Digital Chemical Engineering 20 (2026) 100324

Contents lists available at ScienceDirect

Digital Chemical Engineering

journal homepage: www.elsevier.com/locate/dche

Original Article

Run-to-Run control of Area-Selective Atomic Layer Deposition of SiO_2 on Al_2O_3 and SiO_2 surfaces with intermediate etching steps

Chun-Pei Lin [a], Feiyang Ou [a], Gerassimos Orkoulas [c], Panagiotis D. Christofides [a,b,*]

[a] Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, CA, 90095-1592, USA
[b] Department of Electrical and Computer Engineering, University of California, Los Angeles, CA 90095-1592, USA
[c] Department of Chemical Engineering, Widener University, Chester, PA 19013, USA

ARTICLE INFO

Keywords:
Run-To-Run control
Semiconductor manufacturing processes
CFD simulation
Multiscale simulation
Area-selective atomic layer deposition
Atomic layer etching

ABSTRACT

Area-Selective Atomic Layer Deposition (ASALD) has emerged as a critical bottom-up patterning technique for sub-5 nm semiconductor manufacturing, offering a path to circumvent the alignment challenges of traditional top-down lithography, and prior work involved intermediate Atomic Layer Etching (ALE) to build ASALD-ALE pipeflow to retain selectivity under imperfect inhibition. However, the sensitivity of ASALD-ALE cycles to kinetic disturbances necessitates the development of advanced process control to sustain selectivity. In this work, a Run-to-Run (R2R) control framework based on the Exponentially Weighted Moving Average (EWMA) R2R control algorithm was developed for an integrated multiscale ASALD-ALE process presented in (Ou et al., 2026). The framework was evaluated against a simulated 20% kinetic disturbance in adsorption rates using a multiscale model that couples microscopic surface kinetics with reactor-scale pressure dynamics. Furthermore, a severe 40% kinetic degradation was simulated to validate the robustness of non-rate-limiting steps (oxidation and fluorination), confirming they achieve full saturation within baseline times and justifying their exclusion from the active control loop. A primary innovation of this study is the implementation of a dynamic slope (α) updating scheme – utilizing both pure EWMA and a linear regression hybrid model – to re-characterize the process gain of the adsorption system as the system shifts into a slower kinetic regime. This adaptive approach ensures the controller accurately mirrors the exponential nature of surface reactions, significantly reducing estimation errors compared to static-slope models. Furthermore, a novel sequential protocol was established to pass refined model parameters (α, β) and updated reaction times between interdependent steps. Simulations demonstrate that the connected control architecture effectively mitigates the staircase-like profile caused by discrete time resolutions and eliminates the transient recalibration periods typically observed in isolated control loops. The results show that the integrated sequence requires only 10 cumulative batches to reach target coverages which are 100% saturation for adsorption and 0% for etching, while maintaining high process selectivity. This study provides a robust, high-efficiency digital twin and control methodology for optimizing commercial ASALD operations under realistic manufacturing variability.

1. Introduction

The continued miniaturization of semiconductor devices, driven by the transition to sub-5 nm nodes and complex three-dimensional architectures like Gate-All-Around (GAA) transistors, has placed unprecedented demands on fabrication precision (Mukesh and Zhang, 2022). Traditional top-down manufacturing approaches struggle with cumulative edge placement errors (EPE), which severely degrade device performance. At the process level, inhibitor-assisted Area-Selective Atomic Layer Deposition (ASALD) typically proceeds through an ABC-type cycle, in which a small-molecule inhibitor selectively chemisorbs on the non-growth area (NGA) in Step A, the precursor adsorbs predominantly on the unblocked growth area (GA) in Step B, and the co-reactant converts the chemisorbed precursor into the target film in Step C. Because this chemistry is intrinsically bottom-up and self-aligned, ASALD has emerged as a promising route for mitigating alignment-limited pattern transfer and complementing conventional top-down integration schemes (Mameli et al., 2017; Mackus et al., 2019; Mameli and Teplyakov, 2023). By utilizing small-molecule inhibitors to chemically deactivate non-growth areas (NGA) while allowing deposition on growth areas (GA), ASALD facilitates self-aligned fabrication (Merkx et al., 2020). However, imperfect inhibition often leads to unwanted

* Corresponding author.
E-mail address: pdc@seas.ucla.edu (P.D. Christofides).

https://doi.org/10.1016/j.dche.2026.100324
Received 29 May 2026; Received in revised form 24 June 2026; Accepted 24 June 2026
Available online 30 June 2026

precursor nucleation on the NGA. More generally, selectivity is eventually lost because of pre-existing defects, new defect formation during cycling, and the distribution of molecular reactivities on real surfaces. As a result, recent ASALD strategies increasingly integrate periodic etch-back or atomic layer etching (ALE) correction steps to remove NGA defects and sustain long-term selectivity during thicker growth (Song et al., 2019; Karasulu et al., 2023).

While the fundamental chemistries and multiscale dynamics of ASALD and ALE have been extensively modeled, real-world semiconductor manufacturing environments are inevitably subject to unknown operational disturbances, process drifts, and equipment variations. To ensure consistent film quality and process uniformity in the presence of such variability, advanced process control strategies are essential. In semiconductor manufacturing, Run-to-Run (R2R) control is a supervisory batch-to-batch control strategy in which metrology acquired after one run is used to update the recipe for the next run, making it especially useful when high-value measurements are available only ex situ or with delay (Sachs et al., 2002; Del Castillo and Hurwitz, 1997; Liu et al., 2018). Among R2R methods, the Exponentially Weighted Moving Average (EWMA) controller remains attractive because it recursively updates the model prediction with exponentially decaying memory, thereby giving greater influence to recent runs while preserving a relatively simple control structure. In its classical semiconductor implementation, EWMA mainly updates the model bias/intercept and typically assumes a fixed process gain; this is effective for compensating modest shifts and drifting disturbances, but performance can become sensitive when the disturbed process no longer follows the original gain estimate (Ingolfsson and Sachs, 1993; Del Castillo, 2001; Tom et al., 2023).

Run-to-Run (R2R) control, particularly utilizing the Exponentially Weighted Moving Average (EWMA) algorithm, has proven highly effective in adjusting batch recipes to compensate for process shifts in thermal deposition and etching systems (Liu et al., 2018; Yun et al., 2022c; Wang et al., 2025). Despite these advancements, a significant gap remains in applying R2R control to complex, multi-step ASALD sequences integrated with corrective etching. Prior atomic-layer control studies have generally addressed structurally simpler problems. For example, Tom et al. (2023) considered an ASALD spatial reactor using a single aggregate output (growth per cycle) and one principal manipulated variable (substrate rotation speed), explicitly excluding Step A inhibition from the R2R layer and relying on an EWMA controller with a constant slope (Tom et al., 2023). Wang et al. addressed a standalone ALE process in a discrete-feed reactor, where control focused on the two ALE half-cycle times and leveraged endpoint information extracted from in-run pressure profiles (Wang et al., 2025). These prior formulations are not directly sufficient for the present problem because the combination of ASALD and ALE relies on a delicate balance between sequential adsorption, oxidation, and etching steps, so that a disturbance in one phase changes the surface state inherited by the subsequent phases. Inference from the prior literature with the form of cross-step state coupling violates the weak-coupling, single-output, and fixed-gain assumptions that underpin many earlier EWMA implementations, explaining why a controller that updates only the model bias is unlikely to remain efficient after a large kinetic shift (Del Castillo, 2001; Del Castillo and Rajagopal, 2002; Tom et al., 2023; Yun et al., 2022c; Wang et al., 2025).

In this work, we develop an EWMA-based process control framework for an ASALD process with intermediate etching presented in (Ou et al., 2026). The target process consists of a five-step cycle, as illustrated in Fig. 1, where Step A (inhibitor adsorption), Step B (precursor adsorption), and Step D (etching/conversion) represent the critical rate-limiting stages, while Step C (ozone oxidation) and Step E (fluorination) are sufficiently rapid that they are not assigned independent EWMA time updates. It should be emphasized that this exclusion applies only to the set of actively manipulated reaction times in the R2R controller, not to the physical propagation of surface states. In the multiscale sequence, the terminal surface configuration after each step is still inherited by the following step. Therefore, any incomplete conversion in Step C or Step E, if present, would only modify the downstream surface state and would be reflected in the subsequent coverage or residual-coverage measurements used by the controlled steps. The active control loop is restricted to Steps A, B, and D because Steps C and E reach their target terminal states within the fixed 1.0 s recipe time even under the severe disturbance test discussed in Section 2.6. To evaluate the robustness of the control scheme, we introduce a simulated manual disturbance that decreases the adsorption rate by 20%, which is a kinetic shift significantly slower than the surface reaction rate. The primary control objective is to dynamically tune the process times across batches to ensure that Steps A and B recover to 100% surface coverage, and Step D achieves 0% unwanted coverage under the applied disturbance without losing too much efficiency.

The primary innovation of this study lies in the development of an adaptive, multi-step control algorithm formulated via a Python script. Critically, our framework introduces a dynamic updating scheme for the process slope (α) alongside the traditional model bias (β). Unlike standard EWMA controllers that rely on static baseline slopes, our model re-characterizes the process gain in real time, allowing the controller to adapt to the significantly altered kinetic state caused by manufacturing disturbances. Furthermore, the framework utilizes a novel sequential updating mechanism designed to optimize the entire process holistically. By algorithmically connecting the steps, the control parameters and adjusted reaction times for Step B are initialized based on the results of Step A. This prevents subsequent steps from having to recalibrate from the baseline and drastically reduces the total number of batches required to reach the target. This interconnected logic is further applied to Step D using a scaling parameter to account for its faster kinetic nature. Ultimately, this connected R2R control strategy minimizes the batch-to-batch transient period, providing a robust, highly efficient digital twin and control guideline for commercial ASALD operations facing realistic manufacturing disturbances.

2. Run-to-run control

2.1. Overview

To ensure consistent film quality and mitigate the cascading impacts of process variations in multi-step semiconductor manufacturing, an advanced supervisory process control framework is required. The primary methodological innovations of this work lie in the development of a Python-based engineering process controller (EPC) that introduces two interconnected strategies: an online gain adaptation engine utilizing a dynamic process slope (α) updating scheme that continuously re-characterizes the system sensitivity in real time to handle severe kinetic shifts, coupled with an inter-step sequential relay protocol that acts as an automated feed-forward mechanism to chain distinct processing phases by passing refined parameters (α, β) and updated reaction times sequentially forward to eliminate transient recalibration periods.

To lay out a visual structural guide of this coupled architecture, Fig. 2 displays the physical integration between the microscopic kinetic Monte Carlo (kMC) lattice, the macroscopic Computational Fluid Dynamics (CFD) reactor model, and the supervisory control loop. Alongside this multi-scale plant description, Fig. 2(b) details the exact algorithmic workflow executed by the controller to process data, update adaptive variables, and determine recipes across successive runs.

Run-to-Run (R2R) control operates as a discrete supervisory batch-to-batch strategy where recipe profiles are modified exclusively between individual runs according to a prescribed input–output relationship. Unlike continuous process control loops which adjust parameters within an active run, R2R evaluates the process response only after a batch is completed, computing an optimized input for the subsequent cycle. In advanced atomic layer processing, this structure is especially vital because real-time measurement of key metrology parameters like

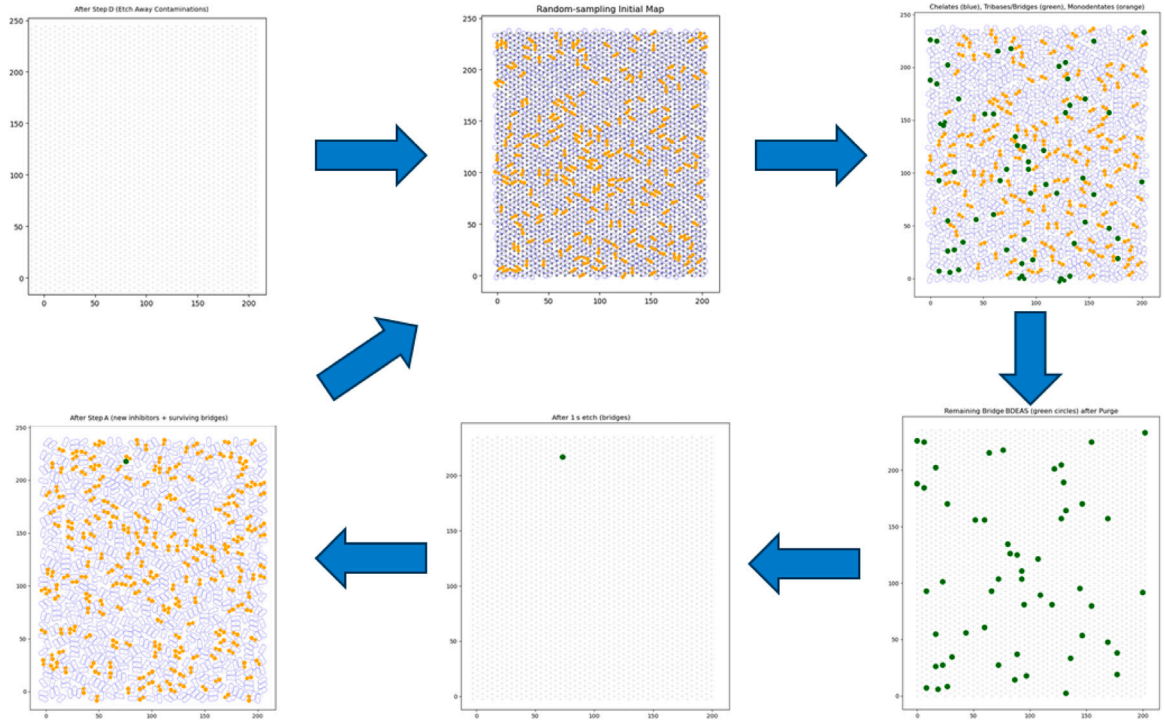


Fig. 1. The ASALD-ALE process flow.

film thickness and edge placement error (EPE) are functionally impossible during execution, meaning control actions must rely on ex situ or post-run metrology updates.

To communicate the methodology directly, the control formulation in this work is organized to progress from high-level architectural innovation down to localized mathematical initialization:

- **Supervisory R2R Architecture with Sequential Relay Logic:** This block establishes the feed-forward communication protocol across interdependent phases. By algorithmically connecting the steps, the control parameters learned in Step A are passed directly to initialize Step B, and scaled by a factor of 0.44 to govern the faster etching kinetics of Step D, optimizing the entire cycle holistically rather than treating half-cycles as isolated loops (Yun et al., 2022c).
- **Real-Time Gain Adaptation (Dynamic α Updating Engine):** This module bypasses the traditional limitation of fixed-slope Exponentially Weighted Moving Average (EWMA) controllers under large process drifts. The script calculates a dynamic α update first – utilizing either a recursive linear regression hybrid model or a secondary EWMA loop – to capture the physical reality of the altered kinetic regime before the intercept is modified, avoiding the double-estimation of process errors (Chien et al., 2014).
- **Local System Identification (First-Order Modeling and Fit):** Acting primarily as a secondary means of obtaining initial controller parameters, this block utilizes baseline, disturbance-free coverage histories to initialize the model. Surface kinetics are locally mapped to a first-order transfer function $G(s) = \frac{K}{\tau s+1}$, where a natural logarithmic transformation is subsequently applied to establish the linear baseline relation $\phi = \alpha(t - t_0) + \beta$ required by the supervisory updating equations (Box, 1957).
- **Step-Specific Boundary Constraints and Data Collection:** This section defines the discrete metrics used to feed the script. Grid configurations are algorithmically isolated exclusively from the rate-limiting Non-Growth Area (NGA), accounting for spatial blocking via a sterically-constrained maximum capacity parameter ($N_{\text{max\_capacity}}$) to handle 100% adsorption saturation targets alongside the reverse log-transformation required for 0% etching defect targets (Ou et al., 2026).

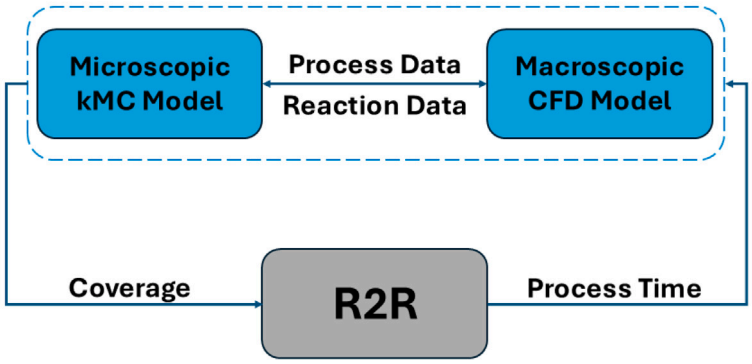
2.2. Supervisory R2R architecture and inter-step sequential relay logic

While the generalized EWMA framework provides a robust foundation, the multiscale nature of ASALD requires specific mathematical adaptations for each half-cycle. Furthermore, rather than operating isolated controllers for each sequence, this framework introduces a sequential connection protocol that feeds control parameters forward, holistically optimizing the entire process.

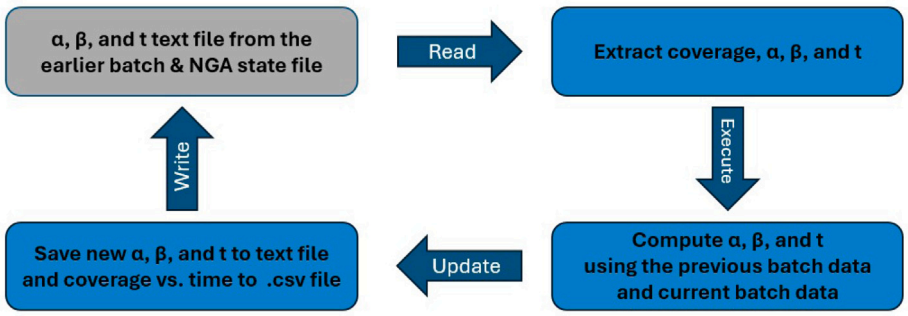
It should be clarified that the term "multivariable" in this work does not refer to a conventional simultaneous MIMO (multi-input multi-output) controller that computes all reaction times from a single vector of aggregate outputs. Such a formulation is impractical for the present ASALD-ALE sequence because final-cycle metrology, such as net film thickness, cannot uniquely separate the contributions of inhibitor adsorption, precursor adsorption, oxidation, etching, and fluorination. For example, the same final thickness may result from different combinations of incomplete Step B deposition and Step D etching, and these steps require corrections with different signs, magnitudes, and target coverages. Therefore, the present framework uses step-specific SISO (single-input single-output) EWMA updates for the rate-limiting phases, while capturing inter-step coupling through sequential transfer of the surface state and the updated control parameters (α, β, t). In this sense, the proposed architecture is better understood as a sequentially connected multi-step R2R control framework with multiple manipulated reaction times and multiple step-specific controlled outputs, rather than as a fully simultaneous MIMO controller.

2.2.1. Step A (Inhibitor adsorption)

As the initiation phase of the cycle, Step A does not require historical batch data from any prior steps. The control sequence defaults the initial batch run time to $t_0 = 2.8$ s. Under the 20% kinetic disturbance, simulations reveal that without active control intervention, the inhibitor coverage on the Non-Growth Area (NGA) falls to 97.9%,



(a)



(b)

Fig. 2. Run-to-run control of the multiscale area selective atomic layer deposition. (a) The overall R2R control interaction loop. (b) R2R control script workflow.

emphasizing the necessity of the R2R parameter update to restore complete surface passivation. Because Step A involves selective adsorption of the inhibitor exclusively on the Non-Growth Area (NGA), there is no physical reaction occurring on the Growth Area (GA). Consequently, the script exclusively extracts grid data for the NGA in Fig. 3 to feed into the EWMA algorithm.

2.2.2. Step B (Precursor adsorption)

Step B utilizes a baseline undisturbed run time of $t_0 = 2.7$ s. A similar trend is observed for Step B under the 20% kinetic disturbance, where the multiscale simulation demonstrates that without active control intervention, the precursor coverage drops significantly to 80.4%, leading to incomplete surface saturation. If simulated as an entirely independent control loop, Step B would utilize the saturated NGA grid data from a completed Step A run but rely strictly on its own independently fitted parameters (K, τ, and intercept) from the data in Fig. 4. However, in the sequentially connected control scheme, Step B directly inherits the final updated intercept (β) and reaction time (t) from the saturated Step A to use as its initial values. Because the undisturbed baseline run times for both adsorption steps are nearly identical ($t_{0A} = 2.8$ s and $t_{0B} = 2.7$ s), this parameter handoff is executed directly without the introduction of an empirical transfer constant or scaling multiplier (Tom et al., 2023). Furthermore, Step B's NGA interactions involve more complex molecular species, making it significantly more susceptible to kinetic disturbances (Ou et al., 2026). Feeding forward Step A's control adjustments without an intermediate scaling

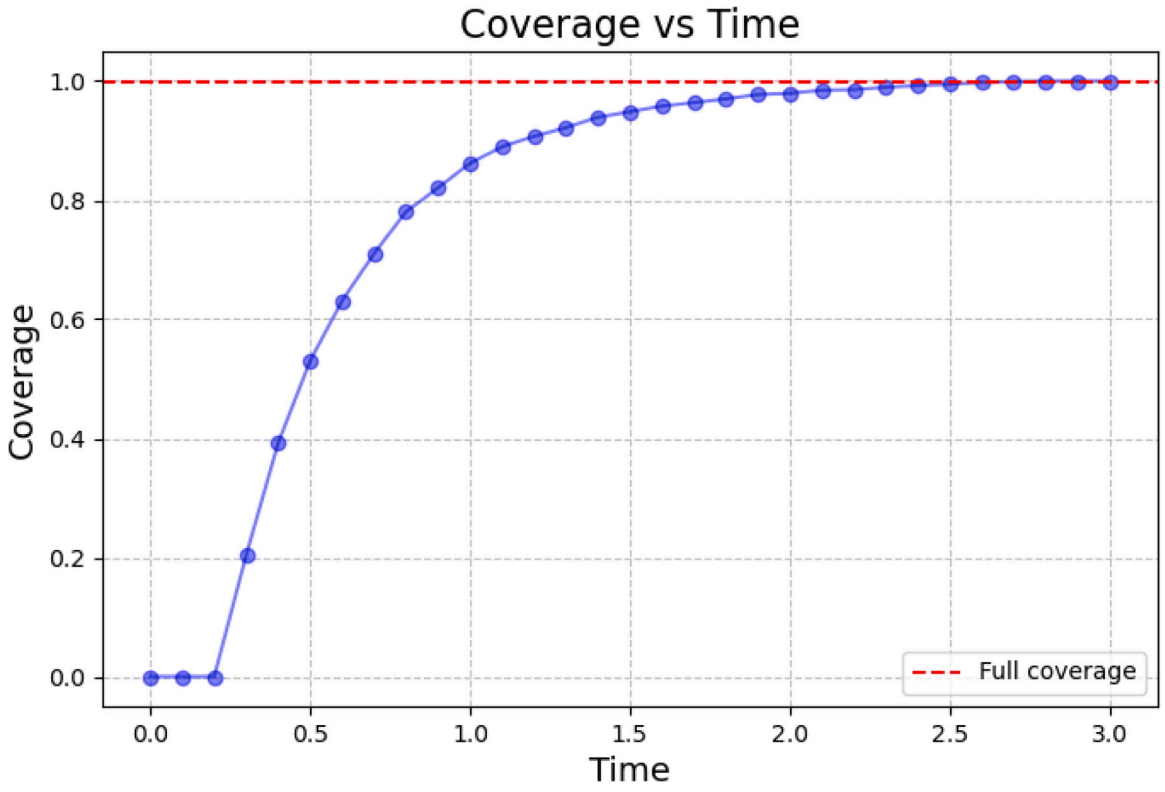


Fig. 3. The coverage vs time evolution in original batch run without disturbance for Step A.

penalty provides Step B with an advanced, pre-calibrated starting point, drastically reducing the cumulative batch count required to achieve complete surface saturation (Yun et al., 2022c).

Following Step B, Step C (ozone oxidation) proceeds normally without EWMA control, as its rapid kinetics natively reach saturation without bottlenecking the process.

2.2.3. Step D (Intermediate etching) and reverse control

As we can see from Fig. 5 Step D introduces a unique control challenge where the primary objective is to reach a target defect coverage of 0% on the NGA, rather than 100% saturation. The baseline recipe establishes an initial batch run time of $t_0 = 1.8$ s for this etching phase. Under the applied 20% disturbance, simulations reveal that without active control intervention, the unwanted precursor nucleation on the NGA increases to 10.2%, severely compromising process selectivity. Because the coverage trajectory decays toward zero, the linear transformation requires a reverse control model. To prevent fatal mathematical errors associated with calculating the natural logarithm of zero, a minuscule constant, ϵ, is introduced. The measured coverage is adjusted to $y_{\text{current}} + \epsilon$, and the target becomes $y_{\text{target}} = \epsilon$. The transformed variable ϕ is subsequently modified to:

$$\phi = \ln\left(\frac{y}{K}\right) \tag{1}$$

The EWMA bias updating equation is correspondingly adjusted to match this reverse logic in Eq. (2), where α_{updated} represents the real-time adjusted process slope computed by the adaptive gain engine discussed in the following section:

$$\beta_{\text{new}} = (1-\lambda)\beta_{\text{old}} + \lambda\left[\ln\left(\frac{y_{\text{current}}}{K}\right) - \alpha_{\text{updated}}(t-t_0)\right] \tag{2}$$

The final time update calculation remains structurally identical, utilizing this new ϕ, β_{new} and α_{updated}.

Unlike the seamless parameter handoff between the adsorption phases, the baseline timescales for Step B ($t_{0B} = 2.7$ s) and Step D ($t_{0D} = 1.8$ s) are substantially different, rendering a direct unscaled handoff of the reaction time physically inappropriate. To connect the control outputs, a rigid scaling factor must be derived based on the physical principle of uniform fractional kinetic degradation across successive stages. Assuming a shared reactor disturbance lengthens the required time for Step B by an operational multiplier x, the updated time becomes $b = x \cdot t_{0B}$. To preserve process symmetry, the proportional lengthening for the intermediate etching phase must be scaled by the ratio of their original time constants, meaning the run time for Step D (d) expands as $d = t_{0D} \cdot (t_{0D}/t_{0B}) \cdot x$. Substituting the operational disturbance multiplier from the preceding step ($x = b/t_{0B}$) yields the following quadratic scaling law:

$$d = b\left(\frac{t_{0D}}{t_{0B}}\right)^2 \tag{3}$$

By inserting our exact baseline parameters ($t_{0D} = 1.8$ s and $t_{0B} = 2.7$ s), this analytical relationship yields a scaling factor of approximately 0.44. Rather than functioning as an empirical tuning variable, this multiplier serves as a mathematically consistent translation rule that safely down-scales the learned time adjustments to match the faster surface kinetics of the etching regime, effectively preventing unaccounted material loss on the growth area.

2.3. Real-time gain adaptation engine (Dynamic α updating)

The mathematical selection of EWMA and recursive linear regression laws for this framework is explicitly predicated on their application to the linearized transformation (ϕ) rather than the raw, unmapped nonlinear surface kinetics (Yun et al., 2022c). Because the actual multiscale process is non-affine, coverage-dependent, and highly sensitive

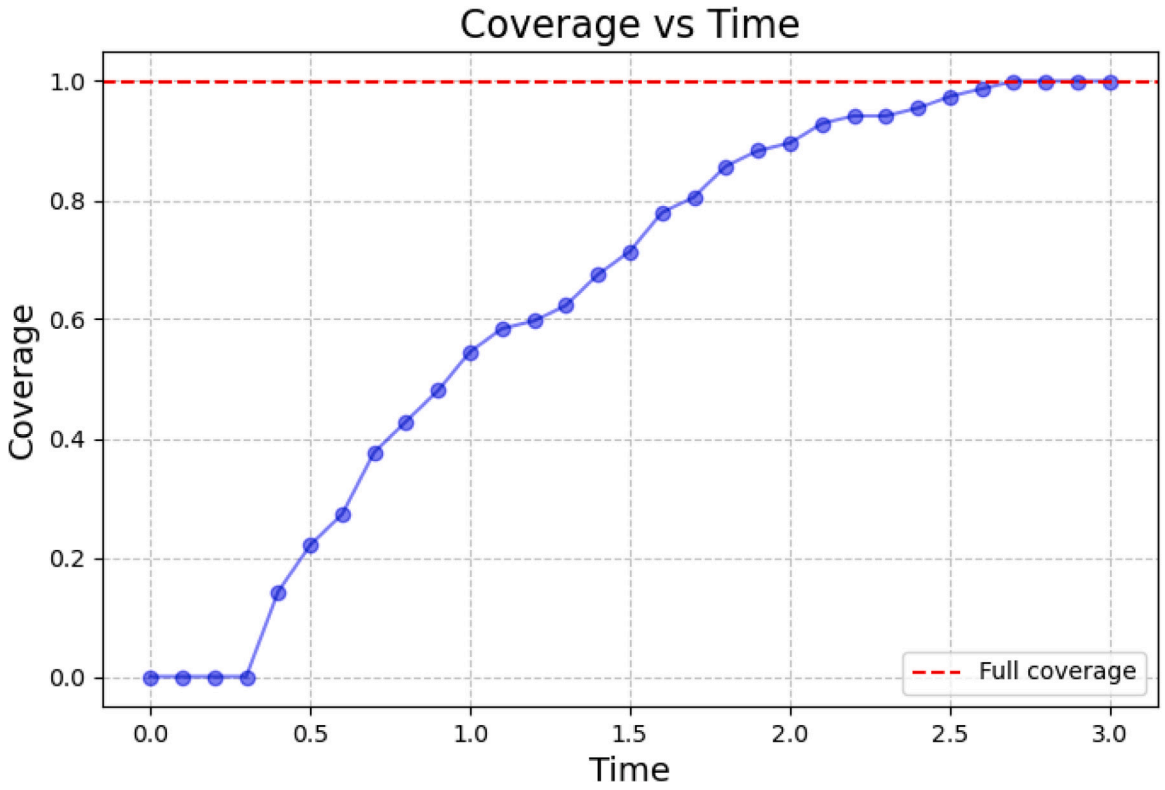


Fig. 4. The coverage vs time evolution in original batch run without disturbance for Step B.

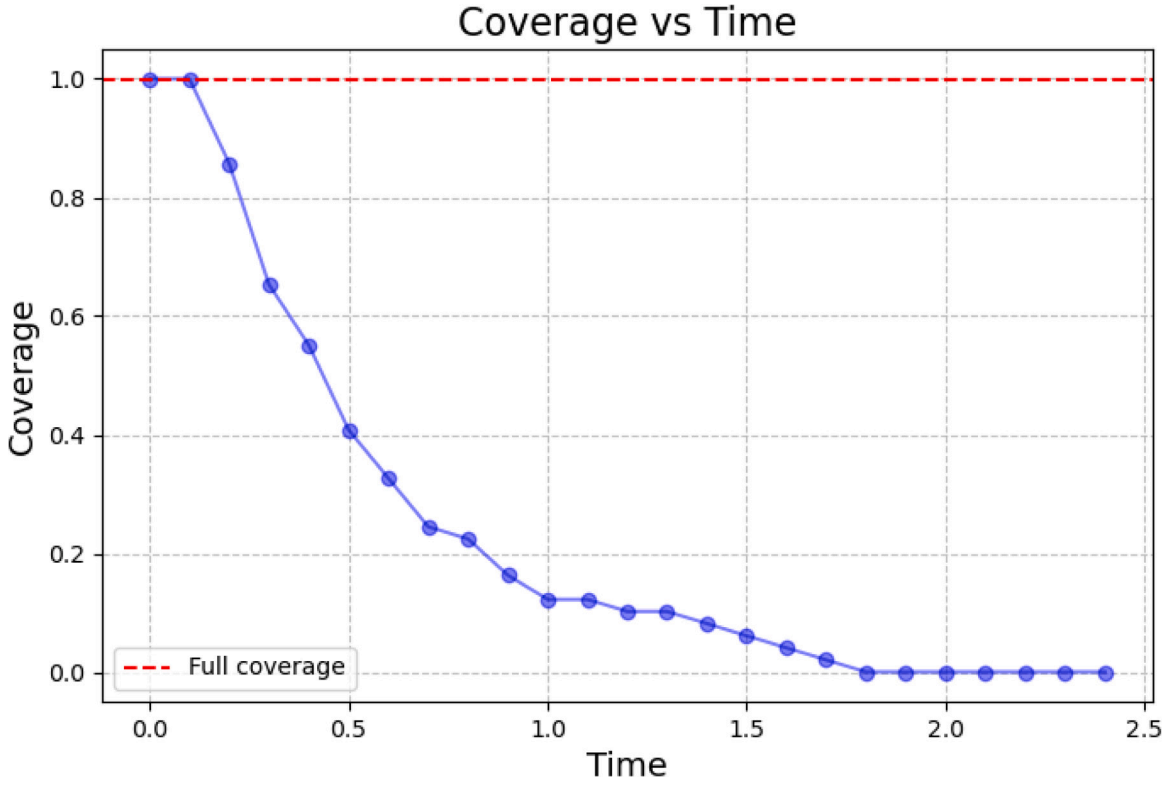


Fig. 5. The coverage vs time evolution in original batch run without disturbance for Step D.

to external drifts, a static linear model inevitably suffers from severe structural mismatch outside its baseline validation point (Tom et al., 2023). Implementing a dual-adaptation scheme to update both α and β effectively treats the controller as a dynamic, online re-linearization engine that continuously aligns the local process gain with the actual physical state of the disturbed reactor (Yun et al., 2022c). Furthermore, managing recurring multi-step variations requires an optimal retention of historical batch memory to separate high-frequency stochastic noise from persistent kinetic shifts (Sachs et al., 2002). The chosen algorithms are uniquely suited for this task where the pure EWMA update leverages recursive exponential filtering to maintain a continuous trend line with minimal computational overhead (Del Castillo, 2001), while the hybrid regression framework explicitly minimizes cumulative tracking error over a localized history of recent runs to balance aggressive system recovery with long-term loop stability (Chien et al., 2014).

In standard semiconductor R2R control applications, a process recipe is modified between discrete "batches" to reject disturbances such as drift, shift, and equipment aging (Ou et al., 2026; Moyne, 2021). The classical exponentially weighted moving average (EWMA) controller therefore serves only as the baseline intercept-update law, with weight parameter $0 < \lambda \leq 1$, through the fixed-slope recursion $\beta_{k+1} = (1-\lambda)\beta_k + \lambda\left(y_k - \alpha x_k\right)$ (Ingolfsson and Sachs, 1993; Del Castillo, 2001; Tom et al., 2023), where β_{k+1} is the updated intercept value, and β_k is the current intercept value.

Under a large kinetic or gain shift, however, the disturbed plant is more appropriately written as $y_k = g\alpha x_k + \beta + \varepsilon_k$, where $g \neq 1$ is a multiplicative gain perturbation and ε_k is the extra disturbance term; substituting this relation into the intercept-only EWMA update gives $\beta_{k+1} = (1-\lambda)\beta_k + \lambda\left[(g-1)\alpha x_k + \beta + \varepsilon_k\right]$, which shows that a scalar bias term can cancel the mismatch only at one operating point $\bar{x}$. Whenever the recipe changes from batch to batch, the remaining mismatch is $(g-1)\alpha(x_k - \bar{x})$, so intercept-only EWMA cannot, in general, remove gain error. For this reason, the present ASALD/ALE controller updates both the slope α and the bias β: unlike the classical intercept-only formulation, dual adaptation can respond simultaneously to additive offset drift and multiplicative kinetic/gain changes in the disturbed process.

Quantitatively, the selected value of the weight parameter λ directly dictates the balance between controller responsiveness and process stability. A larger λ value (typically $\lambda \geq 0.3$, approaching 1.0) yields a highly aggressive control action, enabling rapid recovery from sudden process shifts or large kinetic disturbances because the controller places a heavy mathematical discount on older runs and prioritizes the most recent error feedback. However, this heightened responsiveness comes at the cost of increased sensitivity to high-frequency process noise and false-alarm volatility, which can lead to severe recipe oscillations and even closed-loop instability if the estimated process gain significantly deviates from the actual reactor physics. Conversely, a smaller λ value (typically $0.05 \leq \lambda \leq 0.15$) is less aggressive, functioning as a robust low-pass filter that effectively smooths out random stochastic noise and measurement variations by distributing the update weight across a long memory of historical runs. While this filtered approach maximizes stability and is ideal for tracking slow, persistent equipment drifts, it significantly delays the controller's adaptation rate, resulting in a prolonged transient period and high cumulative defect counts when the system is subjected to sudden, severe process disturbances. The closed-loop instability is largely not considered as an issue.

For the adsorption steps considered in this work, the controlled input is the batch reaction time and the measured output is the final surface coverage. Because the adsorption response is monotonic, bounded, and self-limiting over the operating region of interest, increasing the reaction time cannot decrease the achieved coverage. Therefore, the EWMA R2R update can be interpreted as a discrete integral action correction applied along a known-gain-direction map, rather than as a dynamic feedback loop with oscillatory plant dynamics. Under the maintained sign of the local process gain and bounded EWMA weights, the main tuning tradeoff is not closed-loop instability in the conventional continuous time sense, but the balance between convergence speed, noise sensitivity, and possible recipe overshoot.

The control script requires a suite of parameters to function: the steady-state gain (K), the time constant (τ), the baseline reaction time (t_0), the current batch reaction time (t), the measured coverage (y), and the model tuning parameters (α and β). Parameters K, τ, and t_0 are obtained from the initial baseline fitting of the undisturbed process.

To maintain modularity across discrete multiscale simulation runs, batch-to-batch parameters are not stored in the script's active memory. Instead, the script utilizes a robust file-based handoff system. At the start of each run, the script queries external .txt files to retrieve the reaction time (t), the bias (β), and the slope (α) from the preceding batch. If these files are absent (as in the first run), the script defaults to t_0, the initial fitted intercept, and the baseline slope calculated as $\alpha = -1/\tau$.

To ensure algorithmic reproducibility and clarify the practical implementation of the run-to-run controller, the parameter initialization, moving regression windows, and loop termination criteria are strictly bound by the following execution rules. The recursive linear regression module operates on a dynamic, localized moving window; it initializes immediately upon accumulating a minimum of two active data points (the current and immediate preceding batch runs) and expands to a maximum memory window of three consecutive data points to track the localized trend line without introducing historical lag (Wang et al., 2025). Closed-loop run-to-run control actions continue until the physical stopping criterion is satisfied, defined as the specific batch run where the rate-limiting Non-Growth Area (NGA) achieves its full target milestone (100% saturation for adsorption steps; 0% defect coverage for intermediate etching) (Yun et al., 2022c). Upon reaching this terminal boundary, the script queries high-resolution NGA surface status text files – logged at a spatiotemporal sampling frequency of 0.1 s by the multi-scale CFD digital twin – to quantitatively evaluate post-process recipe time overshoots (ANSYS, Inc., 2017; Yun et al., 2023). For consistency and clarity in performance illustrations, the simulation length for each distinct sensitivity run is automatically fixed to a standardized cumulative batch ceiling.

(i) **Slope (α) Update.** First, the script updates the process slope, α. While the baseline α provides an initial guess, the kinetic rate of a reaction under a 20% disturbance differs significantly from the original state due to the exponential nature of the surface kinetics (Ou et al., 2026). For both the pure EWMA and the linear regression hybrid models, there exists an initial "Batch 0" that is carried out without any controller update, running strictly at the standard baseline time t_0. While the pure EWMA approach can immediately update the slope using its recursive formula from Batch 1, the linear regression hybrid model requires a minimum of two data points to perform a fit. To address this limitation, for Batch 1, the script updates α simply by multiplying the baseline α with the fractional surface coverage resulting from Batch 0, which was run at the standard baseline time. Once Batch 1 is executed, the resulting coverage provides the second data point. Starting from Batch 2, the hybrid controller has the necessary history (from both Batch 0 and Batch 1) to perform a standard linear regression on the coverages and run times of recent runs to calculate the observed values for both α and β, which are then refined through the standard EWMA update law. In the pure EWMA approach, starting from Batch 1, α is dynamically updated using a recursive update law identical in structure to that used for the bias (β) (Wang et al., 2025). This adaptive slope update is given by:

$$\alpha_{\text{updated}} = (1-\lambda_\alpha)\alpha_{\text{old}} + \lambda_\alpha \alpha_{\text{observed}} \tag{4}$$

where $\lambda_\alpha \in (0, 1]$ is the slope weighting factor. In the pure EWMA approach, α_{observed} is the slope from the current batch used as the observed value. Conversely, for the linear regression hybrid model, starting from Batch 2, standard linear regression is performed on the coverages and run times of recent runs to calculate the observed slope

$\alpha_{observed}$, which is then refined through the EWMA update law in Eq. (4) to obtain the final updated slope $\alpha_{updated}$ for the subsequent run. This adaptive structure allows the controller to adjust its sensitivity to time changes ($t - t_0$) in real-time. By continuously refining both α and β, the model bias and gain gradually lose their dependence on the initial baseline and adapt to the specific kinetic reality of the current disturbance. Furthermore, updating α first ensures that the rate of change for the current batch is accurately characterized before the intercept is adjusted to its final value for the subsequent run.

(ii) **Bias (β) Update.** Next, the script calculates the updated model bias, β_{new}. This calculation utilizes the newly updated α rather than the historical value. By integrating the refined slope into the bias update equation, the controller avoids the "double estimation" of error that would occur if both parameters were updated using the same stale information:

$$\beta_{new} = (1-\lambda)\beta_{old} + \lambda\left[\ln\left(1 - \frac{y_{current}}{K}\right) - \alpha_{updated}(t - t_0)\right] \quad (5)$$

This sequential logic allows the model to adapt its gain and intercept in tandem, accurately mirroring the slower kinetic reality of the disturbed ASALD system.

Once the process model bias and slope are updated, the final control action computes the new process recipe (the required reaction time) for the next batch run to compensate for the disturbance. The script calculates the required time update (t_{new}) by rearranging the linear model and inserting the target coverage requirement:

$$t_{new} = \left(\frac{\phi_{target} - \beta_{new}}{\alpha_{updated}}\right) + t_0 \quad (6)$$

Here, $\phi_{target} = \ln(1 - y_{target}/K)$, where y_{target} represents the desired optimal coverage (e.g., saturation for the adsorption steps).

At the end of each calculation loop just before the script triggers the next multiscale simulation batch, t_{new}, β_{new}, and $\alpha_{updated}$ are written back into the parameter-tracking .txt format text file. Alongside this parameter handoff, it logs the current reaction time (t) and the measured Non-Growth Area coverage ($y_{current}$) to a cumulative history file. Consistently saving this batch-to-batch data to files is crucial for comprehensive post-process inspection and serves as the foundational data framework required to interconnect the interdependent steps via the sequential relay protocol outlined in Section 2.2.

2.4. Local system identification and model linearization

To initialize and feed the adaptive supervisory architecture described in Sections 2.2 and 2.3, the baseline, disturbance-free process data must be mapped to a local affine surrogate. Once the surface coverage is successfully extracted from the spatial grid configurations, the next objective is to translate this dynamic behavior into a mathematical model suitable for the Run-to-Run (R2R) controller. To achieve this, the coverage versus time data generated by the multiscale model under normal, disturbance-free conditions is utilized as the baseline process data (Ou et al., 2026). Because surface adsorption and etching kinetics often exhibit exponential saturation behavior, this baseline data is fitted to a standard first-order process control model. The fitting is performed using two complementary approaches: a non-linear regression to establish the overarching process parameters, followed by a linear transformation to isolate the tuning variables for the control script.

Once the NGA histories are available, traditional run-to-run (R2R) controller design is typically formulated through a local affine surrogate of the form $\hat{y}_k = \alpha x_k + \beta$, which in the present manuscript notation is equivalent to the linearized baseline relation in Eq. (9). In conventional semiconductor literature, $\hat{y}_k$ denotes the estimated post-run metrology response, x_k is the manipulated recipe variable, α represents the static process gain or slope, and β is the model bias or intercept (Sachs et al., 2002; Del Castillo and Hurwitz, 1997; Tom et al., 2023). This approximation is mathematically convenient because the classical semiconductor R2R framework operates under the assumption of a single-input/single-output (SISO) representation, fixed process gain over a narrow operating window, and delayed, ex situ metrology (Sachs et al., 2002; Del Castillo and Hurwitz, 1997; Del Castillo, 2001; Tom et al., 2023; Yun et al., 2022c).

However, because the underlying ASALD/ALE coverage dynamics are intrinsically nonlinear, self-limiting, and micro-structurally complex, this linear relation must be interpreted strictly as a local, control-oriented approximation rather than a globally exact kinetic law (Tom et al., 2023). While the affine model successfully captures the dominant trend within a stable operating window, it cannot reproduce the point-wise transient evolution of the true surface state (Tom et al., 2023). Consequently, any deviation between the fitted linear baseline and the simulated coverage trajectory represents highly structured model mismatch – driven by non-linear saturation, steric limitations, inhibitor blocking, defect nucleation, and cascading etch-back coupling – rather than unmodeled measurement noise (Yun et al., 2022c,b). This structural residual becomes particularly problematic when manufacturing disturbances shift the reactor away from its baseline kinetic regime, introducing massive gain errors that invalidate static parameters and directly motivating the real-time dynamic update of both the process slope α and model bias β developed in this framework (Yun et al., 2022b).

This issue is particularly important for ASALD/ALE processes because the underlying kinetics are coverage-dependent, self-limiting, and strongly influenced by inhibitor blocking, defect nucleation, steric constraints, and integrated etch-back pathways. These effects introduce highly nonlinear process behavior and can generate substantial structural residuals, such that direct application of a linear model may lead to poor predictive and control performance (Yun et al., 2022c).

(i) **Non-linear model fitting.** First, the baseline coverage data is fitted directly to the time-domain response of a first-order system subjected to a step input. Assuming a normalized step size magnitude of $M = 1$, the non-linear equation is expressed as:

$$y(t) = K\left(1 - e^{-\frac{t-t_0}{\tau}}\right) \quad (7)$$

where $y(t)$ represents the calculated surface coverage at a given reaction time t, K is the steady-state process gain, t_0 represents the reaction time required to reach full coverage under normal, undisturbed circumstances, and τ is the time constant of the respective reaction phase. This non-linear fit allows us to accurately determine the process gain (K) intrinsic to the reaction dynamics before any control actions are applied. The non-linear regression model results for Step A, B and D are presented in Fig. 6.

(ii) **Linear model fitting.** While the non-linear model accurately describes the physical trajectory of the surface coverage, R2R controllers, particularly EWMA algorithms, rely on linear input–output relationships to efficiently compute batch-to-batch updates (Yun et al., 2022c). To accommodate this, the first-order response equation is algebraically rearranged and transformed using a natural logarithm:

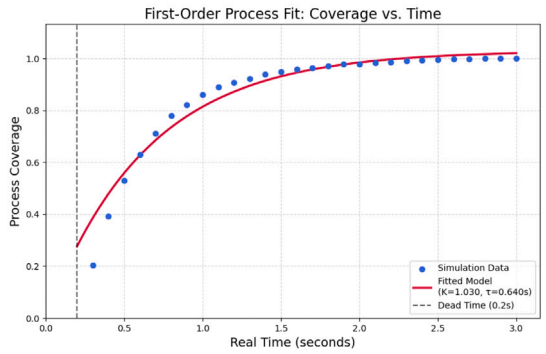
$$\ln\left(1 - \frac{y(t)}{K}\right) = -\frac{1}{\tau}(t - t_0) + \text{Intercept} \quad (8)$$

By plotting $\ln(1 - y/K)$ against the reaction time t, the exponential curve is linearized into a state equation:

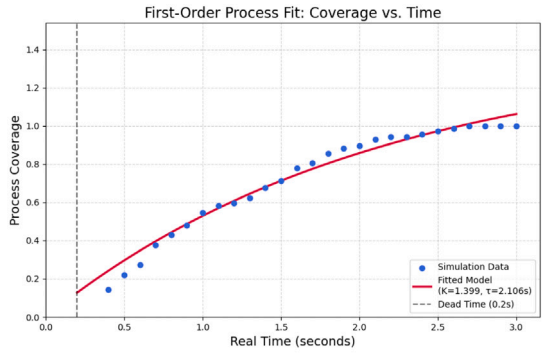
$$\phi = \alpha(t - t_0) + \beta \quad (9)$$

where $\phi = \ln(1 - y/K)$ is the transformed output variable, $\alpha = -1/\tau$ is the linearized process slope defining the sensitivity of coverage to changes in reaction time, and β is the intercept acting as the default model bias ($\beta_{default} = 0$ in the initial undisturbed state).

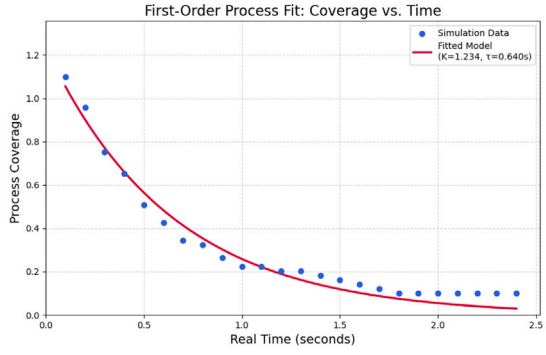
By utilizing these two fitting methods in tandem, the baseline characterization framework successfully maps the highly complex, multiscale kinetics of the ASALD process into a streamlined set of linear parameters (K, τ, and t_0). As demonstrated in Sections 2.2 and 2.3, these initial values serve as the foundational inputs required by the



(a)

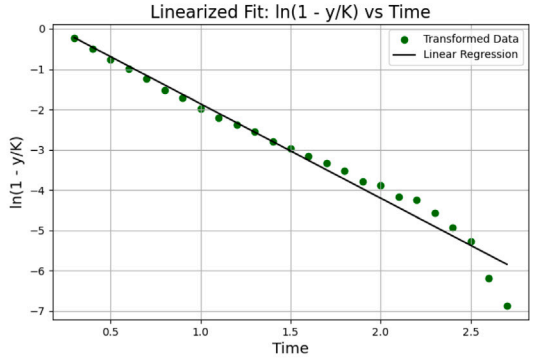


(b)

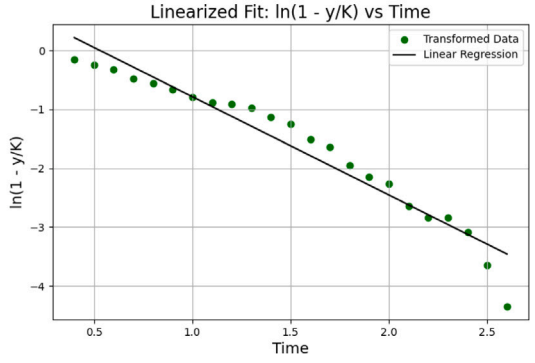


(c)

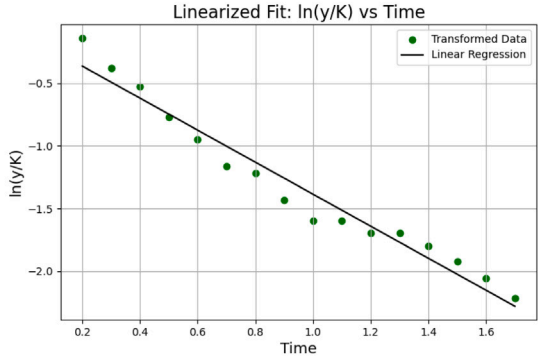
Fig. 6. The relationship between coverage and batch reaction time derived from a standard first order non-linear fitting. (a) Step A. (b) Step B. (c) Step D.



(a)



(b)



(c)

Fig. 7. The relationship between coverage and batch reaction time derived from a standard first order linear fitting. (a) Step A. (b) Step B. (c) Step D.

automated supervisory script to execute precise batch time updates when the system is subjected to manufacturing disturbances (Yun et al., 2022c).

While this affine mapping effectively characterizes the dominant process trends, the pointwise fitting errors observed in Figs. 6 and 7 highlight the structured model mismatch inherent to simplifying non-linear, sterically constrained surface kinetics into a first-order surrogate (Yun et al., 2022c). Physically, the application of a first-order model is justified because the underlying surface reaction rates are vastly faster than the precursor adsorption rates, leaving the overall cycle transiently adsorption-dominated (Yun et al., 2022c). If an alternative regime were present where the subsequent surface reactions were significantly slower, the process would become reaction-limited, strictly restricting the applicability of a first-order approximation.

To ensure numerical stability during the offline linear system identification, a targeted data-preprocessing step is implemented prior to regression: all redundant, flat-tail boundary points ($y = 1.0$ for Steps A and B; $y = 0.0$ for Step D) are removed from the baseline datasets, retaining exclusively the single closest boundary point to eliminate asymptotic trailing biases from skewing the parameter estimation. For the adsorption phases (Steps A and B), ensuring that the unconstrained steady-state process gain resulting from the non-linear regression strictly satisfies $K > 1.0$ provides a vital cushion; because K is strictly greater than unity, the fractional term $y(t)/K$ never reaches exactly 1.0 even at complete surface saturation ($y = 1.0$), thereby preventing the transformation term $\ln(1 - y/K)$ from encountering a numerical singularity ($\ln(0) \to -\infty$) and safely guiding the controller via reaction time extensions back to baseline saturation (Yun et al., 2022c). Conversely, for the etching phase (Step D), driving defect

coverage toward an absolute target of 0% presents a similar risk of singular behavior as the surface approaches complete clearance. As detailed in the reverse control formulation below, the supervisory script actively circumvents this boundary failure mode by embedding a small positive regularizing constant, ϵ, into the measurement and target mappings, bounding the transformation and guaranteeing numerical stability across all operational domains.

2.5. Half-cycle boundary constraints and spatial data extraction

At the end of each simulated batch run, the input to the controller is constructed directly from the transient nongrowth-area (NGA) output of the multiscale ANSYS Fluent framework. In the present workflow, the NGA solution is sampled every 0.1 s, one NGA output file is extracted for each run, the instantaneous NGA surface coverage $\theta_{NGA}(t)$ is evaluated at every stored timestep, and the resulting time-versus-coverage history is assembled separately for each process step and each batch (ANSYS, Inc., 2017; Yun et al., 2023). This procedure is consistent with ANSYS Fluent's report-definition/report-file functionality, which writes transient time-step histories of selected quantities and can include user-defined quantities and flow-time information, and it is also consistent with recent ASALD multiscale CFD studies in which CFD and kMC are integrated to resolve spatiotemporal reactor behavior together with evolving surface coverage (ANSYS, Inc., 2017; Yun et al., 2023).

The script calculates the coverage for a specific step by dividing the total number of sites successfully occupied by the desired species by the maximum theoretical number of sites that can be occupied:

$$\text{Coverage (\%)} = \left(\frac{N_{\text{occupied}}}{N_{\text{max\_capacity}}} \right) \times 100\% \tag{10}$$

where N_{occupied} is determined by algorithmically scanning the extracted grid configuration and counting the sites containing the target molecule for that specific half-cycle (e.g., the inhibitor in Step A or the precursor in Step B) (Yun et al., 2022a). Because both N_{occupied} and $N_{\text{max\_capacity}}$ are derived strictly from physical grid sites, they are always integer values. Consequently, the resulting coverage is not a continuous mathematical function, but rather a discrete sequence of values dictated by the physical addition or removal of individual molecules on the lattice. This physically extracted discrete value serves directly as the numerical input metric (y_{current}) utilized by the R2R parameter updating laws previously detailed in Section 2.3.

It is important to note that in the context of Area-Selective Atomic Layer Deposition (ASALD), this coverage calculation is not applied universally across the entire wafer. The surface is functionally divided into a Growth Area (GA) and a Non-Growth Area (NGA) (Ou et al., 2026). Through simulated observation, the desired surface reactions proceed significantly faster on the GA compared to the NGA. As demonstrated in prior work (Ou et al., 2026), the GA inherently reaches its target coverage ahead of the NGA, the kinetics on the NGA acts as the rate-limiting bottleneck for the overall process. Consequently, this work exclusively isolates and extracts the discrete coverage data from the NGA to serve as the input for the EWMA process control model and subsequent batch time updates. The coverage of the GA is deliberately excluded from the time-updating calculations and is instead monitored purely as a baseline reference to ensure that overall process efficiency does not drop below acceptable operational thresholds.

Furthermore, the denominator, $N_{\text{max\_capacity}}$, is not simply the total absolute number of points on the simulated spatial grid. During physical ASALD, bulky precursor molecules and their associated ligands take up significant spatial volume (Ou et al., 2026). When one of these large molecules adsorbs, it physically blocks adjacent active surface sites, preventing secondary molecules from reacting in those immediate neighboring spaces due to spatial instability and steric hindrance (Yun et al., 2022c). By importing a pre-calculated, sterically-constrained maximum capacity parameter from the previous work (Ou et al., 2026), the script ensures that a calculated discrete coverage of 100% on the NGA accurately reflects a fully saturated, self-limited surface. This provides the EWMA algorithm with physically accurate feedback to tune the updating batch reaction time for the subsequent run.

2.6. Verification of excluded control steps (Step C and Step E)

In the development of the R2R control framework, Step C (ozone oxidation) and Step E (fluorination) were explicitly excluded from the dynamic time-updating loop. To verify the validity of this exclusion, the inherent kinetic robustness of these specific surface reactions was evaluated under severe disturbance conditions.

While the primary focus of this study investigates a 20% kinetic disturbance in the rate-limiting steps, a more extreme 40% reduction in the reaction rate (operating at 60% of the baseline kinetics) was simulated for Steps C and E to test their operational limits. The multiscale simulation revealed that even under this severe 40% degradation, Step C requires only 0.8 s to reach full surface conversion, and Step E achieves completion in just 0.3 s.

The established baseline recipe for the ASALD-ALE cycle already allocates a fixed reaction time of 1.0 s for both of these steps. From the observation of Fig. 8, this standard 1.0-second duration provides a sufficient temporal buffer to fully absorb a 40% kinetic drop without compromising surface saturation, there is no risk of incomplete reactions occurring under the target 20% disturbance investigated in this study. Consequently, applying active R2R time-updates to Steps C and E is mathematically and practically unnecessary. Maintaining these reaction times at their fixed 1.0-second baseline ensures reliable surface conversion while minimizing the computational complexity of the control algorithm, allowing the framework to remain strictly focused on the critical rate-limiting stages (Steps A, B, and D).

3. Results and analysis

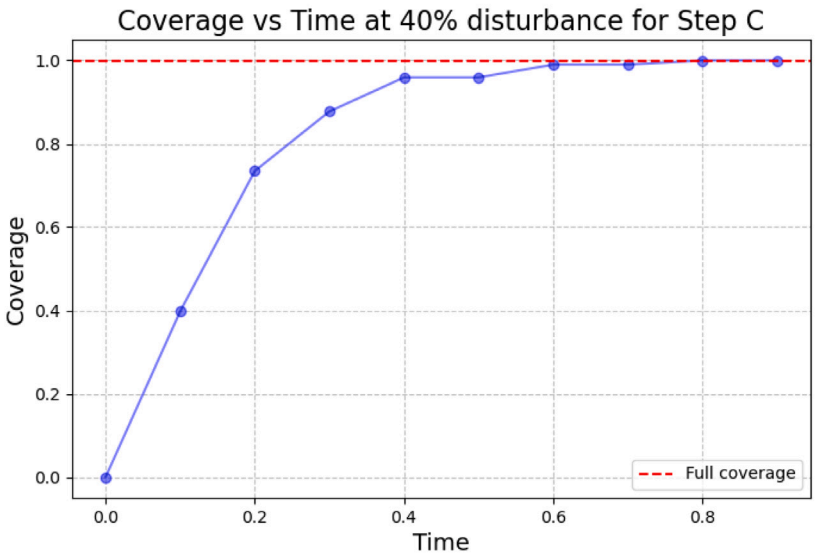
This section presents the performance evaluation and optimization of the proposed EWMA control framework applied to the multiscale ASALD system. The results demonstrate the capability of the dynamic updating scheme to mitigate significant kinetic disturbances and restore target surface coverages across interconnected process steps.

The analysis begins with a detailed investigation of the parameter updating methodologies described before in Section 2.3. Here, the evolution of the process slope (α) and model bias (β) is analyzed under a simulated 20% reduction in the adsorption rate. A critical comparison is made between utilizing recursive linear regression and a secondary EWMA update loop for the dynamic adjustment of α. This evaluation highlights how the sequential updating protocol prevents the double-estimation of process errors and ensures the controller accurately mirrors the slower kinetic trajectory of the disturbed state. The results clarify the trade-offs between responsiveness and stability for each updating model, establishing the most robust configuration for the subsequent integrated runs.

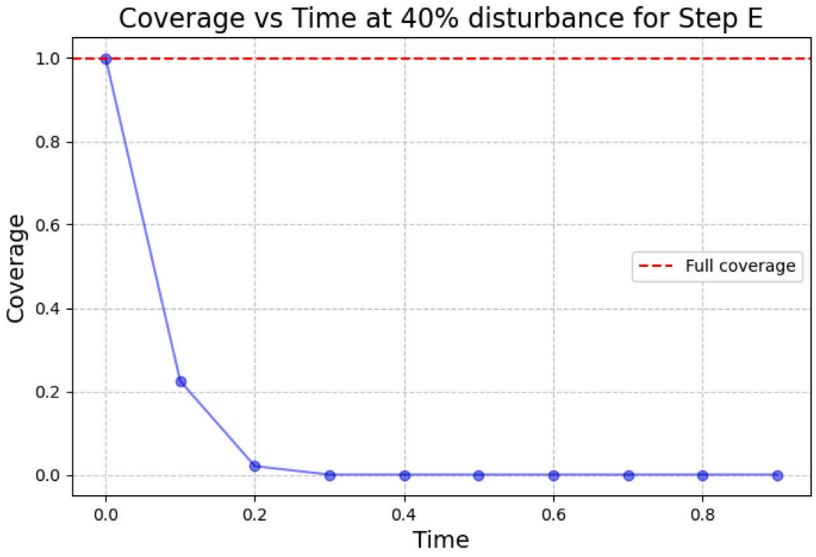
The final part of the results examines the behavior of the complete connected control architecture described before in Section 2.2, which evaluates the sequential execution of Steps A, B, and D. These results demonstrate the effectiveness of the scaling factors and feed-forward mechanisms in facilitating rapid recovery between interdependent cycles. By applying the most suitable parameters identified in the sensitivity study, the multiscale simulation reveals the number of batches required for the Non-Growth Area (NGA) to return to its targets, which are 100% for adsorption and 0% for etching—while maintaining high selectivity. This comprehensive analysis validates the efficiency of the integrated framework in managing multiscale manufacturing cycles subject to realistic operational variability.

3.1. Influences of control parameters on performance

(a) **Update of α.** In standard Run-to-Run (R2R) control applications for semiconductor manufacturing, the process slope (α) is frequently treated as a static parameter derived from initial baseline experiments. However, the validity of a constant α assumes that the local sensitivity of the process output to recipe changes remains uniform across all



(a)



(b)

Fig. 8. Step C and Step E coverage under 40% disturbance. (a) Step C. (b) Step E.

operating states (Wang et al., 2025). In this work, the introduction of a 20% kinetic disturbance significantly alters the adsorption rate, shifting the system into a slower kinetic regime that deviates from the original "fast-state" baseline. As the EWMA controller iteratively lengthens the batch reaction time to restore target coverage, the original α becomes increasingly unsuitable for predicting the system's behavior in this new state. Relying on an outdated slope can lead to persistent estimation errors and slowed convergence.

To maintain an accurate process model throughout the time-update sequence, we implement a dynamic update for α, the detailed procedure for which is explained in Section 2.3. This adaptation ensures that the controller's gain accurately reflects the current physical reality of the reactor rather than its undisturbed history. We evaluate the effectiveness of this adaptation by testing two distinct methodologies: a pure EWMA update utilizing weight factors of $\lambda = 0.3$ and $\lambda = 0.7$, and a linear regression hybrid EWMA approach using the same weight factors. In the pure EWMA method, α is updated using a law identical in structure to the bias update. For the hybrid method, a standard linear regression model is first used to calculate observed values of α and β from recent batch data, which are then refined via the EWMA update law to determine the parameters for the subsequent run. While the pure EWMA updates allow for adjustable sensitivity to recent batch data, the hybrid model provides an alternative that leverages the stability of regression-based observations to track the evolving process slope.

(i) EWMA Control Methods.

The evaluation of the pure EWMA α update for Step A and Step B reveals that the convergence speed is directly governed by the magnitude of the weight factor (λ_α). For clarity, λ_α denotes the λ parameter in Eq. (4) for α update, λ_β denotes the λ parameter in Eq. (2) for β update. Notably, the starting point for the reaction time update immediately following Batch 0 is identical across all tested λ_α values and is only distinguished by different steps. This occurs because the initial slope used to calculate the first time update is a default value derived from the original, undisturbed baseline data. Consequently, regardless of the

chosen λ_α, the controller implements the same initial recipe correction for the first batch under disturbance.

The divergence in performance becomes evident in the subsequent batches, where the controller must re-characterize the process slope to match the 20% slower kinetics. From Fig. 9, the results turn out as expected in a standard EWMA framework, a larger λ_α (e.g., $\lambda_\alpha = 0.7$) allows the script to place more weight on the error observed in the current batch. This higher sensitivity enables the controller to more aggressively adapt the slope from its "fast" baseline to the "slow" disturbed state. As a result, the case with a larger λ_α reaches the target coverage in significantly fewer batches than the lower λ_α cases. This confirms that for the pure EWMA method, the total batch count is directly dependent on the update weight, with higher λ_α values providing the necessary responsiveness to compensate for significant kinetic shifts.

While a larger λ_α provides a more rapid initial approach toward saturation, the recovery trajectory is ultimately modulated by the physical constraints of the multiscale simulation. Specifically, the simulation operates with a minimum time resolution of 0.1 seconds (Ou et al., 2026). Even as the controller accurately identifies the required time adjustments, any calculated increment smaller than this 0.1-second threshold results in no actual change to the implemented reaction time for that batch.

This leads to a pronounced "staircase effect" in the coverage results, where the Non-Growth Area (NGA) coverage remains stagnant for several batches until the cumulative EWMA updates surpass the 0.1-second increment. Although a higher λ_α reduces the total number of batches required to reach full saturation, the discrete nature of the simulation clock means that the long-term approach to the target is characterized by these sudden jumps rather than a continuous curve. Therefore, while the weighting factor determines the overall slope of the recovery, the resolution of the digital twin imposes a fundamental limit on the smoothness of the convergence, regardless of the control model's mathematical sensitivity.

(ii) Linear Regression with EWMA.

In contrast to the pure EWMA approach, the linear regression hybrid EWMA method produces significantly more pronounced time updates. While the pure EWMA model's sensitivity is largely confined to the initial batches, the regression-based model maintains a significant influence over the process slope (α) throughout the entire reaction cycle. As illustrated in Fig. 10, the magnitude of the time updates scales directly with the weight factor, λ_α, indicating that the hybrid model is highly responsive to the evolving kinetics of the disturbed ASALD process.

However, the increased aggressiveness of the hybrid method introduces significant challenges regarding process stability and overshooting. As the controller seeks to bridge the gap caused by the 20% kinetic disturbance, high weight factors can lead to excessive time increments that surpass the physical requirements for surface saturation. This phenomenon is particularly evident in the results for Step B (precursor adsorption). When utilizing a weight factor of $\lambda_\alpha = 0.7$, the controller induces a substantial overshoot of 0.7 s. In the context of self-limiting atomic layer processes, such a large overshoot is considered unacceptable as it risks unnecessary precursor waste and potential secondary nucleation on the Non-Growth Area (NGA).

In comparison, the other tested cases that include the $\lambda_\alpha = 0.7$ setting for Step A demonstrated either zero overshoot or a marginal, manageable overshoot of approximately 0.1 s. This disparity suggests that Step B, characterized by more complex dual-form morphology BDEAS molecule adsorption reactions (Ou et al., 2026) and slower inherent kinetics as demonstrated in longer t_{0B} than t_{0A}, requires a more damped control response than Step A. Consequently, to balance the need for rapid convergence with the requirement for process stability, we have identified the optimal weight factors for the integrated system. In the subsequent analysis of the sequentially connected process Section 3.2, we utilize a weight factor of $\lambda_\alpha = 0.7$ for Step A to maximize its initial characterization speed, and a more conservative $\lambda_\alpha = 0.5$ for Step B to ensure stable recovery without excessive overshooting.

(b) **Update of β.**

While the dynamic adaptation of the slope (α) allows the controller to re-calculate the process gain, the bias (β) update remains the primary mechanism for correcting batch-to-batch offsets caused by disturbances. Here, we want to examine the stability and convergence of the EWMA-based β update when coupled with different α adaptation strategies. The core objective is to identify a tuning configuration that minimizes the transient period after a 20% kinetic disturbance without inducing excessive oscillation in the reaction time.

Following the sequential logic detailed in Section 2.3, the model bias is updated only after the slope has been refined to ensure that the error correction is not erroneously attributed twice to both parameters. We investigate the sensitivity of this interaction by testing the EWMA β-update under two distinct weight factors: $\lambda_\beta = 0.3$ for a more filtered, stable response, and $\lambda_\beta = 0.7$ for aggressive error correction. Throughout these tests, the weight factor for the EWMA-based α update is held constant at $\lambda_\alpha = 0.7$ to provide a responsive baseline for the process gain. Furthermore, we compare these EWMA-driven results against the parameter-free recursive linear regression model for α. This comparison allows us to quantify how the choice of λ_β influences the recovery trajectory of the Non-Growth Area (NGA) coverage when the underlying process model is undergoing continuously successive linearization.

The results presented in Figs. 9, 10 and 11 highlight a complex interplay between the model bias weight factor (λ_β) and the initial behavior of the controller. Observations indicate that λ_β exerts a dual influence on the recovery trajectory: it dictates the magnitude of the starting point for the run-time update immediately following Batch 0, and it simultaneously governs the rate at which the model adapts in subsequent cycles. Specifically, a smaller λ_β leads to a significantly higher starting point for the initial reaction-time update but results in a slower subsequent update rate as the controller places less weight on current error observations.

The impact of this trade-off on the total number of batches required to reach target coverage is not uniform across the ASALD sequence, but rather depends on the specific kinetic state of the step after the first correction. For Step A, employing a smaller λ_β was found to decrease the total batch count. This is attributed to the fact that Step A reaches a high surface coverage of approximately 0.98 immediately following the Batch 0 update. Because the system is so close to saturation, the "higher starting point" provided by the smaller λ_β is sufficient to bridge the remaining gap almost immediately, rendering the slower subsequent update rate irrelevant.

Conversely, for Step B, a smaller λ_β actually leads to a slight increase in the total number of batches required. In this step, the coverage, after the initial batch 0 update, is lower at approximately 0.89. Because Step B is further from the saturation target, the controller requires more significant and frequent updates to the recipe. In this regime, the disadvantage of a slower update rate outweighs the benefit of a higher initial starting point. This indicates that the effect of the λ_β weight factor becomes increasingly important for the "finishing stretch" of the recovery in Step B compared to Step A.

Furthermore, the 0.1-second resolution limit of the simulation interacts with these update rates, where a smaller λ_β may cause the controller to remain stagnant (i.e., do not update its existing value) for longer periods before triggering a detectable time increment. Consequently, the selection of λ_β must be carefully balanced based on the step-specific kinetics: for steps that reach high initial coverage, a smaller λ_β facilitates rapid finalization, while for more resistant steps like Step B, a higher weight factor is necessary to maintain an aggressive and responsive update trajectory.

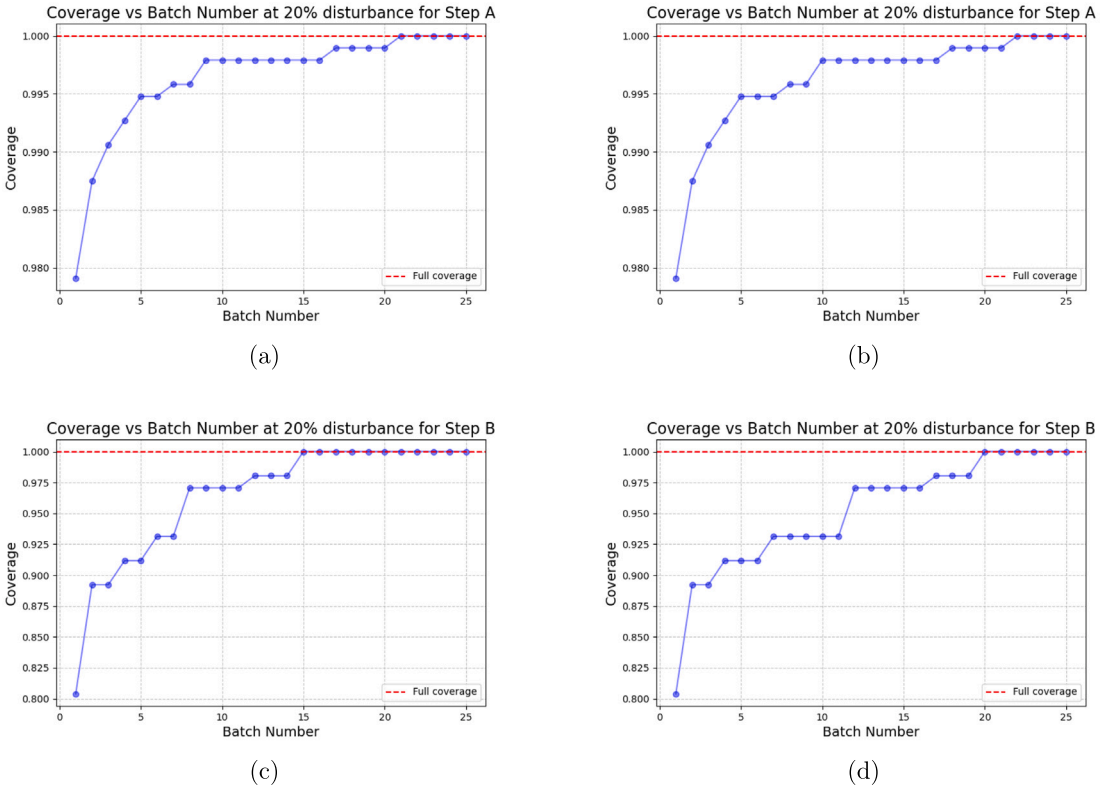


Fig. 9. The performance difference between different EWMA parameters for step A and step B when λ_α update = 0.7 or 0.3 and λ_β update = 0.7. (a) λ_α = 0.7 for Step A. (b) λ_α = 0.3 for Step A. (c) λ_α = 0.7 for Step B. (d) λ_α = 0.3 for Step B.

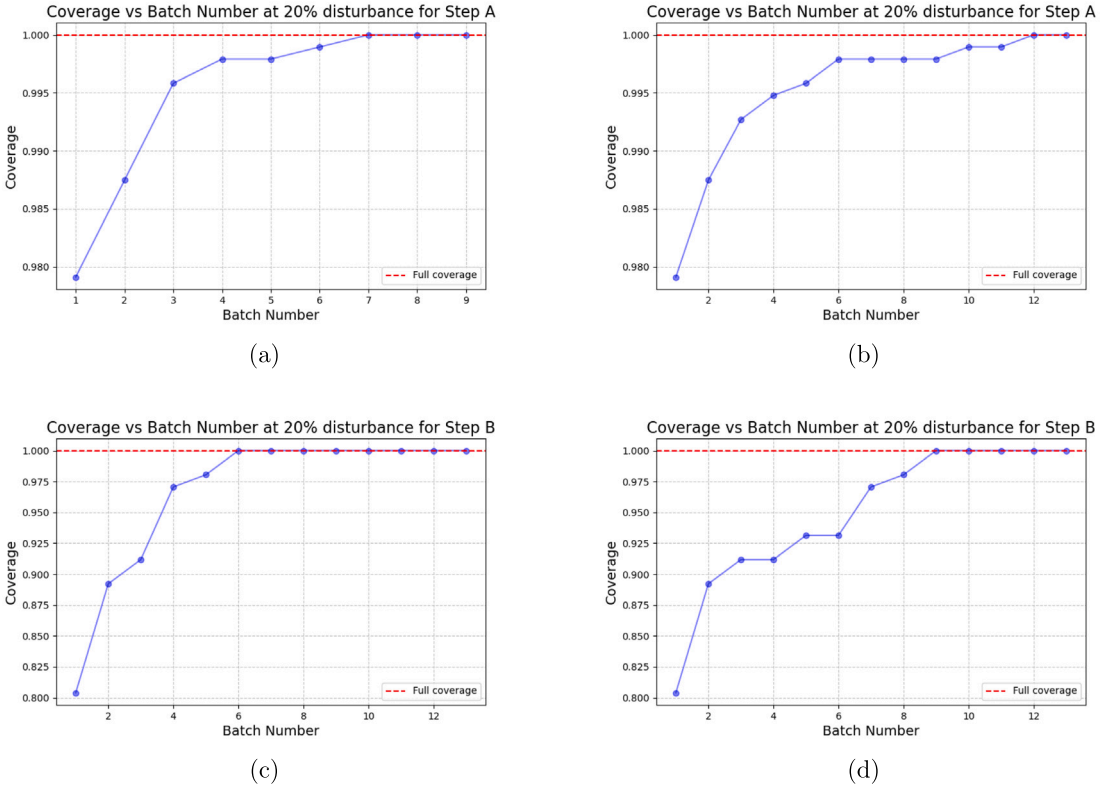


Fig. 10. The performance difference using linear regression between different EWMA parameters for step A and step B when λ_α update = 0.7 or 0.3 and λ_β update = 0.7. (a) λ_α = 0.7 for Step A. (b) λ_α = 0.3 for Step A. (c) λ_α = 0.7 for Step B. (d) λ_α = 0.3 for Step B.

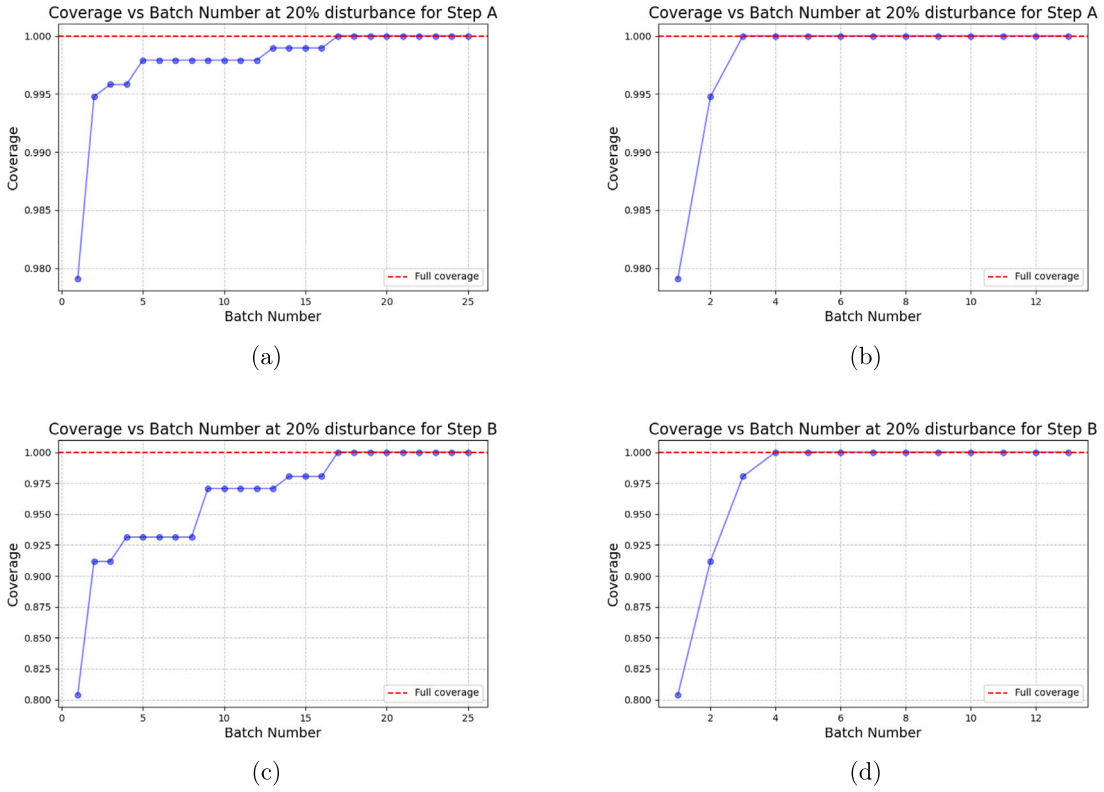


Fig. 11. Performance under λ_β update = 0.3 while using different alpha update model for step A and step B. (a) Step A with EWMA. (b) Step A with linear regression. (c) Step B with EWMA. (d) Step B with linear regression.

3.2. Run-to-run control over multiple ASALD-ALE cycles

The final analysis investigates the system-level performance of the integrated ASALD-ALE control architecture when all process steps are executed in a sequentially connected manner. This section demonstrates how the feed-forward communication protocol between Steps A, B, and D significantly optimizes the overall efficiency of the multi-step cycle under the applied 20% kinetic disturbance (Yun et al., 2022c).

As Step A serves as the foundational stage of the sequence, it does not benefit from prior historical batch data, and its recovery timeline reflects the baseline convergence of an isolated controller. However, the results demonstrate that Step A provides a critical characterization of the reactor's disturbed state, which is leveraged to accelerate the remainder of the cycle. Even though the number of batches required for Step A to reach saturation is not reduced, the convergence of subsequent steps is significantly improved through the parameter handoff mechanism.

By passing the updated reaction time, the refined model bias (β), and the adaptive process slope (α) from Step A to Step B, the controller initializes the precursor adsorption stage with an already-corrected process model. This feed-forward inheritance allows Step B to bypass the initial transient period and achieve target coverage in a substantially reduced number of batch runs. Furthermore, the analysis evaluates the application of these corrections to Step D through time-scaling factors, ensuring that the intermediate etching phase remains tightly controlled. This integrated approach ensures that high selectivity is restored rapidly, minimizing defect accumulation on the Non-Growth Area (NGA) across the entire multiscale simulation.

For the subsequent etching phase (Step D), the control logic is specifically adapted to maintain stability. Unlike the adsorption steps, the dynamic α update is not applied to Step D because its inherent reaction kinetics do not undergo the same magnitude of shift as observed in Steps A and B. Furthermore, Step D requires a more conservative update trajectory to avoid overshooting the 0% coverage target, which would lead to over-etching and loss of the growth area. To achieve this, the inherited correction from Step B is scaled using a factor of 0.44 as calculated in Eq. (3), ensuring the time update is proportional to the faster kinetics of the etching step.

Most importantly, this connected architecture maintains tight process stability across all phases. The reaction time overshoot for both Step A and Step B is restricted to a marginal level of 0.1 s, a deviation that is entirely acceptable within the natural tolerances of self-limiting adsorption chemistries (Ou et al., 2026). Furthermore, for the etching phase (Step D), the calculated time overshoot is less than 0.05 s. Because this value falls well below the 0.1-second physical time-resolution limit of the multiscale process, it effectively guarantees that no unmodeled over-etching occurs on the growth area.

The efficacy of this "relay" logic is illustrated in Fig. 12, which presents the normalized reaction time (t/t_0) against the cumulative batch number across the entire sequence. For Steps A and B, t/t_0 exhibits a steady increase as the controller compensates for the slower adsorption rates. Upon transitioning to Step D, there is a characteristic rapid drop in the t/t_0 value; this reflects the calculated scaling factor, where the required time extension for Step D is approximately 0.44 times the lengthening required for Step B. Ultimately, the integrated framework proves highly efficient, requiring only 10 cumulative batches for all rate-limiting steps (A, B, and D) to reach their respective target coverages and restore process selectivity.

4. Conclusion

This work presented a control methodology for overcoming the inherent variability in next-generation semiconductor patterning by implementing an adaptive Run-to-Run (R2R) control framework on an integrated ASALD-ALE cyclic operation presented in (Ou et al., 2026). By subjecting a multiscale digital twin to a significant 20% reduction

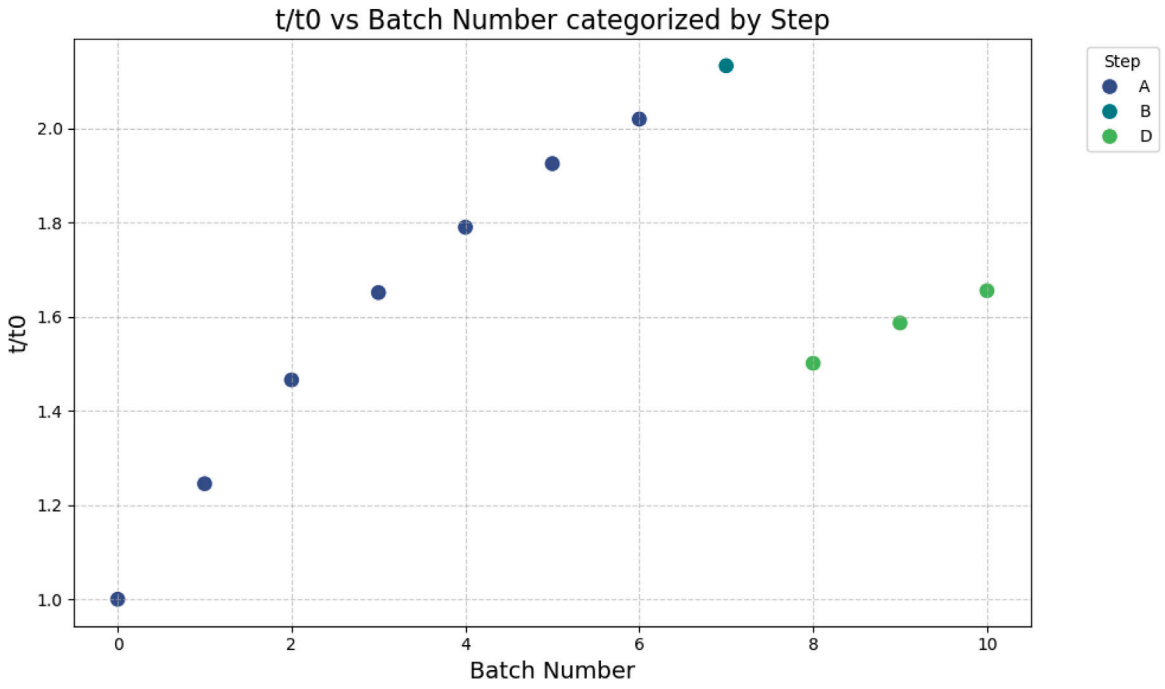


Fig. 12. The batch number required for each step to reach full coverage for step A, step B and step D.

in adsorption kinetics, the study demonstrates the limitations of static control models and highlights the necessity of dynamic parameter estimation to maintain fabrication fidelity.

Central to this control framework is the introduction of a dual-parameter adaptation scheme that refines both the process slope (α) and model bias (β) in real-time. The deployment of a linear regression hybrid EWMA model proved particularly effective, as it successfully navigated the 0.1-second temporal resolution constraints of the simulator to provide smooth, trend-based corrections where pure EWMA models encountered "staircase" stagnation. Sensitivity analyses identified optimal weighting factors; specifically, $\lambda_\alpha = 0.7$ for Step A and 0.5 for Step B, paired with a damping factor of $\lambda_\beta = 0.7$, which balance aggressive recovery with the stability required to prevent unacceptable overshooting during precursor adsorption.

The efficiency of the proposed architecture is further amplified by a sequential "relay" protocol. By enabling subsequent stages to inherit the corrected kinetic state and reaction times from the initial inhibitor phase, the controller bypasses redundant recalibration periods that typically hinder isolated R2R loops. The implementation of a 0.44 scaling factor for the etching phase (Step D) ensures that corrections learned during adsorption are accurately translated across different chemical regimes. Collectively, these innovations reduced the system recovery timeline to just 10 cumulative batches, rapidly restoring the target adsorption saturation and eliminating unwanted residual coverage in the etching step.

From the controller-design perspective, the present framework is based on a local linearized representation derived from an assumed first-order process response. This formulation is appropriate for efficient EWMA-based recipe updating within the operating window considered here, but it may introduce bias when the true coverage dynamics deviate substantially from the first-order approximation or when the process is exposed to disturbance profiles beyond the structured 20% and 40% kinetic reductions examined in this study. Future work can therefore extend the controller by generating simulation data across a wider range of disturbance magnitudes, disturbance types, and operating conditions, and using these data to train machine learning-based surrogate or gain-scheduling models. Such models could provide condition dependent estimates of the local process response and expand the applicability of the R2R controller beyond a single pre-assumed kinetic regime, while preserving the interpretability and batch-to-batch update structure of the EWMA framework.

Ultimately, this work demonstrates that holistic, interconnected control is vital for managing the complex, cascading dynamics of multiscale fabrication sequences. The adaptive framework serves as a versatile blueprint for real-time process regulation, offering a scalable path toward high-yield, self-aligned manufacturing in commercial environments characterized by equipment drift and operational disturbances.

CRediT authorship contribution statement

Chun-Pei Lin: Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Feiyang Ou:** Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Gerassimos Orkoulas:** Writing – review & editing, Conceptualization. **Panagiotis D. Christofides:** Writing – review & editing, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Financial support from the National Science Foundation and the Department of Energy is gratefully acknowledged. This work used computational and storage services associated with the Hoffman2 Shared Cluster provided by the UCLA Office of Advanced Research Computing's Research Technology Group.

References

ANSYS, Inc., 2017. Using report definitions for solution monitoring in fluent. Official ANSYS Fluent documentation, Release 18.0.

Box, G.E.P., 1957. Evolutionary operation: a method for increasing industrial productivity. J. R. Stat. Soc. Ser. C. Appl. Stat. 6, 81–101.

Chien, C.F., Chen, Y.J., Hsu, C.Y., Wang, H.K., 2014. Overlay error compensation using advanced process control with dynamically adjusted proportional-integral R2R controller. IEEE Trans. Autom. Sci. Eng. 11, 473–484.
Del Castillo, E., 2001. Some properties of EWMA feedback quality adjustment schemes for drifting disturbances. J. Qual. Technol. 33, 153–166.
Del Castillo, E., Hurwitz, A.M., 1997. Run-to-run process control: Literature review and extensions. J. Qual. Technol. 29, 184–196.
Del Castillo, E., Rajagopal, R., 2002. A multivariate double EWMA process adjustment scheme for drifting processes. IIE Trans. 34, 1055–1068.
Ingolfsson, A., Sachs, E., 1993. Stability and sensitivity of an EWMA controller. J. Qual. Technol. 25, 271–287.
Karasulu, B., Roozeboom, F., Mameli, A., 2023. High-throughput area-selective spatial atomic layer deposition of SiO_2 with interleaved small molecule inhibitors and integrated back-etch correction for low defectivity. Adv. Mater. 35, 2301204.
Liu, K., Chen, Y., Zhang, T., Tian, S., Zhang, X., 2018. A survey of run-to-run control for batch processes. ISA Trans. 83, 107–125.
Mackus, A.J.M., Merkx, M.J.M., Kessels, W.M.M., 2019. From the bottom-up: Toward area-selective atomic layer deposition with high selectivity. Chem. Mater. 31, 2–12.
Mameli, A., Merkx, M.J.M., Karasulu, B., Roozeboom, F., Kessels, W.E.M.M., Mackus, A.J.M., 2017. Area-selective atomic layer deposition of SiO_2 using acetylacetone as a chemoselective inhibitor in an ABC-type cycle. ACS Nano 11, 9303–9311.
Mameli, A., Teplyakov, A.V., 2023. Selection criteria for small-molecule inhibitors in area-selective atomic layer deposition: fundamental surface chemistry considerations. Acc. Chem. Res. 56, 2084–2095.
Merkx, M.J.M., Sandoval, T.E., Hausmann, D.M., Kessels, W.M.M., Mackus, A.J.M., 2020. Mechanism of precursor blocking by acetylacetone inhibitor molecules during area-selective atomic layer deposition of SiO_2. Chem. Mater. 32, 3335–3345.
Moyne, J., 2021. Run-to-run control in semiconductor manufacturing. In: Encyclopedia of Systems and Control. Springer, pp. 1997–2003.
Mukesh, S., Zhang, J., 2022. A review of the gate-all-around nanosheet FET process opportunities. Electronics 11, 3589.
Ou, F., Alghamdi, A., Lin, C.P., Orkoulas, G., Christofides, P.D., 2026. Multiscale modeling of area-selective atomic layer deposition of SiO_2 on Al_2O_3 and SiO_2 surfaces with intermediate etching steps. Chem. Eng. Sci. 325, 123422.
Sachs, E., Hu, A., Ingolfsson, A., 2002. Run by run process control: Combining SPC and feedback control. IEEE Trans. Semicond. Manuf. 8, 26–43.
Song, S.K., Saare, H., Parsons, G.N., 2019. Integrated isothermal atomic layer deposition/atomic layer etching supercycles for area-selective deposition of TiO_2. Chem. Mater. 31, 4793–4804.
Tom, M., Wang, H., Ou, F., Orkoulas, G., Christofides, P.D., 2023. Machine learning modeling and run-to-run control of an area-selective atomic layer deposition spatial reactor. Coatings 14, 38.
Wang, H., Ou, F., Suherman, J., Orkoulas, G., Christofides, P.D., 2025. Integration of on-line machine learning-based endpoint control and run-to-run control for an atomic layer etching process. Digit. Chem. Eng. 14, 100206.
Yun, S., Tom, M., Luo, J., Orkoulas, G., Christofides, P.D., 2022a. Microscopic and data-driven modeling and operation of thermal atomic layer etching of aluminum oxide thin films. Chem. Eng. Res. Des. 177, 96–107.
Yun, S., Tom, M., Orkoulas, G., Christofides, P.D., 2022b. Multiscale computational fluid dynamics modeling of spatial thermal atomic layer etching. Comput. Chem. Eng. 163, 107861.
Yun, S., Tom, M., Ou, F., Orkoulas, G., Christofides, P.D., 2022c. Multivariable run-to-run control of thermal atomic layer etching of aluminum oxide thin films. Chem. Eng. Res. Des. 182, 1–12.
Yun, S., Wang, H., Tom, M., Ou, F., Orkoulas, G., Christofides, P.D., 2023. Multiscale CFD modeling of area-selective atomic layer deposition: Application to reactor design and operating condition calculation. Coatings 13, 558.