tang X, Jiang J, Sun C, Song K. The Degradation Prediction of Proton Exchange Membrane Fuel Cell Performance Based on a Transformer Model. Energies. 2024;17(12):3050. https://doi.org/10.3390/en17123050
12. Zhang Y, Ma R, Xie R, Liang B, Li Y. A Degradation Prediction Method for PEM Fuel Cell Based on Deep Temporal Feature Extraction and Transfer Learning. IEEE Transactions on Transportation Electrification. 2023. https://doi.org/10.1109/TTE.2023.3262588
13. Izadi MJ, Hassani P, Raeesi M, Ahmadi P. A novel WaveNet-GRU model for PEM fuel cells degradation prediction based on transfer learning. Energy. 2024. https://doi.org/10.1016/j.energy.2024.130602
14. Hu Y, Zhang L, Fu E, Jiang Y, Peng K. Remaining useful life prediction of proton exchange membrane fuel cell based on Wiener process and Bayesian GRU network considering multi-source uncertainties. International Journal of Green Energy. 2024. https://doi.org/10.1080/15435075.2024.2303358
15. Xie Y, Zou J, Peng C, Zhu Y. Degradation prediction and uncertainty quantification for PEMFC using NSGA-optimized deep learning method. 2021 IEEE International Instrumentation and Measurement Technology Conference. 2021. https://doi.org/10.1109/I2MTC50364.2021.9460036
16. Xie Y, Zou J, Peng C, Zhu Y. Performance Degradation Prediction of PEMFC Based on Adaptive Variational Mode Decomposition and Deep Belief Network. 2021. https://doi.org/10.1145/3471985.3472381
17. Tian Q, Chen H, Ding S, Shu L, Wang L, Huang J. Remaining Useful Life Prediction Method of PEM Fuel Cells Based on a Hybrid Model. Electronics. 2023;12(18):3883. https://doi.org/10.3390/electronics12183883
18. Kebede GA, Lo SC, Wang F, Chou JH. Transfer learning-based deep learning models for proton exchange membrane fuel remaining useful life prediction. Fuel. 2024. https://doi.org/10.1016/j.fuel.2024.131461

19. Tang Y, Huang X, Li Y, Ma H, Zhang K, Song K. Degradation Prediction of Proton Exchange Membrane Fuel Cell Based on Multi-Head Attention Neural Network and Transformer Model. Energies. 2025;18(12):3177. https://doi.org/10.3390/en18123177
20. Tang X, Shi L, Li M, Xu S, Sun C. Health State Estimation and Long-Term Durability Prediction for Vehicular PEM Fuel Cell Stacks under Dynamic Operational Conditions. IEEE Transactions on Power Electronics. 2024. https://doi.org/10.1109/TPEL.2024.3502499
21. Li KJ, Hong J, Zhang C, Liang F, Yang H, Ma F, Wang F. Health state monitoring and predicting of proton exchange membrane fuel cells: A review. Journal of Power Sources. 2024. https://doi.org/10.1016/j.jpowsour.2024.234828
22. Li M, Fang L, Tian J, Wang L. Degradation Prediction of Proton Exchange Membrane Fuel Cells Based on ITransformer. 2025. https://doi.org/10.1109/ICPRE67300.2025.11274121
23. Lv L, Pei P, Ren P, Wang H, Geng W. Exploring Performance Degradation of Proton Exchange Membrane Fuel Cells Based on Diffusion Transformer Model. Energies. 2025;18(5):1191. https://doi.org/10.3390/en18051191
24. Xi F, Shen Y, Wang B, Liu X, Hou W, Fu P, Zhao Y, Miao P, Zhang G, Li L, Yin Y, Jiao K. A 4750 hours' durability investigation of the PEMFC stack based on fuel cell hybrid vehicles conditions. International Journal of Green Energy. 2023. https://doi.org/10.1080/15435075.2023.2276157
25. Fu S, Qi H, Zheng C. Experimental research on proton exchange membrane fuel cell durability considering degradation driving cycle and temperature. IEEE Transactions on Transportation Electrification. 2025. https://doi.org/10.1109/TTE.2025.3627243
26. Xiao Y, Zhang J, Ruiz HS. A Physics Informed and Data Fusion Approach for Remaining Useful Life Prediction of PEMFC. 2024. https://doi.org/10.1109/ISEEIE62461.2024.00110
27. Zhou X, Wang W, Cao Y, Yu D, Hou XT. Hybrid Prediction of Fuel Cell Remaining Useful Life with Dynamic Grey Relational Analysis-Bidirectional Long Short-Term Memory Networks and Extended Kalman Filter. Measurement Science and Technology. 2025. https://doi.org/10.1088/1361-6501/ade5c1
28. Cai B, Liu H. Dynamic Prediction of Proton-Exchange Membrane Fuel Cell Degradation Based on Gated Recurrent Unit and Grey Wolf Optimization. Energies. 2024;17(23):5855. DOI: https://doi.org/10.3390/en17235855
29. Borup R, Meyers J, Pivovar B, Kim YS, Mukundan R, Garland N, Myers D, Wilson M, Garzon F, Wood D, Zelenay P, More K, Stroh K, Zawodzinski T, Boncella J, McGrath JE, Inaba M, Miyatake K, Hori M, Ota K, Ogumi Z, Miyata S, Nishikata A, Siroma Z, Uchimoto Y, Yasuda K, Kimijima K, Iwashita N. Scientific Aspects of Polymer Electrolyte Fuel Cell Durability and Degradation. Chemical Reviews. 2007;107(10):3904–3951. https://doi.org/10.1021/cr050182l
30. FCLAB Research Federation. IEEE PHM Data Challenge 2014: PEM fuel cell ageing data. Université Bourgogne Franche-Comté; 2014. https://search-data.ubfc.fr/FR-18008901306731-2021-07-19
31. Reiser CA, Bregoli L, Patterson TW, Yi JS, Yang JD, Perry ML, Jarvi TD. A Reverse-Current Decay Mechanism for Fuel Cells. Electrochemical and Solid-State Letters. 2005;8(6):A273–A276. https://doi.org/10.1149/1.1896466
32. International Organization for Standardization. ISO 14687:2019 Hydrogen fuel quality — Product specification. Geneva: ISO; 2019.
33. Ke C, Han K, Wang Y. A Hybrid Method Combining Degradation Mechanisms and Deep Learning for Lifetime Prediction of Proton Exchange Membrane Fuel Cells under Dynamic Load Cycle Conditions. Journal of Power Sources. 2025. https://doi.org/10.1016/j.jpowsour.2025.236464