Gated Residual End-of-Life Prediction for Variable-Prefix RUL in the PHME 2026 Data Challenge

Giuseppe Mannone[1], Tim Herrmann[1], Martin Dazer[1]

[1] *Institute of Machine Components (IMA), University of Stuttgart, Stuttgart, Germany*
giuseppe.mannone@ima.uni-stuttgart.de
tim.herrmann@ima.uni-stuttgart.de
martin.dazer@ima.uni-stuttgart.de

ABSTRACT

This paper presents a gated residual end-of-life (EOL) method for variable-prefix remaining useful life (RUL) prediction in the PHME 2026 Data Challenge. The task is to estimate the lifetime of partially observed subway-door actuation runs under hidden operating conditions. We frame the problem as EOL estimation: the model first predicts the cycle at which failure occurs and then derives RUL by subtracting the current cycle index. This gives prefixes of different lengths a common lifetime scale. The approach uses engineered prefix features that summarize electrical load, position behavior, shock events, trends, and source-model predictions. Several source models provide an initial EOL prior, while disagreement among them serves as an uncertainty signal. Rather than replacing this prior with one unconstrained predictor, the method refines it through bounded residual corrections. To keep the prediction stable, the method first forms a conservative anchor and calibrates it to the inference setting using released test measurement features. Two bounded specialists then propose residual corrections: one from similar training prefixes and one from shock, position, and current signal evidence. When these specialists disagree, a clipped gate controls how much each correction affects the final EOL estimate. Local diagnostics indicate where the source prior is sufficient, where corrections add value, and where difficult prefixes remain uncertain.

1. INTRODUCTION

Remaining useful life prediction is a core problem in prognostics and health management (PHM), where the objective is not only to identify whether a system is degrading, but to estimate how long it can continue operating before a failure threshold is reached (Jardine, Lin, & Banjevic, 2006; Si, Wang, Hu, & Zhou, 2011). The PHME 2026 Data Challenge instantiated this problem on a controlled subway-door actuation test bench in which opening and closing motions are repeated until position degradation prevents normal operation (PHME 2026 Organizing Committee, 2026a, 2026b). The challenge is practically relevant because door-actuation faults represent a class of electromechanical degradation problems in which observable signals mix controller behavior, motor dynamics, operating condition, gradual drift, and intermittent position shocks.

Recent PHM data challenges have provided benchmark settings for RUL modeling under limited-data and heterogeneous degradation conditions, including filtration-system challenges and broader PHM challenge datasets (Beirami et al., 2020; İnce, Sirkeci, & Genç, 2020; Su & Lee, 2024). Data-driven PHM methods increasingly include deep representation learning and online sequence models for machine health monitoring (Zhao et al., 2019; TV, Gupta, Malhotra, Vig, & Shroff, 2018), but the present setting has only a small number of run-to-failure units and hides the operating regime at inference time. We therefore use a constrained feature-based approach instead of a large end-to-end sequence model. The method is closer to stacked generalization (Wolpert, 1992) and uncertainty-aware residual correction: earlier predictors provide source features, disagreement among them becomes an uncertainty variable, and later stages learn bounded corrections rather than replacing the initial EOL estimate.

The challenge is difficult because several ambiguities occur at the same time. Test units are only partially observed, so the model must infer final lifetime from a prefix before the full degradation trajectory is visible. Two prefixes with similar measured position or current traces can still have different remaining lives if they correspond to different operating regimes or different points on the underlying degradation path. At the same time, the operating condition is hidden at inference time, even though velocity, acceleration, and deceleration affect both the measured signal shapes and the degra-

dation pace. This makes age, operating regime, and health state partly entangled in the released measurements.

The signals are also not ideal monotonic health indicators. The door-actuation process contains gradual drift, controller-dependent electrical behavior, and abrupt position or shock events. A large shock count can indicate severe degradation in one regime, but it is not a complete lifetime proxy across all regimes. Finally, the evaluation rewards a complete RUL trajectory, including early prognostic usefulness, not only the last cycles where failure is visually obvious. A useful model must therefore remain controlled when evidence is weak, while still allowing targeted corrections when source disagreement, retrieval support, or signal-group evidence indicates that the initial EOL estimate is biased.

These properties match the three data challenges highlighted by the organizers: high variability of end-of-life across predictors, uncertainty in the failure amplitude of prognostic indicators, and variable degradation speed (PHME 2026 Organizing Committee, 2026b). Our model addresses the first through an absolute EOL formulation and operating-condition-aware source priors, the second through source-disagreement features and clipped residual envelopes, and the third through trend, shock, retrieval, and signal-group residual evidence that can adjust the EOL estimate without replacing the conservative anchor.

The proposed method treats the task as variable-prefix end-of-life estimation. For a unit observed through cycle t, the model predicts an end-of-life cycle $\hat{E}$ and converts it to RUL by $\widehat{\mathrm{RUL}}(t) = \hat{E} - t$. This formulation makes it possible to combine source predictions, retrieval residuals, signal-group corrections, and uncertainty diagnostics on a common absolute-life scale. The final model is not a single monolithic regressor. Instead, it is a sequence of conservative residual-learning steps over successively calibrated anchors: a strong source prior first locates EOL, and later stages learn only bounded corrections where uncertainty evidence supports them.

We use a small set of method terms throughout the paper. A *prefix* is the observed beginning of a trajectory. A *source prior* is an initial EOL estimate obtained by weighting several source-model predictions. An *anchor* is a deliberately stable baseline EOL estimate used as the reference for later corrections. A *residual correction* is a predicted adjustment to an existing EOL estimate, not a new prediction from scratch. A *specialist* is a branch focused on one type of evidence, and the final *gate* is a bounded router that decides how much each specialist contributes.

The main contributions are:

- a prefix-level EOL formulation and feature representation that combines electrical, energy, position, shock, trend, extrapolated-RUL, and source-prediction evidence;
- source-disagreement features and covariate-shift weighting using released test measurement features;
- bounded retrieval and signal-group residual specialists combined by an uncertainty-aware residual gate;
- a transparent diagnostic protocol, ablation view, and safeguard analysis used to explain the proposed model.

2. CHALLENGE DATASET AND EVALUATION

2.1. System and Failure Definition

The challenge data originate from the PIMSSIS test bench, short for Prognostics of Industrial Machines for System Safety and Integrated Safety, a servomotor-based rotating-machine setup used to study prognostics for a subway ticket-validation door (PHME 2026 Organizing Committee, 2026a). The system contains a three-phase brushless motor and an INGENIA EVEREST XCR controller. The motor moves the door shaft between an open position and a closed position; repeated opening and closing activities are recorded until degradation prevents the closing process from reaching the required position. The data set and related prognostic setting are associated with the subway-door RUL study of (Soualhi et al., 2023).

The failure mode is position degradation during the closing activity. Stochastic degradation events reduce the maximum closing position over time. A failure scenario is reached when the closing process can no longer be executed, operationally defined in the challenge documentation as the closing position falling below 10% of the optimal position. This means that the RUL target is linked directly to the cycle index at which the door can no longer satisfy the closing requirement.

2.2. Data Organization

Each experiment is a run-to-failure trajectory. The official challenge document states that 48 training experiments are released, covering all three operating conditions with 16 training experiments per condition. It also states that 19 test experiments are used for participant evaluation. The test data include operating conditions represented in the training phase, but the operating-condition label is not provided to participants at inference time (PHME 2026 Organizing Committee, 2026b).

Each trajectory folder contains measurement files for opening and closing actions. Each cycle consists of the corresponding opening and closing measurements paired by cycle index. Measurement files are stored as semicolon-separated CSV files without headers, with approximately 600 samples per file and 16 numerical columns. Training folders additionally contain an RUL file whose rows are synchronized with the measurement files. The same RUL value applies to the corresponding opening and closing activities within a cycle.

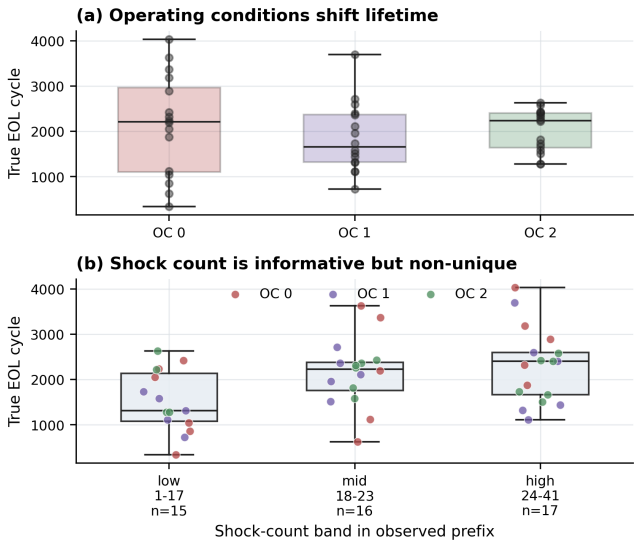

Figure 1. Local diagnostic view of the PHME 2026 RUL problem. (a) True EOL varies substantially across and within operating conditions. (b) Non-overlapping shock-count bands show overlapping EOL ranges, so shock evidence is useful as residual information but not sufficient as a standalone age proxy.

2.3. Operating Conditions

The platform uses three operating conditions that differ in closing velocity, acceleration, and deceleration. The operating-condition (OC) triplets stated in the challenge handout are $(15, 18, 18)$, $(10, 15, 15)$, and $(15, 18, 15)$ (PHME 2026 Organizing Committee, 2026b). In the local diagnostics, we label these as OC 0, OC 1, and OC 2, respectively. The operating condition is available implicitly in the training data distribution but is not provided as a test label. The model therefore needs to infer operating-regime information from measured signals and early-cycle behavior. This hidden-condition setting motivated the use of operating-condition fingerprints, source-prediction ranks within operating condition, and operating-condition-aware retrieval.

Figure 1 provides initial data-level motivation for the modeling choices. The operating conditions have overlapping but different lifetime distributions, and shock count is informative but not a smooth age proxy. The model therefore needs both operating-regime awareness and safeguards against interpreting any single signal as a complete health index.

2.4. Score

The challenge score combines normalized error and temporal prognostic behavior (PHME 2026 Organizing Committee, 2026b). Each submitted RUL trajectory is compared with the hidden true RUL trajectory. The organizer documentation defines a precision-related component based on the fraction of time steps whose relative error remains within a $\pm 10\%$ band; we follow the official scoring definition and refer to the normalized component as Prec_{norm}. The score also includes root mean squared error (RMSE) and prognostic horizon (PH), where PH uses a $\pm 20\%$ tolerance band and measures how early the prediction enters and then remains within the accepted RUL band (Saxena, Celaya, Saha, Saha, & Goebel, 2010).

The RMSE and precision-related raw quantities are mapped to normalized values through the fitted function $f(m)$, with calibration examples $f(0) = 1$, $f(57) = 0.9$, and $f(3000) = 0.1$. The value m is measured in the raw component's units, for example cycles for RMSE and the corresponding raw precision-error quantity for the precision component. The final score is reported by the challenge document as

$$\text{Score} = \frac{\mathrm{Prec}_{norm} + \mathrm{RMSE}_{norm} + \beta\,\mathrm{PH}_{norm}}{2 + \beta},$$

with $\beta = 0.5$ for the PH term. We use local validation metrics for development and diagnostics, but the decisive result is the official organizer evaluation assigned to the submitted file. Official scores are reported only in the Results section; method design is motivated by training-data diagnostics and local validation analyses.

2.5. Data Observations Motivating the Model

The released trajectories suggest a modeling problem that is both physical and statistical. Operating conditions shift lifetime distributions, so a single global age model can become brittle. Degradation is not purely smooth: shock-like position drops and abrupt changes appear in addition to gradual current, energy, and position drift. Source predictions are already strong, but their disagreement grows in difficult regimes and is therefore useful as an uncertainty descriptor. These observations motivated a conservative EOL anchor followed by bounded residual specialists rather than one unconstrained regressor.

Figure 2 shows why the source predictions are treated as both a prior and an uncertainty signal. Higher source disagreement is associated with larger source-prior error, and the bounded residual stages mainly reduce error in high-disagreement regimes. Figure 3 then shows the signal-level evidence behind the specialist branches: open-side position feedback declines as units approach failure, while cumulative shock evidence grows irregularly and with operating-condition-dependent spread. These figures are diagnostic rather than a claim that any single signal is sufficient; together they motivate bounded residual features combined with current, trend, source-disagreement, and retrieval support.

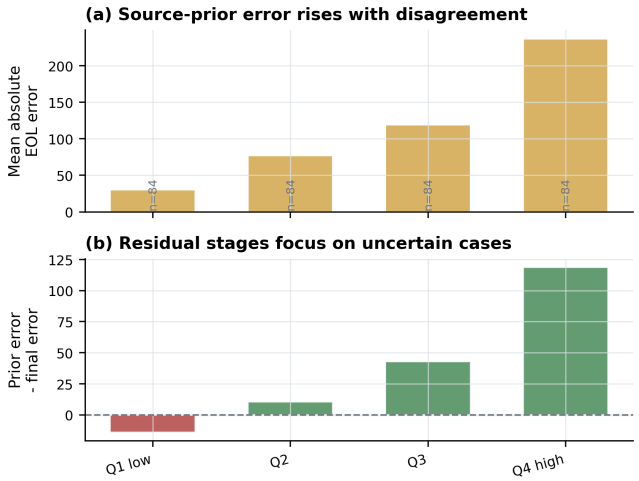

Figure 2. Source-disagreement motivation for residual learning. (a) Mean source-prior EOL error increases across source-disagreement quartiles. (b) The final residual stages mostly reduce error in the high-disagreement regimes, supporting source disagreement as an uncertainty feature rather than only another averaging input.

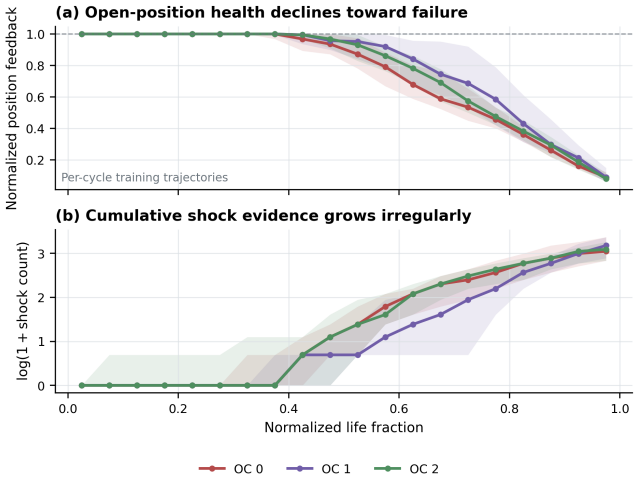

Figure 3. Training-data degradation evidence. (a) Median normalized open-side position feedback over normalized life fraction and operating condition. (b) Median $\log(1 + \text{shock count})$ over the same life axis. Shaded bands indicate interquartile ranges. Normalized life fraction is computed from training EOL for visualization only and is not available as a test label.

3. Feature Extraction and Meta-Dataset Construction

3.1. Latent EOL Supervision from Prefixes

Let unit i be observed through paired opening–closing cycle t, and let E_i be its true end-of-life cycle. The internal supervised label is E_i, while RUL is recovered as

$$\mathrm{RUL}_i(t) = E_i - t.$$

We define E_i as the first cycle index after the last valid cycle used for the RUL trajectory, so the final valid RUL remains positive after integer clipping. The feature vector $x_{i,t}$ summarizes only the observed prefix. Future cycles are used in training only to define the EOL label, not to compute prefix features. This distinction is important because the released test data provide one observed prefix per test unit.

For row-level submission, the model predicts one EOL estimate $\hat{E}_k$ for each test unit k. The predicted RUL for an evaluated row r with cycle index $t_{k,r}$ is then approximately

$$\widehat{\mathrm{RUL}}_{k,r} = \max\left(1, \mathrm{round}(\hat{E}_k - t_{k,r})\right).$$

The implementation applies this as a unit-level EOL correction to the base row-level RUL trajectory, followed by the positive integer clipping required by the submission representation. This keeps the modeling problem focused on the latent failure endpoint.

3.2. Prefix Feature Families

For training, each run-to-failure unit is summarized at observed-life fractions

$$\mathcal{F} = \{0.35, 0.40, 0.50, 0.60, 0.70, 0.80, 1.00\},$$

which yields $48 \times 7 = 336$ local training-prefix rows. For a unit with n_i paired cycles and fraction f, cycles are sorted by paired cycle index and the prefix endpoint is $j_{i,f} = \max(\lceil n_i f \rceil - 1, 0)$. Opening and closing records are paired by their shared cycle index before aggregation, so both motion directions contribute to the same prefix row. After cleaning, clipping, aggregation, and source-prediction enrichment, the final feature representation contained 336 local training-prefix rows, 105 fixed diagnostic rows, and 19 test-summary rows.

The feature design is grouped by role rather than by implementation history. Electrical and energy descriptors capture the early load and controller regime. Position and motion descriptors measure tracking behavior and position loss. Shock and drop descriptors capture abrupt degradation events, including recent position drops and shock density. Trend descriptors convert short histories into degradation-rate evidence using slope windows and extrapolated-RUL candidates. Source-prediction descriptors provide prior EOL estimates and an uncertainty signal through source disagreement. Routing descriptors summarize life proxy, operating-condition fingerprints, retrieval support, branch disagreement, and transfer context.

Health ratios are formed against safe initial denominators so that each prefix can be interpreted relative to its own early state. If x_t denotes an offset from the initial value and x_{init} denotes the initial raw value, a representative health ratio has the form

$$\mathrm{HR}(x_t) = \frac{x_{init} + x_t}{x_{init} + \epsilon},$$

where ϵ is the safe-denominator clipping term. Trend features are computed by fitting a line to $\log(\mathrm{HR})$ over finite windows. For a degradation threshold τ, an extrapolated RUL candidate is formed only when the fitted slope is negative and the window passes reliability checks.

3.3. Source Prediction Features

A central design choice is to include source-model predictions as first-class features. These source models, listed in Table 1, are trained on released training data using deliberately different inductive biases, including rank-based EOL regressors, conservative operating-condition routers, confidence-aware teacher routers, and ensemble predictors over the engineered prefix features. As the table records, the families also differ in role: five of them (S_1–S_5) supply both an EOL estimate and an uncertainty descriptor, whereas the remaining two (S_6 and S_7) contribute uncertainty evidence only.

The weighted source prior used by the downstream stages is

$$\hat{E}_{prior} = 0.32\hat{E}_{S_1} + 0.26\hat{E}_{S_2} + 0.20\hat{E}_{S_3} + 0.14\hat{E}_{S_4} + 0.08\hat{E}_{S_5}.$$

The convex weights were selected on unit-grouped local diagnostics using released training data. The five-source prior and the broader seven-source disagreement set play different roles: the prior provides a stable EOL location estimate, while the broader source set provides source mean, median, standard deviation, minimum, maximum, interquartile range, source gap, rank features, and clipping envelopes. Thus, source disagreement becomes uncertainty evidence rather than only another average.

4. METHOD

4.1. Overview and Notation

The final prediction stack is summarized in Figure 4. The method comprises four conceptual blocks: a source-prior EOL anchor, a single-prefix calibration step, two bounded residual specialists, and a clipped uncertainty-aware gate. Because each block refines the EOL estimate produced by the previous one, the method generates a short chain of intermediate estimates rather than a single output. Table 2 maps the symbol for each of these estimates to its plain meaning, from the initial source-prior estimate $\hat{E}_{prior}$ through to the final gated estimate $\hat{E}_H$, and this notation is used consistently in the remainder of the section. Later stages do not directly re-estimate EOL. They learn residual corrections over an already strong anchor, and those corrections are clipped or gated when support is weak.

4.2. Conservative EOL Anchor

The conservative anchor converts the source-prior estimate and related uncertainty features into a robust baseline EOL predictor. In this subsection, "anchor" means the stable estimate on which later residual corrections are built. It combines conservative residual regression, guarded local memory, and several small models trained on different signal summaries. The anchor is designed to provide a stable EOL surface before applying larger context-dependent residual corrections.

All anchor learners model residuals over the source prior, for example $E - \hat{E}_{prior}$, rather than raw EOL. Residual outputs are regularized and clipped against empirical source envelopes, meaning ranges and quantiles implied by the source-model predictions. A lower bound enforces that the predicted EOL cannot precede the last observed cycle. The guarded local-memory component adds only a low-amplitude correction,

$$\hat{E}_M(x) = \hat{E}_A(x) + g_{mem}(x)\Delta_{mem}(x),$$

where Δ_{mem} is an inverse-distance local residual and g_{mem} decreases when source spread, neighbor dispersion, or operating-condition mismatch indicate weak analog support. In the ablation, this anchor is interpreted as a structural safeguard and residual frame, not as a standalone accuracy improvement over the source prior.

4.3. Single-Prefix Calibration

The transfer-calibration stage addresses the mismatch between multi-prefix training and the official inference setting. During development, each training unit can be summarized at several prefix fractions. The released test data, however, provide one prefix per test unit. To reduce this mismatch, the method first constructs a unit-consensus teacher by combining multiple local-memory predictions for each training unit with a weighted median. The student then learns a residual over the memory anchor,

$$\Delta_i^{teacher} = \hat{E}_i^{teacher} - \hat{E}_M(x_{i,t}),$$
$$\hat{E}_T(x) = \hat{E}_M(x) + f_\theta(x).$$

The student f_θ is a ridge model trained with covariate-shift weights (Hoerl & Kennard, 1970; Shimodaira, 2000). A shallow domain classifier estimates whether a local training row resembles the released test measurement features using source statistics, life proxy, prefix geometry, and uncertainty features. The classifier output is used only as a weighting signal, not as a calibrated probability. If the domain score is

Table 1. Source-prediction families used by the stacked meta layer. The five-source prior estimates EOL location; the seven-source set supplies uncertainty descriptors for residual specialists and gating.

Source	Model family	Main input group	Target	Role
S_1	Phase-aware source router	Life proxy, OC fingerprints, source consensus, shift indicators	EOL	prior and uncertainty
S_2	Confidence-aware teacher router	Teacher predictions, empirical uncertainty, source spread, confidence proxies	EOL	prior and uncertainty
S_3	OC rank-stretch consensus	Consensus EOL, within-OC ranks, life proxy, source spread	EOL	prior and uncertainty
S_4	Spread-phase adaptive ensemble	Prediction spread, life phase, OC estimate, source deltas	EOL	prior and uncertainty
S_5	Weighted conservative consensus	Diverse OC-specific prefix predictors and teacher/fraction anchors	EOL	prior and uncertainty
S_6	Robust source consensus	Robust OC- and spread-aware source predictions	EOL	uncertainty only
S_7	Gated source blend	Source posterior, disagreement, and gating evidence	EOL	uncertainty only

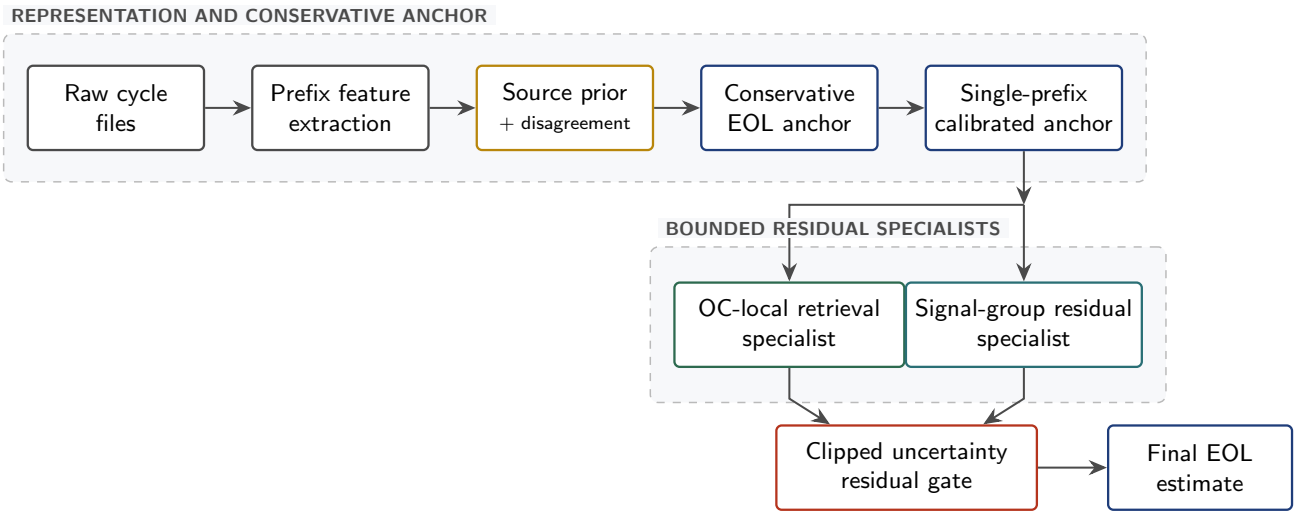

Figure 4. High-level architecture of the gated residual EOL model. Raw cycle files are converted into prefix features and source predictions. The source-prior estimate gives the first EOL location, while source disagreement supplies uncertainty evidence. A conservative anchor and single-prefix calibration define the residual frame. The retrieval and signal-group specialists then provide bounded residual corrections, and the final gate blends them without allowing either branch to dominate completely.

$p_{dom}(x)$, the sample weight is

$$w(x) = 0.25 + 5.0 p_{dom}(x).$$

The constants define a nonzero floor and bounded emphasis on prefixes that resemble the released test measurement features: weights range from 0.25 to 5.25, so highly similar rows can receive up to 21 times the weight of rows far from that measurement-feature region. The values were fixed from local diagnostic sweeps before official scoring. This step uses only released test measurement features as covariates; it does not use hidden RUL labels or official test errors.

4.4. Operating-Condition Retrieval Specialist

The retrieval specialist starts from the calibrated anchor and adds an operating-condition-aware residual correction. It searches a standardized feature space containing source-prior, life-proxy, operating-condition, uncertainty, and health-indicator descriptors. Here, the life proxy is a feature-derived estimate of where the observed prefix lies within the unit's lifetime. Euclidean nearest-neighbor search returns at most 16 neighbors, with inverse-distance weights adjusted by life-proxy similarity and inferred operating-condition compatibility (Eker, Camci, & Jennions, 2014). The retrieval residual is the weighted mean of training residuals $E - \hat{E}_T$; retrieval spread is the corresponding weighted standard deviation.

The branch prediction is

$$\hat{E}_R(x) = \hat{E}_T(x) + \alpha_{ret}(x)\,\mathrm{clip}(\Delta_{ret}(x)).$$

The raw residual is clipped at the 90th percentile of absolute training residuals. The strength α_{ret} increases when neighbor support is strong, retrieval residual variance is low, the operating-condition match is plausible, and the row lies in the predefined life-proxy band. This branch is most useful when a test prefix has reliable analogs in the training memory.

Table 2. Notation map for the main EOL estimates used by the method.

Symbol	Plain meaning
$\hat{E}_{prior}$	initial EOL estimate from weighted source-model predictions
$\hat{E}_A$	stable baseline EOL estimate after conservative residual learning
$\hat{E}_M$	baseline after a small nearest-neighbor memory adjustment
$\hat{E}_T$	baseline adjusted toward the one-prefix test setting
$\hat{E}_R$	EOL estimate from the nearest-neighbor retrieval branch
$\hat{E}_S$	EOL estimate from the shock, position, and current branch
ρ	clipped retrieval-branch weight chosen by the final gate
$\hat{E}_H$	final unit-level EOL estimate after gated blending

4.5. Signal-Group Residual Specialist

The signal-group specialist also starts from the calibrated anchor, but it models residuals through signal groups rather than nearest-neighbor analogy. Three ridge residual experts are trained with covariate-shift-weighted samples: a shock expert using shock counts, recent drops, jerk, and degradation-stability descriptors; a position expert using position-feedback range, drift, and tracking behavior; and a current expert using charging-current, movement-current, bus-voltage, and initialization variability.

At inference, each signal group receives a strength score computed from robust-scaled deviations of its descriptors. The shock, position, and current strengths are normalized into mixture weights, and the weighted residuals form Δ_{sig}. The branch prediction is

$$\hat{E}_S(x) = \hat{E}_T(x) + \alpha_{sig}(x)\,\mathrm{clip}(\Delta_{sig}(x)).$$

The signal-group strength α_{sig} is controlled by life band, total signal-group activity, and uncertainty, and the raw mixture delta is clipped at the 90th percentile of absolute training residuals. This branch is useful when shock, current, or position evidence is clearer than local analog retrieval.

4.6. Bounded Residual Gate

The final gate does not predict EOL directly. It decides how much to trust the retrieval branch relative to the signal-group branch. During development, the retrieval branch is labeled preferable only when its absolute EOL error is at least two cycles smaller than the signal-group error. This two-cycle margin is used as a tie-breaking tolerance: branches whose absolute errors differ by less than two cycles are treated as effectively indistinguishable, preventing noisy local labels from forcing the gate toward one branch. The gate is implemented as a shallow gradient-boosting classifier using retrieval support, retrieval spread, operating-condition match, signal-group activity, branch disagreement, source spread, quantile width, life proxy, and calibrated-anchor context.

The final EOL estimate is

$$\hat{E}_H(x) = \rho(x)\hat{E}_R(x) + (1-\rho(x))\hat{E}_S(x).$$

The retrieval weight $\rho(x)$ is computed from the learned gate probability blended with a conservative retrieval prior and clipped to $[0.05, 0.95]$. The clipping prevents full collapse to either branch and makes the final stage a bounded router rather than an unconstrained selector.

4.7. Safeguards

Several safeguards recur throughout the method. Later stages learn residuals over strong anchors rather than raw EOL. Retrieval and signal-group deltas are clipped. Source prediction envelopes limit uncontrolled drift. Operating-condition and life-proxy gates decide when a correction is plausible. Source standard deviation, source gap, retrieval dispersion, quantile width, and branch disagreement act as uncertainty features. These constraints are important because small local gains can become unstable in the hidden test distribution.

5. EXPERIMENTAL PROTOCOL

All ridge models used standardized inputs. Hyperparameters, clipping thresholds, and gate settings were selected using local diagnostics on the released training data. Nearest-neighbor retrieval used at most 16 neighbors in the standardized feature space. Residual corrections were clipped using empirical training-residual quantiles, and all final EOL estimates were constrained to exceed the last observed cycle.

Candidate models were evaluated using several local views of the released training data. The full local diagnostic view contains 336 rows across multiple observed fractions. A fixed diagnostic view contains 105 rows and is used for trajectory plots and stress testing. Test-prefix diagnostic panels approximate the one-prefix-per-unit inference geometry by sampling one prefix per released training unit, retaining 48 labeled units for stable local error estimates while varying prefix distributions. Table 3 organizes these views by their underlying data, their independence from the fitted models, and their diagnostic purpose. It also makes explicit the distinction we maintain throughout the analysis: the local views characterize error regimes and routing behavior, whereas the hidden-label official evaluation on the 19 test summaries remains the primary challenge result rather than a calibrated alternative to it. When a local view contains prefixes from units also used to fit upstream source or anchor models, we interpret it as a diagnostic stress test rather than as an independent generalization estimate.

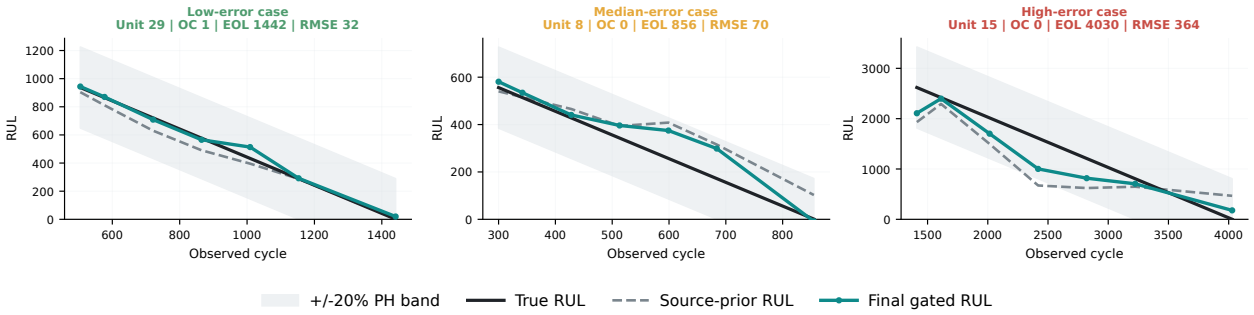

Figure 5. Representative fixed diagnostic RUL trajectories for low-, median-, and high-error diagnostic units. The gray band marks the $\pm 20\%$ prognostic-horizon tolerance around true RUL. Black, dashed gray, and teal curves show true RUL, source-prior RUL, and final gated RUL, respectively.

Model development was driven by local operating-condition analysis, prefix-regime diagnostics, and branch-specific residual behavior. Official evaluations are reported as challenge outcomes for completed submissions; component selection and model interpretation are supported by the local diagnostics reported here. This distinction is important because the official server evaluates the final row-level prediction file, whereas the local views are designed to understand why a model behaves as it does.

Table 3. Diagnostic views used to interpret model behavior. Local views explain failure modes and routing behavior; the hidden-label official evaluation remains the primary challenge result.

View	Data and independence	Purpose
Full-prefix diagnostic	336 rows from 48 units; training-regime diagnostic, not unit-held-out	Residual, source-spread, and routing inspection
Fixed diagnostic trajectories	105 rows from 15 displayed diagnostic units	Qualitative prefix behavior and stress tests
Test-prefix panels	20 panels with 48 labeled units each; one-prefix geometry	Robustness to official-like prefix shapes
Released test measurement features	19 unlabeled test summaries; prediction only	Primary challenge evaluation

For official evaluation, the selected gated residual model converted its unit-level EOL estimates into the required RUL predictions for the released test identifiers.

6. RESULTS

Table 4 combines the official result fields with the fixed diagnostic ablation, while Figures 6 and 7 visualize the corresponding local MAE and calibration behavior. The official challenge result is treated as the external evaluation outcome; the local diagnostics are used only to interpret component behavior and safeguards. For this ablation, the base prefix regressor is a direct EOL baseline trained on the engineered prefix features only, before adding source-prior predictions, residual anchors, specialist corrections, or gating.

Figure 5 first links the unit-level EOL estimates to the submitted RUL curves. The examples were selected by ranking the displayed fixed-diagnostic units by trajectory error, and the source-prior curve is obtained by subtracting the observed cycle from the source-prior EOL estimate. The panels show that the final gated trajectory generally remains bounded around the source-prior trajectory, while the high-error prefix remains difficult. This trajectory view frames the ablation below: Figure 6 should be read as a sequence of modeling roles rather than as a monotonic training trace. The prefix and source-prior stages measure available lifetime information, the anchor stages impose the conservative residual frame, and the specialist and gate stages show where complementary evidence is allowed to move the estimate.

Table 4. Official result summary and fixed diagnostic ablation. MAE and Q90 AE are EOL-cycle errors; lower local diagnostic values are better.

Item		Value
Official aggregate score		0.7741
Test identifiers		19
Fixed diagnostic ablation	**MAE**	**Q90 AE**
Base prefix regressor	360.2	1318.2
Source-prior blend	152.1	419.2
Robust residual anchor	161.3	449.1
Single-prefix calibrated anchor	167.8	441.3
Retrieval specialist	86.6	195.6
Signal-group specialist	86.3	203.1
Final residual gate	85.7	192.5

The ablation is intentionally read as a sequence of modeling roles, not as a monotonic leaderboard. The source-prior blend is the strongest simple local baseline, reducing fixed diagnostic MAE from 360.2 to 152.1 cycles. The robust anchor and single-prefix calibration stages then trade a small local MAE increase for constraints that later specialists rely on: source-envelope clipping, EOL lower bounds, test-prefix alignment, and a bounded residual frame. Figure 6 makes

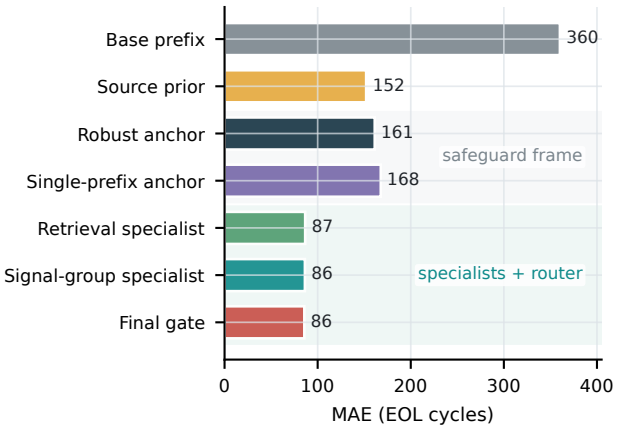

Figure 6. Stage-wise fixed diagnostic MAE. The largest measured reductions occur when the source-prior estimate first replaces the raw prefix baseline and when the retrieval and signal-group specialists are introduced. The shaded anchor stages primarily define the bounded residual frame used by later corrections.

this role explicit by placing the anchor stages between the large source-prior gain and the later specialist gains. Because the fixed diagnostic ablation alone cannot prove tail-risk reduction for those intermediate stages, we interpret them as safeguards that structure later corrections rather than as independent accuracy gains.

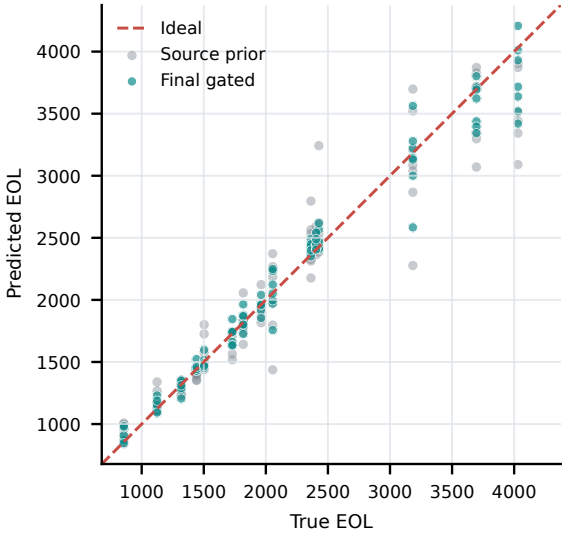

Figure 7. Local EOL calibration view comparing source-prior and final gated EOL estimates against true EOL on the fixed diagnostic view.

Figure 5 and Figure 7 provide complementary behavioral checks. The trajectory panels trace how EOL estimates translate into RUL curves over observed prefixes, while the calibration plot shows that the final gated estimates remain concentrated around the true-EOL diagonal while preserving the source-prior structure. The low- and median-error examples show that the final gated prediction can remain close to the true trajectory across several prefixes. The high-error example highlights a regime in which the available prefix evidence is insufficient to fully correct the EOL estimate. Together, the figures and ablation table show the central result: the major measured local error reduction comes from the retrieval and signal-group residual specialists, while the final gate mainly acts as a bounded router that modestly improves tail behavior.

7. DISCUSSION

The diagnostics suggest that the method benefited from avoiding a single all-purpose correction. The source-prior and conservative anchor locate a plausible EOL region. The single-prefix calibration adapts this surface to the official inference setting. The retrieval specialist exploits analog-based retrieval when neighbor support is trustworthy. The signal-group specialist captures residual evidence when shock, position, or current signals are more informative than memory. The gate combines these behaviors while bounding the effect of either branch.

This structure also explains why the final gate changes only a subset of rows. When the anchor and specialist predictions already agree, large additional corrections are unnecessary. The gate is not intended to alter every row substantially; it makes evidence-dependent routing decisions when the residual branches disagree. The source-disagreement diagnostics in Figure 2 motivate this behavior; additional observed-fraction, transfer-overlap, test-prefix, and grouped operating-condition diagnostics can be produced from the released diagnostic scripts.

Although the implementation was tuned for the PHME 2026 subway-door data, the gated residual strategy is not tied to this actuator. The transferable part is the decomposition: learn or import a strong but imperfect lifetime prior, estimate when that prior is uncertain, train residual specialists that use complementary evidence sources, and let a bounded gate decide how much each residual correction may move the prior. In another PHM dataset, the source prior could come from physics-based degradation models, similarity retrieval, or a neural sequence model; the residual specialists could be defined over domain-specific health indicators such as vibration bands, thermal trends, lubrication features, or operating-regime descriptors. The main requirement is that uncertainty can be represented by measurable disagreement, dispersion, retrieval support, or health-indicator reliability, so that the gate controls corrections instead of becoming another unconstrained lifetime predictor.

8. LIMITATIONS

The system is more complex than a single end-to-end learner. It depends on engineered prefix features, source-prediction features, and several anchor-construction stages. This improves control over small-data uncertainty, but it also makes the approach less compact than a single regressor. The official server released only an aggregate public score for the submitted file, so we cannot decompose the result into official RMSE, precision, and prognostic-horizon components. The local component diagnostics are therefore explanatory rather than official score decompositions. In addition, the model is evaluated only on operating conditions represented in the released training data; behavior under unseen operating-condition values remains untested. With 48 training units, source-disagreement and gate-routing diagnostics are necessarily low-resolution estimates of uncertainty. The gate is therefore analyzed as a routing mechanism rather than claimed as an independently calibrated uncertainty classifier. The transfer weighting uses released test measurement features in a transductive but label-free manner; no hidden RUL labels or official test errors are used. Future work should investigate simpler architectures that preserve the same uncertainty-aware routing principles with fewer feature-defined specialist stages.

9. CONCLUSION

We presented a gated residual EOL method for the PHME 2026 Data Challenge. The method starts from a source-prior EOL anchor, adapts it to the single-prefix inference setting, applies bounded retrieval and signal-group residual corrections, and blends the specialists with a clipped uncertainty-aware gate. The approach was designed for a setting in which test units are only partially observed, operating conditions are hidden, and the available run-to-failure data are limited.

The main lesson from the diagnostics is that useful corrections should remain tied to the uncertainty evidence that motivates them. Source predictions provide a strong initial EOL location, but their disagreement also identifies regimes where a residual correction may be needed. Retrieval support, signal-group activity, and branch disagreement then give the model a way to adjust the prediction while retaining bounds on how far each correction can move the final estimate. In this sense, the proposed stack is not only an accuracy-oriented ensemble, but also a controlled routing mechanism for small-data prognostics.

More broadly, the study suggests that conservative residual architectures can be useful for variable-prefix RUL prediction when a strong prior exists but its errors are regime dependent. Future work should investigate simpler versions of the same uncertainty-aware routing idea, evaluate transfer to unseen operating conditions, and test whether online updates can preserve the safeguards while reducing the dependence on manually engineered specialist stages.

ACKNOWLEDGMENT

The authors thank the PHME 2026 Data Challenge organizers for providing the dataset, evaluation protocol, and official scoring infrastructure.

BIOGRAPHIES

Giuseppe Mannone studied Autonomous Systems at the University of Stuttgart in Germany and received his M.Sc. in October 2024. Since 2024, he has been working as a research assistant in the field of reliability engineering at the Institute of Machine Components. His research focuses on applying artificial intelligence to reliability engineering.

Tim Herrmann studied Mechanical Engineering at the University of Stuttgart in Germany and received his M.Sc. in 2025. Since July 2025, he has been working as a research assistant in the field of reliability engineering at the Institute of Machine Components, University of Stuttgart. He focuses on computational and data-driven methods in reliability engineering.

Martin Dazer works as Head of the Reliability & Driveline Department at the Institute of Machine Components. He received his doctoral degree in reliability engineering from the University of Stuttgart, Germany, in 2019 and completed his habilitation in 2025. His fields of interest include reliability testing, lifetime data analysis, and statistical test planning.

REFERENCES

Beirami, H., Calzà, D., Cimatti, A., Islam, M., Roveri, M., & Svaizer, P. (2020). A data-driven approach for RUL prediction of an experimental filtration system. In *Proceedings of the European Conference of the PHM Society* (Vol. 5). doi: 10.36001/phme.2020.v5i1.1318

Eker, O. F., Camci, F., & Jennions, I. K. (2014). A similarity-based prognostics approach for remaining useful life prediction. In *Proceedings of the European Conference of the PHM Society* (Vol. 2). doi: 10.36001/phme.2014.v2i1.1479

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. doi: 10.1080/00401706.1970.10488634

İnce, K., Sirkeci, E., & Genç, Y. (2020). Remaining useful life prediction for experimental filtration system: A data challenge. In *Proceedings of the European Conference of the PHM Society* (Vol. 5). doi: 10.36001/

phme.2020.v5i1.1317

Jardine, A. K. S., Lin, D., & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*, *20*(7), 1483–1510. doi: 10.1016/j.ymssp.2005.09.012

PHME 2026 Organizing Committee. (2026a). *Data Challenge.* PHME 2026 conference webpage. Retrieved from https://phm-europe.org/data-challenge (Official challenge webpage, accessed May 31, 2026)

PHME 2026 Organizing Committee. (2026b). *Data Challenge PHME 2026.* PHME 2026 challenge handout. Retrieved from https://phm-europe.org/wp-content/uploads/2026/01/Data_Challenge_PHME_2026.pdf (Official challenge handout, accessed May 31, 2026)

Saxena, A., Celaya, J., Saha, B., Saha, S., & Goebel, K. (2010). Metrics for offline evaluation of prognostic performance. *International Journal of Prognostics and Health Management*, *1*(1). doi: 10.36001/ijphm.2010.v1i1.1336

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, *90*(2), 227–244. doi: 10.1016/S0378-3758(00)00115-4

Si, X.-S., Wang, W., Hu, C.-H., & Zhou, D.-H. (2011). Remaining useful life estimation–a review on the statistical data driven approaches. *European Journal of Operational Research*, *213*(1), 1–14. doi: 10.1016/j.ejor.2010.11.018

Soualhi, M., Nguyen, K. T. P., Medjaher, K., Nejjari, F., Puig, V., Blesa, J., Quevedo, J., & Marlasca, F. (2023). Dealing with prognostics uncertainties: Combination of direct and recursive remaining useful life estimations. *Computers in Industry*, *144*, 103766. doi: 10.1016/j.compind.2022.103766

Su, H., & Lee, J. (2024). Machine learning approaches for diagnostics and prognostics of industrial systems using open source data from PHM data challenges: A review. *International Journal of Prognostics and Health Management*, *15*(2). doi: 10.36001/ijphm.2024.v15i2.3993

TV, V., Gupta, P., Malhotra, P., Vig, L., & Shroff, G. (2018). Recurrent neural networks for online remaining useful life estimation in ion mill etching system. In *Proceedings of the Annual Conference of the PHM Society* (Vol. 10). doi: 10.36001/phmconf.2018.v10i1.589

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. doi: 10.1016/S0893-6080(05)80023-1

Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R. X. (2019). Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing*, *115*, 213–237. doi: 10.1016/j.ymssp.2018.05.050