Computers and Chemical Engineering 168 (2022) 108044

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

Machine learning-based run-to-run control of a spatial thermal atomic layer etching reactor

Matthew Tom [a], Sungil Yun [a], Henrik Wang [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:
Semiconductor manufacturing
Spatial thermal atomic layer etching
Run-to-run control
Machine learning
Multiscale modeling
Computational fluid dynamics modeling
Kinetic Monte-Carlo simulation

ABSTRACT

In response to the next technological revolution, atomic layer processes have emerged to produce high-performing, thin-film semiconductor materials. To overcome the long purging times required for conventional atomic layer processes, spatial atomic layer processes have been recognized for their ability to reduce processing times; however, they lack characterization and control. This research aims to construct two novel run-to-run (R2R) control systems using a machine learning model with an artificial neural network (ANN) and an exponentially weighted moving average (EWMA) method for the spatial thermal atomic layer etching (SALE) of aluminum oxide thin films. The two R2R controllers are used in conjunction with a multiscale computational fluid dynamics model of a SALE process with various disturbances to test their effectiveness. Closed-loop simulation results demonstrate that the ANN-based R2R control system reduces etching per cycle variability, maintains the process output within a small region around the setpoint, and outperforms the traditional EWMA-based R2R control system in efficiency.

1. Introduction

Technological innovation in electronic devices has continued to dominate in the global market as high-tech companies maintain their platform by annually introducing technological enhancements to their products. These technological innovations are made possible by the integration of computationally efficient devices that are made possible by semiconductors. The improvements made on semiconductors are consistent with the prediction of Moore's Law (Moore, 1998) as electronics are becoming more densely packed with semiconducting materials; specifically, transistors, which are able to consistently output more computing power while the sizes of these electronics are continually scaling down. For example, the Cerebras company has been able to manufacture wafer-scale chips that are densely occupied with 1.2 trillion transistors (Burg and Ausubel, 2021). As transistors approach near two-dimensional (2D) geometries, transistors are reaching their physical limits (Li et al., 2019) due to the challenges and demands of the fabrication process. One ramification of this stringent fabrication process is the low production rate of these semiconductor devices, which are essential components to most electronic devices, including smart technology and autonomous vehicles. With the rising consumer demand, the semiconductor industry is looking for an efficient and inexpensive method to increase the marketability of these atomic-scale transistors in hopes that conventional, time-consuming processes can be replaced by innovative production methods, thereby increasing their global supply.

The architecture of semiconductor devices has continuously evolved, while semiconductor production specifications become stricter as a consequence of the miniaturization of the transistors. FinFETs, or fin field-effect transistors, became commonly used transistors that were able to improve circuitry speed and preserve charge (Sairam et al., 2007; Jurczak et al., 2009). Nevertheless, they are becoming increasingly difficult to fabricate due to the limitations of the fin dimensions, which are constrained to minimum thicknesses of 7 nm. The fabrication of the fin width to below the 7 nm threshold causes the overall performance of the transistor to deteriorate due to mobility losses and short-channel effects (Razavieh et al., 2019). Following the FinFET era, gate all-around (GAA) transistors emerged to overcome the challenges encountered in the semiconductor miniaturization trend with the desire to maintain the aforementioned properties as well as providing greater power efficiency and electrostatic properties (Guerfi and Larrieu, 2016). Thus, GAA transistors are considered to be superior to FinFETs due to their versatile design and capability of densely

* 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.compchemeng.2022.108044
Received 11 July 2022; Received in revised form 13 September 2022; Accepted 23 October 2022
Available online 28 October 2022

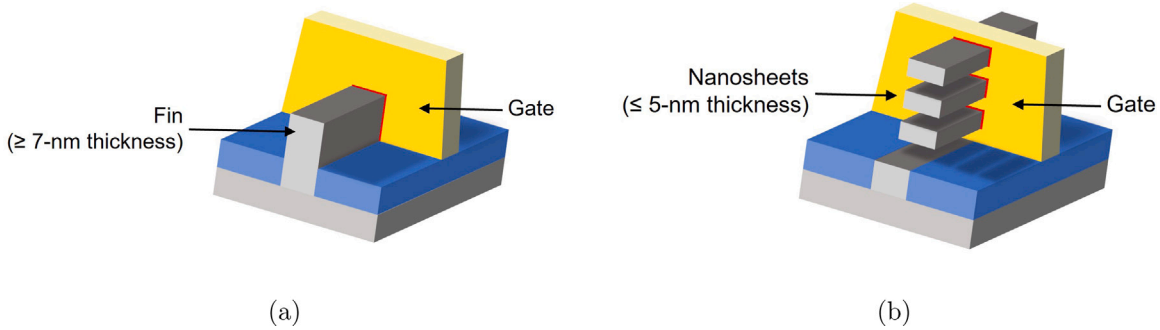

Fig. 1. Illustrations of a (a) FinFET (fin field-effect transistor) and a (b) GAA (gate all-around) transistor. FinFETs are bounded by a thickness of 7 nm as opposed to the nanosheets in GAA transistors, which are able to continue the trend of thickness minimization. However, nanosheet thicknesses are dependent on the oxide film layer, which must be introduced in precise dosages to maintain minimal thickness specifications.

occupying semiconductor wafers, which are attributed to their vertically stacked channels that are the so-called nanosheets or nanowires (also known as nanoribbons). These nanosheets are capable of reaching dimensions below the 5 nm threshold, which offers an extension of Moore's Law. A comparison of the FinFET and GAA architectures is illustrated in Fig. 1, which presents the GAA design that allows nanosheet thicknesses below 5 nm. Despite the potential to overcome the 5 nm threshold, it is difficult to manufacture these GAA transistors because of the stringent quality specifications on the nanosheet dimensions. Over the past decade, much experimentation has been delegated to the precise control of nanosheet thin-film substrate thicknesses through a series of synergistic and self-limiting reactions through atomic layer processes.

Thermal atomic layer etching (ALE), a process that is used in semiconductor fabrication, etches active layers of a thin film substrate with reactive species known as "precursors" that are introduced in sequential pulses and provides greater control of the thin film thickness. Thermal ALE is a subsequent processing solution of thermal atomic layer deposition (ALD), which often encounters deposition growth on non-growth areas causing the misalignment of transistors during the stacking process as well as edge placement errors (Faraz et al., 2015; Kanarik et al., 2015; Oehrlein et al., 2015). Current industrial applications for semiconductor fabrication employ thermal ALE to improve the localization (selectivity) of thin-film deposition on growth areas due to conventional ALD methods lacking a selective mechanism (Merkx et al., 2020). When the reactive environment is combined with high operating temperatures, this process allows for feasible etching rates (Kanarik et al., 2015). Thermal ALE introduces two precursors in sequential steps known as "half-cycles" that are separated by purging steps. The purging steps use an inert gas to exhaust unreacted precursor species and byproducts that may deteriorate the reaction progression and the conformity of the thin film. These purging steps are required for the removal of monolayers of surface material in each cycle to exemplify a "self-limiting" process (Kanarik et al., 2018) by preventing precursor intermixing, and in conventional ALE processes, the purging steps are time-consuming. This self-limiting behavior is made possible by purposefully selecting bulky precursor species that produce steric hindrance during the adsorption process that inhibits the diffusion of these species beyond the surface of the substrate, thereby exemplifying as an impermeable wall (Keuter et al., 2015). Also, metal oxides including Al_2O_3 act as diffusion barriers that also create an impermeable boundary layer (Carrasco et al., 2004; Hirvikorpi et al., 2010). Thus, the proper selection of the precursor species is a critical factor in ensuring that the reactions produce conformal thin films with half-cycles that are spontaneous in nature.

As mentioned previously, the fabrication of the oxide thin film is difficult to accomplish and depends on the precursor and oxide film selection. Several oxide films have already been studied and are summarized by Fortunato et al. (2012), Faraz et al. (2015) and Sheng et al. (2018). An ideal oxide film is characterized by an ability to produce ultra smooth surfaces (Fortunato et al., 2012) without requiring high operating temperatures and costs (Ye et al., 2017; Ding and Wu, 2020) and to produce high reaction rates. The oxide films provide insulation between the semiconductor channels and the gate, which is illustrated in Fig. 1 and serves to reduce current and electron losses to a minimum due to the chemical bonding of oxide compounds that inhibit electron mobility (Sang and Chang, 2020). The thickness of the oxide films also depends on the quantum capacitance, which increases with decreasing oxide film thickness (Sinha and Chaudhury, 2013). The selection of the oxide thin film varies depending on the semiconductor device; however, the oxide film is generally chosen such that the film can be produced with great feasibility from a kinetics and thermodynamics perspective while minimizing the operating temperature and the amount of precursor needed. The procedures required to fabricate aluminum oxide (Al_2O_3) thin films are characterized by the aforementioned characteristics due to the properties of aluminum, which has a higher electropositivity and hence a stronger adherence to strongly electronegative atoms such as fluorine and oxygen (Lee et al., 2016). Thus, Al_2O_3 is selected for this work.

Despite thermal ALE being well-suited for the precise control of the thicknesses of thin film substrates, these processes are also time-consuming due to the long purging steps required to ensure that the ideal reaction environment is reached so that a self-limiting nature is obtained. To resolve this issue, the emergence of rapid atomic layer deposition (ALD) (Zywotko et al., 2018) and spatial ALD (Poodt et al., 2012; Faraz et al., 2015) processes have significantly shortened the total process time while preserving thin film quality and conformity. The most notable difference between rapid ALD and spatial ALD processes is the inclusion of purging steps. Rapid ALD does not require purging steps; rather, it introduces the two precursors sequentially, which results in some intermixing of the precursors. Thus, rapid ALD requires significantly higher precursor pressures in order to maintain the self-limiting nature due to this intermixing (Zywotko et al., 2018). On the other hand, spatial ALD has been proven to be more dependable in producing conformal thin films (Levy et al., 2009). Spatial ALD (SALD) does this by continuously introducing precursors and purging species in physically isolated regions that are adjacent to one another. The substrate travels through each zone for an exposure time that depends on the substrate velocity (Muñoz-Rojas et al., 2019). Fig. 2 illustrates the differences in the reactor model and processes for conventional and spatial ALD/ALE reactors. Specifically, conventional ALD/ALE processes inject precursor and purge species into the reaction chamber in sequential steps, while spatial ALD/ALE processes allow the substrate to move between reaction and purge zones while the precursor and purge species are continuously introduced. In particular, the sheet-to-sheet spatial ALD/ALE reactor (Freeman et al., 2010) is presented in Fig. 2(b). This work aims to integrate the sheet-to-sheet (S2S) spatial ALD/ALE reactor model for the thermal ALE of aluminum oxide thin films.

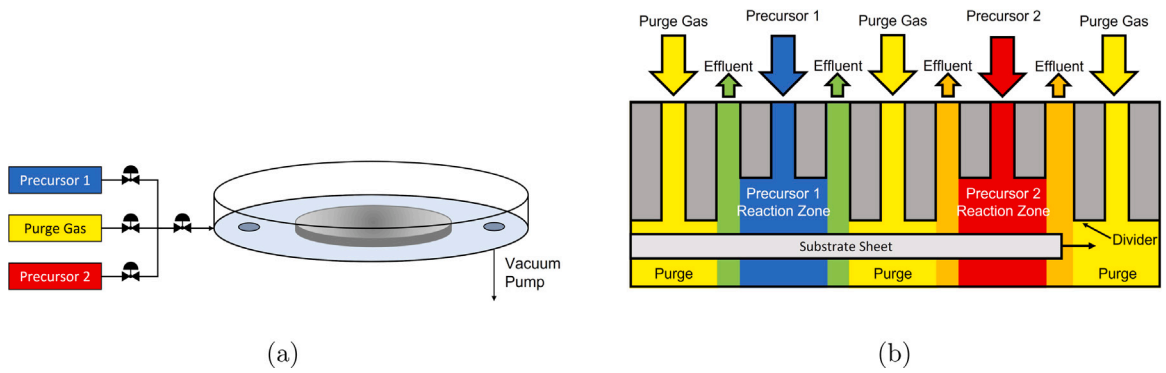

Fig. 2. Conventional (a) and spatial (b) ALD/ALE reactor diagrams illustrating their structural and operational differences. Conventional ALD/ALE reactors contain a stationary wafer substrate that is exposed to the precursor and purge gases in sequential steps. Spatial ALD/ALE reactors feed precursor and purge species continuously in isolated zones while the wafer substrate is driven through each zone at a constant velocity on a conveyor belt.

Although ALE reactions have been integrated into industry and are abundantly used for semiconductor fabrication (Faraz et al., 2015; Kanarik et al., 2015; Oehrlein et al., 2015), there is a limited understanding of how these reactions behave under a range of operating conditions, and generally most experimental modeling has produced limited data sets (Lee et al., 2016). It is difficult to develop procedures that produce optimal results in a laboratory setting due to the reactions being characterized as instantaneous (Rahman, 2021; Lill, 2021); thus, this research adopts a computational modeling approach to overcome the aforementioned challenges. This approach connects a microscopic domain that consists of surface reaction kinetics and thermodynamics to a macroscopic domain that consists of transport phenomena behavior in the fluid phase. The microscopic model establishes a statistically sporadic simulation of the thin film etching process through a kinetic Monte Carlo (kMC) algorithm, while the macroscopic model numerically calculates spatiotemporal parameters through computational fluid dynamics (CFD). The combination of the microscopic kMC simulation and the macroscopic CFD model establishes the so-called "multiscale computational fluid dynamics" model, which becomes a unique algorithm that simultaneously computes the microscopic surface and macroscopic fluid fields. This work will continue from prior work on a two-dimensional (2D) multiscale CFD modeling of spatial thermal ALE of a sheet-to-sheet reactor for aluminum oxide thin films (Yun et al., 2022b) by establishing a control system to ensure that the process operating conditions and output quality are maintained.

It is desired to implement a control system that ensures that the process operation and control will be sustainable in the presence of disturbances that impede the desired reaction rate and conformity of the thin film substrate. The control system would minimize the number of fluctuations observed in the etching per cycle (EPC) by adjusting the input variables until the desired setpoint is reached. For etching processes, there are many sources of shift and drift disturbances. For instance, a wafer substrate that receives an etching process may contribute to variable results in adjacent cycle runs or batches, which can be caused by changes in the condition of the equipment used or unknown irregularities (Moyne et al., 2018). Also, SALE processes can deposit byproducts on the sidewalls of the reactor where the reaction rate is controlled by the temperature of the chamber. The deposited byproducts may affect the heat flux into the chamber, and thus, reduce the reaction rate. With hundreds of process cycles being operated daily, the process would drift as accumulated byproducts are deposited on the walls of the reaction chamber. However, with a robust control system in place, the manipulated inputs of the process would be continuously updated to account for the changes in the reaction chamber environment that could contribute to these disturbances while conforming to product specifications, reducing variability, and maintaining process control.

Due to the high spontaneity of the reaction, it is difficult to archive in-batch control of the real-time operation of the process. Thus, the use of run-to-run (R2R) or batch-to-batch control systems is employed in this work to develop a methodology for updating the input variables based on the target EPC. R2R control systems are typically used to compensate for processes that naturally drift away from their setpoint and processes that may have large variances in between runs. In response to equipment fatigue and changes in input composition, a R2R control system is structured to improve process stability for better process performance and productivity by decreasing the variance in between runs by optimizing input variables based on the previous batch run data (Moyne et al., 2018). By implementing a R2R control system, the variance of the output in between runs will decrease, which will allow the process to stay within operable limits (Kotz and Johnson, 2002). If the process has a naturally occurring shift or drift disturbance, then a R2R control system will also allow the process to operate for longer times in between input adjustments while maximizing production time and improving the overall throughput.

The present work will perform input updates for multiple input variables, including the precursor flow rates and substrate velocity in response to a single output variable, EPC; thus, this R2R control system is modeled based on a multiple-input-single-output (MISO) model. The adjustment of the precursor flow rates and the substrate velocity provides a greater understanding of the adjustments for each input variable and their relation to the computed output variable after each adjustment. Two separate R2R controllers are implemented in conjunction to the multiscale CFD simulation with different methods including the conventionally used exponentially weighted moving average (EWMA) for linear systems and a machine learning (ML) method that is applicable for both linear and nonlinear systems. These two methods will construct the so-called EWMA-based and artificial neural network (ANN) based R2R controllers to determine their response to various disturbances and to conclude which R2R controller is more effective at mitigating these disturbances. The EWMA-based R2R controller, which requires the specification of deterministic weights, will also be implemented with two different weight parameters to determine the effect of the weight parameters on the magnitude of the adjustment on the input variables. It is also worth mentioning that prior research has been accomplished in establishing a R2R control system with a multiscale CFD model for conventional thermal ALE under the presence of a pressure and kinetic shift disturbance (Yun et al., 2022d) to maintain the output variable within a high-precision range as well as ensuring product conformance. However, the aforementioned work utilizes a two-loop single-input single-output (SISO) model to simulate individual input and output variables independently without regard to the coupling between the input and output variables. The present work overcomes this issue and accounts for the coupling between multiple inputs and a single output by developing both an ANN-based and an EWMA-based R2R control system that adopts a robust MISO model for control by accounting for the relations of the multiple input variables and their influence on the output.

2. Spatial ALE modeling

2.1. Microscopic surface domain

2.1.1. Surface kinetics description

Generally, in conventional atomic layer etching (ALE), two half-reactions occur in a sequential manner so that the two precursors do not intermix with one another within the reactor chamber. Effective precursor separation is guaranteed by a purge cycle that uses an inert gas such as N_2 between each precursor half-cycle. Unlike conventional ALE, the two precursors are separately and continuously dosed in distinct locations at constant flow rates in a spatial atomic layer etching (SALE) reactor and thus, isolating the half-reactions spatially. The substrate alternates back and forth under a precursor injection assembly, resulting in a significant reduction of purge time. Fig. 2(b) shows a schematic illustration of a SALE process in which there are two half-reaction zones that are situated between N_2-enriched zones. A vacuum port is located between each half-reaction and N_2 zone to exhaust residual precursors or byproducts produced from the half-reactions. These vacuum ports are imperative for film quality and thickness control due to their purpose in achieving self-limiting behavior.

The spatial atomic layer etching (SALE) of Al_2O_3 utilizes two precursors as reactants: hydrogen fluoride (HF) and trimethylaluminum (TMA), $Al(CH_3)_3$. First, when a substrate moves into the HF reaction zone, HF fluorinates the top surface of Al_2O_3 thin films to yield a modified layer of AlF_3. Theoretically, it is assumed that only a single layer (monolayer) of Al_2O_3 is exposed to HF and is modified, which exemplifies self-limiting behavior. Next, unconsumed species such as HF, byproduct, H_2O, and N_2, are discharged through an adjacent vacuum port when the substrate reaches the vacuum zone using N_2 gas as the carrier gas that is supplied in the N_2-enriched zone. Then, the substrate is introduced to the second precursor, TMA, when it arrives at the adjoining TMA zone. TMA adsorbs onto the modified surface layer and converts the AlF_3 layer into a volatile layer of dimethylaluminum fluoride (DMAF), $AlF(CH_3)_2$. The volatile DMAF is spontaneously desorbed from the surface, resulting in the etching of the AlF_3 layer. Finally, the next vacuum port exhausts residual materials consisting of TMA, DMAF, and N_2. An additional adjacent N_2 zone sweeps any remaining traces of TMA and DMAF through the vacuum zone, thus preparing the substrate for another cycle of the ALE process.

The aforementioned procedure proceeds by transferring the substrate back and forth until the desired thickness is achieved. The overall chemical reaction is described as follows:

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

Furthermore, each half-cycle is described by the following overall reactions:

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

$$AlF_3\ (s) + 2Al(CH_3)_3\ (g) \rightarrow 3AlF(CH_3)_2\ (g)$$

It is necessary to examine all possible intermediate reaction pathways for microscopic modeling; however, there are challenges in identifying the infinite number of possible reaction pathways. Thus, the total number of reaction steps can be simplified by identifying critical intermediate reaction pathways, also known as rate-determining steps, which have a significant impact on the overall reaction time and are described by Lee et al. (2016), Natarajan and Elliott (2018) and Yun et al. (2022a,b) who proposed an elementary reaction pathway for the half-cycles. Due to a lack of experimental data, the reaction pathways were probed on the microscopic level by using density functional theory (DFT) and electronic structure calculations, which were computed by the open-source package, Quantum ESPRESSO (QE). A greater explanation of the kinetic mechanism and its properties are detailed by Yun et al. (2022a). Variables and their definitions for the microscopic model discussion are summarized in Table 1.

Table 1
Definitions of variables used in the microscopic model.

Variable	Definition
A	Pre-exponential factor
A_{site}	Surface area of an active reaction site
E_a	Activation energy
EPC	Etching per cycle
f_c	Coverage fraction
f_e	Etching fraction
h	Planck constant
k	Reaction rate constant
k_B	Boltzmann constant
$k_{d,ads}$	Reaction rate constant of adsorption reaction
k_i	Reaction rate constant of reaction, i
k_{sum}	Sum of the reaction constants
m	Molar mass of adsorption species, HF and TMA
N	Number of reaction pathways
p	Reaction index number
P	Operating pressure
Q	Partition function for the reactants
$Q^{\ddagger}$	Partition function for the transition state
R	Universal gas constant
t	Reaction time progress
T	Operating temperature
Z	Coordination number
Δt	Time interval
Γ_1, Γ_2	Random numbers where $\Gamma_1, \Gamma_2 \in (0, 1]$
σ	Sticking coefficient for HF and TMA

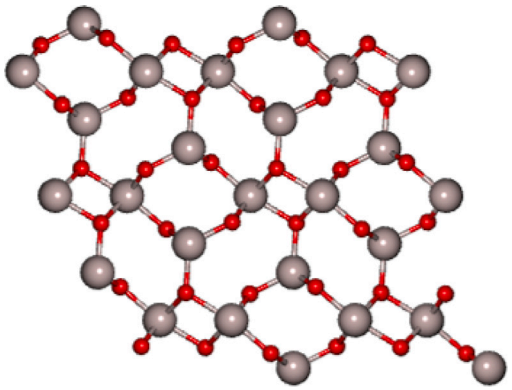

Fig. 3. θ-Al_2O_3 ($\bar{2}$ 0 1) crystal structure.

2.1.2. Lattice model of aluminum oxide

There are various crystal structures of Al_2O_3 that depend on the ambient operating conditions. A proper structure should be constructed, since crystal structures considerably affect the quantum properties and parameters of surface kinetics. Broas et al. (2017) reported that θ-Al_2O_3 ($\bar{2}$ 0 1) was discovered on Si (1 0 0) after the atomic layer deposition (ALD) of Al_2O_3. Therefore, in this work, θ-Al_2O_3 ($\bar{2}$ 0 1), which is visualized in Fig. 3, is used as a substrate for SALE. θ-Al_2O_3 ($\bar{2}$ 0 1) is approximated onto a lattice model that comprises a 300 × 300 grid to simulate the surface kinetics; thus, the lattice model has 90,000 reaction sites, Al and O for Step A and Al and F for Step B, for each monolayer, and the reaction sites are visualized in Yun et al. (2022a). It is notable that increasing the size of the lattice grid would dramatically increase the computation time, therefore, this work utilizes a 300 × 300 grid to improve the efficiency of the simulation of the surface reaction kinetics while possibly sacrificing the accuracy of the computation. To overcome this issue, a stochastic approach to exemplify the randomness of surface kinetics is elucidated in the following section.

2.1.3. Kinetic Monte Carlo simulation

In general, surface reactions are described by macroscopic rate equations such as reaction rate expressions and mass balances that define the surface kinetics. However, this approach fails to integrate

the stochastic behavior of surface kinetics at the microscopic level. Therefore, this paper will use the kinetic Monte Carlo (kMC) method to model the surface kinetics in great detail due to its ability to accurately describe surface kinetics from a microscopic perspective. The kMC method is a computing method that is used to simulate surface reactions based on a fundamental principle that chemical reactions exhibit stochastic behavior. Specifically, the modeling of surface kinetics in various reaction mechanisms has been performed in a great deal of research with kMC methods (Lou and Christofides, 2003; Fu et al., 2008; Shirazi and Elliott, 2014; Weckman et al., 2018; Ding et al., 2019, 2020b; Yun et al., 2022a). In this work, the variable step size method, also known as the Bortz–Kalos–Lebowitz (BKL) algorithm or the n-fold way, is used to model surface reactions at the molecular level.

To employ the kMC method, all reaction rate constants must be specified and computed using statistical thermodynamics and quantum mechanics principles. Various methods are selected based on the type of reaction that is performed. The rate constants of surface reactions, diffusion, and desorption are commonly computed with transition-state theory (TST) as follows (Jansen, 2012):

$$k = A \exp\left(\frac{-E_a}{RT}\right) \tag{1}$$

$$A = \frac{k_B T}{h} \frac{Q^{\ddagger}}{Q} \tag{2}$$

where k represents the reaction rate constant, A indicates the pre-exponential factor, E_a denotes the activation energy, R is the gas constant, T is the temperature, k_B is the Boltzmann constant, h is the Planck constant, $Q^{\ddagger}$ is the quantum partition function of the transition state, and Q is the quantum partition function of the reactants. The ratios of the transition state and reactant quantum partition functions are simplified to unity (Jansen, 2012). Previously, the activation energies of all reaction pathways were computed with the nudged elastic band (NEB) calculation method (Yun et al., 2022a). The NEB method locates a minimum energy path between the reactants and products to identify saddle points (i.e., the transition state of reactants) and calculates the activation energies of reactions (Berne et al., 1998). Additionally, the reaction rate constants for adsorption reactions can be calculated through collision theory (CT), which is derived from Maxwell–Boltzmann statistics (Jansen, 2012).

$$k_{d,ads} = \frac{2PA_{site}\sigma}{Z\sqrt{2\pi m k_B T}} \tag{3}$$

where Z represents the coordination number, σ is the sticking coefficient, A_{site} is the area of an active reaction site, and m is the mass of the adsorption species. In this work, the sticking coefficients of HF and TMA are selected as 0.15 (Fontaine et al., 2012) and 0.02 (Schwille et al., 2017), respectively.

The kMC simulation is conducted on the lattice model of Al_2O_3 as discussed in Section 2.1.2 in accordance with the following procedures:

1. The sum of all reaction constants, which is an integral parameter to formulate the kMC algorithm, is first calculated to simulate the surface reactions on the lattice model.

$$k_{sum} = \sum_{i=1}^{N} k_i \tag{4}$$

where k_{sum} is the sum of all reaction constants, k_i is the rate constant of the reaction i, and N is the total number of the possible reaction pathways. k_{sum} is calculated whenever the surface reactions proceed.

2. A reaction is then selected by picking a random number, $\Gamma_1 \in (0, 1]$, that satisfies the following algorithm:

$$\sum_{i=1}^{p-1} k_i \leq \Gamma_1 k_{sum} \leq \sum_{i=1}^{p} k_i \tag{5}$$

where p is the reaction index. If the chosen Γ_1 holds for Eq. (5), the reaction p is selected on the reaction site of the lattice model.

3. The time interval, Δt, is computed to evolve the system clock according to a random number, $\Gamma_2 \in (0, 1]$ after the kMC algorithm has been performed on all of the reaction sites of the lattice model.

$$\Delta t = \frac{-\ln \Gamma_2}{k_{sum}} \tag{6}$$

4. The system clock evolves, $t \rightarrow t + \Delta t$.

Steps 1 through 4 (above) are repeated in a loop until a termination condition is met. The kMC algorithm is governed by k_{sum}, which is continuously updated. Thus, the reaction pathways that occur are enabled and those that are not observed are ignored across the lattice model as k_{sum} is updated. From the kMC model, the fraction of substrate surface that has been modified by HF, f_c, and the fraction of substrate surface that is etched by TMA, f_e, are calculated based on the occurrence of reactions from Step 2. The etching per cycle (EPC) in Å/cycle is calculated by multiplying the coverage and etching fractions by a coefficient of 0.46 Å/cycle, which resembles the maximum amount of EPC that can be produced and has been observed by Lee et al. (2016), which is described by the following equation:

$$\text{EPC} = 0.46 \times f_c \times f_e \tag{7}$$

where EPC $\in [0, 0.46]$ such that $f_c, f_e \in [0, 1]$.

Remark 2.1. The microscopic model may produce a disparity between data results. One of the disadvantages of the microscopic kMC model is that it utilizes a random number simulation such as a pseudo-random number generator that may produce intrinsic variability in the results with no way to determine the best and worst-case scenarios (Raychaudhuri, 2008). As a result, the variance in the time evolution and the calculation of the coverage and etching fractions will contribute to fluctuations in the etching per cycle calculation. However, there is no deterministic approach that can predict the outcome of the fluctuations prior to initiating the simulation. The statistical error generated from the kMC simulation is largely dependent on the input parameters, such as the size of the lattice and the domain of the random number generator, which must be specified carefully.

2.2. Macroscopic fluid domain

In the study of the macroscopic domain, it is necessary to simulate the effects of the fluid distribution within the reactor and determine their effect on the operation of the reactor. The design of the two-dimensional (2D) spatial ALE reactor is first constructed using the computer-aided design (CAD) software, Ansys SpaceClaim, and then discretized into finite elements through meshing software via Ansys Workbench. To simulate the macroscopic fluid domain, a finite element difference method is used through Ansys Fluent to numerically simulate the transport phenomena effects spatiotemporally.

The reactor assembly is modeled after a sheet-to-sheet (S2S) spatial reactor consisting of injection and exhaustion ports that are assembled adjacently, which is exemplified in Fig. 2(b). The spatial reactor depicted in the schematic is designed through Ansys SpaceClaim and meshed through Ansys Workbench, which is pictured in Fig. 4. The mesh is produced by 2D triangular cells with the surface of the substrate consisting of 18 nodal regions. A greater discussion of the characteristics of the reactor design and mesh is provided in Yun et al. (2022b), whose work proposed an optimal reactor design for preventing the intermixing of precursor species. Thus, the reactor model is constructed with a 0.25 mm gap distance between the substrate and the injection dividers, a reactor length of 160 mm, and 9 injection ports of 10 mm width each.

The mesh of the reactor geometry has a significant role in determining the precision of the spatial distribution calculations in the numerical simulation, especially for regions situated near the boundary

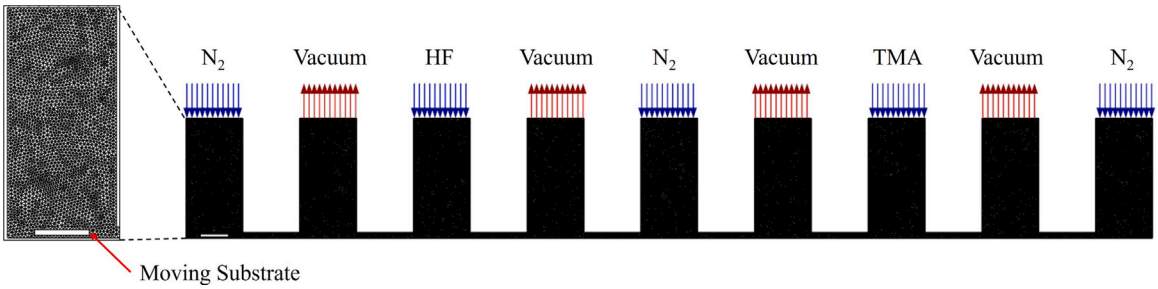

Fig. 4. A 2D lateral view of the dynamic mesh for the spatial reactor design with 0.25 mm gap distance.

Table 2
Definitions of variables used in the macroscopic model.

Variable	Definition
d	Normalized boundary distance
E	Internal energy of the system
$\vec{F}$	External body force
$\vec{g}$	Acceleration due to gravity of Earth
h_j	Sensible enthalpy for species, j
J_j	Diffusion flux of species, j
P	System pressure
S_h	Source term for heat transfer
S_m	Source term for mass transfer
t	Time
$\vec{u}$	Substrate or mesh displacement velocity
$\vec{v}$	Fluid velocity
α	Diffusion parameter
γ	Diffusion coefficient
ρ	Density of the fluid mixture
$\bar{\bar{\tau}}$	Symmetric rank-two stress tensor

conditions. Finer discretized elements or cells produce more accurate calculations with a trade-off of increasing the computational load in comparison to that of coarser meshes. Also, to overcome the possibility of boundary conditions diverging and to reduce inaccuracies, the parameters of the reactor mesh quality, which consist of skewness, orthogonal quality, and aspect ratio, are calculated through Ansys Workbench so that they fall within reasonable criteria ranges as recommended by ANSYS (2021). The mesh resolution, also called the distribution of cells, has a substantial role in the calculations around boundary conditions. A higher resolution generates meshes that are densely composed of smaller-sized cells for better convergence and accuracy.

The computational fluid dynamics (CFD) software, Ansys Fluent, contains a coupled, pressure-based solver with steady-state and transient modes that are used to numerically calculate the momentum, energy, and mass transfer equations, which are described by the following formulas:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \left(\rho \vec{v}\right) = S_m \tag{8}$$

$$\frac{\partial \left(\rho \vec{v}\right)}{\partial t} + \nabla \cdot \left(\rho \vec{v}\vec{v}\right) = -\nabla P + \nabla \cdot \left(\bar{\bar{\tau}}\right) + \rho \vec{g} + \vec{F} \tag{9}$$

$$\frac{\partial}{\partial t}\left(\rho E\right) + \nabla \left(\vec{v}\left(\rho E + P\right)\right) = -\nabla \left(\Sigma h_j J_j\right) + S_h \tag{10}$$

where ρ represents the fluid mixture density, $\vec{v}$ is the fluid mixture velocity, S_m is the mass transfer source generation or consumption term, P is the static pressure, $\bar{\bar{\tau}}$ is the symmetric rank two stress tensor, $\rho\vec{g}$ is the gravitational body force of the fluid, $\vec{F}$ is the external body force acting on the fluid, E is the internal energy of the system, h_j is the sensible enthalpy of the species j, J_j is the diffusion flux of species j, and S_h is the heat transfer source generation or consumption term. A summary of the complete list of variables and their definitions is defined in Table 2.

To simulate the dynamic substrate in the spatial ALE reactor configuration, a dynamic mesh is applied. Diffusion-based smoothing and remeshing are two methods that are used to update the mesh at each time step while maintaining the mesh quality and calculation accuracy on the surface boundary conditions caused by the motion of the substrate. When the substrate moves, the diffusion-based smoothing method modifies the original mesh of the previous time step with the following equation:

$$\nabla \cdot \left(\gamma \nabla \vec{u}\right) = 0 \tag{11}$$

where γ is the diffusion coefficient and $\vec{u}$ is the mesh displacement velocity. This method adjusts the mesh based on the boundary distances, which effectively restores the mesh quality around the surface boundary layer of the moving substrate. The boundary distance-based diffusion coefficient is calculated with the following equation:

$$\gamma = \frac{1}{d^{\alpha}}, \quad \alpha \in [0, 2] \tag{12}$$

where d is the normalized boundary distance and α is the diffusion parameter, which is defined with a value of 1.5 in this work. The remeshing process is another mesh adjustment method that is used in conjunction with diffusion-based smoothing, which updates the surroundings of the dynamic mesh to maintain an acceptable mesh skewness criterion.

The specified boundary conditions include a no-slip, laminar surface boundary layer condition on the surfaces of the substrate and the walls of the reactor. The inlet precursor flow is characterized by homogeneous flow as the precursor compound is assumed to be perfectly mixed before being introduced into the reactor at a constant temperature of 573 K. The operating temperature inside the reactor is defined to be 573 K and the reactor wall is defined to have zero heat flux. Also, the operating pressure of the reactor is assumed to remain constant at 300 Pa.

One of the assumptions made about the reactor model is that the operating conditions of the reactor at time $t = 0$ are at steady state such that all precursor and purge gases are being continuously fed until a steady-state operating condition is reached. A steady-state simulation without substrate movement is conducted initially to simulate the steady-state operation environment. Following the steady-state solver, the dynamic model is then simulated with a transient, coupled, pressure-based solver with a constant substrate velocity specified with a time step size of 5×10^{-4} s with a maximum of 200 iterations possible for each time step. To make the interconnection between macroscopic and microscopic domains, source terms for each species are specified through user-defined functions (UDFs). From the CFD simulation, area-averaged surface pressures of the species are collected and transferred to the kMC model, where the kMC model calculates the source terms. The generation and consumption source terms are calculated based on the fraction of lattice sites that are modified or etched, which are computed by microscopic simulation for a number of time steps during the time evolution. A complete description of the architecture of this interconnection between the macroscopic and microscopic domains is discussed in Section 2.3.

Remark 2.2. The presented macroscopic model uses a first-order numerical transient solver method (the finite element method) that simultaneously calculates the evolution of time and space-dependent parameters. Numerical error is generated as a consequence of the spatial and temporal discretizations. In addition, the macroscopic model utilizes a simpler two-dimensional (2D) mesh as opposed to a three-dimensional (3D) mesh to reduce computational demand and simulation time, which may influence the calculation of the area-averaged precursor pressures, which are spatially dependent. Thus, the macroscopic model contributes to the intrinsic variability of the coverage and etching fractions as well as the etching per cycle. As such, the definitions of the boundary conditions as well as the specification of the solver methods and mesh will contribute to the overall modeling error.

2.3. Two-dimensional multiscale computational fluid dynamics model structure

The integration of the microscopic substrate surface domain and the macroscopic gas-phase domain establishes the multiscale computational fluid dynamics (CFD) model that simulates the evolution of spatial–temporal dynamics and microscale reaction kinetics, simultaneously. Previous research has been dedicated to two-dimensional (2D) and three-dimensional (3D) multiscale CFD simulations for atomic layer processes (Zhang et al., 2019; Ding et al., 2019; Zhang et al., 2020; Yun et al., 2021); however, the aforementioned works were limited in the scope of combining the microscopic and macroscopic domains such that the consumption and generation of species in the macroscopic domain were not specified as a boundary condition for the mass transfer source term. Yun et al. (2022c) were able to improve the connection between the microscopic and macroscopic domains by specifying a constant consumption term while neglecting the generation of byproducts of the precursor species as a surface boundary condition on the substrate but was limited in integrating a precursor pressure-dependent consumption term due to the fineness of the substrate mesh. Recently, Yun et al. (2022b) were able to construct a 2D multiscale CFD model that provides the precursor pressure-dependent consumption and byproduct generation terms to provide an accurate pressure profile of the macroscopic fluid domain. The aforementioned 2D multiscale CFD model is presented in Fig. 5, which describes the sequential path taken to calculate the pressures of byproduct, inert, and precursor species from the macroscopic CFD model and export the pressures to the kMC microscopic model to calculate the updated source generation, consumption, and time evolution terms in the simulation loop for the macroscopic CFD model, which are specified by user-defined functions (UDFs) in Ansys Fluent.

A complete discussion of the multiscale results as well as the impact of operating conditions (substrate velocity, precursor and inert gas flow rates, vacuum pressure) and the reactor design (gap distance) on the etching rate and intermixing of species, respectively, are elaborated by Yun et al. (2022b).

The connection between the microscopic and macroscopic domains is achieved by coupling Ansys Fluent's CFD modeling software with a Python kMC programming script with a Linux Bash Shell script with 36 compute cores and 192 GB of random access memory (RAM). This present work will continue to use the 2D multiscale model in Yun et al. (2022b) to continue the simulation work with varying input parameters to generate a diverse data set, which is desired for achieving an accurate and robust regression model and this model will serve as the foundation for the run-to-run control system for the spatial ALE reactor.

3. Run-to-run control system

In practice, there are many external factors that can influence the quality of the wafers produced in a fabrication plant. From process drifts caused by equipment fatigue to noise caused by variance and other process disturbances, identifying the sources of error and implementing corrections are time-consuming and costly. A more industrially relevant solution is to develop control systems that can make adjustments that compensate for these changes and minimize statistical variation in between runs. Due to the spontaneous nature of the half-cycle thermal ALE reactions, there are challenges associated with establishing real-time process control; however, ex situ process controllers have emerged to resolve the latter issue.

Run-to-run (R2R) or batch-to-batch controllers are an example of such control systems and are commonly used to improve closed-loop stability and decrease process variability by using data from the previous run to modify the inputs ex situ. The adjustments made to the inputs between successive batch runs depend on the method used. The exponentially weighted moving average (EWMA) method interpolates between two bias terms, a weighted bias computed from the previous batch run and a bias evaluated from a linear regression model, using a deterministic weight that requires extensive process knowledge. The limitations of the EWMA algorithm, such that the method is constrained to linear regression models and requires a user-defined weight, can be resolved by adopting an artificial neural network algorithm that is applicable for nonlinear systems and does not require the specification of a weight parameter. Therefore, in this work, two control architectures are employed in conjunction with the multiscale CFD model discussed in Section 2.3: the exponentially weighted moving average (EWMA) based R2R controller and the artificial neural network (ANN) based R2R controller. Both methods provide greater insight into controller selection for linear and nonlinear systems and raise questions about the influence of weights on the controller response to disturbances.

The architectures of the EWMA-based and ANN-based R2R controllers are illustrated in Fig. 6. The EWMA-based R2R controller adjusts the inputs based on an exponentially weighted bias term, while the ANN-based R2R controller modifies the inputs based on the deviation of the output from the target value. The output, which is the etching per cycle, EPC, is typically measured by the amount of mass loss in real-time using a sensitive weighing equipment such as a Quartz Crystal Microbalance (QCM). However, for the spatial reactor configuration, the QCM cannot be used to measure mass loss in real-time while the substrate is in motion. The QCM is instead used to measure the amount of mass loss off-line (outside the spatial reactor environment and when the substrate is stationary) in order to perform the necessary correction to the inputs. The input parameters consist of the precursor (HF and TMA) flow rates and the substrate velocity, which are manipulated using valves and an actuator, respectively. It is also assumed that all input parameters are constant for the entire duration of the batch run. A summary of the list of variables and their definitions in this section is provided in Table 3. This section will be organized first by discussing the EWMA-based R2R controller and followed by an analysis of the ANN-based R2R controller.

3.1. EWMA-based R2R controller

The EWMA-based R2R controller is a statistical process control method that is used widely for run-to-run control. An EWMA-based R2R control system operates with the aid of a linearized model by adjusting a process input based on the output of prior batch runs. The controller manipulates each input by monitoring changes in the output from the setpoint specification. This modification of the inputs is conducted by purposefully favoring newer process run data over older batch data to monitor more recent changes in the process environment through a weighing parameter. The algorithm used to update the inputs can be determined analytically, which reduces the need for numerical solvers,

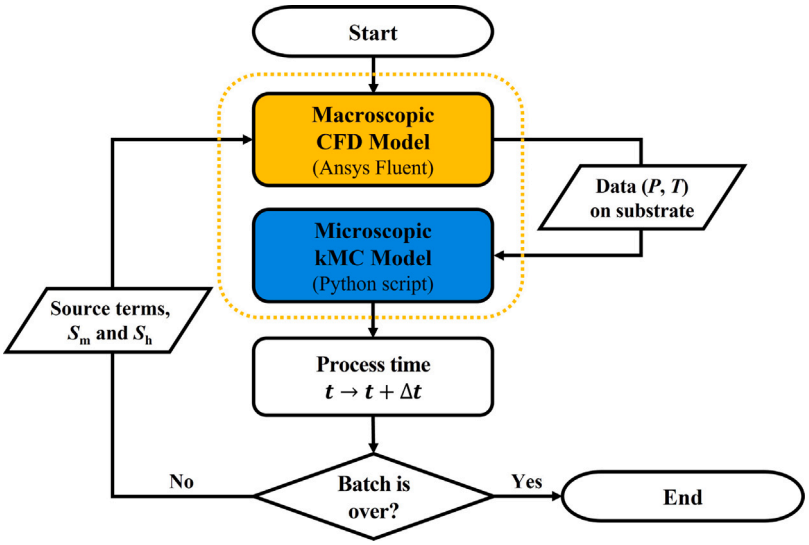

Fig. 5. A process diagram depicting the multiscale CFD model that connects the macroscopic and microscopic domains to simultaneously compute pressure and temperature spatially in the macroscopic domain and reaction kinetics in the microscopic domain through each kMC time update. The connection of the macroscopic and microscopic models is made possible through a Linux Bash Shell script.

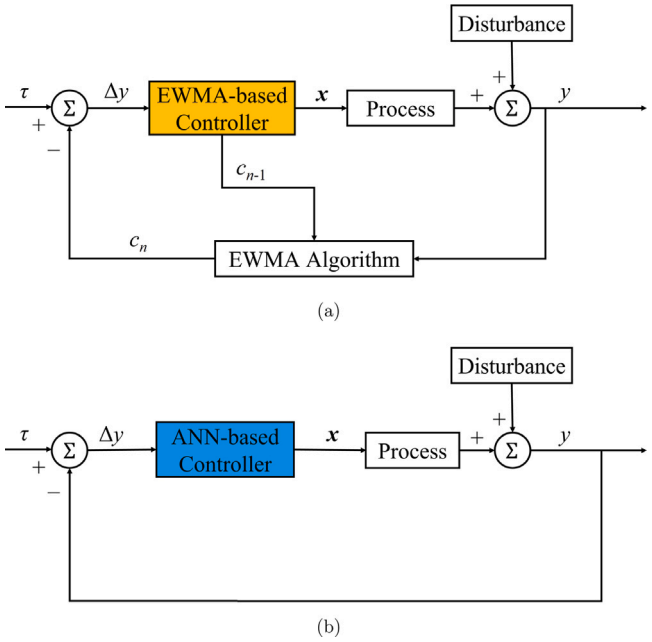

Fig. 6. Schematic R2R control diagrams of an (a) EWMA-based controller and an (b) ANN-based controller. The "process" is the multiscale CFD model, **x** denotes the input vector, y represents the output variable, τ is the setpoint or target value, c is the bias for the previous batch run, $n-1$, and the current batch run, n, and Δy is the deviation of the output computed from the multiscale CFD simulation from the setpoint, τ.

Table 3
Definitions of variables used in the R2R control system.

Variable	Definition
$\mathbf{a}$	Vector of coefficients for the linear regression model
$\mathbf{b}$	Bias for the input layer to the hidden layer
c_n	Intercept or bias of the EWMA model for the nth batch number
$\mathbf{E}$	Vector of residuals
f	Activation function for the ANN
$\mathbf{I}$	Square identity matrix
$\mathbf{J}$	Column vector of ones
l	Number of input parameters
L	Lagrange multiplier function $L(\mathbf{x}_n, \boldsymbol{\Lambda})$ for the nth batch number
m	Total number of batch runs
MSE	Mean square error
n	Batch run number
p	Total number of neurons in the hidden layer
S	Residual sum of squares
$\mathbf{w}_1$	Vector of weight of neurons from input layer to the hidden layer
$\mathbf{w}_2$	Vector of weight of neurons from hidden layer to the output layer
$\mathbf{x}$	Vector of the inputs
$\mathbf{X}$	Aggregate vector of the offline input data
y	Output computed by the multiscale CFD model
$\hat{y}$	Predicted output computed by the multiscale CFD model
$\mathbf{Y}$	Aggregate vector of the offline output data
z	Sample size of data points
$\Delta\mathbf{x}$	Input deviation
Δy	Output deviation
ε	Bias for output layer
λ	EWMA weight factor
Λ	Lagrange multiplier
$\boldsymbol{\sigma}$	Gradient vector of the input-hidden layer activation function
τ	Etching per cycle target or setpoint

and is simple to integrate into R2R control systems. However, the selection of the fixed weight parameter is a tedious task that requires a diverse data set to determine the optimal weighing parameter that reduces the output variability generated by oscillatory input adjustments and that offers output stability around the setpoint. In addition, the EWMA weight is constrained to particular disturbances based on their impact on the product conformity. Also, the EWMA-based R2R controller is a linear control system; thus, it is of paramount importance to ensure the data regression is linear.

Previous work has been conducted in the analysis of EWMA-based R2R controllers for atomic layer deposition and etching processes (Crose et al., 2017, 2019; Ding et al., 2020a; Zhang et al., 2020; Yun et al., 2021, 2022d); however, the aforementioned works have not conducted R2R control with an advanced multiscale CFD model that simulates spatio-temporally in conjunction with time and pressure-dependent surface kinetics. Also, prior works have not investigated multivariate input and single input R2R control systems. This research aims to construct an EWMA-based R2R controller for a previously developed multiscale CFD model (Yun et al., 2022b) for spatial thermal atomic layer etching using a multiple-input-single-output (MISO) system that consists of three input variables, precursor (HF and TMA) flow rates and substrate velocity, and one output variable, etching per cycle (EPC). This section on the EWMA-based R2R controller formulation is organized by discussing the MISO multiple linear regression development followed by an overview of the EWMA method and lastly, the

procedures of the input model update are elucidated and the analytical derivation is provided in detail.

3.1.1. MISO linear regression model

The EWMA method utilizes a linear model that relates the input arguments to the output term. A linear regression is desired for the EWMA model, which updates the bias or intercept of the linear regression equation through each batch run. It is notable that the EWMA-based controller is restricted to linear regression models for updating the bias term through successive batch runs and for reducing computational demand. The use of a linear regression model also is beneficial for obtaining explicit, analytical solutions for updating the inputs for the next batch run without requiring tedious numerical approximation methods such as the Newton–Raphson Method to evaluate the updated inputs. For the EWMA-based R2R controller, a multiple linear regression model will be developed to ensure an analytical solution is obtainable for the input update.

For multiple linear regression fitting, a generic linear equation is defined as follows:

$$\hat{y}_n = \mathbf{a}^T \mathbf{x}_n + c_n \tag{13}$$

where $\hat{y} \in \mathbb{R}$ is the scalar-valued output that is evaluated by the linear regression model for batch number n, $\mathbf{a} \in \mathbb{R}^l$ is the coefficient vector $\begin{bmatrix} a_1 & a_2 & \cdots & a_l \end{bmatrix}^T$ that is independent of the batch run and is therefore constant, l is the number of input variables, $\mathbf{x}_n \in \mathbb{R}^l$ is the input vector $\begin{bmatrix} x_{1,n} & x_{2,n} & \cdots & x_{l,n} \end{bmatrix}^T$ at a batch run of n, and $c_n \in \mathbb{R}$ is the intercept or bias for a batch run of n. It is notable that the initial regression model produced from offline data is for $n = 0$, thus the initial bias determined by multiple regression modeling will be designated as c_0. The output, $\hat{y}$ of the regression model is the etching per cycle (EPC) in Å/cycle. Also, in this regression model, a total of $l = 3$ scalar inputs in the input vector, $\mathbf{x}_n$ will be used and are constrained to be in the following ranges:

$$x_{1,n} \in [20, 100]\,,\ x_{2,n} \in [20, 100]\,,\ x_{3,n} \in [10, 100]$$

where $x_{1,n}$ is the HF flow rate in sccm, $x_{2,n}$ is the TMA flow rate in sccm, and $x_{3,n}$ is the substrate velocity in mm/s for a batch number of n. This regression model is composed of multiple inputs and a single output, which establishes the so-called multiple-input-single-output (MISO) system. For the remainder of this work, the convention of subscript n is used to denote the batch number. All vectors of the form $\mathbb{R}^l$ or $\mathbb{R}^p$ are column vectors unless otherwise noted. Vectors that have a superscript of T, for instance, $\mathbf{a}^T$ in Eq. (13), refers to the transpose of the vector.

The generation of the linear regression model utilizes a sample of $z = 270$ data points that are collected for a combination of inputs, which are defined in the multiscale CFD model to compute each successive output. The outputs are evaluated offline for HF flow rates, $x_{1,n}$, from 20 to 100 sccm in intervals of 20 sccm, TMA flow rates, $x_{2,n}$, from 20 to 100 sccm in intervals of 10 sccm, and substrate velocities from 10 to 100 mm/s in intervals of 20 mm/s for the range of 20 to 100 mm/s. Results of the multiscale CFD simulation data set are presented in Fig. 7 in iso-contours of the HF flow rate of 20 sccm, TMA flow rate of 40 sccm, and substrate velocity of 80 mm/s, which illustrates the effects of these manipulated variables on the etching per cycle.

With the data collection defined, the linear regression model is optimized using the least squares linear multiple regression method that determines the ideal regression that minimizes the distance or residual between the predicted outputs evaluated from the regression model and the actual outputs evaluated from the multiscale CFD model to construct a best fitting regression line. In other words, it is desired to select the vector of coefficients, $\mathbf{a}$, such that $\mathbf{E}$, which represents the magnitude of the deviation of the linear model prediction from the multiscale simulation data, is minimized. Hence, the following minimization problem on the sum of squares of the data residuals is introduced:

$$\min \|\mathbf{E}\|^2 \tag{14a}$$

$$\text{s.t.}\quad \mathbf{E} = \mathbf{Y} - \mathbf{X}\mathbf{a} - c_0\mathbf{J} \tag{14b}$$

where $\|\cdot\|$ is the l_2-norm, $\mathbf{E} \in \mathbb{R}^z$ is the residual data sample vector $\begin{bmatrix} e_1 & e_2 & \cdots & e_z \end{bmatrix}^T$, $\mathbf{Y} \in \mathbb{R}^z$ is the data sample output vector $\begin{bmatrix} y_1 & y_2 & \cdots & y_z \end{bmatrix}^T$ computed by the multiscale CFD simulation, $\mathbf{X} \in \mathbb{R}^{z\times l}$ is the data sample input vector,

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1l} \\ x_{21} & x_{22} & \cdots & x_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ x_{z1} & x_{z2} & \cdots & x_{zl} \end{bmatrix}$$

where l is the number of inputs, $\mathbf{a} \in \mathbb{R}^l$ is the coefficient vector $\begin{bmatrix} a_1 & a_2 & \cdots & a_l \end{bmatrix}^T$, c_0 is the initial bias or intercept for batch $n = 0$, and $\mathbf{J} \in \mathbb{R}^z$ is a column vector of z ones $[1 \ \ 1 \ \cdots \ 1]^T$. The sum of the squares of the residuals, $\|\mathbf{E}\|^2$ is expanded and then minimized to determine the coefficient vector of the inputs, $\mathbf{a}$, and the initial intercept, c_0, which are as follows:

$$\mathbf{a} = 10^{-3}\begin{bmatrix} 0.0121 & 0.346 & -1.84 \end{bmatrix}^T$$

$$c_0 = 0.478$$

The coefficient matrix, $\mathbf{a}$, and the intercept, c_0, are substituted into Eq. (13) to produce the multiple linear regression model of the offline data:

$$\hat{y} = 10^{-3}\left(0.0121x_{1,n} + 0.346x_{2,n} - 1.84x_{3,n} + 478\right) \tag{15}$$

The accuracy of the linear regression model, Eq. (15) is presented in Fig. 8 in iso-contours of HF flow rate of 20 sccm, TMA flow rate of 40 sccm, and substrate velocity of 80 mm/s, which illustrates the magnitude of the deviation of the linear regression model from the multiscale CFD results in Fig. 8. The linear regression model has a mean squared error (MSE) that is computed using Eq. (23) and is determined to have an MSE of 4.236×10^{-4} Å/cycle, which suggests that the deviation of 270 data points estimated by the linear regression model from the multiscale CFD model simulation results is low. Fig. 8(b) reveals the possibility that the linear model strongly deviates from the multiscale CFD data results in regions of high substrate velocity and low HF flow rate. Fig. 8(a), which is presented by an iso-contour of HF flow rate at 20 sccm, confirms the observation that lower HF flow rates produce greater deviation from the multiscale CFD data results. Generally, lower TMA flow rates in Fig. 8(c) produce greater deviation from the multiscale CFD data results. Thus, the linear regression model produces a regression that accurately reflects the trend of the actual multiscale CFD data results; however, one may deduce that there is disagreement between the predicted and multiscale CFD model results for lower TMA flow rates and higher substrate velocities, which may affect the reliability of the results and input update. To overcome this issue, a large domain for each input that is previously defined in this section was purposefully chosen in the event that large input changes are required to correct major disturbances; however, the formulation in Section 3.1.3 highlights a methodological approach to ensure that input changes between successive batch runs are minimized. Also, a greater discussion on the limitations of the disturbances for EWMA-based R2R controllers will be discussed in Section 4.

3.1.2. Bias update using the EWMA method

Disturbances in atomic layer processes are attributed to several factors such as continual equipment fatigue and byproduct formation during the reaction half-cycles. As a result, the measured outputs of a process become dependent on the results of the prior run and some of these characteristics are inherited in the subsequent batch run. Although older data is useful, newer data would need to be weighed more to produce a greater influence on the upcoming batch run. To resolve this issue, the exponentially weighted moving average (EWMA) statistical model is integrated into the R2R controller. An advantageous property of the EWMA approach is through the application of a weight parameter that continuously decreases exponentially as the data ages.

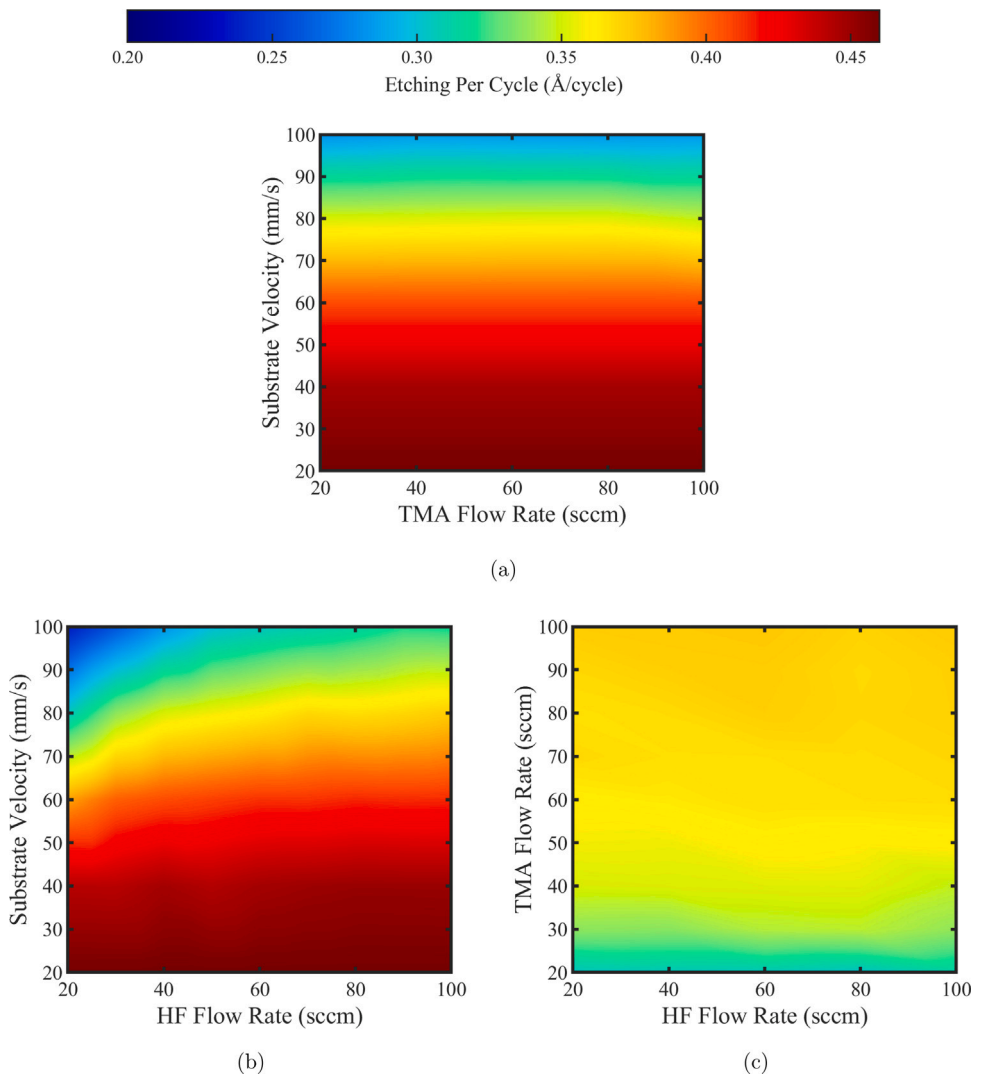

Fig. 7. Multiscale CFD simulation results of the etching per cycle represented by iso-contours of (a) HF flow rate of 20 sccm, (b) TMA flow rate of 40 sccm, and (c) substrate velocity of 80 mm/s.

However, the selection of the fixed weight parameter is of utmost importance as the magnitude of the weight parameter dictates the amount of noise generated by modifying the inputs, which concurrently affects the measured output. Process control becomes susceptible to the overadjustment of process inputs as well as the measured output deviating greatly from the setpoint subject to larger weights. Although self-tuning methods have been proposed to modify weights ex situ (Del Castillo and Hurwitz, 1997; Da et al., 2002; Su and Hsu, 2004), these proposals are computationally demanding for the multiscale CFD simulation and require an abundance of data to understand the relationship between the EWMA weight and the level of stability and product conformance that is achieved for particular disturbances. Thus, this work aims to study the influence of the weight on the stability of the input adjustment and the recorded output for EWMA-based R2R controller models that have the same controller architecture through two weight parameters.

One of the disadvantages of the EWMA statistical approach is that the data is constrained to linearized regression models. For a nonlinear input–output relationship, the severity of the nonlinearity may result in non-smooth control actions and greater deviations from the setpoint. Several areas of research are dedicated to resolving the challenges associated with nonlinear processes. For instance, Del Castillo and Yeh (1998) developed an optimizing adaptive quality controller (OAQC) using quadratic-based models to achieve comparable controller response with linear-based controllers. Yun et al. (2022d) proposed a normalization method to improve linearity using a sigmoidal-like function to achieve stronger linear regression fitting for a linear EWMA-based controller. Artificial neural networks (ANN) based R2R controllers (Wang and Mahajan, 1996) have also emerged to regress nonlinear systems. Nonetheless, a linear regression model will characterize this EWMA-based controller to reduce computational demand and to facilitate the process for evaluating the analytical solution for multiple inputs.

The EWMA-based controller model is formulated by weighing the bias term, c_n, in Eq. (13). The next bias term is weighed based on the prior weighted batch run bias, c_{n-1}, and the computed bias from the linear regression model in Eq. (13) upon the rearrangement of terms to solve for c_n. The resulting EWMA expression for calculating the updated bias, c_n is described as follows:

$$c_n = \lambda \left(y_n - \mathbf{a}^T \mathbf{x}_{n-1} \right) + (1 - \lambda)\, c_{n-1}, \quad \text{where} \quad \lambda \in (0, 1] \tag{16}$$

where $c_n \in \mathbb{R}$ is the weighted bias term at batch run n, $y_n \in \mathbb{R}$ is the output of the multiscale CFD model corresponding to the input vector, $\mathbf{x}_{n-1}$, $\mathbf{a} \in \mathbb{R}^l$ is the coefficient vector $\begin{bmatrix} a_1 & a_2 & \cdots & a_l \end{bmatrix}^T$ of the linear regression model from Eq. (13), l is the number of inputs, $\mathbf{x}_{n-1} \in \mathbb{R}^l$ is the input vector $\begin{bmatrix} x_{1,n-1} & x_{2,n-1} & \cdots & x_{l,n-1} \end{bmatrix}^T$, and $\lambda \in \mathbb{R}$ is the weighting coefficient. The EWMA of the bias term, c_n, which is updated across consecutive batch runs, has a substantial role in the calculation of the

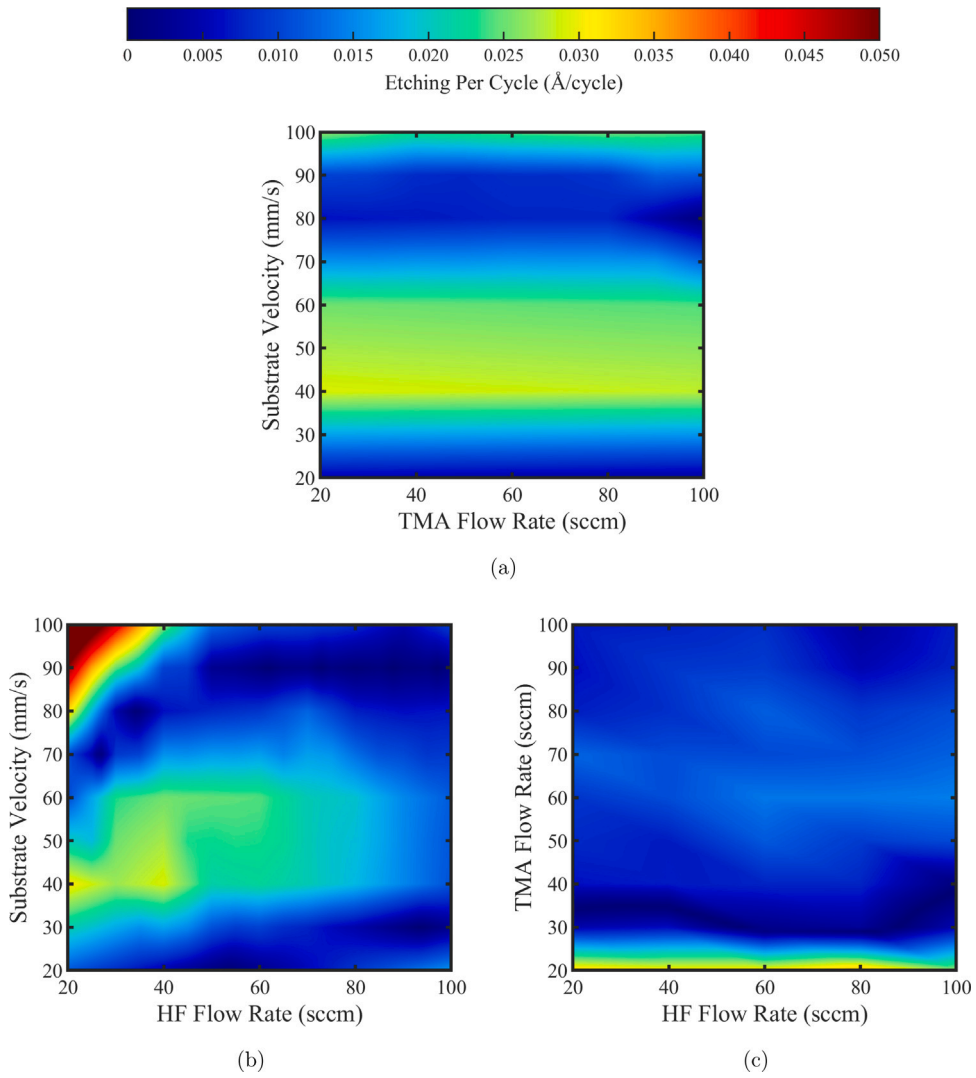

Fig. 8. Linear regression model etching per cycle deviation from the multiscale CFD data in Fig. 7 represented by iso-contours of (a) HF flow rate of 20 sccm, (b) TMA flow rate of 40 sccm, and (c) substrate velocity of 80 mm/s. The MSE of the linear regression model is 4.236×10^{-4} Å/cycle.

inputs of the next batch through the EWMA-based R2R control system, which is discussed in Section 3.1.3.

The EWMA equation, Eq. (16), can also be expanded into a non-recursive form (Montgomery, 2013) by replacing c_{n-1} with the equation for the previous iteration and repeating this until all instances of c have been replaced with y and $\mathbf{x}$ on the right-hand side of the equation. Then, c_n will be a function of only λ, y, and $\mathbf{x}$ with the following form:

$$c_n = \sum_{i=0}^{n-1} \left[(1-\lambda)^i \lambda \left(y_{n-i} - \mathbf{a}^T \mathbf{x}_i \right) + (1-\lambda)^n c_0 \right], \quad \text{where} \quad \lambda \in (0, 1] \tag{17}$$

where n represents the current batch number. The non-recursive EWMA equation illustrates that $(1-\lambda) < 1$ and that the older biases are assigned exponentially smaller weights in comparison to the more recent biases, which have larger weights. This preferential weighing of more recent data allows the controller to adapt to more recent changes in the process while maintaining some knowledge of the older batch runs. To understand the effects of the EWMA weight, two weights will be specified to the same EWMA-based R2R controller and run in parallel to determine the differences in the adjustments made to the inputs and their effect on the output. For this work, a lower weight of $\lambda = 0.3$ and a larger weight of $\lambda = 0.7$ will be specified for two separate simulations for the same EWMA-based R2R controller that are run in parallel simulations. The following section will describe the implementation of the bias update for an underdetermined system of linear equations to determine the ideal input update.

3.1.3. Input update of EMWA

The established linear model in Eq. (15) illustrates the need for a formulation to determine an optimal input update such that the change to the inputs to overcome the effects of the disturbance is minimized as discussed in Section 3.1.1 while ensuring that the resulting update will produce an output that is proximal to the setpoint. This particular linear system is characterized by an underdetermined system that consists of more dependent variables (3) than equations (1). As a result, a minimum least squares summation method is proposed on the deviation of the next batch, n, and previous batch, $n-1$, inputs while constrained by the new input resulting in the desired setpoint, as follows:

$$\min_{\mathbf{a}^T \mathbf{x}_n = \beta_n} \left\| \mathbf{x}_n - \mathbf{x}_{n-1} \right\|^2 \tag{18a}$$

$$\text{s.t.} \quad \beta_n = \tau - c_n \tag{18b}$$

where $\|\cdot\|$ is the l_2-norm, $\mathbf{a} \in \mathbb{R}^l$ is the vector $\begin{bmatrix} a_1 & a_2 & \cdots & a_l \end{bmatrix}^T$ of coefficients for the inputs, l is the number of inputs, $\mathbf{x}_n \in \mathbb{R}^l$ is the vector $\begin{bmatrix} x_{1,n} & x_{2,n} & \cdots & x_{l,n} \end{bmatrix}^T$ for the following batch run n, $\mathbf{x}_{n-1} \in \mathbb{R}^l$ is the vector $\begin{bmatrix} x_{1,n-1} & x_{2,n-1} & \cdots & x_{l,n-1} \end{bmatrix}^T$ for the preceding batch run $n-1$, $\tau \in \mathbb{R}$ is the setpoint or target of the output, and $c_n \in \mathbb{R}$ is the exponentially

weighted bias evaluated from Eq. (16). This minimization problem is then solved by utilizing a Lagrange function, $L(\mathbf{x}_n, \Lambda)$, defined as

$$L(\mathbf{x}_n, \Lambda) = \left(\mathbf{x}_n - \mathbf{x}_{n-1}\right)^T \left(\mathbf{x}_n - \mathbf{x}_{n-1}\right) + \Lambda(\mathbf{a}^T\mathbf{x}_n - \beta_n) \tag{19a}$$

where Λ is a Lagrange multiplier. First, the partial derivatives of L with respect to $\mathbf{x}_n$ and Λ are taken and equated to zero to yield, respectively, the equations,

$$\nabla_{\mathbf{x}_n} L = 2\left(\mathbf{x}_n - \mathbf{x}_{n-1}\right) + \Lambda\mathbf{a} = 0 \tag{19b}$$

$$\nabla_{\Lambda} L = \mathbf{a}^T\mathbf{x}_n - \beta_n = 0 \tag{19c}$$

where the systems of equations are solved to yield the following equation to perform the input vector update, $\mathbf{x}_n$:

$$\mathbf{x}_n = \mathbf{x}_{n-1} - \mathbf{a}\left(\mathbf{a}^T\mathbf{a}\right)^{-1}\left(\mathbf{a}^T\mathbf{x}_{n-1} - \beta_n\right) \tag{19d}$$

Eq. (19d) is the fundamental and analytical solution for an underdetermined linear system that will be integrated with the input modification process. The procedural steps for conducting the EWMA-based R2R controller input update are summarized as follows:

1. A linear regression model of the general form in Eq. (13) is constructed with offline data to determine the vector of input coefficients, **a**, and the initial bias, c_0.
2. The EPC output from the multiscale CFD simulation is evaluated and substituted into Eq. (16) to determine an exponentially weighted bias, c_n, for the upcoming batch run, n.
3. The upcoming weighted bias, c_n, is substituted into β_n in Eq. (18b) to calculate the vector of input parameters in Eq. (19d), $\mathbf{x}_n$, for the next batch run n.
4. Steps 2 and 3 are repeated for m number of batch runs where **a** is constant and c_n is updated through each batch run.

3.2. ANN-based R2R control

Overcoming nonlinear behavior is challenging as some atomic layer processes have nonlinear relationships (Smith and Boning, 1997). Several methods can be accomplished to integrate nonlinear input–output relationships into R2R control models. For instance, Yun et al. (2022d) were able to linearize a nonlinear data set generated offline with a sigmoidal-like, median-effect equation, which transformed the nonlinear data into an apparent linear model. This section entails a machine learning approach for R2R control of nonlinear systems.

Recently, machine learning (ML) algorithms, in particular artificial neural networks (ANNs), have emerged as promising tools for process modeling, optimization, and control in various engineering fields. Owing to the development of high-performance computing resources, ML has been widely applied in diverse research subjects. For example, ML was used to build a real-time controller (Wu et al., 2021; Alhajeri et al., 2022), to model engineering processes (Ding et al., 2021; Abdullah et al., 2021a,b; Yun et al., 2022a; Abdullah et al., 2022), and to improve an empirical model for an electrochemical reactor (Luo et al., 2022). In particular, ANNs have been extensively employed to lessen the computational demand and to improve the accuracy of model approximation, especially for nonlinear systems. Wang and Mahajan (1996) first suggested using an ANN-based run-to-run (R2R) controller, and their results illustrated that the ANN-based R2R controller had a comparable performance to that of an EWMA-based R2R controller. An ANN model for R2R control as a distinct control algorithm from the conventional EWMA algorithm to determine a new input at every batch run under a disturbance was trained. However, due to a lack of computing resources, the ANN-based filter of the R2R controller was limited to a single neuron.

The ANN-based R2R controller is an alternative solution to the EWMA-based R2R controller, which is limited to linear systems. One of the advantages of the ANN-based R2R controller is that this controller formulation does not require the specification of a deterministic weight factor in the case of the EWMA-based R2R controller. The weight factors are instead obtained through the process of training a feed-forward neural network (FNN) model. Also, ANN-based R2R controllers can operate under nonlinear systems, which are regressed during the data training process and do not require bias updates. After the predictive model is established, the update of the inputs is conducted using an optimization technique to tune the input variables minimally through each successive batch run.

In this work, the framework for the ANN-based R2R controller is adopted from Wang and Mahajan (1996) and the algorithm is then further expanded with more neurons to achieve the best fit model. First, the same data set described in Section 3.1, which was used to formulate a multiple-input-single-output (MISO) regression model for the EWMA-based R2R control, is used to train an ANN model with the lowest mean square error. Then, a numerically approximated input update is constructed in accordance with the trained ANN model, which updates the inputs from the deviation of the output from the target. Finally, the developed ANN-based R2R controller is integrated with the multiscale computational fluid dynamics (CFD) model to deal with disturbances such as shift and drift disturbances. The schematic flow diagram is illustrated in Fig. 6(b), and the following sections elucidate the ANN-based R2R method in greater detail.

3.2.1. ANN model training

Figs. 7 and 8 illustrate the possibility of nonlinearity in the linear regression model for the EWMA-based R2R controller due to slight variations in output prediction for lower TMA flow rates and higher velocities. Thus, a separate ANN-based R2R controller is developed for establishing a nonlinear predictive model for the offline data set. The generation of the nonlinear predictive model first begins with a feed-forward neural network (FNN), which handles nonlinear data efficiently and reliably through a robust process of training and testing the data. Despite FNNs being practical for nonlinear systems, FNNs may come at a cost when defining an optimal number of neurons and hidden layers, which will affect the accuracy of the predictive model and may promote the curse of dimensionality or overfitting. Also, the FNN structure also depends on the type of activation function, which will decide the usefulness of particular data to the predictive model. The specification of the delegation of data for training and testing will also have a substantial role in the accuracy of the model. However, despite the aforementioned disadvantages in using deterministic specifications to characterize the FNN, the testing and training of the predictive model resolves the latter to ensure the reliability of the results. The training and testing of this FNN are conducted through TensorFlow's Keras, an established application programming interface (API) through a Python script.

The ANN-based R2R control model comprises three types of layers, the input, hidden, and output layers, which are illustrated in Fig. 9. A single hidden layer is specified to prevent the overfitting and mapping of multiple inputs to a single output, which is expressed as follows:

$$\hat{y} = F(\mathbf{x}) = f\left(\mathbf{w}_2{}^T\mathbf{h} + \varepsilon\right) \quad \text{where} \quad \mathbf{h} = f\left(\mathbf{w}_1\mathbf{x} + \mathbf{b}\right) \tag{20a}$$

$$= f\left(\mathbf{w}_2{}^T \cdot f\left(\mathbf{w}_1\mathbf{x} + \mathbf{b}\right) + \varepsilon\right) \tag{20b}$$

where $\hat{y} \in \mathbb{R}$ is the predicted output computed by the ANN regression model, F is the artificial neural network mapping function of the input vector, f is the activation function, $\mathbf{w}_2 \in \mathbb{R}^p$ is the weight vector $\begin{bmatrix} w_{1,2} & w_{2,2} & \cdots & w_{p,2} \end{bmatrix}^T$ from the hidden layer to the output layer, p is the total number of neurons in the hidden layer, $\varepsilon \in \mathbb{R}$ is the bias for the output $\hat{y}$, $\mathbf{x} \in \mathbb{R}^l$ is the input vector $\begin{bmatrix} x_1 & x_2 & \cdots & x_l \end{bmatrix}^T$, l is the number of input variables, $\mathbf{w}_1 \in \mathbb{R}^{p\times l}$ is the weight matrix

$$\mathbf{w}_1 = \begin{bmatrix} w_{11,1} & w_{12,1} & \cdots & w_{1l,1} \\ w_{21,1} & w_{22,1} & \cdots & w_{2l,1} \\ \vdots & \vdots & \ddots & \vdots \\ w_{p1,1} & w_{p2,1} & \cdots & w_{pl,1} \end{bmatrix} = \begin{bmatrix} \mathbf{w}_{1,1}^T \\ \hline \mathbf{w}_{2,1}^T \\ \hline \vdots \\ \hline \mathbf{w}_{p,1}^T \end{bmatrix} \tag{21}$$

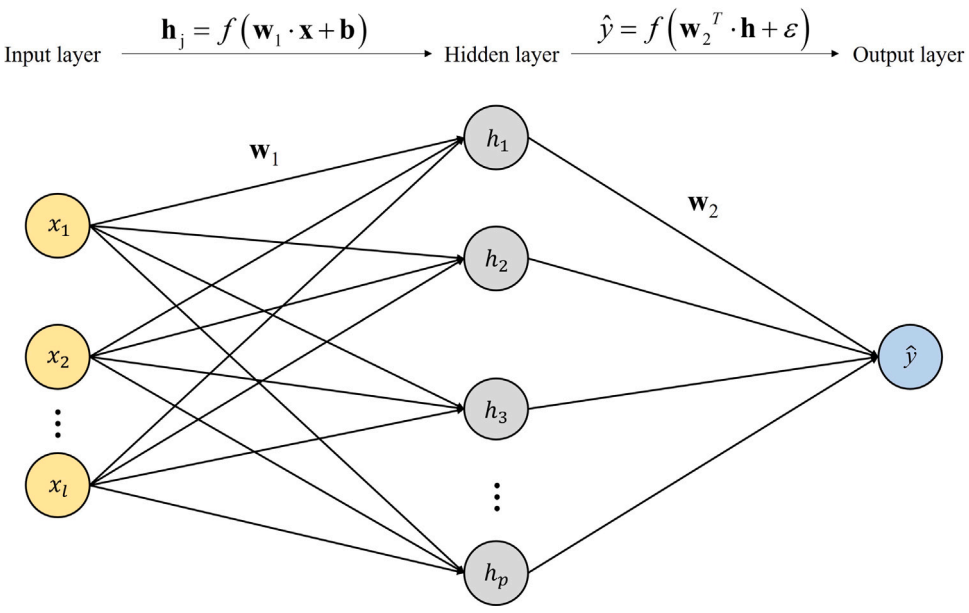

Fig. 9. An artificial neural network architecture consisting of an input, output, and hidden layer. In this work, the input layer has $l = 3$ inputs (x_1, x_2, and x_3), the hidden layer is defined with $p = 20$ neurons (h_j for $j = 1, 2, \ldots, p$), and the output layer consists of one output ($\hat{y}$).

from the input layer to the hidden layer, and $\mathbf{b} \in \mathbb{R}^p$ is the bias vector $\begin{bmatrix} b_1 & b_2 & \cdots & b_p \end{bmatrix}^T$ from the input layer to the hidden layer. The weight matrix, $\mathbf{w}_1$ can be divided into p row vectors where $\mathbf{w}_{i,1}^T \in \mathbb{R}^l$ for the space $\begin{bmatrix} w_{1i,1} & w_{2i,1} & \cdots & w_{li,1} \end{bmatrix}$ for $i = 1, 2, \ldots, p$ is the ith row vector. To validate the trained ANN model, $z = 270$ data points collected offline from the multiscale CFD model are divided into a training set consisting of 80% and a testing set of 20% of the total data points. The input layer has 3 input nodes ($l = 3$), which are the HF flow rate (x_1) in sccm, TMA flow rate (x_2) in sccm, and substrate velocity (x_3) in mm/s. The hidden layer consists of $p = 20$ neurons in which interconnect computations are processed. The output layer has one output node, $\hat{y}$, which is the etching per cycle (EPC) in Å/cycle. The activation function, f, will use the rectified linear unit (ReLU), which facilitates the process for updating the input in Section 3.2.2 and is generalized as follows:

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases} \tag{22}$$

Thus, the activation function f passes the summed weight-adjusted input variables to yield the value of each neuron in the hidden layer. Likewise, the activation function f also passes the summed weight-adjusted node variables in the hidden layer to calculate the output in the same manner.

The ANN model is trained using 20 neurons in the hidden layer until a low mean square error (MSE) is achieved. The MSE is expressed as

$$MSE = \frac{1}{z}\sum_{j=1}^{z}(y - \hat{y})^2 \tag{23}$$

where z is the number of offline data points, which is 270 points, $y \in \mathbb{R}$ is the reference output variable obtained from the multiscale CFD simulation, and $\hat{y} \in \mathbb{R}$ is the predicted output variable from the trained ANN model. From the FNN training, a MSE of 2.251×10^{-4} Å/cycle is observed, which indicates that the average deviation of the predicted ANN results from the actual multiscale CFD results is low in magnitude. From the low MSE value, it is demonstrated that the ANN model predicts the trends of the multiscale CFD results reasonably and improves the regression from the linear model in Fig. 8. The predicted results of the ANN model are presented in Fig. 10 with predicted EPC results being similar to the multiscale CFD EPC results in Fig. 7

3.2.2. Input update from ANN

An artificial neural network (ANN) method has been extensively used to predict an output based on a trained data set due to its ability to tolerate nonlinear input–output relationships. However, Wang and Mahajan (1996) devised an approach to adopt an ANN model for run-to-run (R2R) control using the process sensitivity of the output deviation, which is the derivative of the mapping function, $F(\mathbf{x}) = \hat{y}$ with respect to the inputs, $\mathbf{x}_n$, at batch run of n. To navigate around the nonlinearity of the predicted model, a first-order Taylor Series expansion about the previous input vector, $\mathbf{x}_{n-1}$ on the mapping function, $F(\mathbf{x}) = \hat{y}$, in Eq. (20b) is used as follows:

$$\hat{y}_n = \hat{y}_{n-1} + \nabla_{\mathbf{x}_n}\hat{y}\Big|_{x_n = x_{n-1}} \cdot \left(\mathbf{x}_n - \mathbf{x}_{n-1}\right) \quad \text{where} \quad \hat{y}_n = \tau,\ \hat{y}_{n-1} = y_{n-1} \tag{24a}$$

$$y_{n-1} - \tau = \Delta y = \nabla_{\mathbf{x}_n}\hat{y}\Big|_{x_n = x_{n-1}} \cdot \left(\mathbf{x}_{n-1} - \mathbf{x}_n\right) = \hat{y}\Big|_{x_n = x_{n-1}} \Delta\mathbf{x} \tag{24b}$$

where $\Delta y = y_{n-1} - \tau,\ \Delta\mathbf{x} = \mathbf{x}_{n-1} - \mathbf{x}_n$

where $y_n \in \mathbb{R}$ and $y_{n-1} \in \mathbb{R}$ are the outputs evaluated by the multiscale CFD simulation for the next batch run, n, and the preceding batch run $n-1$, respectively, $\hat{y}_n \in \mathbb{R}$ and $\hat{y}_{n-1} \in \mathbb{R}$ are the predicted outputs evaluated by the ANN model for the next batch run, n, and the preceding batch run $n-1$, respectively, $\mathbf{x}_n \in \mathbb{R}^l$ is the updated input vector $\begin{bmatrix} x_{1,n} & x_{2,n} & \cdots & x_{l,n} \end{bmatrix}^T$ for the next batch run n, $\mathbf{x}_{n-1} \in \mathbb{R}^l$ is the preceding input vector $\begin{bmatrix} x_{1,n-1} & x_{2,n-1} & \cdots & x_{l,n-1} \end{bmatrix}^T$ for the previous batch run $n-1$, l is the number of inputs, τ is the setpoint or target value of the output, Δy is the output deviation term from the setpoint, and Δx is the input deviation term. Eq. (24b) sets $y_n = \tau$ as a constraint by assuming that the input update is sufficient to produce the setpoint for the following batch run. For the ReLU activation function in Eq. (22), $\nabla_{\mathbf{x}_n}\hat{y}_{n-1}\Big|_{x_n = x_{n-1}}$ equates to the following:

$$\nabla_{\mathbf{x}_n}\hat{y}\Big|_{x_n = x_{n-1}} = \nabla_{\mathbf{x}_n} f\left(\mathbf{w}_2{}^T \cdot f\left(\mathbf{w}_1\mathbf{x}_n + \mathbf{b}\right) + \varepsilon\right)\Big|_{x_n = x_{n-1}} \tag{25a}$$

$$= \nabla_{\mathbf{x}_n} f\left(\mathbf{w}_1\mathbf{x}_n + \mathbf{b}\right)\Big|_{x_n = x_{n-1}} \cdot \nabla_{\mathbf{x}_n} f\left(\mathbf{w}_2{}^T\mathbf{h} + \varepsilon\right)\Big|_{\mathbf{h} = f(\mathbf{w}_1\mathbf{x}_{n-1} + \mathbf{b})} \tag{25b}$$

The input term for the hidden to the output layer must always be greater than zero in order to generate a predictive model; thus, the derivative of the hidden-output layer ReLU function must be equal to $\mathbf{w}_2$.

$$\nabla_{\mathbf{x}_n}\hat{y}\Big|_{x_n = x_{n-1}} = \boldsymbol{\sigma}^T \cdot \mathbf{w}_2 \quad \text{where} \quad \boldsymbol{\sigma}^T = \nabla_{\mathbf{x}_n} f\left(\mathbf{w}_1\mathbf{x}_n + \mathbf{b}\right)\Big|_{x_n = x_{n-1}} \tag{25c}$$

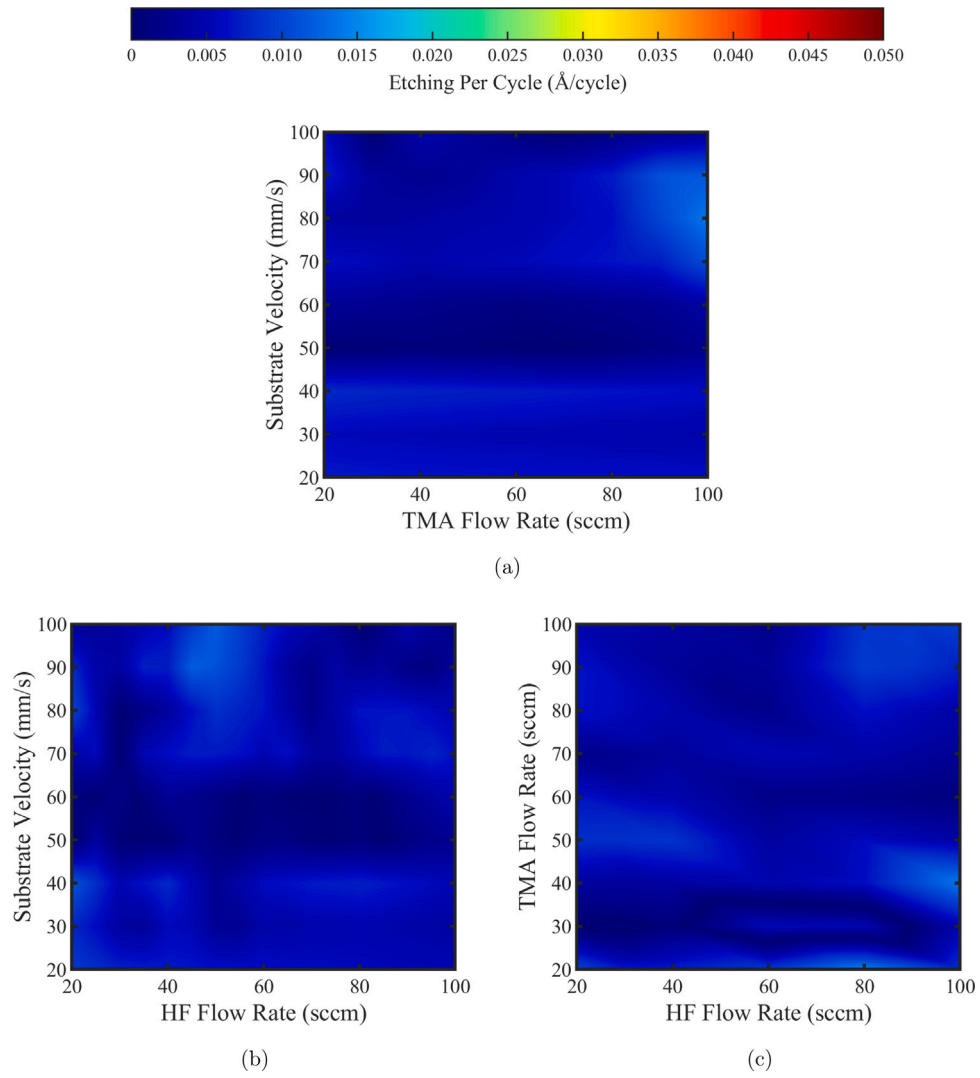

Fig. 10. Artificial neural network model etching per cycle deviation from the multiscale CFD data in Fig. 7 represented by iso-contours of (a) HF flow rate of 20 sccm, (b) TMA flow rate of 40 sccm, and (c) substrate velocity of 80 mm/s. The MSE of the ANN model is 2.251×10^{-4} Å/cycle.

The input term for the hidden to the output layer must always be greater than zero in order to generate a predictive model; thus, the derivative of the hidden-output layer ReLU function must be equal to $\mathbf{w}_2$. $\boldsymbol{\sigma} \in \mathbb{R}^{p \times l}$ is the derivative matrix of the activation function from the input to the hidden layer and is defined as the following:

$$\boldsymbol{\sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1l} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pl} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\sigma}_1^{\,T} \\ \hline \boldsymbol{\sigma}_2^{\,T} \\ \hline \vdots \\ \hline \boldsymbol{\sigma}_p^{\,T} \end{bmatrix}$$

computed from $\nabla_{\mathbf{x}_n} f\left(\mathbf{w}_1 \mathbf{x}_n + \mathbf{b}\right)\Big|_{x_n = x_{n-1}}$ since the activation function, f, is ReLU; thus, the ith row vector, $\boldsymbol{\sigma}_i^{\,T} \in \mathbb{R}^{1 \times l}$ for the space $\begin{bmatrix} \sigma_{1i} & \sigma_{2i} & \cdots & \sigma_{li} \end{bmatrix}$ is calculated by:

$$\boldsymbol{\sigma}_i^{\,T} = \begin{cases} \mathbf{0} & \text{if} \quad \mathbf{w}_{i,1}^{\,T} \cdot \mathbf{x}_{n-1} + \mathbf{b} < 0 \\ \mathbf{w}_{i,1}^{\,T} & \text{if} \quad \mathbf{w}_{i,1}^{\,T} \cdot \mathbf{x}_{n-1} + \mathbf{b} \geq 0 \end{cases} \quad \text{for} \quad i = 1, 2, \ldots, p \tag{26}$$

where p is the number of the neurons in the hidden layer, $\mathbf{0} \in \mathbb{R}^l$ is a vector of zeros, and $\mathbf{w}_{1,i}^{\,T} \in \mathbb{R}^l$ is the ith row vector of the weight matrix, $\mathbf{w}_1$, as shown in Eq. (21). Eq. (25) represents the process sensitivity of the output deviation and depends on the weights computed by the FNN where $\mathbf{w}_1 \in \mathbb{R}^{p \times l}$ and $\mathbf{w}_2 \in \mathbb{R}^p$. From Eq. (24b), it is desired to find $\mathbf{x}_n$ that has the setpoint value as an output.

Following the first-order Taylor Series expansion of the predicted model, a minimal input update is needed to ensure that deviations from the standard operating conditions are minimized. This optimization is conducted using the projection theorem. Essentially, a projection of the coordinate input axes $\left(x_1, x_2, \ldots, x_l\right)$ in the $\mathbb{R}^{l+1}$ Euclidean space for l inputs is made onto the solution plane generated in Eq. (24b), which calculates the deviation of the next input, $\mathbf{x}_n$, from the prior input, $\mathbf{x}_{n-1}$ and is expressed as follows:

$$\Delta \mathbf{x} = \left(\nabla_{\mathbf{x}_n} \mathbf{x}_n\right)^T \nabla_{\mathbf{x}_n} \Delta y\Big|_{x_n = x_{n-1}} \left[\left(\nabla_{\mathbf{x}_n} \Delta y\Big|_{x_n = x_{n-1}}\right)^T \nabla_{\mathbf{x}_n} \Delta y\Big|_{x_n = x_{n-1}}\right]^{-1} \Delta y \tag{27a}$$

where $\nabla_{\mathbf{x}_n} \Delta y \in \mathbb{R}^l$ represents the unit vector that is orthogonal to the solution plane, Δy, and $\nabla_{\mathbf{x}_n} \mathbf{x}_n$ is the unit vector of the coordinate axes of the inputs. The gradient of the coordinate axes $\left(\nabla_{\mathbf{x}_n} \mathbf{x}_n\right)^T \in \mathbb{R}^{l \times l}$ produces a square identity matrix, $\mathbf{I} \in \mathbb{R}^{l \times l}$. Combining Eqs. (25c) and (27a) yields the following expression for updating the inputs for the next batch run:

$$\mathbf{x}_n = \mathbf{x}_{n-1} - \mathbf{I}\left(\boldsymbol{\sigma}^T \mathbf{w}_2\right)\left[\left(\boldsymbol{\sigma}^T \mathbf{w}_2\right)^T \left(\boldsymbol{\sigma}^T \mathbf{w}_2\right)\right]^{-1} \left(y_{n-1} - \tau\right) \tag{27b}$$

where $y_{n-1} \in \mathbb{R}$ is the output computed by the multiscale CFD simulation from the previous batch run $n-1$, τ is the setpoint value for

the output, $\mathbf{w}_1 \in \mathbb{R}^{p\times l}$ is the weight vector from the input layer to the hidden layer in the FNN, $\mathbf{w}_2 \in \mathbb{R}^p$ is the weight vector from the hidden layer to the output layer in the FNN, p is the number of neurons in the hidden layer, l is the number of inputs, $\mathbf{x}_n \in \mathbb{R}^l$ is the updated input vector for the next batch n, and $\mathbf{x}_{n-1} \in \mathbb{R}^l$ is the previous input vector for batch run $n-1$. Also, the unit vector of Δy in Eq. (24b) is equal to the gradient of $\hat{y}$ in Eq. (25c) with respect to $\mathbf{x}_n$ at $\mathbf{x}_{n-1}$. Eq. (27b) is an optimal procedure for computing the update for the next inputs. It is notable that the coefficient of $y_{n-1}-\tau$ in Eq. (27b) represents the process sensitivity made to the adjustment of the input vector. The magnitude of the adjustment is strictly related to the weights of the neurons computed by the ANN model and does not require a specification of a deterministic weight, namely, in the case of the EWMA-based R2R controller. In summary, the input parameter update for the ANN-based R2R controller is as follows:

1. A collection of offline data is separated into two groups delegated to the training and testing of the FNN model in Eq. (20b), which is defined by a particular number of neurons, hidden layers, and the activation function.
2. Step 1 is repeated until a low MSE is evaluated in Eq. (23) is observed to produce a predictive model.
3. A first-order Taylor Series expansion is performed on the predictive model to determine an equation that relates the deviation of inputs from the previous and next batch runs in Eq. (24b).
4. The projection theorem in Eq. (27b) is used to evaluate the input update for the next batch run.
5. Steps 3 and 4 are repeated for m total number of batch runs.

4. Multiscale CFD and R2R controller performance analysis

The semiconductor manufacturing processes are routinely subjected to occasional shifts and deterministic or nondeterministic drifts, which have an impact on the product quality. The loss of product quality is costly compared to the costs of an established control system. Therefore, in general, run-to-run (R2R) control is used to regulate process variables to maintain product quality under disturbances including variability, shift, and drift (Moyne et al., 2018). Variability of the results are attributed to stochastic behavior if the input parameters are not specified correctly and to the definitions made on the boundary and mesh in the numerical CFD solver. Sudden shifts or offsets are attributed to a maintenance operation or a change of production specifications. Furthermore, equipment aging effects or systematic deterioration cause equipment-related drifts, leading to a deviation from the conformal output quality. Therefore, an appropriate controller is crucial to achieve an ideal product quality. In this work, three different external disturbances, referred to as "kinetic" disturbances, due to the disturbances being multiplied by reaction rate constants, are considered to evaluate the developed control strategies: mild shift, severe shift, and nondeterministic drift. The shift and drift disturbances are constrained to particular situations in order to adopt the EWMA and ANN-based R2R algorithms for the R2R control system such that the process cannot drift rapidly and that the process can experience occasional drifts (Del Castillo and Hurwitz, 1997). Thus, two shifts are introduced by multiplying shift factors to the reaction rate constants in the microscopic model to produce a time lag in the surface kinetics. A drift is introduced by multiplying the reaction rate constant with a drift factor, resulting in a gradual reduction in the reaction rate for each batch. In general, drifts are attributed to the aforementioned immeasurable external influences, such as equipment aging, which slow down the process over time from a kinetics perspective. A kinetic disturbance, analogous to the defined shifts, can be regarded as a drift that steadily reduces the reaction rate from batch to batch. As a result, the selected disturbances serve as excellent examples of process disturbances in industry.

4.1. Intrinsic variance

The multiscale CFD model simulates the thermal ALE reactions using a stochastic approach that resembles the occurrence of reaction kinetics in the real world. However, this simulation may generate intrinsic variance, a principle that chemical processes produce some variability as a consequence of this stochastic property. For the multiscale CFD model, there are two sources of intrinsic variance: different solutions to the CFD reactor configuration and the stochastic behavior of the kMC simulation. With the two sources of variance stemming from these connected macroscopic and microscopic models, the sources of error can combine to yield results that fluctuate if the boundary conditions and input parameters are not specified carefully. When comparing two otherwise identical runs, the natural variance in the CFD calculations will affect the precursor concentrations on the surface of the substrate, which will subsequently affect the reaction rate calculated by the kMC model. This reaction rate is also specified by the user-defined function (UDF) substrate surface boundary condition in the CFD simulation, which is a source of error that is retained through each time step. Additionally, the simplicity of the CFD model is caused by the first-order numerical simulation method that can compute parameters that deviate from one another. As a result, the calculated precursor pressures from the CFD simulation, which are used to calculate the time evolution, coverage and etching fractions, etching per cycle (EPC), and the surface consumption and generation rates will also fluctuate. In this manner, the two models interact, which can cause each simulation run to have differing results.

In order to guarantee the acceptability of the computed results from the multiscale CFD simulation, an error calculation is needed to measure the variability and stochastic behavior of the EPC results. This stochastic behavior is attributed to the numerical approach to solve the governing equations, which are shown in Eqs. (8) through (10). The equations are solved using numerical methods such as the finite element method for the spatial and temporal discretization. Inevitably, accompanied numerical errors are generated within the defined convergence criteria, resulting in the bounded (but small) stochastic behavior that we observe in our calculations. To quantify the numerical error of the calculations from the multiscale CFD model is tedious; therefore, a statistical analysis is performed. The multiscale CFD model was simulated with constant reactor operation settings to collect a sample size of multiscale CFD results for constant input conditions to determine the percentage of data points that deviate from the mean value of the sample size. A collection of 50 sample multiscale CFD computed EPC data results were collected with a HF flow rate of 20 sccm, TMA flow rate of 40 sccm, and substrate velocity of 80 mm/s without the influence of any kinetic disturbances. Histograms of the aggregate results of the coverage fraction, etching fraction, and EPC are presented in Fig. 11, which illustrates the number of data points that deviate in factors of standard deviation, σ, from the mean value, μ, of the sample size.

Results from Fig. 11 show an asymmetric distribution of data results, which illustrates the stochastic behavior of both the kMC random number generator and the numerical method of the CFD simulation. All histograms produce a single outlier point that is more than 3 standard deviations away from the mean value. Despite the inherent stochastic behavior, 70% to 74% of the 50 data points fall within 1 standard deviation of the mean value; thus, these data points exemplify statistical significance and illustrate the consistency of the results produced from the multiscale CFD model. Therefore, the effects of intrinsic variation on the multiscale CFD model are negligible in comparison to the effects of any shift or drift disturbances and will not influence the response of the run-to-run controller.

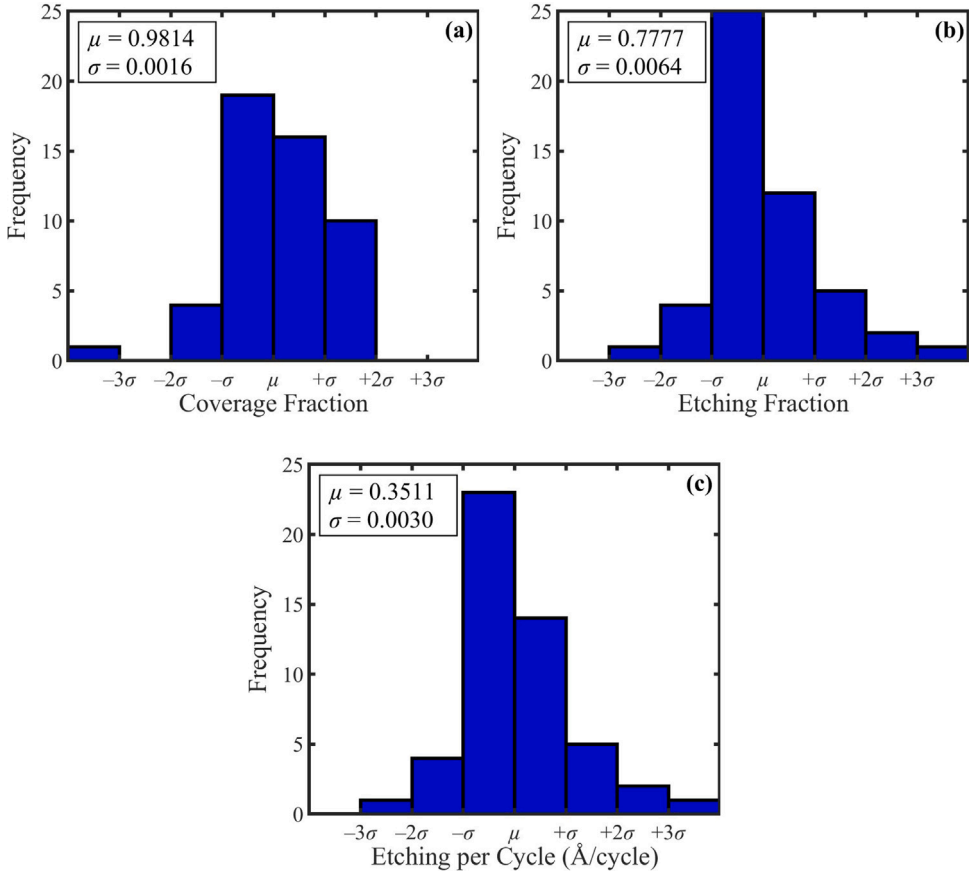

Fig. 11. Histograms depicting the distribution of the coverage fraction (a), etching fraction (b), and etching per cycle (c) for a sample size of 50 multiscale CFD data points for 20 sccm HF flow rate, 40 sccm TMA flow rate, and 80 mm/s substrate velocity without the influence of disturbances. Each bar represents a standard deviation, σ, from the mean, μ, of the data set.

4.2. Shift disturbances

As mentioned in Section 3, R2R controllers equipped with a MISO (multiple-input-single-output) model and an ANN (artificial neural network) model are simulated under Shift-1 (smaller offset) and Shift-2 (larger offset). In addition, the EWMA (exponentially weighted moving average) based R2R controller is simulated with two different weights of 0.3 and 0.7. The inputs and output changes in accordance with the controllers under Shift-1 are outlined in Fig. 12. The result with the higher EMWA weight ($\lambda = 0.7$) displays large overshoots and undershoots in the output response and causes the output to oscillate rigorously around the target, but the oscillation is moderated after the fourth batch number. In addition, the result of the higher EMWA weight ($\lambda = 0.7$) exhibits a larger noise, in contrast to the results of both the lower EMWA weight ($\lambda = 0.3$) and the ANN-based R2R controller, despite approaching the target line as shown in Fig. 12(d). It is observed that the EWMA-based R2R controller with a lower EWMA weight demonstrates a better controller response than that with a higher one such that there is no severe overshoot or undershoot. However, the R2R controller with the lower EWMA weight displays some variance, but the variance is relatively small around the target. Meanwhile, the ANN-based R2R demonstrates the most robust controller response among the controllers since it reaches the target in lesser batch runs and generates minimal variance. To quantitatively evaluate the controller performance, the mean square error function is employed, which is expressed as follows:

$$MSE = \frac{1}{z}\sum_{k=1}^{z}\left(y_k - \tau\right)^2 \tag{28}$$

where MSE denotes the mean square error, z is the number of data samples, y_k is the output of the sample k, and τ is the output setpoint or target. The mean square errors for the three controllers with various disturbances are summarized in Table 4. The MSE values of the EMWA weight of 0.3 and 0.7, and the ANN are 1.22×10^{-5} Å^2/cycle^2, 1.12×10^{-4} Å^2/cycle^2, and 6.21×10^{-6} Å^2/cycle^2, respectively. It is concluded that the higher EMWA weight offsets exceedingly from the setpoint against a small shift disturbance and produces a larger amount of variance. On the other hand, the ANN-based controller outperforms under Shift-1 by producing a robust response within 2 batch runs while minimizing variance. This initial response is reflected by the predictive model to compensate for the nonlinear behavior in Fig. 7, which illustrates that the EPC has a greater dependency on the HF and TMA flow rates in contrast to the linear regression model presented in Fig. 8. The compensation for the effects of the disturbance using the first-order Taylor Series method suggests that the rate of convergence to an output proximal to the setpoint occurs much faster than that of the EWMA-based method. The progression of input updates through each batch run for the R2R control response is shown in Fig. 12(a) through Fig. 12(c). The simulation results of Shift-2 are illustrated in Fig. 13. The higher EMWA weight ($\lambda = 0.7$), as with the results under the smaller shift (Shift-1), intensively reacts to the disturbance and generates oscillatory behavior in the first few batches. However, the lower EMWA weight ($\lambda = 0.3$) yields a small overshoot at the second batch number and smoothly reaches the target line. It was observed that the results of the ANN-based R2R controller marginally fluctuates around the target and approaches closer to the target line than that of the EWMA-based R2R controllers for both weights. As shown in Table 4, the MSE values of the EMWA weight of 0.3 and 0.7, and the ANN are 1.70×10^{-5} Å^2/cycle^2, 7.46×10^{-4} Å^2/cycle^2, and 9.56×10^{-5} Å^2/cycle^2, respectively. The EWMA-based controller with the lower EWMA weight demonstrates the most robust performance, whereas the

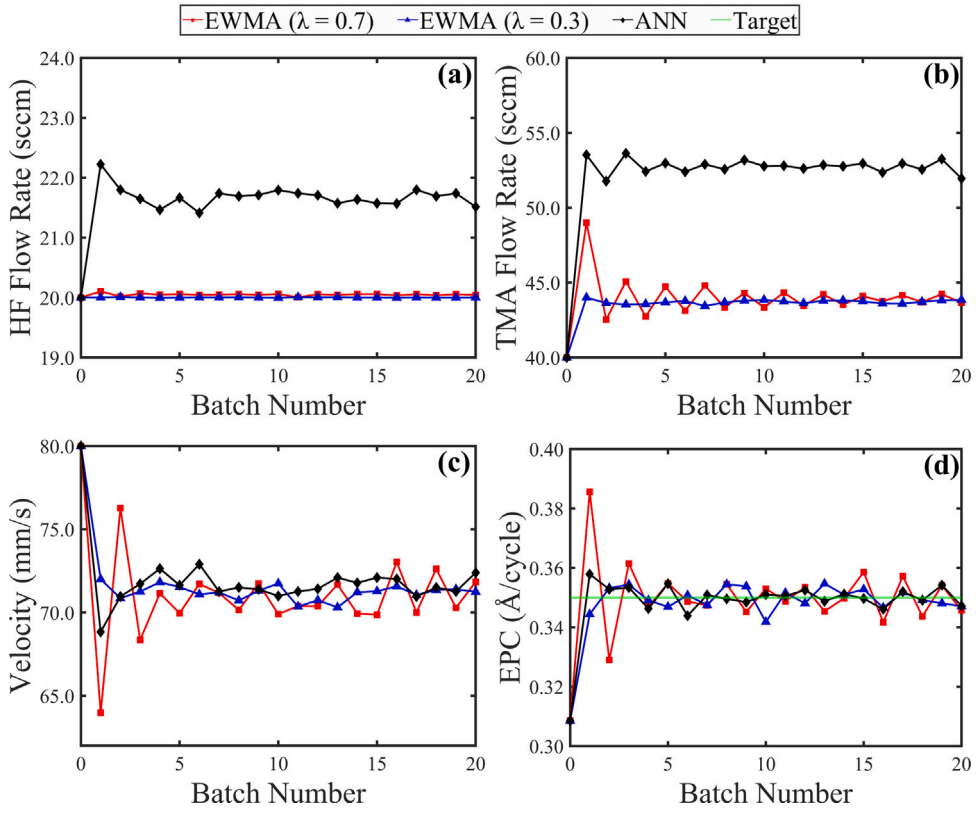

Fig. 12. Graphical illustrations depicting the influence of various R2R controller algorithms on an environment that is introduced to a mild kinetic shift disturbance of 0.8 on the input parameters including the (a) HF flow rate, (b) TMA flow rate, (c) substrate velocity, which are computed from the output variable, (d) etching per cycle (EPC).

ANN-based R2R controller response is comparable to the results of the lower EWMA weight controller.

For both shift disturbances, the EWMA-based R2R controllers for both EWMA weights of 0.3 and 0.7 update the HF and TMA flow rates in a small range. A consequence of the minimization problem in Eq. (18a) limits the magnitude of the input change from the previous batch run, providing substantial control of the process environment. As opposed to the EWMA-based R2R controller, the ANN-based R2R controller makes a larger correction to the precursor flow rates causing a substantial inflation of the standard precursor flow rates compared to that of the adjustments made by the EWMA-based R2R controller. This robust response is made possible by the algorithm used to determine the minimum deviation from the prior run, which is accomplished through a projection problem that is first constructed from a first-order Taylor Series approximation that appears to overestimate the initial response. However, the corrective action to reduce the magnitude of the initial response demonstrates a reduction in the rate of convergence, which illustrates that the ANN-based R2R controller locates convergent solutions in lesser batch runs compared to that of the EWMA-based R2R controllers, which fluctuate greatly with no apparent reduction in the rate of convergence. As pictured in Fig. 8, the EWMA model does not capture the effect of the precursor flow rates such that the EPC for each slice is almost the same. However, given that the ANN model considerably correlates the CFD data set to conform to nonlinear relationships, there is a stronger relationship between the precursor flow rates and the EPC, which is presented in Fig. 10. Therefore, despite the EWMA-based and the ANN-based R2R controllers successfully compensate for both shifts, the lower weight of 0.3 for the EWMA-based R2R controller optimally performs under a severe shift, while the ANN-based R2R controller functions reasonably under a mild shift.

The effects of the spatiotemporal behavior of the production of DMAF when subjected to a shift disturbance of factor 0.60 is illustrated in Fig. 14, respectively, for the sheet-to-sheet (S2S) spatial reactor with and without the integration of a R2R control system at batch number 10. The generation of DMAF is representative of the consumption of AlF_3 surface sites where higher DMAF concentrations are indicative of greater etching rate. Fig. 14(a) illustrates that the integration of the R2R control system produces greater DMAF generation on the substrate surface in contrast to results without a control system in Fig. 14(b), which suggests that the control system increases the etching rate and overcomes the influence of shift disturbances, maintaining the desired etching per cycle rate.

Table 4
Mean square error (Å^2/cycle^2) comparison for various disturbances and R2R control systems.

Algorithm	Shift-1	Shift-2	Drift
EWMA ($\lambda = 0.3$)	1.22×10^{-5}	1.70×10^{-5}	6.18×10^{-5}
EWMA ($\lambda = 0.7$)	1.12×10^{-4}	7.46×10^{-4}	6.47×10^{-5}
ANN	6.21×10^{-6}	9.56×10^{-5}	2.94×10^{-5}

4.3. Drift disturbance

A drift disturbance is also introduced to the microscopic model to monitor the effectiveness of various R2R control algorithms as described in Section 4. The drift disturbance is simulated by multiplying the reaction rate constant by a factor that changes through each batch number. The progression of the factor is calculated with the expression $1 - 0.02 \times n$, where n is the batch number, to provide a gradual reduction of the reaction rate constant in compliance to the suggestions from Del Castillo and Hurwitz (1997). The controller response to the drift disturbance is presented in Fig. 15 where Fig. 15(a) to Fig. 15(c) illustrate the updates to the inputs and Fig. 15(d) shows the output, which is calculated from the multiscale CFD model. From Fig. 15(d), the EWMA-based R2R controller with two different weights and the ANN-based R2R controller situate around the target line and never follow

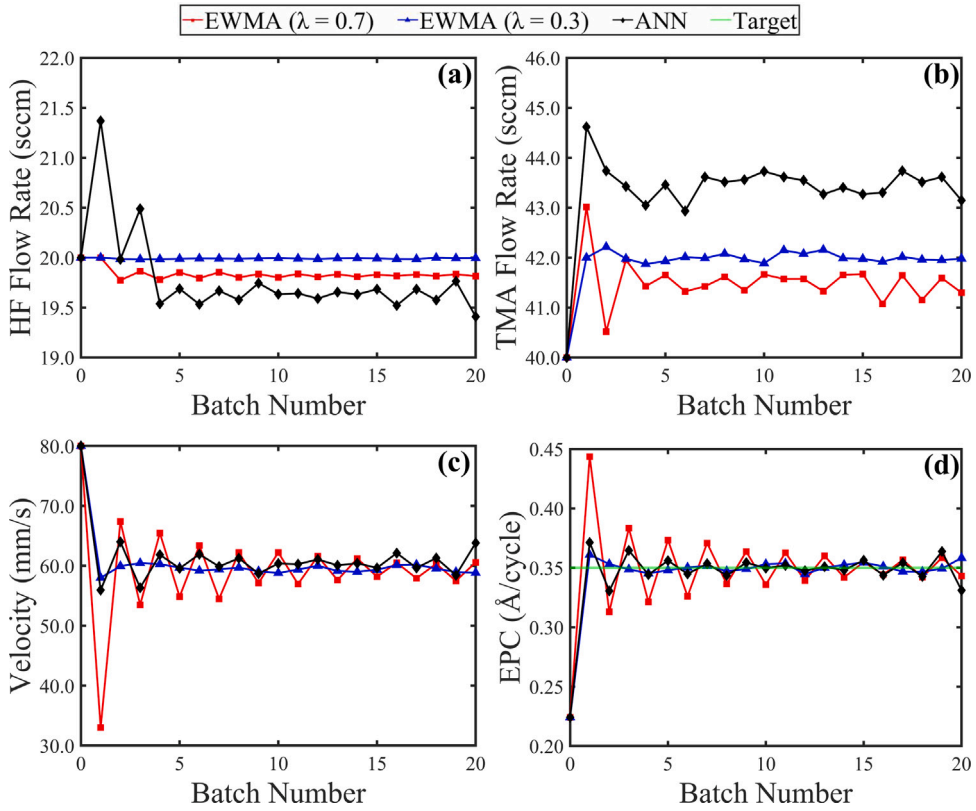

Fig. 13. Graphical illustrations depicting the influence of various R2R controller algorithms on an environment that is introduced to a severe kinetic shift disturbance of 0.6 on the input parameters including the (a) HF flow rate, (b) TMA flow rate, and (c) substrate velocity, which are computed from the output variable (d) etching per cycle (EPC).

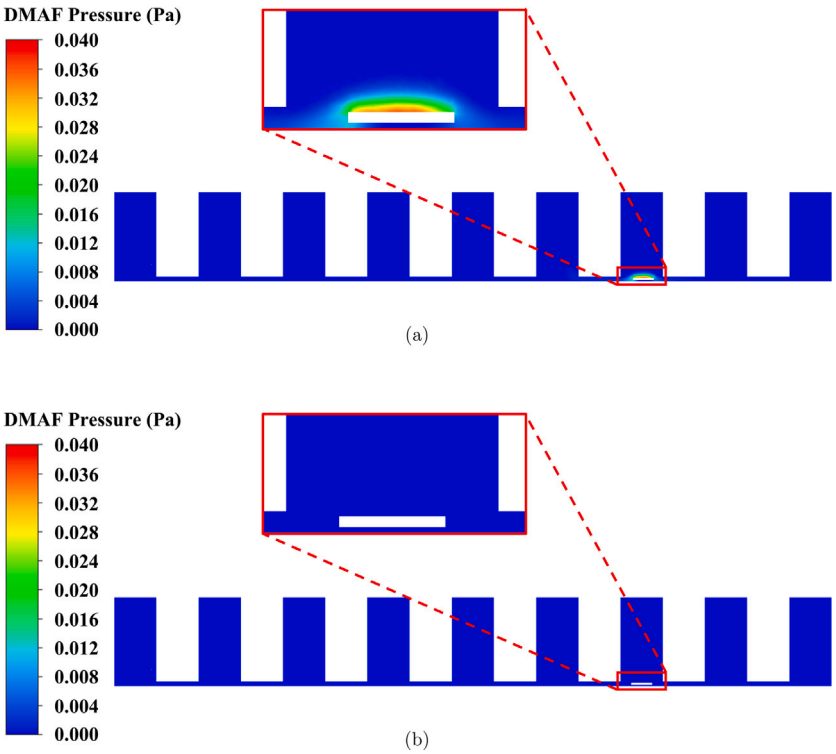

Fig. 14. Multiscale CFD pressure data of the generation of DMAF in the TMA injection region (a) with and (b) without an EWMA-based R2R control system for $\lambda = 0.3$ when subjected to a shift disturbance of 0.6 at the 10th batch run.

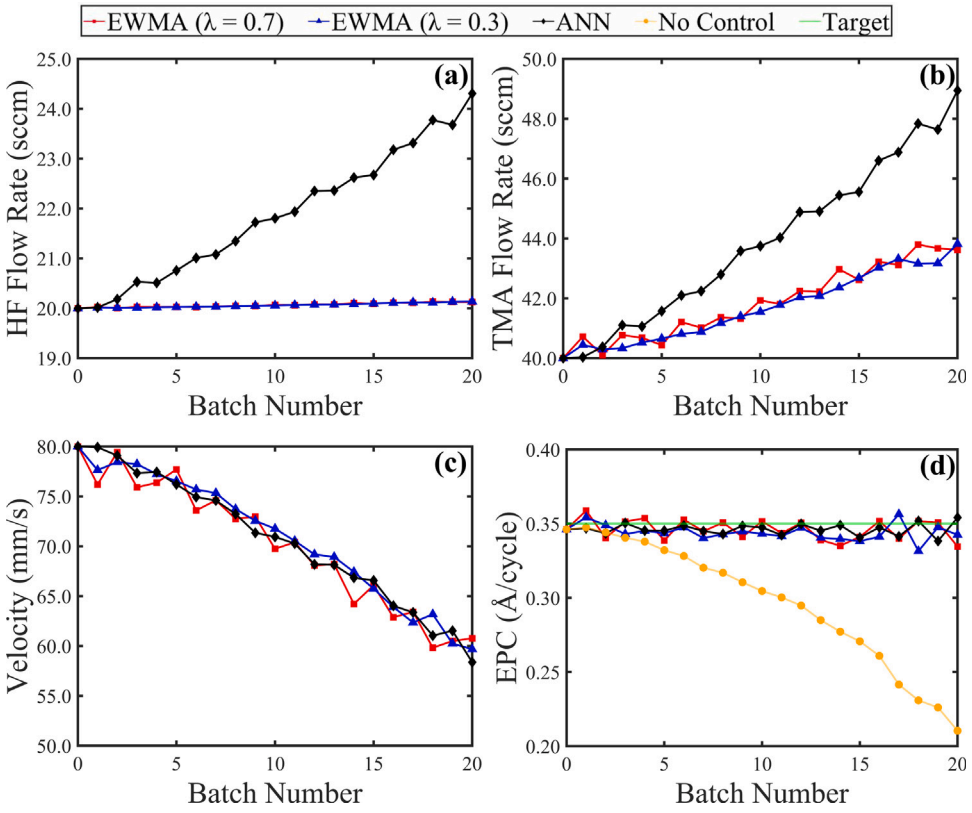

Fig. 15. Graphical illustrations depicting the influence of various R2R controller algorithms on an environment that is introduced to a kinetic drift disturbance on the input parameters including the (a) HF flow rate, (b) TMA flow rate, and (c) substrate velocity, which are computed from the output variable (d) etching per cycle (EPC).

the path of the "no control" line, which indicates that the adjustments made to the input variables maintain the quality conformance for the amount of EPC on the substrate. Table 4 reveals that the ANN-based R2R controller produces the lowest MSE compared to the EWMA-based R2R with the weight of 0.3 and 0.7, which suggests that the ANN-based R2R controller is more robust at reducing fluctuations around the target line. However, the results from Fig. 15(a) and Fig. 15(b) indicate that the ANN-based R2R controller produces greater weights for the precursor flow rates compared to that of the EWMA-based R2R with the weight of 0.3 and 0.7 while the weights for the velocity are relative to the EWMA-based R2R controller with the weight of 0.3. Contrary to the shift disturbance, the deviation of the ANN-based R2R controller inputs for the precursor flow rates from the standard operating conditions expands through each batch run, which is suggestive that minimal changes in the disturbance cause dramatic overshooting caused by the first-order Taylor Series approximation, which will rapidly overestimate the update. From an optimization perspective, the ANN model does not optimize the input parameters but performed the control actions on the basis of the ANN-based model as discussed in Section 4.2.

5. Conclusion

A novel R2R control design was developed using an artificial neural network (ANN) model. This model was trained from the data of a multiscale computational fluid dynamics (CFD) simulation for the spatial thermal atomic layer etching process consisting of one hidden layer with of 20 nodes. The ANN model has three input variables in the input layer, which are the HF and TMA precursor flow rates and the substrate velocity, 20 nodes in the hidden layer, and the output variable in the output layer, which is the etching per cycle (EPC). To evaluate the performance of the ANN-based R2R controller in comparison with existing approaches, an EWMA-based R2R controller, which has been widely used in the semiconductor industry, was simulated under realistic disturbances. The results revealed that the performance of the ANN-based R2R controller was superior to that of the EWMA-based R2R controller. Specifically, a lower EWMA weight worked efficiently under the shift disturbances by minimizing noise and variance but failed to reach the target specification when simulated with a nondeterministic drift disturbance. Meanwhile, the ANN-based R2R controller successively eliminated the effects of mild shift, severe shift, and drift disturbances. In contrast with the EWMA-based R2R controller using a fixed tuning parameter, the ANN-based R2R utilizes a self-determining tuning approach that routinely weights the parameter (i.e., the process sensitivity). This benefit makes the ANN-based R2R controller perform robustly in comparison with the EWMA-based R2R controller that requires the selection of the tuning parameter. Therefore, the capability of the proposed ANN-based R2R controller to various disturbances has been assessed and verified through the multiscale CFD simulation for the spatial atomic layer etching of Al_2O_3. It was also demonstrated that the developed ANN-based R2R controller is readily implementable due to the self-determined tuning unlike the EWMA-based R2R controller.

CRediT authorship contribution statement

Matthew Tom: Conceptualization, Methodology, Software, Writing – original draft. **Sungil Yun:** Conceptualization, Methodology, Software, Writing – original draft. **Henrik Wang:** Methodology, Writing – original draft. **Feiyang Ou:** Methodology, Writing – original draft. **Gerassimos Orkoulas:** Advising, Writing – review & editing. **Panagiotis D. Christofides:** Advising, Writing – review & editing.

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.

Data availability

Data will be made available on request.

Acknowledgments

Financial support from the National Science Foundation is gratefully acknowledged. This work used computational and storage services associated with the Hoffman2 Shared Cluster provided by UCLA Institute for Digital Research and Education's Research Technology Group. The authors would like to thank the comments from Fahim Abdullah and Aisha Alnajdi (University of California, Los Angeles) for their support in the organization and review of the derivation of the algorithms used in this work and Dr. Yi Ming Ren (Meta Platforms) in the development of the machine learning model.

References

Abdullah, F., Wu, Z., Christofides, P.D., 2021a. Data-based reduced-order modeling of nonlinear two-time-scale processes. Chem. Eng. Res. Des. 166, 1–9.

Abdullah, F., Wu, Z., Christofides, P.D., 2021b. Sparse-identification-based model predictive control of nonlinear two-time-scale processes. Comput. Chem. Eng. 153, 107411.

Abdullah, F., Wu, Z., Christofides, P.D., 2022. Handling noisy data in sparse model identification using subsampling and co-teaching. Comput. Chem. Eng. 157, 107628.

Alhajeri, M.S., Luo, J., Wu, Z., Albalawi, F., Christofides, P.D., 2022. Process structure-based recurrent neural network modeling for predictive control: A comparative study. Chem. Eng. Res. Des. 179, 77–89.

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

Berne, B.J., Ciccotti, G., Coker, D.F., 1998. Classical and Quantum Dynamics in Condensed Phase Simulations. World Scientific, London, pp. 385–404.

Broas, M., Kanninen, O., Vuorinen, V., Tilli, M., Paulasto-Kröckel, M., 2017. Chemically stable atomic-layer-deposited Al_2O_3 films for processability. ACS Omega 2, 3390–3398.

Burg, D., Ausubel, J.H., 2021. Moore's law revisited through intel chip density. PLoS One 16, e0256245.

Carrasco, J., Lopez, N., Illas, F., 2004. First principles analysis of the stability and diffusion of oxygen vacancies in metal oxides. Phys. Rev. Lett. 93, 225502.

Crose, M.G., 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.G., 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.

Da, L., Kumar, V., Tay, A., Al Mamun, A., Ho, W.K., See, A., Chan, L., 2002. Run-to-run process control for chemical mechanical polishing in semiconductor manufacturing. In: Proceedings of the IEEE International Symposium on Intelligent Control. Vancouver, Canada, pp. 740–745.

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., Yeh, J.-Y., 1998. An adaptive run-to-run optimizing controller for linear and nonlinear semiconductor processes. IEEE Trans. Semicond. Manuf. 11, 285–295.

Ding, S.-J., Wu, X., 2020. Superior atomic layer deposition technology for amorphous oxide semiconductor thin-film transistor memory devices. Chem. Mater. 32, 1343–1357.

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

Ding, Y., Zhang, Y., Chung, H.Y., Christofides, P.D., 2021. Machine learning-based modeling and operation of plasma-enhanced atomic layer deposition of hafnium oxide thin films. Comput. Chem. Eng. 144, 107148.

Ding, Y., Zhang, Y., Kim, K., Tran, A., Wu, Z., Christofides, P.D., 2019. Microscopic modeling and optimal operation of thermal atomic layer deposition. Chem. Eng. Res. Des. 145, 159–172.

Ding, Y., Zhang, Y., Orkoulas, G., Christofides, P.D., 2020b. Microscopic modeling and optimal operation of plasma enhanced atomic layer deposition. Chem. Eng. Res. Des. 159, 439–454.

Faraz, T., Roozeboom, F., Knoops, H.C.M., Kessels, W.M.M., 2015. Atomic layer etching: what can we learn from atomic layer deposition? ECS J. Solid State Sci. Technol. 4 (6), N5023–N5032.

Fontaine, H., Veillerot, M., Danel, A., 2012. Deposition behavior of volatile acidic contaminants on metallic interconnect surfaces. Mater. Sci. 103–104, 365–368.

Fortunato, E., Barquinha, P., Martins, R., 2012. Oxide semiconductor thin-film transistors: A review of recent advances. Adv. Mater. 24, 2945–2986.

Freeman, D.C., Levy, D.H., Cowdery-Corvan, P.J., 2010. Method for producing compound thin films. US Patent 7, 858, 144 B2.

Fu, K., Fu, Y., Han, P., Zhang, Y., Zhang, R., 2008. Kinetic Monte Carlo study of metal organic chemical vapor deposition growth dynamics of GaN thin film at microscopic level. J. Appl. Phys. 103 (10), 103524.

Guerfi, Y., Larrieu, G., 2016. Vertical silicon nanowire field effect transistors with nanoscale gate-all-around. Nanoscale Res. Lett. 11, 210.

Hirvikorpi, T., Vähä-Nissi, M., Mustonen, T., Iiskola, E., Karppinen, M., 2010. Atomic layer deposited aluminum oxide barrier coatings for packaging materials. Thin Solid Films 518, 2654–2658.

Jansen, A.P.J. (Ed.), 2012. An Introduction to Kinetic Monte Carlo Simulations of Surface Reactions, Vol. 1. Academic Press, London, pp. 38–119.

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

Kanarik, K.J., Lill, T., Hudson, E.A., Sriraman, S., Tan, S., Marks, J., Vahedi, V., Gottscho, R.A., 2015. Overview of atomic layer etching in the semiconductor industry. J. Vac. Sci. Technol. A 33, 020802.

Kanarik, K.J., Tan, S., Gottscho, R.A., 2018. Atomic layer etching: rethinking the art of etch. J. Phys. Chem. Lett. 9, 4814–4821.

Keuter, T., Menzler, N.H., Mauer, G., Vondahlen, F., Vaßen, R., Buchkremer, H.P., 2015. Modeling precursor diffusion and reaction of atomic layer deposition in porous structures. J. Vac. Sci. Technol. A 33, 01A104.

Kotz, S., Johnson, N.L., 2002. Process capability indices-a review, 1992–2000. J. Qual. Technol. 34 (1), 2–19.

Lee, Y., DuMont, J.W., George, S.M., 2016. Trimethylaluminum as the metal precursor for the atomic layer etching of Al_2O_3 using sequential, self-limiting thermal reactions. Chem. Mater. 28, 2994–3003.

Levy, D.H., Nelson, S.F., Freeman, D., 2009. Oxide electronics by spatial atomic layer deposition. J. Disp. Technol. 5, 484–494.

Li, M.-Y., Su, S.-K., Wong, H.-S.P., Li, L.-J., 2019. How 2D semiconductors could extend Moore's law. Nature 567, 169–170.

Lill, T., 2021. Atomic Layer Processing: Semiconductor Dry Etching Technology. Wiley-VCH, Weinheim.

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

Luo, J., Canuso, V., Jang, J.B., Wu, Z., Morales-Guio, C.G., Christofides, P.D., 2022. Machine learning-based operational modeling of an electrochemical reactor: Handling data variability and improving empirical models. Ind. Eng. Chem. Res. 61, 8399–8410.

Merkx, M.J.M., Vlaanderen, S., Faraz, T., Verheijen, M.A., Kessels, W.M.M., Mackus, A.J.M., 2020. Area-selective atomic layer deposition of TiN using aromatic inhibitor molecules for metal/dielectric selectivity. Chem. Mater. 32 (18), 7788–7795.

Montgomery, D.C., 2013. Introduction to Statistical Quality Control, seventh ed. John Wiley & Sons, Hoboken.

Moore, G.E., 1998. Cramming more components onto integrated circuits. Proc. IEEE 86, 82–85.

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

Muñoz-Rojas, D., Nguyen, V.H., de la Huerta, C.M., Jiménez, C., Bellet, D., 2019. Spatial atomic layer deposition. In: Chemical Vapor Deposition for Nanotechnology. IntechOpen, pp. 1–25.

Natarajan, S.K., Elliott, S.D., 2018. Modeling the chemical mechanism of the thermal atomic layer etch of aluminum oxide: a density functional theory study of reactions during HF exposure. Chem. Mater. 30, 5912–5922.

Oehrlein, G.S., Metzler, D., Li, C., 2015. Atomic layer etching at the tipping point: An overview. ECS J. Solid State Sci. Technol. 4, N5041.

Poodt, P., Cameron, D.C., Dickey, E., George, S.M., Kuznetsov, V., Parsons, G.N., Roozeboom, F., Sundaram, G., Vermeer, A., 2012. Spatial atomic layer deposition: a route towards further industrialization of atomic layer deposition. J. Vac. Sci. Technol. A 30 (1), 010802.

Rahman, F., 2021. Atomic layer processes for material growth and etching—A review. IEEE Trans. Semicond. Manuf. 34, 500–512.

Raychaudhuri, S., 2008. Introduction to Monte Carlo simulation. In: 2008 Winter Simulation Conference. IEEE, Miami, USA, pp. 91–100.

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.

Sairam, T., Zhao, W., Cao, Y., 2007. Optimizing finfet technology for high-speed and low-power design. In: Proceedings of the 17th ACM Great Lakes Symposium on VLSI. Association for Computing Machinery, Stresa-Lago Maggiore, Italy, pp. 73–77.

Sang, X., Chang, J.P., 2020. Physical and chemical effects in directional atomic layer etching. J. Phys. D: Appl. Phys. 53, 183001.

Schwille, M.C., Schössler, T., Schön, F., 2017. Temperature dependence of the sticking coefficients of bis-diethyl amionsilane and trimethylaluminum in atomic layer deposition. J. Vac. Sci. Technol. A 35, 01B119.

Sheng, J., Lee, J.-H., Choi, W.-H., Hong, T., Kim, M., Park, J.-S., 2018. Review article: Atomic layer deposition for oxide semiconductor thin film transistors: Advances in research and development. J. Vac. Sci. Technol. A 36, 060801.

Shirazi, M., Elliott, S.D., 2014. Atomistic kinetic Monte Carlo study of atomic layer deposition derived from density functional theory. J. Comput. Chem. 35 (3), 244–259.

Sinha, S.K., Chaudhury, S., 2013. Impact of oxide thickness on gate capacitance—a comprehensive analysis on MOSFET, nanowire FET, and CNTFET devices. IEEE Trans. Technol. 12, 958–964.

Smith, T., Boning, D., 1997. A self-tuning EWMA controller utilizing artificial neural network function approximation techniques. IEEE Trans. Compon. Pack. Manuf. Technol. C 20 (2), 121–132.

Su, C.-T., Hsu, C.-C., 2004. A time-varying weights tuning method of the double EWMA controller. Omega 32 (6), 473–480.

Wang, X., Mahajan, R., 1996. Artificial neural network model-based run-to-run process controller. IEEE Trans. Compon. Pack. Manuf. Technol. C 19 (1), 19–26.

Weckman, T., Shirazi, M., Elliott, S.D., Laasonen, K., 2018. Kinetic Monte Carlo study of the atomic layer deposition of zinc oxide. J. Phys. Chem. C 122, 27044–27058.

Wu, Z., Rincon, D., Luo, J., Christofides, P.D., 2021. Machine learning modeling and predictive control of nonlinear processes using noisy data. AIChE J. 67 (4), e17164.

Ye, Z., Yuan, Y., Xu, H., Liu, Y., Luo, J., Wong, M., 2017. Mechanism and origin of hysteresis in oxide thin-film transistor and its application on 3-D nonvolatile memory. IEEE Trans. Electron. Devices 64, 438–446.

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., 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. Multiscale computational fluid dynamics modeling of thermal atomic layer etching: Application to chamber configuration design. Comput. Chem. Eng. 161, 107757.

Yun, S., Tom, M., Ou, F., Orkoulas, G., Christofides, P.D., 2022d. Multivariable run-to-run control of thermal atomic layer etching of aluminum oxide thin films. Chem. Eng. Res. Des. 182, 1–12.

Zhang, Y., Ding, Y., Christofides, P.D., 2019. Multiscale computational fluid dynamics modeling of thermal atomic layer deposition with application to chamber design. Chem. Eng. Res. Des. 147, 529–544.

Zhang, Y., Ding, Y., Christofides, P.D., 2020. Multiscale computational fluid dynamics modeling and reactor design of plasma-enhanced atomic layer deposition. Comput. Chem. Eng. 142, 107066.

Zywotko, D.R., Faguet, J., George, S.M., 2018. Rapid atomic layer etching of Al_2O_3 using sequential exposures of hydrogen fluoride and trimethylaluminum with no purging. J. Vac. Sci. Technol. A 36, 061508.