Available online at www.sciencedirect.com

Chemical Engineering Research and Design

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

Multivariable run-to-run control of thermal atomic layer etching of aluminum oxide thin films

Sungil Yun[a], Matthew Tom[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

Article history:
Received 3 March 2022
Received in revised form 22 March 2022
Accepted 23 March 2022
Available online 29 March 2022

Keywords:
Semiconductor manufacturing
Thermal atomic layer etching
Multiscale modeling
Computational fluid dynamics modeling
Multivariable run-to-run control
Exponentially weighted moving average controller design

ABSTRACT

With the growing scarcity of semiconducting devices stemming from volatile prices, shortened supplies, and increased demand that are attributed to the Covid-19 pandemic, manufacturers are looking for efficient ways to facilitate the production of nanoscale semiconducting devices. Thermal atomic layer etching (ALE) is a promising method that can overcome the obstacles encountered during the production of semiconducting devices via conventional approaches by delivering precise dosages of reagent to etch monolayers of substrate surface material in a cyclic operation. However, thermal ALE has not been extensively studied and characterized to become fully embraced by the semiconductor manufacturing industry. Recent work by our group has led to the development of a multiscale computational fluid dynamics modeling framework that was used to optimally design a desirable reactor configuration and operating conditions for the thermal ALE process. Despite this progress, additional research is needed to ensure that the film quality is maintained in the presence of operational disturbances. Therefore, the present work is focused on the development of a multivariable run-to-run (R2R) control system to mitigate the impact of critical operational disturbances. It is demonstrated that the developed multivariable R2R control system can efficiently overcome the negative effects of unknown disturbances that may impact film uniformity by regulating input variables within a minimal number of batch runs.

1. Introduction

Recently triggered by the Covid-19 pandemic, there has been a transition from the real world to the digital world or the so-called “metaverse.” As telecommuting has been increasing, the demand for a wider range of electronic devices has surged. This trend requires robust and reliable computing power, which are obtained from high-performance semiconductors. In addition, the autonomous vehicle market is greatly expanding and more vehicles are integrating additional safety and convenience features that would require additional semiconducting materials. The combination of these factors is resulting in a growing demand for these valuable ultra-high-performance semiconductors from different industries during this Fourth Industrial Revolution. However, the semiconductor fabrication speed has not been able to sustain this growing demand and with the uncertainty of the Covid-19 pandemic, shortages are becoming

[*] Corresponding author at: Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, CA 90095-1592, USA.
E-mail address: pdc@seas.ucla.edu (P.D. Christofides).

https://doi.org/10.1016/j.cherd.2022.03.039

more prevalent in the semiconductor market (Voas et al., 2021). Thus, a more effective and efficient method for producing these semiconductors must be established to meet the consumer demand for these materials. Despite the demand for the miniaturization of semiconductors, the continued shrinking of the fin dimensions of the fin field-effect transistors (FinFETs) has been obstructed by an obstacle that is associated with the 5 *nm* node (Razavieh et al., 2019). Gate-all-around (GAA) transistors have become the most promising competitor to replace FinFET technology, which are able to enter the sub-5 *nm* era.

Atomic layer etching (ALE), considered one of the most advanced etching techniques in semiconductor fabrication, is anticipated to overcome this miniaturization issue (Lu et al., 2018). ALE is an etching process that uses sequential precursor pulses in between purge steps to achieve self-limiting behavior by etching monolayers of substrate surface. Due to this behavior, ALE has received great attention as a promising etching process that is able to evoke the sub-5 *nm* era. Many researchers have investigated and demonstrated the technical viability of ALE processes with various materials such as Si (Abdulagatov and George, 2018), SiO_2 (Metzler et al., 2014), MoS_2 (Kim et al., 2017), Si_3N_4 (Li et al., 2016), and Al_2O_3 (Lee et al., 2016). In particular, high-*k* materials, which increase computing speed with lower current leakage (Jurczak et al., 2009), have received tremendous attention. Recently, Lee et al. (2016) has validated the self-limiting behavior of the thermal ALE of aluminum oxide. Thermal ALE processes are particularly noticeable since they can ensure uniformity and conformity of thin films, thus resulting in ultra-smooth thin films. Nevertheless, the thermal ALE of aluminum oxide has not been extensively studied, and thus, the process has not been fully characterized. The thermal ALE of Al_2O_3 has been investigated microscopically (Yun et al., 2022a) and from a multiscale modeling framework (Yun et al., 2022b) that provides a multiscale perspective on the process, which was validated from the experimental results of Lee et al. (2016).

Despite the advancements made in ALE modeling, an optimal control scheme and guideline for performing process control for the ALE system are lacking. For example, prior research has been conducted in the feedback control of silicon (Crose et al., 2019). However, one of the most notable control methods in the semiconductor industry is run-to-run (R2R) control, which is a form of a batch system where a process recipe is modified between "batches" according to the given input-output relation. Despite the fact that R2R control has been gradually used in semiconductor fabrication facilities, the lack of quantitative and qualitative process information and nonlinear dynamic behaviors are some reasons why semiconductor manufacturers are skeptical about integrating R2R control to ALE processes. In addition, a single R2R algorithm may not be sufficient to cover all possible disturbances such as process drift, shift, and other kinds of variability (Ning et al., 1996). To overcome this issue, a combination of batch process control and feedback control has recently emerged in the semiconductor manufacturing industry (Campbell et al., 2002). The conjunctive strategy of R2R and feedback control has been studied by Yun et al. (2021) and Zhang et al. (2020). As another combined strategy, effective multi-algorithms for R2R control have emerged as robust, stable, and optimal control schemes in the semiconductor manufacturing industry (Moyne et al., 2018). In this research, multiple algorithms for multivariable R2R control are developed to ensure conformity and uniformity of thin films and to adapt to disturbances that impede the standard operating conditions in the context of thermal ALE of aluminum oxide thin films.

2. Multiscale CFD modeling of thermal ALE

Thermal atomic layer etching (ALE) for aluminum oxide (Al_2O_3) thin films is simulated using multiscale computational fluid dynamics (CFD) modeling, which is an integrated form of microscopic and macroscopic modeling. This multiscale model is conducted similarly to prior research by Kwon et al. (2015) and Crose et al. (2017). Based on the kinetic Monte Carlo (kMC) method, microscopic modeling composes of a general simulation of the kinetic components of the reaction under ideal conditions and is simulated through randomness to generate a kinetic model of the surface. For the gas transport domain, the previously developed inclined plate reactor is used and CFD simulations are performed using Ansys Fluent 2021R2, which is a widely used commercial CFD software. For the surface domain, the process data from the CFD model is exported and transferred to the microscopic model to calculate the coverage or etching fraction in the atomistic point of view.

In the previous research (Yun et al., 2022b), the multiscale CFD simulation included the influence of transport phenomena and kinetic effects to various reactor models, which were examined for their effectiveness in delivering a uniform precursor flow pattern to the substrate surface and for the time to achieve complete coverage and etching. It was revealed that the inclined plate reactor delivered the most optimal results when comparing the types of reactors that were studied. Therefore, in this work, the multiscale CFD model of the inclined plate reactor design is adopted to develop and optimize a multivariable R2R control system.

2.1. *Microscopic surface modeling*

The thermal atomic layer etching (ALE) of Al_2O_3 uses two precursor pulses in a sequential manner. Thus, there are two reaction steps in between the purge steps: Step A and Step B. In Step A, the wafer is exposed to hydrogen fluoride (HF), which fluorinates the aluminum oxide on the surface to produce an AlF_3 layer. Next, a purge step sweeps the remaining HF in the reactor to avoid undesired reactions and to guarantee self-limiting behavior. Then, the fluorinated surface (AlF_3) undergoes a ligand-exchange and is converted into a volatile species, dimethylaluminum fluoride [DMAF, $AlF(CH_3)_2$], when trimethylaluminum (TMA), $Al(CH_3)_3$, is supplied during Step B. The volatile layer, DMAF, is desorbed and thus, a monolayer of the substrate surface is removed. Lastly, another purge step removes the remaining TMA and residual products. A cyclical operation consists of the aforementioned steps, which are repeated until the desired film thickness is achieved. The overall reaction is described by

$$Al_2O_3\ (s) + 6HF\ (g) + 4Al(CH_3)_3\ (g) \rightarrow 6AlF(CH_3)_2\ (g) + 3H_2O\ (g)$$

A microscopic surface model for the etching of Al_2O_3 thin films was described in the previous work (Yun et al., 2022a). A brief description of the microscopic model is presented here. θ-Al_2O_3 $(\bar{2}\ 0\ 1)$ was found on Si(1 0 0) through the atomic layer deposition (ALD) process under annealing (Broas et al., 2017). Thus, θ-Al_2O_3 $(\bar{2}\ 0\ 1)$ is used as the preferred lattice structure. A 300×300 lattice is applied to the microscopic model in which etch reactions take place. The surface kinetics in the atomistic level is modeled using the kinetic Monte-Carlo (kMC) method, in particular the variable step size method (VSSM) that is often called the *n*-fold way or BKL, which refers to the algorithm developed by Bortz, Kalos, and Lebowitz (Jansen, 2012). The kMC algorithm was popularized by Gillespie (1976) who integrated the Monte Carlo algorithm into chemical kinetics. The kMC method is widely used to simulate individual reactions on the microscopic scale including the dependence on the lattice structure (Lou and Christofides, 2003). In the kMC algorithm, the total rate constant, k_{total}, is an important parameter that selects a reaction on the reaction site and calculates the time progression, which is computed as follows:

$$k_{\text{total}} = \sum_{i=1}^{N} k_i \tag{1}$$

where k_i is the reaction rate constant of the reaction i, and N is the number of reactions. After all reactions are defined and the total reaction rate constant is calculated, a random number, $\gamma_1 \in (0, 1]$, is chosen to determine the reaction on the site by using the criterion as follows:

$$\sum_{i=1}^{j-1} k_i \le \gamma_1 k_{\text{total}} \le \sum_{i=1}^{j} k_i \tag{2}$$

where j represents the reaction j. Finally, an additional random number, $\gamma_2 \in (0, 1]$, is generated to compute the time interval defined as:

$$\Delta t = \frac{-\ln(\gamma_2)}{k_{\text{total}}} \tag{3}$$

Detailed kinetic mechanisms for the fluorination and ligand-exchange reaction as well as their kinetic parameters, which were calculated using electronic structure optimization methods and Density Functional Theory, can be found in the research of Yun et al. (2022a).

2.2. *Macroscopic modeling*

In the previous work, as shown in Fig. 1a, a 3D computational fluid dynamics (CFD) model for the inclined plate reactor design was developed and evaluated by Yun et al. (2022b). The optimized inclined plate reactor showed the uniform precursor distribution while preserving the fastest cycle time among other geometries. Thus, in this paper, the inclined plate reactor is adopted to formulate the control scheme. The reactor configuration consists of a cylindrical-shaped chamber with a 500 *mm* diameter and 10 *mm* height, a round-shaped inlet of 20 *mm* diameter, and a round-shaped outlet of 40 *mm* diameter. A substrate of 300 *mm* is placed at the center of the reactor. To optimize the precursor distribution, an arch-shaped inclined plate with 2 *mm* thickness and 5° angle from the horizontal is equipped near the inlet. Meshing Mode, a feature of Ansys Fluent 2021R2, was utilized to generate the mesh for the plate reactor. Tetrahedral cells were adopted to reduce simulation time while maintaining the accuracy of the numerical calculations. Mesh quality criteria ranges recommended from ANSYS (2021) including orthogonality, skewness, aspect ratio, and resolution were integrated into the development of the plate mesh, which is illustrated in Fig. 1b. The characteristics of the plate mesh as well as the quality criteria calculated from Ansys Fluent 2021R2 are discussed in greater detail by Yun et al. (2022b).

Ansys Fluent 2021R2 is utilized to conduct numerical computational fluid dynamics (CFD) calculations of the fluid flow and to simulate the surface kinetics of the HF and TMA half-cycles in the ALE process. The pressure-based solver under transient mode in Ansys Fluent is used with a time step of 0.025 s and 200 maximum iterations per time step. The coupled algorithm is used to decrease the computation time. The pressure-based solver in Ansys Fluent solves the mass and momentum conservation equations, which are expressed as follows:

$$\frac{\partial \rho}{\partial t} + \nabla\cdot(\rho\vec{v}) = S_m \tag{4}$$

$$\frac{\partial(\rho\vec{v})}{\partial t} + \nabla\cdot(\rho\vec{v}\vec{v}) = -\nabla p + \nabla\cdot()\bar{\bar{\tau}}) + \rho\vec{g} + \vec{F} \tag{5}$$

where ρ is the density of the mixture, $\vec{v}$ is the velocity of the mixture, S_m is the mass transfer source term, p is the static pressure, $\bar{\bar{\tau}}$ is a symmetric rank two stress tensor, $\rho\vec{g}$ is the gravitational body force, and $\vec{F}$ is the external body force. In

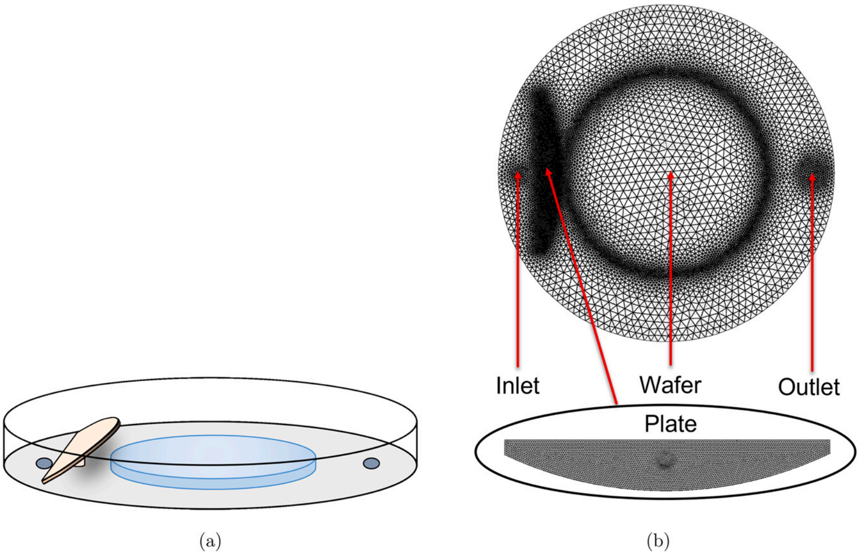


Fig. 1 – A schematic diagram of the inclined plate reactor (a) and the inclined plate reactor mesh generated from Ansys Fluent (b) are illustrated from Yun et al. (2022b).

addition, Ansys Fluent solves the conservation of energy equation which is described by:

$$\frac{\partial}{\partial t}(\rho E) + \nabla(\vec{v}(\rho E + p)) = -\nabla(\Sigma h_j J_j) + S_h \tag{6}$$

where E is the internal energy, h_j is the sensible enthalpy of species j, J_j is the diffusion flux of species j, and S_h is the heat transfer source term. To ensure more accurate results in the pressure-based coupled solver, the second-order upwind scheme is used for spatial discretization, leading to a higher order accuracy through a Taylor series expansion. It is also required to calculate gradients to compute the values of scalars at the cells on the mesh. To calculate the gradients, the Incomplete Lower Upper (ILU) method is employed, which requires more computing power but offers better smoothing properties for the pressure-based coupled solver. Based on these solver settings, the aforementioned equations are solved iteratively at a given time step. Once the solution satisfies the convergence criteria, the solver advances to the next time step.

To generate a more realistic profile, the consumption of precursors and generation of products are included in the CFD model by determining the reaction rate from the modified Arrhenius equation that includes the temperature dependence of the pre-exponential factor, which is defined as follows:

$$k_j = A_j T^{\beta_j} e^{-E_{A,j}/RT} \tag{7}$$

In the above equation, k_j is the reaction rate constant for reaction j, T is the temperature on the surface, β_j is the temperature exponent for reaction j, $E_{A,j}$ is the activation energy for reaction j, and R is the ideal gas constant. Both half-cycle reactions are developed in Ansys Fluent, which are described below:

$$Al_2O_3 \quad (s) + 6HF \quad (g) \rightarrow 2AlF_3 \quad (s) + 3H_2O \quad (g)$$

$$2AlF_3 \quad (s) + 4Al(CH_3)_3 \quad (g) \rightarrow 6AlF(CH_3)_2 \quad (g)$$

The monitoring of the surface temperature in real time is necessary to ensure that the temperature does not decrease, especially for Step A, which requires higher operating pressures for HF at lower temperatures (Yun et al., 2022a). Typically, the surface temperature is maintained to guarantee film quality, and thus, a PI (proportional-integral) controller is assumed to work appropriately in this process. In addition, the cyclical operation is carried out through a user-defined function (UDF) in which operating conditions and boundary conditions are specified. The detailed description for the macroscopic model can be found in the previous work of Yun et al. (2022b).

3. Multivariable R2R control formulation

Various algorithms have been employed in existing run-to-run (R2R) control algorithms for chemical vapor deposition (CVD) and chemical mechanical polishing (CMP) processes including the exponentially weighted moving average (EWMA), predictor-corrector control (PCC), and optimizing adaptive quadratic controller (OAQC) methods. For example, Yun et al. (2021) utilized the EWMA and PCC methods on PEALD of HfO_2 thin films to compare their effectiveness and Crose et al. (2017) developed an R2R controller using the EWMA algorithm for thin film Si-H solar cells. The EWMA controller, which is a linear approximation model-based controller, is widely utilized and integrated into semiconductor fabrication processes due to its versatility to compensate for process shift, drift, and noise among other disturbances (Chien et al., 2014). Run-to-Run controllers are adjusted through multiple batches by using statistical process control (SPC) and by tuning the controllers using engineering process control (EPC) to provide a new recipe (i.e., input) for the next batch run (Fan et al., 2002).

In this work, an EWMA-based multivariable R2R control scheme is developed using two inputs (the process time and the precursor flow rate) and two outputs (the coverage or etching fraction and the precursor partial pressure). In order to tune an EWMA-based R2R controller, the input-output relationship must be established first. A multi-input-multi-output (MIMO) model may be used to formulate the multivariable R2R controller; however, in this work, the process is locally approximated by two single-input-single-output (SISO) linear regression models, with each SISO model being defined by an equation of the following form:

$$y_t = \alpha + \beta u_{t-1} \tag{8}$$

where y_t is the output of the process for batch run t, α is the bias, β is the process gain, and u_{t-1} is the recipe (or input variable) in between batch run $t-1$ and batch run t. The process parameters, α and β, are determined using the standard least squares method where α is the y-intercept and β is the slope of the linear regression model. R2R controllers are modeled under an evolutionary operation mode (Box, 1957) that combines statistical results to improve efficiency and increase productivity. In other words, the control system works to adapt to changes in the process environment and tune the manipulated variables to sustain standard output conditions from the R2R

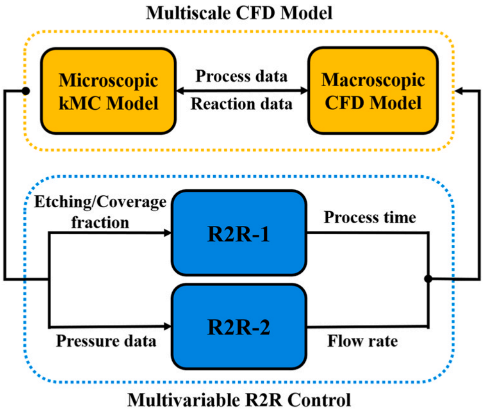


Fig. 2 – Multivariable run-to-run control of the inclined plate reactor.

Table 1 – Standard operating conditions for the multiscale CFD simulation.

Standard Condition	Step A	Step B
Precursor Flow Rate (sccm)	150	70
Operating Pressure (Pa)	133	133
Temperature (K)	573	573
N_2 Flow Rate (sccm)	150	150
Process Time (s)	1.1	2.0

Table 2 – A comparison of the standard linear, piecewise, and modified median-effect regression model parameters that are calculated from the standard least squares method for Steps A and B.

R2R	Half-Cycle	Regression Model	R^2	Process gain	Bias
R2R-1	Step A	Linear Model	0.9251	1.1807	– 0.1350
		Piecewise-1	0.9366	1.4446	– 0.2335
		Piecewise-2	0.8688	0.1820	0.8069
		Modified Median-Effect	0.9740	7.5727	1.5184
	Step B	Linear Model	0.8019	0.7011	– 0.1728
		Piecewise-1	0.9520	1.3256	– 0.7072
		Piecewise-2	0.8531	0.1222	0.7719
		Modified Median-Effect	0.9854	7.3701	0.8523
R2R-2	Step A	Linear Model	0.9651	0.8286	–145.6438
	Step B	Linear Model	0.9398	1.2388	– 74.8118

controller. The first step of the control work is to update the model, which is expressed below:

$$a_t = \lambda(y_t - bu_{t-1}) + (1 - \lambda)a_{t-1} \tag{9}$$

where λ represents the weight factor that is responsible for the translation of the regression model, a_t is the updated bias, and b is the process gain that is the same as β in Eq. (8). Typically, λ is chosen to be 0.1 ~ 0.3 (Campbell et al., 2002). After the model is updated, the new recipe for the next batch run is computed by the following expression:

$$u_t = \frac{T - a_t}{\beta} \tag{10}$$

where u_t is the new recipe for the next batch run and T is the target or desired value of the output variable.

The stability and robustness of the EWMA-based R2R controllers have been studied; however, it has been reported that the EWMA and PCC (also known as double-EWMA) methods underperform for nonlinear systems due to the linear model-based R2R algorithm (Ning et al., 1996; Moyne et al., 2018). Nevertheless, the EWMA method is able to be used for nonlinear systems if linear regression models are well-developed for fitting nonlinear responses, which is discussed in this work. In this paper, two different R2R controllers are formulated and integrated, resulting in the Multivariable R2R control system in which each controller is modeled using a single-input-single-output (SISO) regression model as shown in Fig. 2. The first R2R (denoted as R2R-1) tracks the reaction progression (reflected by the coverage or etching fraction) to adjust the process time. Typically, a quartz crystal microbalance (QCM) is used to monitor the etching rate through mass changes along the surface of the thin film substrate (Lee and George, 2015) in real time, which allows the etching and coverage fractions to be computed. The second R2R (denoted as R2R-2) manipulates the flow rate of the precursors by monitoring the precursor partial pressure deviation from the desired partial pressure at the standard operating conditions for the developed inclined plate reactor outlined in Table 1. The following sections discuss the tuning methodology of R2R-1 and R2R-2 in greater detail (Table 2).

3.1. R2R-1 modeling and tuning

The schematic process diagram shown in Fig. 3 illustrates the Multivariable R2R control system. R2R-1 is implemented to adjust the valve opening time (i.e., process time) for precursor injection to the reactor to achieve complete AlF_3 coverage and etching for Steps A and B, respectively. Multiscale computational fluid dynamics (CFD) simulation data are collected at standard operating conditions for Steps A and B for the inclined plate reactor, which are summarized in Table 1. Process or valve opening times of 1.1 s and 2.0 s in Table 1 reflect the ideal time for the Al_2O_3 thin film substrate to reach complete coverage or etching, respectively, under the standard operating conditions. In this work, it is assumed that the time for the valve to fully opens is negligible, and thus, the dynamics of the valve do not interfere with the process dynamics. In other words, R2R-1 only adjusts the process time; however, the flow rate is only adjusted by the upstream valve as shown in Fig. 3, which is

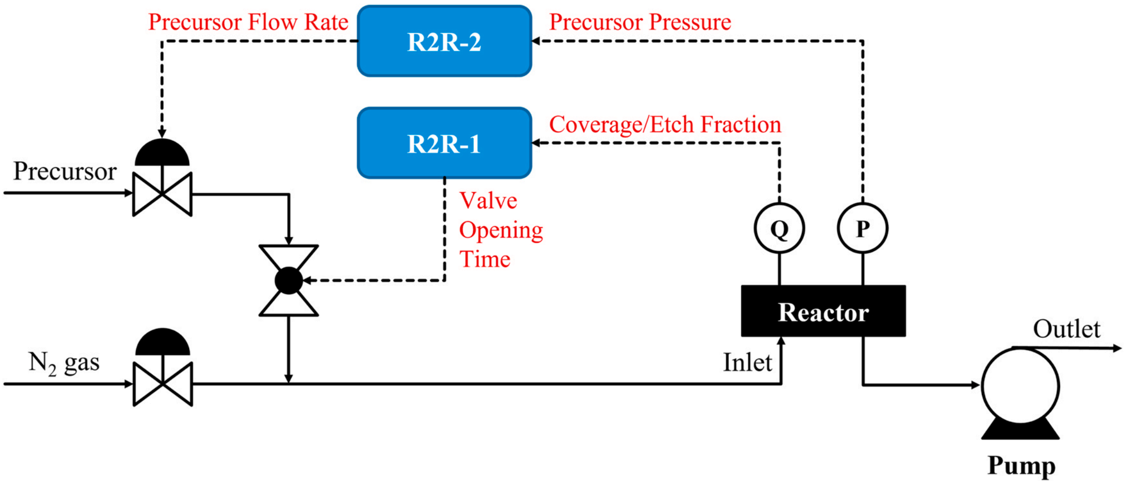


Fig. 3 – The Multivariable Run-to-Run control system.

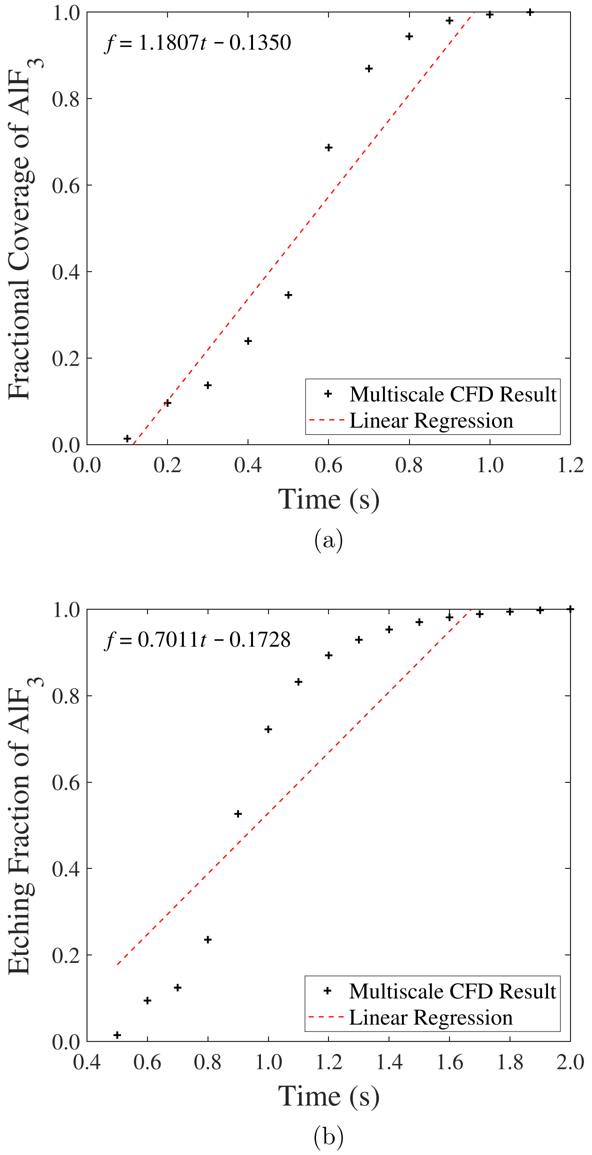


Fig. 4 – The input-output relationship between the fractional coverage of AlF_3 and the precursor valve opening time (a) for Step A and the etching fraction of AlF_3 and the precursor valve opening time (b) for Step B, which is derived from a standard linear regression model for the EWMA-based R2R controller, R2R-1. For Steps A and B, the R^2 values from Table 2 indicate marginal linear behavior for Step A and a lack of linear behavior for Step B.

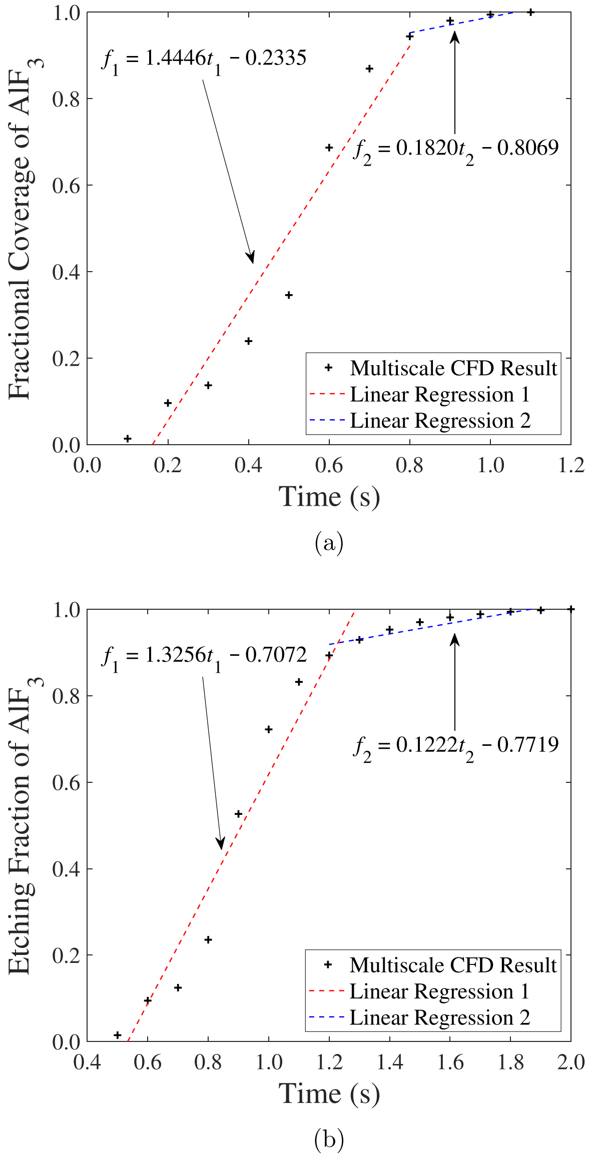


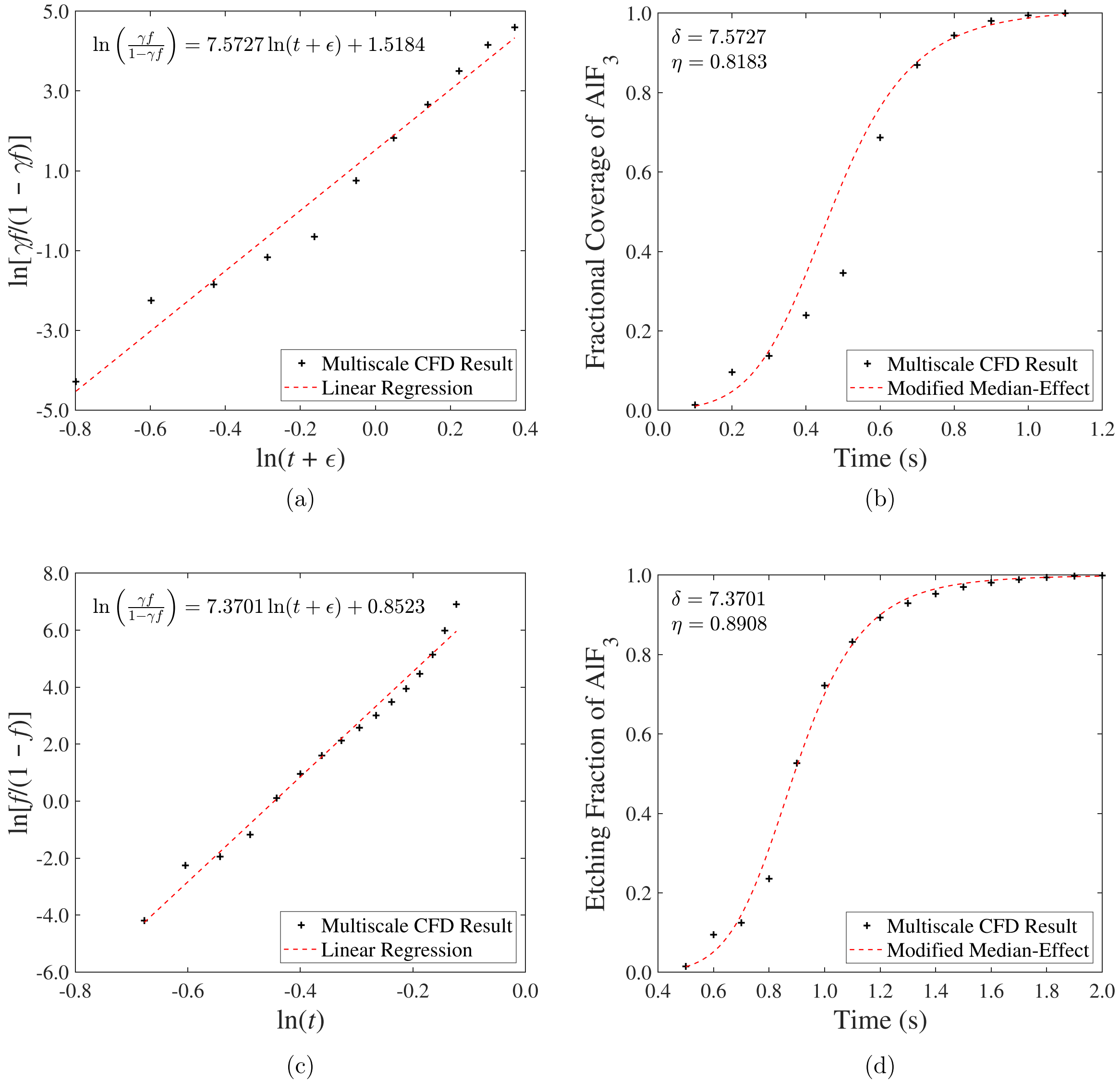
Fig. 5 – The input-output relationship between the fractional coverage of AlF_3 and the precursor valve opening time (a) for Step A and the etching fraction of AlF_3 and the precursor valve opening time (b) for Step B, which is derived from a standard linear regression model divided into linear piecewise functions for the EWMA-based R2R controller, R2R-1. The R^2 values from Table 2 indicate a marginal to moderate linear relationship of multiscale CFD data.

controlled by R2R-2. The R2R-1 controller is approximated to a single-input-single-output (SISO) model in which the process time (t) for the precursor injection serves as the recipe ($u = t$ in Eqs. (12 and 13)) and the etch or coverage fraction (f) of AlF_3 serves as the output variable ($y = f$ in Eqs. (12 and 13)). It is of paramount importance to determine a reasonable solution to the SISO regression model with low variance along the regression line since the solution has a great impact on the control work (Moyne et al., 2018). In this work, multiple regression models for R2R-1 are developed and evaluated using the standard least squares method between the result data and the solution of the SISO model. First, the standard linearization of the multiscale CFD simulation results for Steps A and B are presented in Fig. 4.

Fig. 4 reveals that the linear regression model does not fit the simulation data for Steps A and B due to their nonlinear process responses and the larger variance of the data from the regression line, thus the linear regression of the entire data set may be difficult to accomplish. To improve the least squares values and thereby strengthen the fit of the regression model to the multiscale CFD results, several techniques are used to achieve a better correlation of the model data: piecewise and modified median-effect. For the piecewise regression model, the data curve is divided into two groups to generate two linear piecewise plots that have an intersection point (t_p,f_p) at a time of, t_p, which is expressed as follows:

$$f_p = \frac{\alpha_2 - \alpha_1}{\beta_1 - \beta_2} \tag{11}$$

where α_1 and α_2 are the biases for the two plots, respectively, and β_1 and β_2 are the process gains, respectively. A



(a)

(b)

(c)

(d)

Fig. 6 – The input-output relationship with the logarithmized multiscale CFD data and logarithmized time from the modified median-effect equation for Step A (a) with $\gamma = 0.99$ and $\epsilon = 0.35$ and for Step B (c) with $\gamma = 1.00$ and $\epsilon = 0$ for the EWMA-based R2R controller, R2R-1. The R^2 values in Table 2 indicate a strong linear relationship for Steps A and B. The multiscale CFD results with the standard modified median-effect equation are presented in (b) and (d) for Steps A and B, respectively.

conditional loop is necessary to ensure the correct piecewise regression model to use, which is based on the intersection point, f_p. The piecewise regression models are presented in Fig. 5 and displays a marginal improvement of linearity for both precursor injection steps.

As another alternative technique to improve the curve fitting of the multiscale CFD results, the median-effect equation is adopted from prior research (Chou, 1976) and is applied to the formulation for the coverage and etching fraction progression with time. The median-effect equation is suitable for the EWMA method because of its logistic-like function behavior, which exemplifies the multiscale CFD data trend and its ability to transform both the coverage (or etching) fraction and the process time into logarithmic forms. The median-effect method can also be modified into a linearized function in the form of Eq. (8), which will be discussed later in this section. The median-effect equation is defined as follows:

$$\frac{f}{1-f} = \left(\frac{t}{\eta}\right)^{\delta} \tag{12}$$

where f is the coverage or etching fraction, t is the process time, and the constants, η and δ, are obtained by regressing the median-effect equation into a linearized form to obtain the slope and y-intercept. It is also notable that f can be solved algebraically, which demonstrates the practicality of integrating the median-effect equation into the EWMA model. In this research, additional tuning constants (γ and ϵ) are introduced to reduce the variance of the data to generate a “modified” median-effect equation, which is defined below.

$$\frac{\gamma f}{1-\gamma f} = \left(\frac{t+\epsilon}{\eta}\right)^{\delta} \tag{13}$$

In particular, γ and ϵ are adjustable parameters for modifying the median-effect equation in Eq. (12) and are determined by tuning the factors until a desirable R^2 value of the linear regression is obtained. The original median-effect equation from Chou (1976) is obtained by declaring $\gamma = 1$ and $\epsilon = 0$ in Eq. (13). The parameter, γ, is bounded such that $\gamma \in (0, 1]$, and is employed to shift the upper horizontal asymptote of the median-effect regression while ϵ is utilized for translating

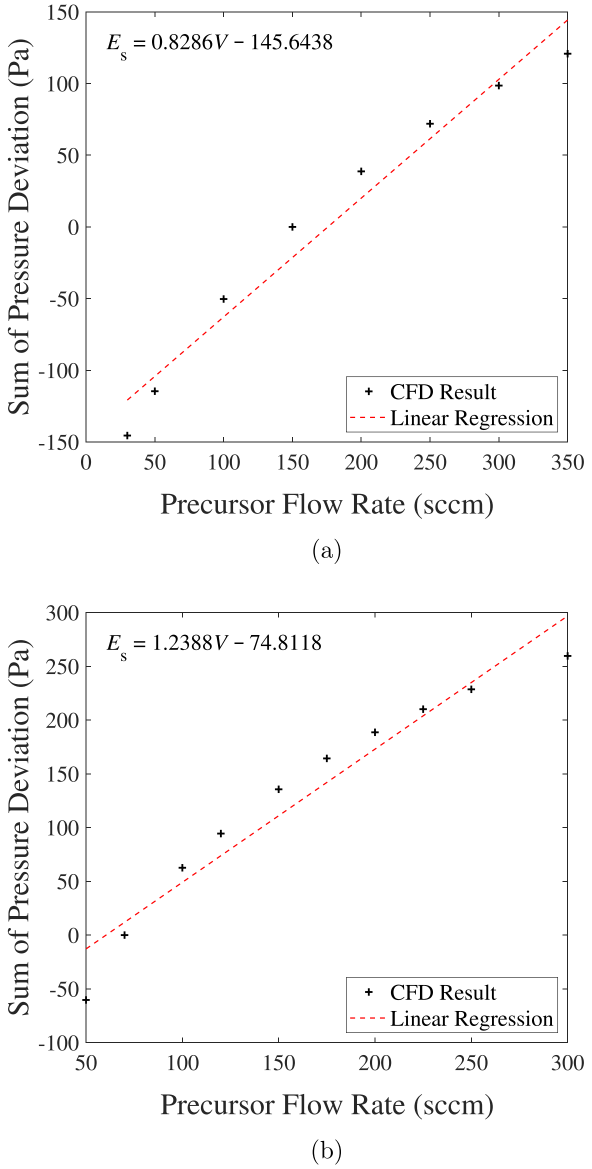


Fig. 7 – The input-output relationship between the sum of partial pressure deviations and the flow rate for Step A (a) and Step B (b), which is derived from a standard linear regression model for R2R-2. The R^2 values from Table 2 indicate a moderate linear relationship.

the regression line along the ordinate direction. The linearized form of the modified median-effect equation exemplified by Chou (2011) is derived below:

$$\ln\left(\frac{\gamma f}{1-\gamma f}\right) = \delta \ln(t+\epsilon) - \delta \ln(\eta) \tag{14}$$

Plotting the left-hand side of Eq. (14) as a function of $\ln(t+\epsilon)$ can generate a linearized plot to determine the values of η and δ through linear regression where $\beta = \delta$ and $\alpha = -\delta \ln(\eta)$ from Eq. (8). The regression models of the modified median-effect equations for Steps A and B are presented in Fig. 6. For Step A, values of $\gamma = 0.99$ and $\epsilon = 0.35$ are used and for Step B, values of $\gamma = 1.0$ and $\epsilon = 0$ are used, and hence, Step B is represented by the original median-effect model. The linearized modified median-effect regression model in Fig. 6a shows that the transformed multiscale CFD dataset exhibits a more linear relationship compared to that of the linear and piecewise regression models. The regression model also correlates the data from the multiscale CFD simulation reasonably, which is observed in Fig. 6b. The results from the modified median-effect regression model for Step B indicate a strong correlation of the multiscale CFD data and are displayed in Fig. 6c and d. The transformation of the multiscale CFD data improves the linearization of both the linear (Fig. 4b) and piecewise (Fig. 5b) regression models, which is validated by an R^2 value of 0.9854 and exhibits a strong least squares fit. In the simulations, a given output variable is logarithmized to compute a logarithmized recipe for the next batch run, while a logarithmized recipe is calculated from Eq. (14) and the antilogarithmic form is used to simulate the process.

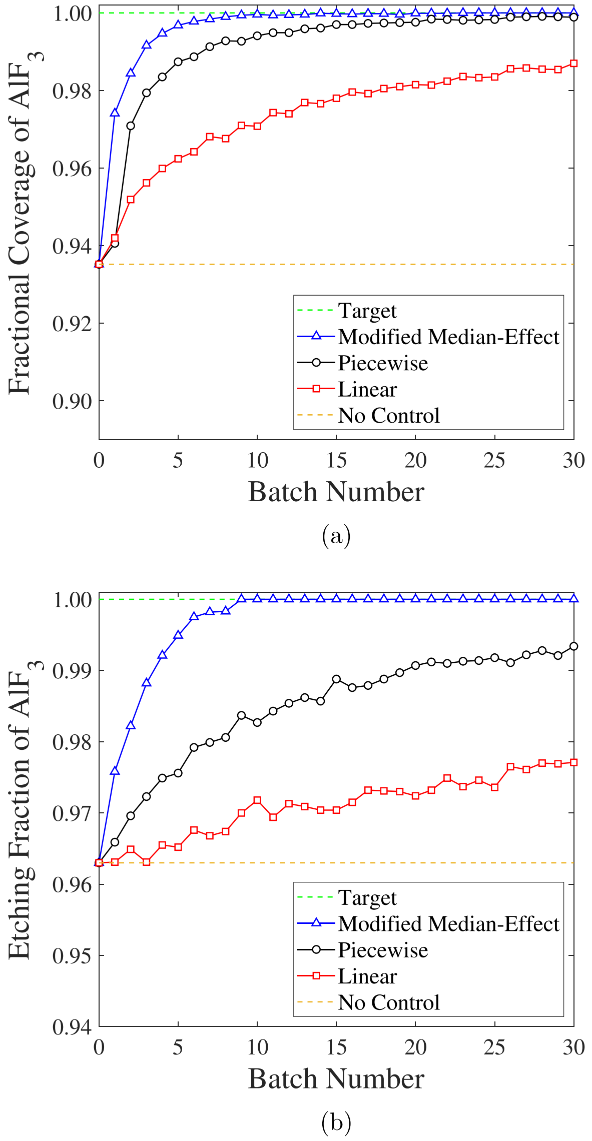
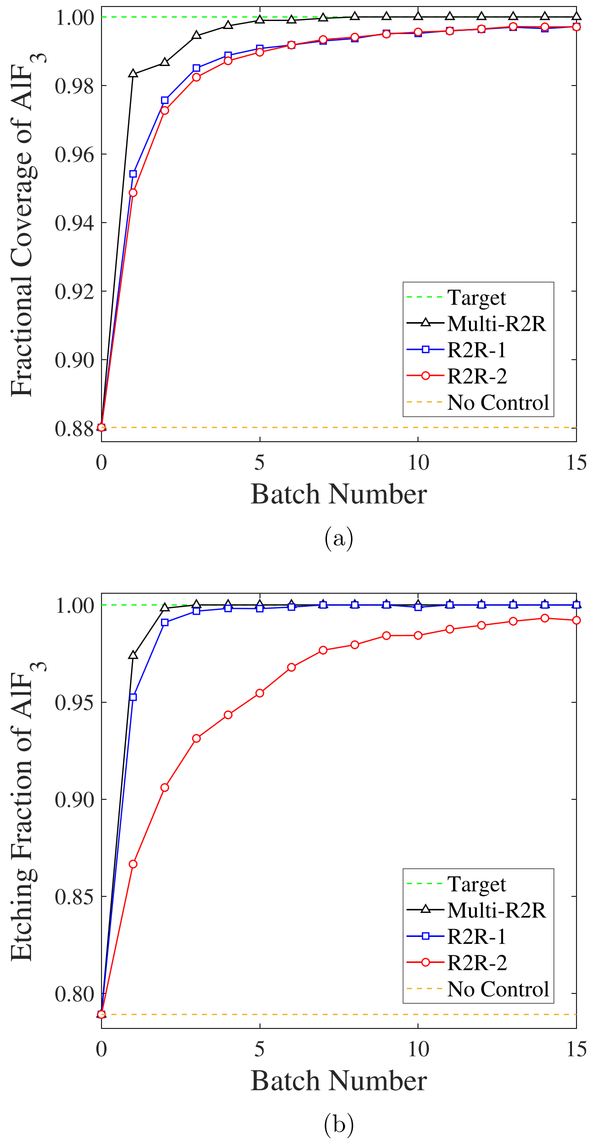


Fig. 8 – Comparison of the responses for various regression methods of R2R-1 under the presence of a kinetic disturbance for Steps A (a) and B (b). The weight factors (λ) of 0.3 and 0.1 are used for Steps A and B, respectively.

In conclusion, the SISO regression models of R2R-1 controllers for both half-cycles are obtained using three different regression methods and the parameters are presented in Table 2. As a result, the modified median-effect method provides better linear models for the SISO regression and overcomes the nonlinear relationship of the multiscale CFD data. The performances of the three models are evaluated in Section 4.



(a)

(b)

Fig. 9 – Comparison of the responses of various R2R control systems under the presence of a kinetic and pressure disturbance for Steps A (a) and B (b). A weight factor (λ) of 0.3 is chosen for all case studies. R2R-1 and the multivariable R2R system are simulated with the modified median-effect regression model.

3.2. *R2R-2 modeling and tuning*

R2R-2 is developed to minimize the precursor partial pressure deviation from the ideal partial pressure at the standard operating condition within the reactor by adjusting the precursor flow rate to the reactor. The R2R-2 controllers for both half-cycles are also approximated to SISO models with the precursor flow rate, V, serving as the recipe ($u = V$), and the sum of the partial pressure deviation at various times, E_s, is defined as the output variable ($y = E_s$). The controlled variable for R2R-2 is expressed as the following:

$$E_s = \sum_{i=1}^{4} C_i \left[P_m(t_i) - P_d(t_i)\right] \qquad (15)$$

where $P_m(t_i)$ refers to the measured partial pressure at time t_i for four times of 0.1 s, 0.2 s, 0.4 s, and 0.6 s, and $P_d(t_i)$ refers to the desired partial pressure obtained from the standard operating pressure condition that is outlined in Table 1 without any disturbances. C_i represents a coefficient that is selectively chosen to improve the linearity of the recorded pressure deviation at the four times. A pressure measuring device such as a pressure sensor can be implemented to monitor the pressure within the reactor at various times to ensure that the sum of the partial pressure deviations is reduced to prevent further loss of control of the system and degradation of the substrate materials. Therefore, R2R-2 can effectively function with the sum of pressure deviations as the output parameter, y_t, and as the precursor (HF and TMA) flow rates serving as the recipe, u_t. As illustrated in Fig. 7a and b, the SISO regression models fit the multiscale CFD data well using the coefficients of 1.0 for all times for Step A and 1.0 for all times for Step B except for 2.0 for the first timestep ($C_1 = 2.0$) for Step B. The solutions of the SISO model have R^2 values of 0.9651 and 0.9398 for Steps A and B, respectively.

4. Simulation results and discussion

Semiconductor manufacturing processes can be subject to some unmeasurable drifts or shifts from wall deposition and equipment aging in the semiconductor industry. In this work, two different disturbances are simulated to evaluate the performance of the multivariable R2R control system. The first disturbance is a shift (referred to as a “kinetic disturbance”), which is able to be driven by cyclical operations, machine maintenance or changes in process settings (Moyne et al., 2018). It can be implemented simply by multiplying the reaction constants by a factor that is less than 1. The other disturbance (referred to as a “pressure disturbance”) is simulated by reducing the operating pressure, which can be caused by a malfunction of the vacuum pump. In particular, R2R-2 is designed to deal with any pressure deviation from the standard conditions. A threshold value of 0.999 is set as the target for the fractional coverage for Step A and of the etching fraction for Step B. The multiscale computational fluid dynamics (CFD) simulation is performed by using 24 parallel computing processors with 384 GB memory on a compute cluster. Simulation time for a half-cycle takes half an hour on average. The batch-to-batch calculation of the control action is not demanding compared to the multiscale CFD simulation, and thus, it is considered negligible.

Initially, the R2R-1 controllers for both half-cycles are simulated under a kinetic disturbance using the multiscale CFD model to evaluate the regression models discussed in Section 3.1. After a regression model with a better control performance is selected, the Multivariable R2R control system is simulated, evaluated, and then compared with the single R2R control system (R2R-1 and R2R-2) under the two disturbances. As shown in Fig. 8, the modified median-effect regression model outperforms the linear and piecewise regession models under a kinetic disturbance. The kinetic disturbance is modeled such that the reaction rate constants are multiplied by a shift factor of 0.7 in the kMC model when the process time is updated, leading to a reduction in the reaction rates. The modified median-effect regression models reach the target for Step A and Step B at batch runs of 8 and 9, respectively. However, the individual R2R models do not reach the target value until a batch run of 30. One of the major concerns of process control is the possibility that the controllers may dramatically overshoot the target threshold. However, Fig. 8 reveals that the modified median-effect models not only reach the target in a reduced number of batch runs, but also when they approach the target, the controllers add sufficiently small increments to the new recipe to prevent the possibility of overshooting the target. In other words, the controller response overshoots and rapid control actions are not allowed in the process due to the best curve fit, and thus, the modified median-effect regression model is selected for the R2R-1 controllers.

The results of the Multivariable R2R and individual R2R control systems under the two shift disturbances for Steps A and B, respectively, are presented in Fig. 9. To model the kinetic disturbance, shift factors of 0.55 and 0.7 are multiplied to the reaction rate constants for Step A and Step B, respectively, and the pressure disturbance is simulated by reducing the operating pressure of the reactor to 40 Pa from 133 Pa. For Step A, both individual R2R configurations reveal that the effects of the disturbances are not able to be removed within 15 batch runs as shown in Fig. 9a. For Step B, as illustrated in Fig. 9b, R2R-1 is able to mitigate the impact of the disturbances within 7 batch runs. However, R2R-2 is not able to eliminate the effects of the disturbances within 15 batch runs. In contrast to individual R2R controllers, the multivariable R2R control system is able to successfully mitigate the effect of the disturbances within batch number 5 for Step A in Fig. 9a and within 3 batch runs for Step B in Fig. 9b. As a result, the multivariable R2R configuration performs more efficiently since a lesser number of batch runs is required to approach complete coverage or etching, which is represented by the target line.

The multivariable R2R control algorithm is also efficient in regard to adjusting the input variables: process time and precursor flow rate. Fig. 10 shows the corresponding input values from the control work in Fig. 9. For both half-cycles, R2R-1 continuously increases the process time to compensate for the effect of the disturbances, as can be seen in Fig. 10a and c. However, In Step B, R2R-1 does not update the recipe after batch run 9 since it achieves a full etching fraction. R2R-2 also continues to increase the flow rate to reach the target as shown in Fig. 10b and d. In this work, there is no limit to the precursor flow rate even though there are valve opening constraints in the semiconductor industry. As predicted, the multivariable R2R algorithm manipulates the two input values such that both parameters (process time and flow rate) are significantly less than those of the R2R-1 and R2R-2. Therefore, it is demonstrated that the multivariable R2R control system exhibits a stronger performance compared to that of the individual R2R-1 and R2R-2 control systems.

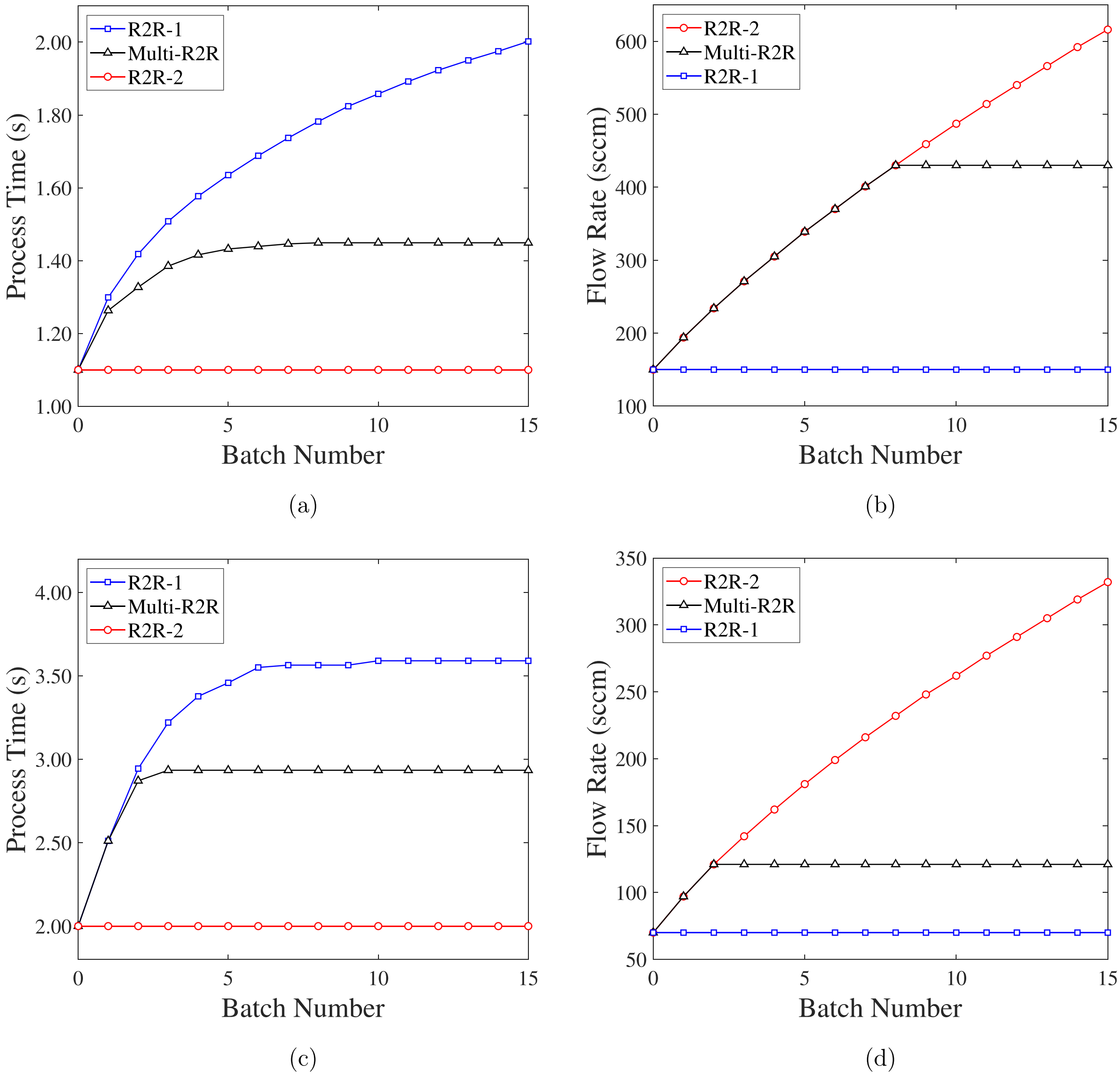


Fig. 10 – Progression of the adjustments made to the recipes (process time, precursor flow rate) in the presence of a kinetic and pressure disturbance through various EWMA-based R2R control systems for Steps A (a-b) and B (c-d).

5. Conclusion

In this work, the previously developed multiscale computational fluid dynamics (CFD) model for the thermal atomic layer etching (ALE) of aluminum oxide thin films was employed to build a multivariable R2R control system using an integration strategy. First, two individual R2R controllers (R2R-1 and R2R-2) were formulated based on the single-input-single-output (SISO) regression method using the exponentially weighted moving average (EWMA) algorithm. To significantly increase the linearity of the SISO model, a novel regression method, the modified median-effect, was implemented and compared to standard linearization and piecewise linearization. The modified median-effect method outperformed the other regression models and demonstrated the best fit of the multiscale CFD data using the EWMA method for the nonlinear system. R2R-1 was designed to adjust the valve opening time of precursor release to the reactor by measuring the coverage or etching fraction on the wafer, while R2R-2 was formulated to maintain the desired partial pressure of the precursor by manipulating the precursor flow rate into the reactor. Kinetic and pressure disturbances, which are industrially-relevant disturbances, were introduced to the system to determine the effectiveness of various R2R control algorithms. Consequently, this study substantiates that the multivariable R2R control scheme is able to successfully achieve complete coverage and etching and overcome the effects of disturbances in the least number of batch runs compared to that of the individual R2R control schemes implemented only one at the time.

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 is gratefully acknowledged. The authors would also like to thank Mr. Junwei Luo and Mr. Yi Ming Ren for their collaboration and assistance during phases of this project.

References

Abdulagatov, A.I., George, S.M., 2018. Thermal atomic layer etching of silicon using O_2, HF, and $Al(CH_3)_3$ as the reactants. Chem. Mater. 30, 8465–8475.

ANSYS, 2021. Ansys Fluent Theory Guide. ANSYS Inc., Canonsburg, PA.

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

Broas, M., Kanninen, O., Vuorinen, V., Tilli, M., Paulasto-Kröckel, M., 2017. Chemically stable atomic-layer-deposited Al_2 O_3 films for processability. ACS Omega 2 (7), 3390–3398.

Campbell, W., Firth, S., Toprac, A., Edgar, T., 2002. A comparison of run-to-run control algorithms. In: Proceedings of the 2002 American Control Conference, 2150–2155, Anchorage, AK.

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.

Chou, T.C., 1976. Derivation and properties of michaelis-menten type and hill type equations for reference ligands. J. Theor. Biol. 59, 253–276.

Chou, T.C., 2011. The mass-action law based algorithms for quantitative econo-green bio-research. Integr. Biol. 3, 548–559.

Crose, M., Kwon, J.S.I., Tran, A., Christofides, P.D., 2017. Multiscale modeling and run-to-run control of PECVD of thin film solar cells. Renew. Energy 100, 129–140.

Crose, M., Zhang, W., Tran, A., Christofides, P.D., 2019. Run-to-run control of PECVD systems: application to a multiscale three-dimensional CFD model of silicon thin film deposition. AIChE J. 65, e16400.

Fan, S.K.S., Jiang, B.C., Jen, C.H., Wang, C.C., 2002. SISO run-to-run feedback controller using triple EWMA smoothing for semiconductor manufacturing processes. Int. J. Prod. Res. 40, 3093–3120.

Gillespie, D.T., 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J. Comput. Phys. 22, 403–434.

Jansen, A.P.J. (Ed.), 2012. An Introduction to Kinetic Monte Carlo Simulations of Surface Reactions. Academic Press, London, pp. 1.

Jurczak, M., Collaert, N., Veloso, A., Hoffmann, T., Biesemans, S., 2009. Review of FINFET technology. In: 2009 IEEE International SOI Conference, Foster City, CA, USA, 1–4.

Kim, K.S., Kim, K.H., Nam, Y., Jeon, J., Yim, S., Singh, E., Lee, J.Y., Lee, S.J., Jung, Y.S., Yeom, G.Y., Kim, D.W., 2017. Atomic layer etching mechanism of MoS_2 for nanodevices. ACS Appl. Mater. Interfaces 9, 11967–11976.

Kwon, J.S.I., Nayhouse, M., Christofides, P.D., 2015. Multiscale, multidomain modeling and parallel computation: application to crystal shape evolution in crystallization. Ind. Eng. Chem. Res. 54, 11903–11914.

Lee, Y., George, S.M., 2015. Atomic layer etching of Al_2 O_3 using sequential, self-limiting thermal reactions with $Sn(acac)_2$ and hydrogen fluoride. ACS Nano 9, 2061–2070.

Lee, Y., DuMont, J.W., George, S.M., 2016. Trimethylaluminum as the metal precursor for the atomic layer etching of Al_2 O_3 using sequential, self-limiting thermal reactions. Chem. Mater. 28, 2994–3003.

Li, C., Metzler, D., Lai, C.S., Hudson, E.A., Oehrlein, G.S., 2016. Fluorocarbon based atomic layer etching of Si_3 N_4 and etching selectivity of SiO_2 over Si_3 N_4. J. Vac. Sci. Technol. A 34, 041307.

Lou, Y., Christofides, P.D., 2003. Feedback control of growth rate and surface roughness in thin film growth. AIChE J. 49, 2099–2113.

Lu, W., Lee, Y., Murdzek, J., Gertsch, J., Vardi, A., Kong, L., George, S.M., delAlamo, J.A., 2018. First transistor demonstration of thermal atomic layer etching: InGaAs finfets with sub-5 nm fin-width featuring in situ ale-ald. In: 2018 IEEE International Electron Devices Meeting (IEDM), 39.1.1–39.1.4, San Francisco, CA.

Metzler, D., Bruce, R.L., Engelmann, S., Joseph, E.A., Oehrlein, G.S., 2014. Fluorocarbon assisted atomic layer etching of SiO_2 using cyclic Ar/C_4 F_8 plasma. J. Vac. Sci. Technol. A 32, 020603.

Moyne, J., Del Castillo, E., Hurwitz, A.M., 2018. Run-to-Run Control in Semiconductor Manufacturing. CRC Press.

Ning, Z., Moyne, J., Smith, T., Boning, D., DelCastillo, E., Yeh, J.Y., Hurwitz, A., 1996. A comparative analysis of run-to-run control algorithms in the semiconductor manufacturing industry. In: IEEE/SEMI 1996 Advanced Semiconductor Manufacturing Conference and Workshop. Theme-Innovative Approaches to Growth in the Semiconductor Industry. ASMC 96 Proceedings, 375–381.

Razavieh, A., Zeitzoff, P., Nowak, E.J., 2019. Challenges and limitations of CMOS scaling for FinFET and beyond architectures. IEEE Trans. Nanotechnol. 18, 999–1004.

Voas, J., Kshetri, N., DeFranco, J.F., 2021. Scarcity and global insecurity: the semiconductor shortage. IT Prof. 23, 78–82.

Yun, S., Ding, Y., Zhang, Y., Christofides, P.D., 2021. Integration of feedback control and run-to-run control for plasma enhanced atomic layer deposition of hafnium oxide thin films. Comput. Chem. Eng. 148, 107267.

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., Ou, F., Orkoulas, G., Christofides, P.D., 2022b. Multiscale computational fluid dynamics modeling of thermal atomic layer etching: application to chamber configuration design. Comput. Chem. Eng. 161, 107757.

Zhang, Y., Ding, Y., Christofides, P.D., 2020. Integrating feedback control and run-to-run control in multi-wafer thermal atomic layer deposition of thin films. Processes 8 (18), 2020.