ALDTWIN – A HYBRID PHYSICS-BASED AND DATA-DRIVEN DIGITAL TWIN OF ATOMIC LAYER DEPOSITION PROCESS FOR SEMICONDUCTOR INNOVATION AND MANUFACTURING

Anonymous authors
Paper under double-blind review

ABSTRACT

Atomic layer deposition (ALD) is essential for conformal thin-film growth in advanced semiconductor fabrication, where equipment monitoring and maintenance are critical for atomic-scale precision and process consistency. However, in shared cleanroom environment, monitoring equipment health and predicting process outcomes are challenging because the same tools support diverse recipes, users, materials and operating conditions, while reliable fault and maintenance labels are scarce. We present ALDTwin, a hybrid physics-based and data-driven digital twin with an LLM agent-coordinated workflow for monitoring ALD processes under weakly labeled conditions. The methodology distinguishes intentional recipe-induced chamber conditioning effects from unintended equipment-state changes, including drift, contamination and malfunctions. We demonstrate ALDTwin using sensor data from a shared lab-scale cleanroom ALD system. The LLM agent coordinates data preparation and qualification before invoking simulation and anomaly models. Validated sensor logs are transformed into process-relevant features, after which an authoritative monitoring layer evaluates tool state, run completeness and data quality. Forty-four process metrics are compared with adaptive recipe-specific baselines to suppress expected recipe-dependent variation and identify equipment-state deviations. Qualified runs are passed to a physics-based feature scale simulation for estimating deposition outcomes, while incomplete or abnormal runs bypass simulation but remain part of longitudinal tool-state assessment, with abnormal runs escalated for engineer inspection. In parallel, one LSTM autoencoder detects abrupt deviations within multivariate sensor traces, while a second identifies gradual equipment drift across successive runs using process-relevant metrics. Finalized metric-based anomaly scores are incorporated into a Kalman forecasting layer to project near-term tool behavior. By separating probabilistic LLM coordination from deterministic qualification and physics-based evaluation, ALDTwin integrates safeguards, unsupervised learning, simulation and temporal forecasting within an interpretable, grounded, self-verifying framework for trustworthy process monitoring and decision support.

1 INTRODUCTION

Atomic Layer Deposition (ALD) has become an enabling thin film deposition technology for semiconductor fabrication because it offers a level of thickness control, conformality and material precision that is difficult to achieve with conventional deposition methods (George, 2010). Achieving that precision requires consistent precursor delivery, purge behavior, temperature and gas flow. Changes in equipment condition can therefore compromise an experiment even when the programmed recipe remains unchanged. For researchers, recognizing such changes matters beyond the deposition itself; an unreliable film may be carried into subsequent fabrication steps inside the cleanroom, resulting in wastage of samples, time and resources while also making device results harder to interpret. Sensor logs offer an opportunity to identify suspect runs before the sub-sequent fabrication effort continues.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

On a shared ALD tool in a lab-scale cleanroom, different users alternate among materials, cycle counts, pulse durations and purge settings, producing changes in pressure and thermal trajectories. At the same time, chamber history, component aging and servicing can alter the response of the equipment. Recipe transitions are already known to shift sensor distributions and degrade fault-detection models in semiconductor manufacturing (Shim et al., 2021). In lab-scale cleanrooms, sparse fault labels and incomplete maintenance records further complicate interpretation. Pooling the history can make an ordinary recipe change appear anomalous; accommodating too much variation can obscure equipment drift. The central challenge is therefore to distinguish recipe-induced variation from deviations associated with changes in equipment state when reliable fault labels are scarce. Researchers therefore need evidence that is comparable across repeated uses of a recipe and specific enough to guide review before further experiments proceed.

Statistical process monitoring detects departures from historical behavior (Qin, 2012), while recurrent autoencoders extend this capability to multivariate temporal patterns (Malhotra et al., 2016). Virtual metrology supports product-quality prediction (Maitra et al., 2024) and ALD digital twins enable process optimization with in-situ characterization (Yanguas-Gil et al., 2022). However, these efforts generally treat monitoring, process prediction and tool-state assessment separately. In shared lab-scale facilities, this distinction can determine whether an experiment should proceed: an unnoticed change in equipment condition may compromise a deposition despite an unchanged recipe, propagating an unreliable sample through costly downstream fabrication and obscuring the source of poor device performance. The challenge is therefore not merely detecting unusual data but determining whether a deviation reflects expected recipe-dependent behavior or a meaningful change in tool condition when fault labels are scarce.

We present ALDTwin, a hybrid physics-based and data-driven digital twin for near-real time post-run monitoring. Once an ALD process run is completed & that particular run's logs become available, an LLM agent-integrated workflow extracts key process metrics and determines whether the completed run deviates from the active historical operating envelope for the same recipe. When no abnormality is detected, the extracted metrics serve as inputs to a physics-based simulation calibrated against experimental data. The simulation predicts ALD process outcomes such as film thickness and quality before subsequent fabrication steps proceed. Missing or out-of-bound values trigger engineer inspection and suspend subsequent processing until disposition. Incomplete runs are recorded and bypass simulation. The LLM-driven agent acts as the user-facing orchestration layer that translates the deterministic monitoring and simulation outputs into interpretable feedback for experimentalists, helping them inspect relevant sensor trends and decide whether follow-up characterization, recipe adjustment or maintenance action is required.

Two recipe-specific LSTM autoencoders provide complementary views of equipment behavior at different temporal scales. For the raw sensor model, each completed run is represented as a 300×14 matrix obtained by resampling 14 synchronized sensor channels - chamber pressure, Heaters 6–17 and MFC1 to 300 time points. This representation preserves the detailed within-run waveform and is used to identify abrupt deviations in an individual process run. In parallel, the same synchronized run is reduced to a 39-element vector of physically interpretable process metrics. Ten consecutive same-recipe metric vectors form a 10×39 input window for the metric-based LSTM autoencoder. This metric-based representation is better suited to identifying gradual shifts in the tool-state baseline associated with effects such as chamber conditioning, component aging or unrecorded servicing while also providing physically interpretable descriptors of how the process behavior is changing over time. These scores are then passed to a robust Kalman forecasting model that tracks their temporal evolution and estimates near-term tool-drift trends, thus providing an additional indication of whether the tool is moving away from its established recipe-specific operating baseline. By combining recipe-specific models, physically meaningful process metrics and unlabeled historical data, this approach supports both run-level abnormality detection and longer-term tool-state assessment without requiring a single global definition of normal behavior or a labeled fault history.

Following the emphasis on synchronization and state in digital twin research (Jones et al., 2020), ALDTwin links each run to measurements, model outputs, simulation and engineer dispositions. Unlike conventional tool-native alarms and interlocks that respond to predefined conditions, ALDTwin provides a comprehensive recipe-aware workflow that uses unsupervised learning for multivariate historical behavior from unlabeled data, performs near-real time inference using the trained models against evolving same-recipe histories and integrates learned evidence with authoritative monitoring and physics-based simulation.

2 Related Work

Semiconductor Equipment Monitoring: Unsupervised methods span one-class, isolation, reconstruction, recurrent, and attention-based formulations (Qin, 2012; Schölkopf et al., 2001; Liu et al., 2008; Ruff et al., 2021; Malhotra et al., 2016; Xu et al., 2022; Tuli et al., 2022). Semiconductor studies address recipe transitions, process drift, maintenance effects, and tool health (Chen et al., 2020; Fan et al., 2020; Hwang et al., 2023; Kim et al., 2023; Shim et al., 2021; Chen & Blue, 2009).

ALD and Digital Twins: Related work covers virtual metrology, ALD diagnostics, hybrid optimization, synchronized digital twins, self-driving laboratories, and scientific-agent orchestration (Maitra et al., 2024; Morales et al., 2023; Tom et al., 2024; Navabi et al., 2025; Kritzinger et al., 2018; Jones et al., 2020; Yanguas-Gil et al., 2022; Abolhasani & Kumacheva, 2023; Boiko et al., 2023).

Research gap and ALDTwin: Current equipment monitoring models identify abnormal tool behavior, while digital twins optimize processes & predict deposition outcomes. Shared lab-scale facilities require both capabilities in one operational system. ALDTwin addresses this gap by integrating recipe-aware monitoring, deterministic qualification, calibrated physics-based prediction, temporal forecasting and LLM-coordinated reporting into a single auditable framework.

3 Problem Setting and Data

We use data from two commonly used thermal recipes on a shared ALD system: `Thermal_HfO2_200C` and `Thermal_Al2O3_200C`. The two recipes are modeled separately because their normal pressure and equipment-level responses are different. This recipe-specific separation provides the reference needed to distinguish an unusual tool response from an intentional change in process chemistry.

3.1 Recorded sensor data and run synchronization

For every ALD run, the ALD system generates separate pressure and heater log files. The pressure file records pressure time, chamber pressure, recipe identity and number of remaining cycles. The heater file has its own time base and records Heaters 6-17, MFC1 carrier gas flow, MFC time and the corresponding recipe/run-state information. The pressure trace therefore captures the repeated pressure response associated with precursor, purge and oxidant steps, while the heater and MFC channels describe the thermal and gas delivery behavior of the tool during the same run.

The two files are first paired using the run timestamp and recipe encoded in their filenames, so the pressure and heater files must correspond to the same run and the same supported recipe. Their sensor values are then synchronized because the two files are recorded on different time grids. Pressure time is used as the reference grid; pressure and run-state fields are retained directly, Heaters 6-17 are linearly interpolated onto the pressure timestamps, and MFC1 is aligned from its own MFC time using nearest-neighbor matching. This produces one synchronized multivariate record for each run which is used by the downstream feature extraction and anomaly detection workflow.

Table 1: Data used for the two recipe-specific LSTM monitoring tasks.

Recipe	Logged Sensor Data	Training History
HfO_2 200°C	Pressure, Heater 6–17, MFC1, Cycles Remaining, Recipe State	2021–2025
Al_2O_3 200°C	Pressure, Heater 6–17, MFC1, Cycles Remaining, Recipe State	2021–2025

3.2 Run eligibility and historical training data

Not every logged run is treated as an eligible model input. A pressure/heater pair must first be available for the same timestamped run and must correspond to one of the two supported recipes. Run completion is checked using the 'Cycles Remaining' remaining column; a completed run must

have a minimum recorded value of 1. Incomplete or manually stopped runs are retained in the record but they do not enter the same-recipe baseline used for normal completed depositions and will not be included in the deposition simulation.

For the experiments in this work, the recipe-specific LSTM autoencoders are trained on the available historical data from 2021–2025 for the two recipes in Table 1. Separate models are trained for HfO_2 and Al_2O_3 so that normal recipe-dependent differences are learned within each recipe rather than being interpreted as anomalies. After training, the model weights and preprocessing transforms are frozen and used as reference for subsequent runs.

3.3 Data from incoming new runs

After the historical models are frozen, every subsequently received run is treated as a new incoming run. It passes through the same file pairing, synchronization and preprocessing steps, after which the frozen LSTM autoencoders are used only for inference; their weights are not changed during routine monitoring. The resulting processed data are then passed to the downstream qualification, anomaly-scoring and forecasting stages described in Section 4. Each incoming run therefore contributes new evidence to the recipe-specific monitoring history without modifying the deployed models, preserving consistency between historical and prospective evaluation.

4 The ALDTwin Framework

ALDTwin combines three complementary functions within a single process-monitoring architecture: unsupervised learning from historical runs to characterize tool drift, near-real-time anomaly assessment of newly completed runs and short-horizon forecasting of the evolving tool state. Figure 1 summarizes these three roles. For the ALD system considered here, sensor logs become available after a run finishes, so the 'online' stage is implemented as immediate post-run monitoring rather than in-situ control during deposition. The workflow is coordinated by an LLM agent, while the underlying qualification, anomaly scoring, forecasting and simulation remain governed by explicit computational modules.

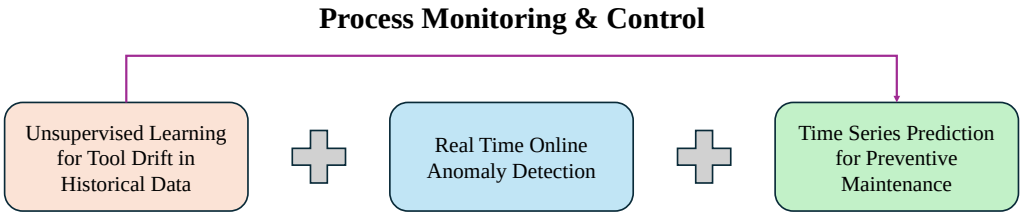


Figure 1: Overview of the Framework for Process Monitoring & Control

4.1 LLM-agent coordinated post-run workflow

For every incoming run, the agent first invokes the data processing pipeline described in Section 3: the pressure and heater/equipment logs are paired, synchronized to a common time base and transformed into process-relevant metrics. The synchronized run record then feeds the monitoring branches shown in Figure 2. The central operational decision is made by the authoritative rule-based run qualification layer. This layer checks run completion and data quality and compares 44 monitored process metrics with the active same-recipe baseline. In this way, expected differences between HfO_2 and Al_2O_3 operation are separated from deviations in the response of the equipment itself.

A run that satisfies the active checks is marked as qualified and can proceed to the physics-based simulation. A run with missing or out-of-bound monitored values is escalated for engineer review, while incomplete runs are recorded but do not proceed to simulation. The LLM agent does not alter these decisions. Its role is to coordinate the prescribed modules, preserve the resulting evidence and present the outcome in a form that is useful to the experimentalist. This separation keeps the safety-

relevant decision path auditable while still allowing learned models and language-based interaction to provide additional context.

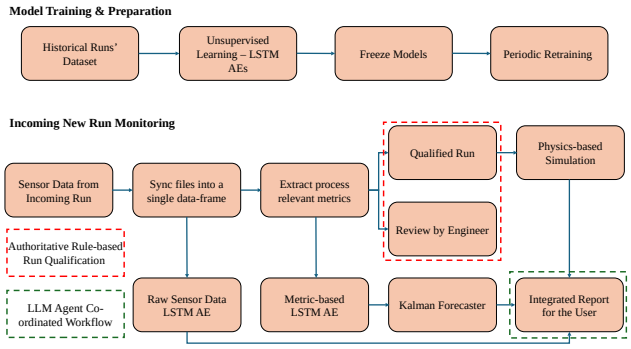


Figure 2: ALDTwin Model-Training and Incoming-Run Monitoring Workflow

4.2 Complementary LSTM autoencoder models

Two recipe-specific LSTM autoencoders provide complementary unsupervised evidence at different temporal scales. The raw sensor data LSTM autoencoder operates directly on the multivariate sensor waveform of an individual completed run. Each eligible run is resampled to a fixed sequence of 300 time points containing pressure, Heater 6–17 and MFC1, giving 14 sensor channels. The reconstruction error of the frozen model provides an immediate run-level anomaly score and is intended to capture abrupt deviations that occur within a single process run.

The metric-based LSTM autoencoder instead operates on compact process descriptors extracted from consecutive runs. Each input contains 39 process-relevant features across a fixed-run for same recipe window, including pressure-response, heater, MFC and inter-run timing information. This representation is designed to capture slower changes in the operating baseline that develop across multiple runs. Because the windows overlap, tool-drift scores for the most recent runs remain provisional until sufficient later same-recipe context becomes available; a tool-drift score is finalized after nine subsequent valid runs. The raw and metric models are therefore not competing fault classifiers; they provide two views of equipment behavior, one emphasizing isolated within-run deviations and the other emphasizing longer term tool drift. The raw-model reconstruction error is reported as a run-level anomaly score, whereas the metric-based reconstruction error is reported as a tool-drift score; both serve as supporting evidence rather than authoritative declarations of a physical fault.

4.3 Kalman forecasting

Finalized tool-drift scores from the metric-based LSTM are passed to a robust Kalman forecaster. The forecaster tracks the temporal evolution of the metric-based tool-drift signal and projects its near-term trajectory over 1-, 5- and 10-finalized-run horizons. These forecasts indicate whether the tool is moving toward or away from its recent recipe-specific operating baseline, providing experimentalists with insight into the evolving tool condition so that upcoming experiments can be planned with greater awareness of the current tool state.

4.4 Physics-Based Simulation

In parallel, a qualified run is passed to the physics-based simulation. The simulator uses the extracted run conditions, including cycle count, purge timing and pressure-related metrics, to rapidly estimate deposition outcomes such as film thickness and growth per cycle. Recent qualified runs of the same recipe provide process context for the simulation when available. The simulator provides a computationally efficient physics-based representation of ALD growth, linking measured process conditions to predicted deposition behavior and film evolution within the integrated monitoring framework. Runs that are incomplete or require engineer inspection bypass this branch.

4.5 Integrated reporting and periodic model refresh

The final stage combines the authoritative run verdict with the physics-based outputs, raw sensor anomaly score, metric-based tool-drift evidence and Kalman tool-health trend into an integrated report for the user. The LLM agent coordinated layer acts as the interface across these modules; it organizes the evidence and communicates the state of the run but it cannot override the rule-based qualification, modify numerical model outputs or independently infer the root cause of an equipment fault. Engineer review therefore remains the controlling action whenever the deterministic layer identifies an abnormal condition.

During routine monitoring, the deployed LSTM models and preprocessing transforms remain frozen so that successive incoming runs are evaluated against a consistent reference. New observations accumulate in the monitoring history and after 100 finalized post-training tool-drift scores are available for a recipe, the workflow opens a periodic retraining review. Retraining and validation occur offline and an updated model is promoted only as a new frozen version after review rather than being modified automatically from individual incoming runs. Detailed metric definitions, model architectures, scoring equations, forecasting equations and simulator formulation are elaborated upon in the Appendix.

5 Results

This section evaluates the principal outputs of the monitoring modules introduced in Sections 4.2–4.5. The results are organized by evidence layer rather than combined into a single anomaly decision because each module serves a different role as outlined in the overall ALDTwin workflow.

5.1 Run-Level Anomaly Detection and Tool-Drift Analysis

Figure 3 shows the historical anomaly score trajectories produced by the raw sensor LSTM autoencoder for the two monitored recipes. For each run, the score is the reconstruction MSE of the normalized sensor data representation. Lower values indicate closer agreement with learned recipe-specific behavior, whereas elevated values reflect sensor trajectories that are increasingly difficult for the recipe-specific model to reconstruct. The circled points therefore indicate comparatively anomalous within-run behavior relative to the surrounding run history.

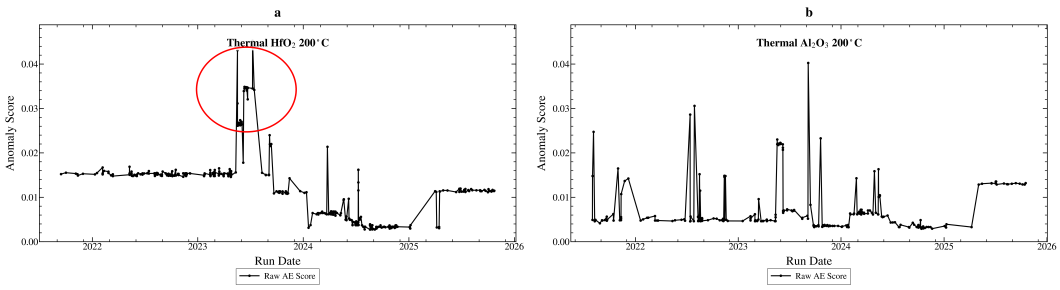


Figure 3: Historical raw sensor LSTM autoencoder anomaly scores for (a) Thermal HfO_2 at 200 °C and (b) Thermal Al_2O_3 at 200 °C

The metric-based LSTM provides a different view of the same tool history. As shown in Figure 4, the tool drift score evolves more smoothly across consecutive runs and exhibits sustained shifts between operating regimes rather than only isolated spikes. This behavior is consistent with the model construction in Section 4.2; each tool-drift score is derived from a sliding window of process-relevant metrics spanning multiple same-recipe runs, so the representation emphasizes changes that persist across the recent operating history. The change in tool drift score between two consecutive points provides an estimate of how much the equipment condition has evolved from one run to the next, while local minima and maxima in the score trajectory can serve as candidate transition points at which the tool condition may have changed, for example following maintenance, repair, chamber

servicing or another intervention. In the absence of complete maintenance labels, these extrema are interpreted as indicators of possible equipment-state transitions rather than definitive maintenance events. The raw sensor and metric-based histories should therefore be interpreted jointly but not numerically compared across models; the former emphasizes abrupt within-run departures, whereas the latter provides a slower multi-run view of evolving equipment behavior.

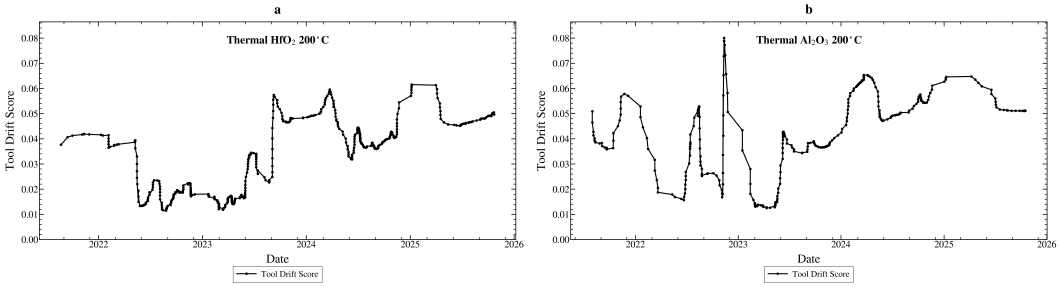


Figure 4: Historical metric-based LSTM autoencoder tool-drift scores for (a) Thermal HfO_2 at 200 °C and (b) Thermal Al_2O_3 at 200 °C

5.2 KALMAN FORECASTING OF FINALIZED TOOL-DRIFT SCORES

The local level Kalman forecaster was evaluated on five previously unseen incoming runs per recipe in two operating modes (Figure 5). Five consecutive unseen runs were retained as a compact chronological holdout consistent with the near-term forecasting objective, while preserving strict temporal separation between the historical filter state and prospective evaluation. Open-loop forecasts were generated without assimilating held-out finalized scores, whereas rolling one-step operation predicted each next finalized score and then updated the filter once that observation became available. Open-loop MAPE is 0.517% for Thermal HfO_2 and 0.200% for Thermal Al_2O_3 at 200 °C; rolling one-step operation reduces these values to 0.310% and 0.134%, respectively. Across all five runs, point-wise absolute percentage errors remain below 1% for both recipes. The results therefore demonstrate low-error short-horizon forecasting in both modes, with improved tracking when newly finalized observations are sequentially assimilated.

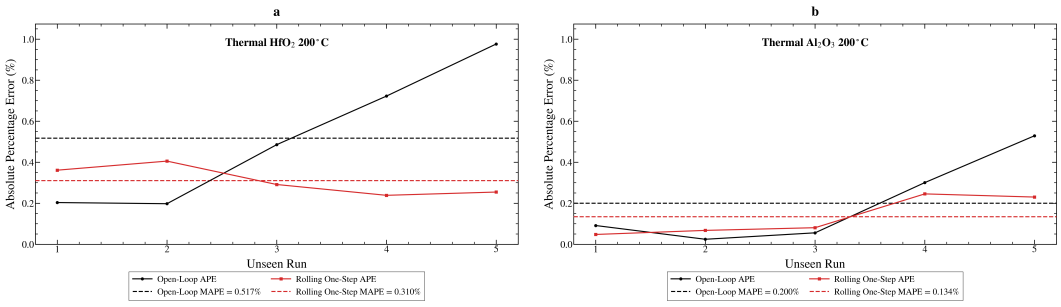


Figure 5: Absolute percentage error for the local-level Kalman forecaster in open-loop and rolling one-step modes over five unseen runs for (a) Thermal HfO_2 at 200 °C and (b) Thermal Al_2O_3 at 200 °C

5.3 PHYSICS-BASED SIMULATION RESPONSE

Table 2 summarizes the physics-based simulation results for the two 200,°C ALD recipes. The simulation represents the cyclic ALD chemistry through the time-dependent evolution of surface coverage during the precursor-pulse, purge, co-reactant-pulse and second-purge stages, with chemistry-specific kinetic parameters calibrated against experimental deposition behavior. For the HfO_2 and Al_2O_3 processes considered here, the corresponding TDMAH/H_2O and TMA/H_2O reaction systems

are parameterized separately so that the simulated growth response remains anchored to the experimentally observed behavior of each chemistry. Within ALDTwin, process conditions extracted from an incoming run are supplied to this physics-based simulation to estimate the expected deposition response, including growth per cycle and film thickness. Because the simulation explicitly relates the observed process conditions to the underlying surface-reaction kinetics, it provides a physically grounded estimate of the post-deposition film response before ex-situ characterization is performed. Such rapid estimates can reduce the need for routine, time-consuming and expensive metrology after every deposition, while also providing early guidance for planning subsequent fabrication or characterization steps.

Table 2: Physics-based Simulation Results for the two 200 °C ALD Recipes

Recipe	Purge Time (s)	Growth per Cycle (nm/cycle)	Total Thickness (nm)
HfO_2, 200 °C	5	0.121	6.05
HfO_2, 200 °C	10	0.111	5.55
HfO_2, 200 °C	20	0.103	5.15
HfO_2, 200 °C	30	0.100	5.00
Al_2O_3, 200 °C	5	0.119	5.95
Al_2O_3, 200 °C	10	0.109	5.45
Al_2O_3, 200 °C	20	0.102	5.10
Al_2O_3, 200 °C	30	0.100	5.00

5.4 Integrated workflow output and model refresh

The component-level results above are combined only after the authoritative run qualification has been established. For a qualified incoming run, the integrated report presents the deterministic verdict together with the raw sensor anomaly score, metric-based tool drift evidence, the Kalman trend forecast and the physics-based simulation output. Runs requiring inspection retain the learned model evidence for longitudinal tool-state assessment but do not proceed through the simulation branch. The deployed preprocessing and LSTM bundles remain frozen during routine monitoring, with retraining handled as the reviewed periodic update described in Section 4.5 rather than as an automatic per-run model change. A worked end-to-end representation of the integrated report, evidence flow and refresh/governance logic is provided in Appendix I.

6 Discussion

6.1 Operational significance and supporting validation

The results demonstrate how complementary evidence can be combined within a stateful digital twin without forcing heterogeneous signals into a single anomaly decision. The deterministic qualification layer provides an auditable assessment against recipe-specific operating limits, while the raw sensor LSTM identifies unusual within-run behavior, the metric-based LSTM captures longer-term changes in equipment state, the Kalman model projects the near-term evolution of this state and the physics-based simulator translates measured process conditions into estimates of deposition outcome. Disagreement between these layers can itself be informative; for example, a run may remain within deterministic limits while exhibiting an unusual sensor trajectory or occurring during a longer-term shift in the learned tool state. ALDTwin therefore preserves these signals as complementary evidence for experimental decision support rather than treating any individual learned score as proof of a physical fault.

An important supporting observation is provided by the Thermal HfO_2 200,°C history in Figure 3. The raw sensor LSTM identifies a prominent concentration of high anomaly scores around mid-2023, including the June 2023 period during which users of the tool independently reported several deposition runs with unsatisfactory experimental outcomes. These retrospective reports do not constitute a complete or formally labeled failure dataset and consequently cannot be treated as conventional supervised validation. Nevertheless, the temporal agreement provides an independent operational reference supporting the practical relevance of the learned anomaly signal and motivates prospective collection of run-level experimental dispositions in future deployments.

6.2 Limitations, safety and human oversight

The present study remains limited to two recipes on a single ALD tool and operates in a setting without dense fault, maintenance or film quality labels. Reconstruction errors should therefore be interpreted as relative indicators of deviation rather than calibrated fault probabilities. Deterministic limits are presently defined from process knowledge and historical behavior, the Kalman analysis supports only short-horizon trend forecasting. These constraints position ALDTwin as a post-run monitoring and decision-support framework rather than an autonomous closed-loop process control system. Accordingly, learned scores, simulation results and LLM-generated explanations cannot override the authoritative qualification layer and runs requiring inspection remain subject to expert review. Because existing published approaches generally address individual tasks such as equipment anomaly detection, wafer-quality prediction or process optimization rather than the complete end-to-end, recipe-aware ALDTwin workflow, they are not directly comparable as system-level baselines under the same data, supervision and evaluation assumptions; controlled component-level comparisons therefore remain an important direction for broader validation.

6.3 Generalization and future work

Future work will extend the framework through multi-tool and multi-site validation, prospectively collected run outcomes and maintenance records, independent film metrology, uncertainty-aware fusion of the complementary evidence streams, higher fidelity physics models and controlled evaluation of model refresh policies. A particularly important extension is an ontology-backed knowledge graph connecting runs, recipes, process steps, materials, sensors, equipment subsystems, maintenance events, anomaly evidence and physics-model predictions. Such a representation could provide the agent with structured process and equipment context, enable provenance-aware reasoning and root cause analysis and help distinguish recipe-induced variation from genuine equipment-state changes across heterogeneous process histories. More broadly, this knowledge graph could provide a semantic foundation for extending ALDTwin beyond a single ALD tool toward interoperable digital twins spanning multiple semiconductor fabrication processes and characterization systems.

Extending ALDTwin from post-run monitoring on a single tool to deployment in high-volume semiconductor fabs would additionally require real-time, event-driven telemetry and integration with factory-level infrastructure. High-rate equipment data could be acquired through the SEMI Equipment Data Acquisition (EDA/Interface A) standards alongside established SECS/GEM interfaces, enabling streaming telemetry to feed low-latency fault detection and classification (FDC) and advanced process control (APC) services for rapid excursion detection and intervention (SEMI, 2026; Goodlin et al., 2003; Moyne et al., 2007). Building on the semantic representation described above, anomaly and drift information could be integrated with tool, chamber, recipe, wafer, lot, maintenance and yield context through MES/YMS-connected knowledge graphs, supporting contextual reasoning, model evolution and manufacturing decision support (Meyers et al., 2022; Su et al., 2024; Somma et al., 2026). Within such a deployment, the LLM-based agentic layer could support workflow orchestration, root-cause analysis and disposition recommendations, while deterministic and validated automation would retain authority over production-critical control actions.

7 Conclusion

This work introduced ALDTwin, a hybrid physics- & data-driven digital twin for monitoring ALD processes in a heterogeneous, label-sparse lab-scale cleanroom. The framework combines deterministic run qualification, complementary raw sensor and metric-based LSTM autoencoders, short-horizon Kalman forecasting, experimentally calibrated physics-based simulation and LLM-agent coordinated reporting while preserving explicit human oversight. Results across two widely used ALD recipes demonstrate the ability to identify unusual run behavior, track long-term changes in equipment state, forecast near-term tool drift trends and estimate deposition outcomes from observed process conditions. Together, these capabilities establish ALDTwin as an auditable and stateful foundation for post-run process monitoring and experiment decision support in shared lab-scale facilities.

AI USE STATEMENT

Generative AI tools were used to assist with programming and implementation, code refinement and debugging, literature and information search, and for language editing, rephrasing and improvement during manuscript preparation. All AI-assisted code and text were carefully and thoroughly reviewed and verified by the authors, who take responsibility for the final implementation, scientific content, results and claims presented in this work. API key of Gemini 3.8 Flash was additionally used as the language generation component of the ALDTwin reporting workflow described in I.

ETHICS STATEMENT

All experimental data analyzed in this work consist of authentic process and equipment measurements collected during actual operation of an ALD system in a shared lab-scale semiconductor fabrication facility; no synthetic data are substituted for the experimental sensor records used to evaluate the monitoring framework. Because these records contain facility-specific operational information and are subject to privacy and data-governance constraints, the raw dataset cannot be released publicly. Access to appropriately shareable data may be considered upon specific reasonable request to the authors, subject to applicable facility and institutional restrictions.

REPRODUCIBILITY STATEMENT

ALDTwin was developed and evaluated for an ALD system operating in an actual shared lab-scale semiconductor fabrication facility. The paper describes the data, framework and evaluation methodology, while also providing additional details on data preprocessing and eligibility, process-metric extraction, LSTM autoencoder formulation and scoring, Kalman forecasting, physics-based ALD simulation and integrated workflow outputs. The raw facility data cannot be publicly released because of the privacy and data-governance constraints described above; appropriately shareable data may be provided upon reasonable request subject to applicable restrictions. The authors are also prepared to share the complete implementation codebase or individual components of the ALDTwin workflow, upon specific reasonable request to facilitate reproduction and extension of the reported results.

REFERENCES

Milad Abolhasani and Eugenia Kumacheva. The rise of self-driving labs in chemical and materials sciences. *Nature Synthesis*, 2:483–492, 2023. doi: 10.1038/s44160-022-00231-0.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624:570–578, 2023. doi: 10.1038/s41586-023-06792-0.

Argon Chen and Jakey Blue. Recipe-independent indicator for tool health diagnosis and predictive maintenance. *IEEE Transactions on Semiconductor Manufacturing*, 22(4):522–535, 2009. doi: 10.1109/TSM.2009.2028215.

Chieh-Yu Chen, Shi-Chung Chang, and Da-Yin Liao. Equipment anomaly detection for semiconductor manufacturing by exploiting unsupervised learning from sensory data. *Sensors*, 20(19): 5650, 2020. doi: 10.3390/s20195650.

Shu-Kai S. Fan, Chia-Yu Hsu, Chih-Hung Jen, Kuan-Lung Chen, and Li-Ting Juan. Defective wafer detection using a denoising autoencoder for semiconductor manufacturing processes. *Advanced Engineering Informatics*, 46:101166, 2020. doi: 10.1016/j.aei.2020.101166.

Steven M. George. Atomic layer deposition: An overview. *Chemical Reviews*, 110(1):111–131, 2010. doi: 10.1021/cr900056b.

Brian E. Goodlin, Duane S. Boning, Herbert H. Sawin, and Barry M. Wise. Simultaneous fault detection and classification for semiconductor manufacturing tools. *Journal of The Electrochemical Society*, 150(12):G778–G784, 2003. doi: 10.1149/1.1623772.

Rakhoon Hwang, Seungtae Park, Youngwook Bin, and Hyung Ju Hwang. Anomaly detection in time series data and its application to semiconductor manufacturing. *IEEE Access*, 11:130483–130490, 2023. doi: 10.1109/ACCESS.2023.3333247.

Richard W. Johnson, Adam Hultqvist, and Stacey F. Bent. A brief review of atomic layer deposition: From fundamentals to applications. *Materials Today*, 17(5):236–246, 2014. doi: 10.1016/j.mattod.2014.04.026.

David Jones, Chris Snider, Aydin Nassehi, Jason Yon, and Ben Hicks. Characterising the digital twin: A systematic literature review. *CIRP Journal of Manufacturing Science and Technology*, 29:36–52, 2020. doi: 10.1016/j.cirpj.2020.02.002.

Youngju Kim, Hoyeop Lee, and Chang Ouk Kim. A variational autoencoder for a semiconductor fault detection model robust to process drift due to incomplete maintenance. *Journal of Intelligent Manufacturing*, 34(2):529–540, 2023. doi: 10.1007/s10845-021-01810-2.

Werner Kritzinger, Matthias Karner, Georg Traar, Jan Henjes, and Wilfried Sihn. Digital twin in manufacturing: A categorical literature review and classification. *IFAC-PapersOnLine*, 51(11):1016–1022, 2018. doi: 10.1016/j.ifacol.2018.08.474.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422, 2008. doi: 10.1109/ICDM.2008.17.

Varad Maitra, Yutai Su, and Jing Shi. Virtual metrology in semiconductor manufacturing: Current status and future prospects. *Expert Systems with Applications*, 249:123559, 2024. doi: 10.1016/j.eswa.2024.123559.

Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. LSTM-based encoder–decoder for multi-sensor anomaly detection. In *ICML 2016 Workshop on Anomaly Detection*, 2016. arXiv:1607.00148.

Bart Meyers, Johan Van Noten, Pieter Lietaert, Bavo Tielemans, Hristo Hristov, Davy Maes, and Klaas Gadeyne. Knowledge graphs in digital twins for manufacturing—lessons learned from an industrial case at Atlas Copco Airpower. *IFAC-PapersOnLine*, 55(10):13–18, 2022. doi: 10.1016/j.ifacol.2022.09.361.

Carlos Morales, Ali Mahmoodinezhad, Rudi Tschammer, Julia Kosto, Carlos Alvarado Chavarin, Markus Andreas Schubert, Christian Wenger, Karsten Henkel, and Jan Ingo Flege. Combination of multiple operando and in-situ characterization techniques in a single cluster system for atomic layer deposition: Unraveling the early stages of growth of ultrathin Al2O3 films on metallic Ti substrates. *Inorganics*, 11(12):477, 2023. doi: 10.3390/inorganics11120477.

James R. Moyne, Hazem M. Hajj, Kevin Beatty, and Russell Lewandowski. SEMI E133—the process control system standard: Deriving a software interoperability standard for advanced process control in semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing*, 20(4):408–420, 2007. doi: 10.1109/TSM.2007.907617.

Pouyan Navabi, Remya Ampadi Ramachandran, Harshdeep Bhatia, Majid Jaberi-Douraki, Urmila Diwekar, Cortino Sukotjo, and Christos G. Takoudis. Artificial intelligence in atomic layer deposition. *Journal of Vacuum Science & Technology A*, 43(6):060801, 2025. doi: 10.1116/6.0004881.

S. Joe Qin. Survey on data-driven industrial process monitoring and diagnosis. *Annual Reviews in Control*, 36(2):220–234, 2012. doi: 10.1016/j.arcontrol.2012.09.004.

Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021. doi: 10.1109/JPROC.2021.3052449.

Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001. doi: 10.1162/089976601750264965.

SEMI. Equipment data acquisition (EDA/Interface A) standards suite: SEMI E120, E125, E128, E132, E134, E138, and E164. Semiconductor Equipment and Materials International, 2026. URL https://www.semi.org/en/about/SEMI_Standards. Standards information page, accessed September 2026.

Jaewoong Shim, Sungzoon Cho, Euiseok Kum, and Suho Jeong. Adaptive fault detection framework for recipe transition in semiconductor manufacturing. *Computers & Industrial Engineering*, 161: 107632, 2021. doi: 10.1016/j.cie.2021.107632.

Alessandra Somma, Domenico Amalfitano, Alessandra De Benedictis, and Patrizio Pelliccione. TwinArch: A digital twin reference architecture. *Journal of Systems and Software*, 231:112613, 2026. doi: 10.1016/j.jss.2025.112613.

Chang Su, Yong Han, Xin Tang, Qi Jiang, Tao Wang, and Qingchen He. Knowledge-based digital twin system: Using a knowlege-driven approach for manufacturing process modeling. *Computers in Industry*, 159–160:104101, 2024. doi: 10.1016/j.compind.2024.104101.

Matthew Tom, Henrik Wang, Feiyang Ou, Gerassimos Orkoulas, and Panagiotis D. Christofides. Machine learning modeling and run-to-run control of an area-selective atomic layer deposition spatial reactor. *Coatings*, 14(1):38, 2024. doi: 10.3390/coatings14010038.

Shreshth Tuli, Giuliano Casale, and Nicholas R. Jennings. TranAD: Deep transformer networks for anomaly detection in multivariate time series data. *Proceedings of the VLDB Endowment*, 15(6): 1201–1214, 2022. doi: 10.14778/3514061.3514067.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=LzQQ89U1qm_.

Angel Yanguas-Gil, Steven Letourneau, Noah Paulson, and Jeffrey W. Elam. Experimental platform and digital twin for AI-driven materials optimization and discovery for microelectronics using atomic layer deposition. In *NeurIPS 2022 Workshop on AI for Accelerated Materials Design (AI4Mat)*, 2022. URL https://openreview.net/forum?id=2rwjFfJzu8J.

APPENDIX

A ATOMIC LAYER DEPOSITION PROCESS FUNDAMENTALS

Atomic Layer Deposition (ALD) is a cyclic, surface-controlled thin film deposition process in which gaseous reactants are introduced sequentially rather than simultaneously. The defining feature of ALD is that each exposure is designed to proceed through a self-limiting surface reaction: once the available reactive surface sites have been consumed, additional precursor exposure ideally does not continue to increase the amount deposited during that half-cycle. Repeating these self-limiting reactions enables precise control of film thickness through the number of ALD cycles while maintaining conformal coverage over planar and high-aspect-ratio structures (George, 2010; Johnson et al., 2014).

A typical thermal ALD cycle contains four sequential stages, as shown in Figure 6. First, the precursor pulse introduces the metal-containing precursor into the chamber, where precursor molecules chemisorb onto available surface sites until the surface approaches saturation. Second, the precursor purge removes excess precursor molecules and volatile byproducts from the gas phase so that the two reactants do not mix directly in the chamber. Third, the co-reactant or oxidant pulse introduces the second reactant, which reacts with the adsorbed precursor-derived surface species to form the desired solid material and regenerate surface groups for the next cycle. Fourth, the co-reactant purge removes the remaining reactant and byproducts before the next precursor exposure begins. The cycle is repeated until the target film thickness is reached.

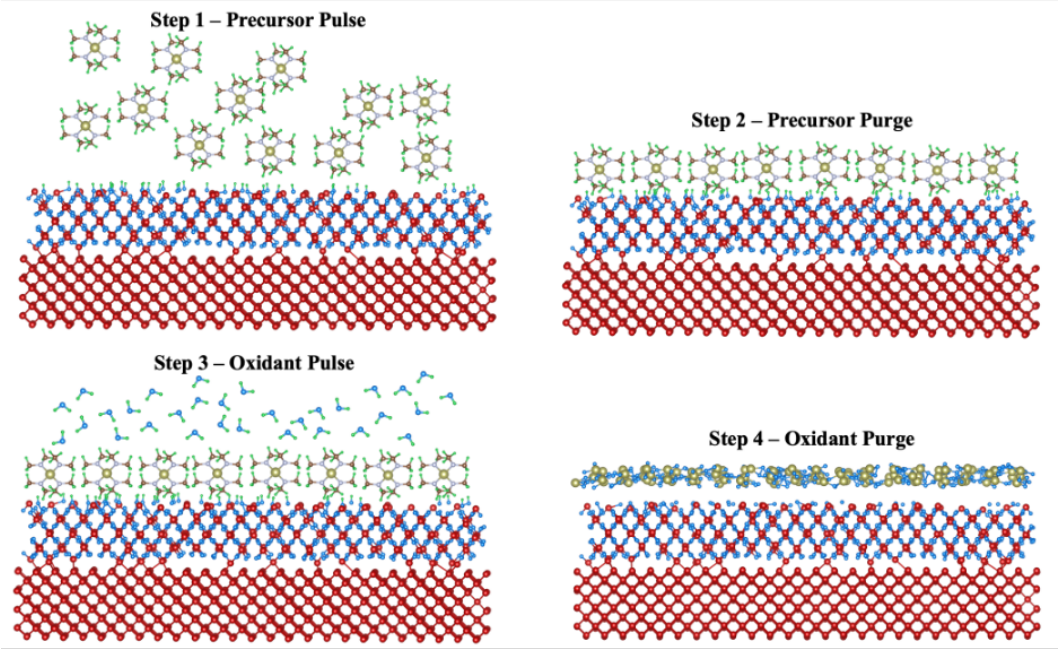


Figure 6: Four sequential steps involved in a typical thermal ALD Process cycle: precursor pulse, precursor purge, oxidant pulse and oxidant purge.

Although the programmed sequence is simple, the resulting film depends on the coupled behavior of precursor delivery, chamber pressure, carrier gas transport, surface saturation, purge efficiency, substrate temperature, byproduct removal and the evolving condition of the deposition chamber. Insufficient purge time can leave residual reactants or byproducts in the chamber, whereas unstable precursor delivery, pressure response or temperature can alter the degree of surface saturation reached during a cycle. Chamber history can also influence subsequent runs because prior materials and process conditions may introduce contaminants and alter the chemical state of internal surfaces. These effects make pressure, heater and mass-flow-controller traces useful indicators of whether the physical execution of a recipe is consistent with its expected behavior.

For the two thermal recipes examined in this work, the cyclic pressure signature therefore contains physically meaningful process information rather than merely serving as a generic time-series input. Repeated precursor and H_2O responses, the pressure level between pulses, purge-related timing and the stability of heater and carrier gas signals provide a compact view of how the process was executed. ALDTwin uses this structure explicitly: recipe-specific preprocessing and metric extraction preserve the relationship between the recorded sensor response and the underlying ALD cycle, while the raw sensor model retains the detailed within-run temporal behavior.

B Shared Lab-Scale Fabrication versus High-Volume Manufacturing

The monitoring problem considered in this work is shaped by the operating environment in which the ALD tool is used. A shared lab-scale cleanroom is not simply a smaller version of a high-volume manufacturing fab. The same equipment may support process development, exploratory materials synthesis, device prototyping and user training, with different user groups alternating among materials, substrates and recipe conditions. High-volume manufacturing, by contrast, is generally organized around highly standardized process flows and repeated production recipes. Table 3 summarizes the operational differences that are most relevant to process monitoring in this study.

Table 3: Operational Differences Between Lab-Scale and High-Volume Fabrication

Feature	Shared Lab-Scale Fabrication	High-Volume Fabrication
Equipment Usage Model	Shared/Multi-User - Tools are shared across diverse research groups. Different groups use different combinations of materials & process recipes	Streamlined - Equipment is often dedicated to specific process steps or layers. Strict segregation exists to prevent cross-contamination between incompatible materials.
Equipment State & History	Discontinuous - The chamber condition is unpredictable due to the previous users' different recipe.	Steady State - Tools run the same recipe continuously, maintaining a consistent thermal and chemical equilibrium. Statistical Process Control (SPC) constantly monitors drift
Process Control & Recipe	High Variability - Recipes are frequently adjusted or reinvented for every experiment. There is no standard flow; parameters are tuned manually for unique experimental requirements.	Fixed Standard Operating Procedure - Processes are locked. Variations are strictly minimized to ensure repeatability and yield.
Wafer Handing	Manual - Experimentalists load samples manually. Transfer between tools is done by hand, thus introducing human error and contamination risks.	Automated - Fully automated material handling systems. Human contact with wafers is non-existent.

This distinction motivates the recipe-aware design of ALDTwin. In a production environment with a tightly repeated recipe, a global or long-lived process baseline can often be meaningful. In the shared lab-scale setting considered here, however, normal operating behavior must first be conditioned on recipe context before changes associated with equipment state can be interpreted. The same consideration also explains why unlabeled historical sensor data, explicit run qualification and human review are central to the framework rather than secondary implementation details.

C Data Organization, Run Pairing, and Sensor Alignment

This section provides additional implementation details for the data organization and preprocessing steps summarized in Section 3. The objective of these steps is to convert the separate equipment logs generated for each ALD process run into a recipe-specific, time-aligned representation that can be used consistently by both the raw-sensor and metric-based monitoring branches. No global baseline is formed across recipes; the HfO_2 and Al_2O_3 histories remain separated throughout data preparation and subsequent modeling.

C.1 Raw ALD Log Structure

Each ALD process run is represented by two raw equipment log files in text format (`.txt`) with complementary information. The pressure log contains the pressure time base, chamber pressure, recipe identity, run-state information and the 'cycles remaining' field. The repeated pressure response reflects the cyclic precursor, purge and oxidant sequence of the ALD recipe, while 'cycles remaining' provides a direct record of progression through the programmed deposition cycles. The heater/equipment log is recorded on a separate time base and contains the Heater 6-17 channels, MFC1 carrier-gas-flow information, the corresponding MFC time and recipe/run-state information. The heater channels describe the thermal behavior of the tool during the run, whereas MFC1 captures the carrier gas delivery response.

These two files are treated as complementary records of the same physical process rather than as independent samples. Only the two supported recipes considered in this work, `Thermal_HfO2_200C` and `Thermal_Al2O3_200C`, are included in the recipe-specific monitoring datasets. The two recipe histories are kept separate because differences in chemistry and programmed process conditions produce different nominal pressure and equipment responses. Con-

sequently, a sensor pattern that is normal for one recipe is not used to define normal behavior for the other.

C.2 Run Pairing, Cleaning and Temporal Alignment

Before any model input or run-level metric is generated, the pressure and heater/equipment files are paired using the run timestamp and recipe encoded in their filenames. The implementation requires the two files to have the same filename stem and to resolve to the same supported recipe. A missing counterpart, a filename mismatch or a recipe mismatch is therefore rejected before temporal alignment. This pairing step prevents sensor channels from different depositions from being combined when files from many users and experiments are present in the historical archive.

The paired tables are cleaned before alignment by trimming the recorded column names, standardizing the pressure-side `Pressure Time` and heater-side `Heater Time` fields to a common `Time` label and removing the redundant `Loop` field. The heater-side column names are kept distinct when the two tables are combined so that identically named metadata fields cannot overwrite one another. These operations standardize the log structure while preserving the measured sensor values and the pressure-side recipe and cycle-state information used by the downstream workflow.

The cleaned logs are then synchronized because their sensor channels are not recorded on a common time grid. Pressure time is used as the reference time base for the synchronized run record. Chamber pressure is copied directly on this grid, and `Cycles Remaining` and recipe information are carried from the pressure side. Heaters 6-17 are linearly interpolated from the heater/equipment time base onto the pressure timestamps and rounded to two decimal places. MFC1 is aligned from its own MFC time using nearest-neighbor matching and is likewise rounded to two decimal places. The result is one synchronized multivariate record in which pressure, heater, MFC and run-state information refer to a common sequence of time points.

This alignment is performed before the two downstream representations are formed. For the raw sensor LSTM branch, the synchronized pressure, Heater 6-17 and MFC1 signals provide the multivariate within-run waveform that is subsequently resampled to the fixed model input length described in Section 4. For the metric-based branch, the same synchronized record is reduced through deterministic signal processing routines into physically interpretable run level descriptors. Using the same paired and aligned source record for both branches ensures that their anomaly evidence is derived from the same underlying execution of the ALD run even though the two models operate at different temporal scales.

C.3 Run Eligibility and Dataset Construction

A logged file is not automatically treated as an eligible completed run sample. First, a corresponding pressure/heater log pair must be available for the same timestamped run and must belong to one of the two supported recipes. Run completion is then determined from the number of remaining cycles field in the pressure log. In the data used in this study, a run is considered completed when the minimum recorded value of number of remaining cycles reaches 1. A run for which this condition is not satisfied is treated as incomplete or manually stopped rather than as a normal completed deposition.

Incomplete runs are retained as part of the operational record rather than being interpreted as equipment faults solely because they ended early. However, they are not included in the same recipe baseline population used to characterize normal completed depositions and they do not proceed to the physics-based deposition outcome simulation. This distinction is important in a shared lab-scale facility, where a user may intentionally terminate an experiment for reasons that are unrelated to tool health.

After pairing, synchronization and the completion check, eligible runs are organized chronologically within their respective recipe histories. The historical modeling dataset used in this work spans 2021-2025 for both Thermal HfO_2 at 200°C and Thermal Al_2O_3 at 200°C. Separate recipe-specific histories are then used to construct the raw sensor and metric-based model inputs described in Section 4. This organization preserves the chronological evolution of each recipe while preventing intentional recipe-to-recipe differences from being treated as anomalous equipment behavior.

The historical archive extends into 2025, but model training is truncated at a fixed chronological cutoff before the November 2025 evaluation sequence described in Section 5.2. The five incoming runs used for prospective evaluation are the first five chronologically subsequent eligible runs after this cutoff and are not included in model training.

D PROCESS-RELEVANT METRIC EXTRACTION FROM ALD SENSOR LOGS

The synchronized run record described in Appendix C.2 is converted into a compact set of physically interpretable descriptors by deterministic signal processing routines. The purpose of this stage is not to assign a learned anomaly label but to summarize the pressure, thermal, carrier gas and cycle completion behavior of each run in a reproducible form. Each aligned run therefore produces one run-level feature row that can be used by the deterministic monitoring layer, the metric-based LSTM autoencoder and, for qualified runs, the physics-based simulation. Figure 7 illustrates the principal quantities extracted from a representative pressure trace together with the simultaneously aligned heater and MFC signals.

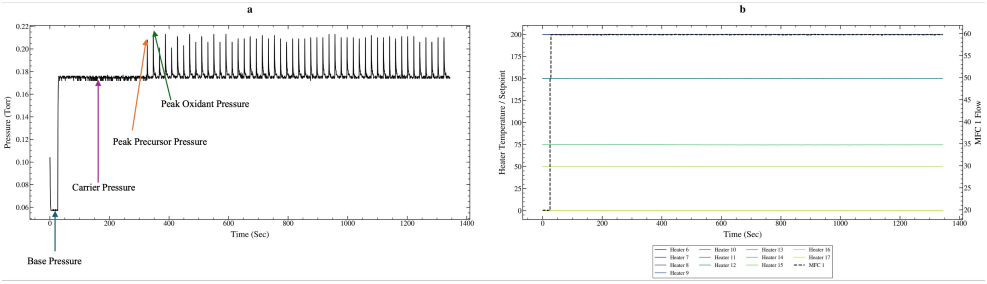


Figure 7: Example of (a) pressure data and (b) heater and MFC data from the ALD tool. The pressure illustrates the extracted base pressure, carrier pressure, precursor peak and oxidant peak quantities.

D.1 PRESSURE-RESPONSE METRICS

The pressure extraction routine first converts `Time`, `Pressure` and `Cycles Remaining` to numerical values and partitions the run whenever `Cycles Remaining` changes. This segmentation follows the actual cyclic process sequence without assuming a fixed sampling interval or a fixed segment duration. Let the duration of internal cycle segment s be T_s. A run-specific characteristic separation between the two dominant pressure responses is estimated as

$$\Delta T = \frac{1}{N_s} \sum_{s=1}^{N_s} \frac{T_s}{2}, \tag{1}$$

where N_s is the number of valid internal segments used in the estimate.

Within each segment, sample i is treated as a candidate local pressure maximum when

$$p_i > p_{i-1}, \qquad p_i \geq p_{i+1}. \tag{2}$$

The candidates are ranked by pressure amplitude. Beginning with the largest candidate, the algorithm searches for a second remaining maximum whose temporal spacing is consistent with the characteristic separation,

$$|(t_j - t_i) - \Delta T| \leq \max(0.25\Delta T,\ 0.50\,\mathrm{s})\,. \tag{3}$$

For an accepted pair, the earlier response is assigned to the precursor pulse and the later response to the oxidant pulse. In the two recipes analyzed here, the precursor is TDMAH for HfO_2 and TMA for Al_2O_3, while H_2O is the oxidant. An unmatched final peak is not used to calculate the precursor-to-oxidant interval. Across the valid cycle pairs, the routine reports the mean and median precursor peak pressure, the mean and median H_2O peak pressure and the mean precursor-to-oxidant time

interval. The latter is retained as the purge-related interval (`Mean Purge Time`) supplied to the physics-based simulation.

Base pressure is extracted from the initial low pressure portion of the trace before the first sustained pressure rise. A centered 0.20 s rolling median, denoted $\widetilde{p}(t)$, suppresses isolated noise. If $P_{90} - P_{10}$ is the run-level pressure range and Δt_{med} is the median sampling interval, the adaptive derivative threshold is

$$\gamma = \max\left[0.010,\ 0.10\frac{P_{90} - P_{10}}{\Delta t_{\mathrm{med}}}\right]. \tag{4}$$

The first point at which the derivative of the filtered trace exceeds γ and the pressure is at least 0.002 Torr above the preceding median level marks the end of the base pressure region. If a reliable rise is not found, the algorithm uses the first five samples or the first 1% of the trace, whichever is larger. `Base Mean` and `Base Median` are then calculated from the original pressure values within this region.

Carrier pressure is obtained from the open intervals between consecutive detected precursor and oxidant peaks. A mean and median are calculated for each interval and then combined over the run to obtain `Carrier Mean` and `Carrier Median`. To reduce sensitivity to a uniform shift in absolute chamber pressure, three dimensionless descriptors are also formed using the carrier median as the reference:

$$R_{\mathrm{base}} = \frac{P_{\mathrm{base,med}}}{P_{\mathrm{carrier,med}}}, \qquad R_{\mathrm{ox}} = \frac{P_{\mathrm{ox,med}}}{P_{\mathrm{carrier,med}}}, \qquad R_{\mathrm{pre}} = \frac{P_{\mathrm{pre,med}}}{P_{\mathrm{carrier,med}}}. \tag{5}$$

These pressure and dimensionless descriptors preserve the physical relationship among pumpdown, carrier flow and the two ALD dose responses while compressing a long pressure trace into run-level descriptors.

D.2 Heater, MFC and cycle-completion metrics

For Heater 6 through Heater 17, the mean, median, first valid value and last valid value are calculated over the synchronized run. These statistics capture both the central thermal state and any net change from the beginning to the end of the deposition. The full set is retained in the processed metrics table, while the metric-LSTM model uses the recipe-validated subset specified in Appendix E.2.

The MFC1 trace is summarized around its principal upward transition. Let m_i be the aligned MFC signal and $\Delta m_i = m_i - m_{i-1}$. The largest positive difference is accepted as a genuine jump only if

$$\max_i(\Delta m_i) > \max[3\,\sigma_{\Delta m},\ 0.02\,(m_{\max} - m_{\min})]\,, \tag{6}$$

where $\sigma_{\Delta m}$ is the standard deviation of the first differences. When a valid jump is detected, the routine reports the mean and median MFC values before and after the transition together with the first and last MFC values. When no jump satisfies Eq. 6, the full trace is treated as the pre-jump region.

E Complementary LSTM Autoencoder Models

The two unsupervised models use the same general encoder-decoder principle but operate on different representations and temporal scales. As summarized in Figure 8, the metric model reconstructs windows of consecutive runs to expose gradual changes in tool behavior, whereas the raw-sensor model reconstructs the within-run waveform of one run to expose abrupt run-level departures. Both models are trained separately for Thermal HfO_2 at 200°C and Thermal Al_2O_3 at 200°C so that intentional recipe-dependent differences do not define the anomaly score.

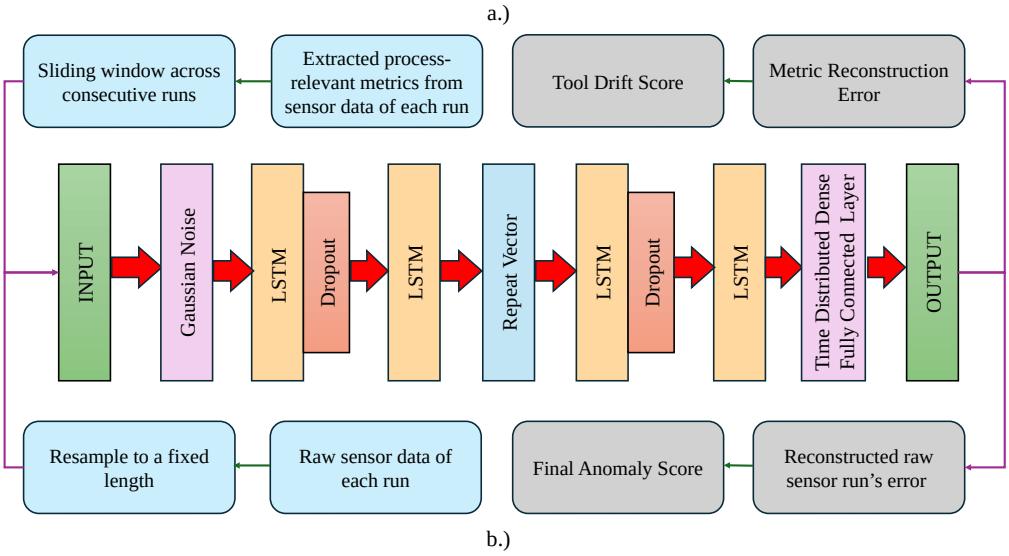


Figure 8: Complementary LSTM autoencoder strategies used in ALDTwin: (a) the metric-based sliding-window model maps reconstruction error to a tool-drift score across consecutive runs (b) the raw sensor model reconstructs a fixed length multivariate trace to obtain an isolated run-level anomaly score.

E.1 LSTM AUTOENCODER FORMULATION

For an input vector $\mathbf{x}_t$ at sequence position t, an LSTM cell maintains a hidden state $\mathbf{h}_t$ and memory state $\mathbf{c}_t$. Using input, forget and output gates $\mathbf{i}_t$, $\mathbf{f}_t$ and $\mathbf{o}_t$, the recurrence can be written as

$$\mathbf{i}_t = \sigma(W_i\mathbf{x}_t + U_i\mathbf{h}_{t-1} + \mathbf{b}_i)\,, \tag{7}$$
$$\mathbf{f}_t = \sigma(W_f\mathbf{x}_t + U_f\mathbf{h}_{t-1} + \mathbf{b}_f)\,, \tag{8}$$
$$\mathbf{o}_t = \sigma(W_o\mathbf{x}_t + U_o\mathbf{h}_{t-1} + \mathbf{b}_o)\,, \tag{9}$$
$$\widetilde{\mathbf{c}}_t = \tanh(W_c\mathbf{x}_t + U_c\mathbf{h}_{t-1} + \mathbf{b}_c)\,, \tag{10}$$
$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \widetilde{\mathbf{c}}_t, \tag{11}$$
$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t), \tag{12}$$

where $\sigma(\cdot)$ is the logistic sigmoid and $\odot$ denotes element-wise multiplication. The encoder maps an input sequence X to a latent representation and the decoder reconstructs it,

$$\widehat{X} = D(E(X)). \tag{13}$$

For a sequence with T positions and F features, the reconstruction objective is mean-squared error (MSE),

$$\mathcal{L}_{\text{MSE}}(X, \widehat{X}) = \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F}\left(X_{t,f} - \widehat{X}_{t,f}\right)^2. \tag{14}$$

A high reconstruction error therefore indicates that the input is poorly represented by the temporal structure learned from the historical same recipe data. It is interpreted as anomaly evidence rather than as a confirmed equipment fault.

E.2 METRIC-BASED LSTM AUTOENCODER FOR TOOL-DRIFT SCORING

Each completed same recipe run is first represented by 39 features. The feature vector consists of (i) elapsed time since the previous same recipe run, (ii) seven pressure/ratio descriptors – median precursor peak, median H_2O peak, base median, carrier median, base ratio, H_2O ratio and precursor ratio, (iii) the median, first and last values for Heater 6, 7, 8, 9, 10, 11, 14, 16 and 17, and (iv)

four MFC descriptors – median before the detected jump, median after the jump, first value and last value. Thus,

$$F_{\text{metric}} = 1 + 7 + (9 \times 3) + 4 = 39. \tag{15}$$

Missing feature rows are excluded from model training. The remaining runs are sorted chronologically within a recipe and split chronologically into an 80% training partition and a 20% validation partition. A MinMax scaler is fitted only on the training partition and is then frozen. A fixed number of run windows are formed separately inside the training and validation partitions, giving model inputs $X_i^{(m)} \in \mathbb{R}^{10\times 39}$.

The metric autoencoder applies Gaussian noise with standard deviation 0.01, followed by an 8-unit LSTM, 30% dropout, a 4-unit latent LSTM, a ten-step repeat vector, a 4-unit decoder LSTM, 30% dropout, an 8-unit decoder LSTM and a time-distributed dense output layer with 39 units. The LSTM layers use $\tanh$ activation and L_2 kernel and recurrent regularization of 10^{-3}. Training uses Adam with learning rate 10^{-3}, MSE loss, batch size 32 and a maximum of 150 epochs; early stopping monitors validation loss with patience 10 and restores the best weights.

One scalar reconstruction error is first calculated for each ten-run window using Eq. 14. Let e_i denote the error of window i. Because adjacent windows overlap, a run can occur in several windows. The metric anomaly score assigned to run r is therefore

$$S_r^{(m)} = \frac{1}{|\mathcal{I}_r|} \sum_{i\in\mathcal{I}_r} e_i, \qquad \mathcal{I}_r = \{i : \text{window } i \text{ contains run } r\}. \tag{16}$$

This mapping makes the score sensitive to how a run fits within the neighboring same recipe temporal context rather than only to the run in isolation. A recent run does not yet participate in all windows that will eventually contain it, so its score is marked provisional. For a ten-run window, the score becomes final only after nine later valid same-recipe runs have arrived. Once final, it is protected from revision for the same immutable model version and becomes eligible for the forecasting layer in Appendix F.

E.3 Raw sensor LSTM autoencoder for isolated run-level abnormalities

The raw model operates directly on the synchronized sensor waveform and does not use the 39-feature summary. Its 14 input channels are chamber pressure, Heater 6 through Heater 17 and MFC1. Number of remaining cycles is used only as an eligibility check and is not supplied as a learned feature. A run is eligible for raw model scoring when the first recorded cycle value is at least 5 and the final cycle value is 1.

For each eligible run, time is converted to a numeric sequence, invalid time points are removed, the data are sorted and duplicate time points are removed. The run time is then mapped to a normalized interval and every sensor channel is converted to numeric form and internally filled/interpolated when valid samples exist. Each channel is finally linearly resampled onto a common 300-point normalized grid, producing $X_r^{(s)} \in \mathbb{R}^{300\times 14}$. The chronological 80/20 training/validation split is applied at the run level. A MinMax scaler is fitted only to the flattened time points of the training runs and is then reused unchanged for validation and later incoming runs.

The raw autoencoder applies Gaussian noise with standard deviation 0.01, followed by a 32-unit LSTM, 30% dropout, a 16-unit latent LSTM, a 300-step repeat vector, a 16-unit decoder LSTM, 30% dropout, a 32-unit decoder LSTM and a time-distributed dense layer with 14 outputs. As in the metric model, the LSTM layers use $\tanh$ activation with L_2 kernel and recurrent regularization of 10^{-3}. Training uses Adam with learning rate 10^{-3}, MSE loss, batch size 16 and at most 150 epochs with validation-loss early stopping (patience 10); chronological sequence order is preserved during fitting by disabling shuffle.

For an eligible run r, the final raw-sensor anomaly score is the reconstruction MSE across all 300 time points and 14 sensor channels,

$$S_r^{(s)} = \frac{1}{300\times 14} \sum_{t=1}^{300} \sum_{f=1}^{14} \left(X_{r,t,f}^{(s)} - \widehat{X}_{r,t,f}^{(s)} \right)^2. \tag{17}$$

Channel-wise reconstruction errors can also be retained for inspection. Unlike the metric score, this run-level score is final immediately after a valid run is reconstructed because it does not depend on

later overlapping windows. The workflow does not impose a calibrated fault threshold on this score; it is supporting evidence that complements, but does not override, the authoritative deterministic run qualification.

F Local Level Kalman Forecasting of Tool Drift Scores

The forecasting layer operates only on finalized metric-based LSTM tool drift scores. Provisional scores are excluded because the overlapping window attribution in Eq. 16 can still change until nine later same-recipe runs have arrived. Forecast state is also isolated by recipe and immutable metric-model version so that scores produced by different model versions are never combined into one filter state.

Let $s_t \geq 0$ denote the finalized metric score of the tth eligible same-recipe run received by the forecaster. To stabilize the scale, the observation is

$$y_t = \log(s_t + \epsilon), \qquad \epsilon = 10^{-12}. \tag{18}$$

The implemented model is a one-state local-level Kalman filter,

$$\ell_t = \ell_{t-1} + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, Q), \tag{19}$$

$$y_t = \ell_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, R), \tag{20}$$

where ℓ_t is the latent log-score level. No slope or acceleration state is propagated. This choice intentionally avoids extrapolating a persistent trend from the weakly labeled anomaly score sequence and instead tracks the evolving local level as new finalized observations arrive.

The recipe-specific noise scale is initialized robustly from historical finalized scores. For $z_t = \log(s_t + \epsilon)$ and first differences $d_t = z_t - z_{t-1}$, the robust scale is

$$\widehat{\sigma}_d = \max\left[1.4826\,\mathrm{median}(|d_t - \mathrm{median}(d)|)\,,\ 10^{-3}\right]. \tag{21}$$

With $V = \widehat{\sigma}_d^2$, the implementation uses

$$R = \max(V, 10^{-6}), \qquad Q = \max(100V, 10^{-8}), \qquad P_0 = \max(V, 10^{-4}), \tag{22}$$

where P_t denotes the uncertainty of the filtered level.

For a new observation after initialization, the one-step prediction and update are

$$\ell_t^- = \ell_{t-1}, \tag{23}$$

$$P_t^- = P_{t-1} + Q, \tag{24}$$

$$\nu_t = y_t - \ell_t^-, \tag{25}$$

$$K_t = \frac{P_t^-}{P_t^- + R}, \tag{26}$$

$$\ell_t = \ell_t^- + K_t \nu_t, \tag{27}$$

$$P_t = (1 - K_t) P_t^-. \tag{28}$$

The first finalized score initializes the level directly. For an h-run open-loop forecast from the current state,

$$\widehat{y}_{t+h|t} = \ell_t, \qquad \mathrm{Var}(y_{t+h} | \mathcal{F}_t) = P_t + hQ + R. \tag{29}$$

Thus the point forecast of a local level model is constant over the open-loop horizon while uncertainty increases with the number of unseen finalized score arrivals. The reported 95% interval is constructed on the log scale as $\widehat{y} \pm 1.96\sqrt{\mathrm{Var}}$ and mapped back to anomaly-score space using

$$\widehat{s} = \max\left(e^{\widehat{y}} - \epsilon, 0\right). \tag{30}$$

ALDTwin exposes two interpretations of this same filter state. In the open-loop mode, the model reports fixed 1-, 5- and 10-finalized-run horizons without assimilating future scores. In rolling one-step mode, it predicts the next finalized score before that score is observed, records the prediction and its interval and then applies Eq. 28; repeating this procedure gives sequential one-step-ahead forecasts. These forecasts are interpreted as short horizon projections of the metric-based LSTM tool drift scores.

G PHYSICS-BASED ALD KINETIC SIMULATION

The physics-based component of ALDTwin uses a calibrated surface-kinetic formulation written generically in terms of a metal-containing precursor, denoted precursor A, and a co-reactant/oxidant, denoted species B. In the two thermal ALD recipes considered in this work, precursor A corresponds to tetrakis(dimethylamido)hafnium (TDMAH) for HfO_2 deposition and trimethylaluminum (TMA) for Al_2O_3 deposition, while H_2O serves as species B in both cases. Each chemistry is represented by the same four-stage ALD sequence - precursor pulse, first purge, co-reactant pulse and second purge, with a surface state that evolves continuously through the stages and from one cycle to the next. The kinetic parameters are calibrated separately for the material/chemistry under consideration, allowing the same modeling methodology to be applied to both recipes in this study and, after chemistry-specific parameter identification, to other precursor/co-reactant combinations.

G.1 SURFACE STATE AND CHEMISTRY-SPECIFIC PARAMETERIZATION

The model tracks the fractional coverage of metal-fragment-occupied surface sites,

$$\theta_M \in [0, 1], \tag{31}$$

with surface site density s_0. Across the four stages, the same surface balance is written in the general form

$$s_0 \frac{d\theta_M}{dt} = \text{gain} - \text{loss}. \tag{32}$$

The parameterization combines four sources of information. Pulse time, purge time, cycle count and nominal temperature/pressure settings are obtained directly from the process recipe. The incident molecular fluxes, flux_A and flux_B, are derived from the measured pressure responses associated with precursor A and species B. The surface site density s_0 and growth-per-cycle calibration are obtained from ex-situ metrology using post-deposition film measurements together with the corresponding material properties. Finally, the reaction/sticking probabilities β_A and β_B and the effective evaporation/removal fluxes $\Phi_{\text{ev},A}$ and $\Phi_{\text{ev},B}$ are determined from time-resolved in-situ metrology. These quantities are chemistry dependent and are therefore calibrated independently for the HfO_2 and Al_2O_3 process systems rather than transferred unchanged from one precursor/material combination to another.

G.2 GENERIC FOUR-STAGE PRECURSOR/CO-REACTANT SURFACE-COVERAGE MODEL

During the precursor A pulse, precursor molecules adsorb on the fraction of surface sites that remain unoccupied while already adsorbed species may be removed. The governing equation is

$$s_0 \frac{d\theta_M}{dt} = \text{flux}_A \, \beta_A (1 - \theta_M) - \Phi_{\text{ev},A} \theta_M. \tag{33}$$

Here flux_A is the incident molecular flux of the selected precursor, β_A is its sticking probability per collision, and $(1 - \theta_M)$ is the fraction of open surface sites. This open-site factor produces the self-limiting adsorption behavior characteristic of ALD. In the present study, this stage corresponds to a TDMAH pulse for HfO_2 or a TMA pulse for Al_2O_3.

During the first purge, the precursor valve is closed and no additional precursor flux arrives. The surface coverage therefore obeys

$$s_0 \frac{d\theta_M}{dt} = -\Phi_{\text{ev},A} \theta_M, \tag{34}$$

with the analytical decay

$$\theta_M(t) = \theta_0 \exp\left(-\frac{\Phi_{\text{ev},A} t}{s_0} \right). \tag{35}$$

The recipe-defined purge duration determines how far this removal process proceeds before the co-reactant pulse begins.

During the species-B pulse, the co-reactant reacts with the precursor-derived surface species and consumes the metal-fragment-occupied sites. For the oxidant-driven conversion used in the present thermal oxide processes, the governing equation is

$$s_0 \frac{d\theta_M}{dt} = -\text{flux}_B \, \beta_B \theta_M, \tag{36}$$

where flux_B is the incident co-reactant molecular flux and β_B is the corresponding reaction probability per collision. In both recipes studied here, species B is H_2O; the chemistry-specific reaction converts the adsorbed precursor-derived surface species toward the corresponding metal oxide.

During the second purge, the co-reactant valve is closed and the remaining co-reactant-related surface species are removed according to

$$s_0 \frac{d\theta_M}{dt} = -\Phi_{\mathrm{ev},B}\theta_M. \tag{37}$$

The resulting θ_M is carried directly into the precursor pulse of the next cycle; the surface state is not reset at a cycle boundary. Sequential integration of the four stages therefore preserves the physical memory of incomplete adsorption, reaction, or purge removal from one stage and cycle to the next.

G.3 Sensor-derived molecular flux and kinetic calibration

The measured pressure traces provide the connection between equipment sensor data and the incident molecular fluxes used in the kinetic equations. For either gas species $i \in \{A, B\}$, with measured pressure-bump area $A_{P,i}$, pulse duration $t_{p,i}$, molecular mass M_i, process temperature T, and Boltzmann constant k_B, the Hertz–Knudsen relation used in the formulation is

$$\mathrm{flux}_i = \frac{(A_{P,i}/t_{p,i})\,133.322}{\sqrt{2\pi M_i k_B T}}, \qquad i \in \{A, B\}, \tag{38}$$

where the factor 133.322 converts Torr to Pa. The precursor and co-reactant pressure responses are therefore converted separately into flux_A and flux_B, with the appropriate molecular mass and pulse conditions for the selected chemistry.

The remaining kinetic parameters are obtained through chemistry-specific metrology calibration. Saturated endpoint ex-situ measurements provide the calibration for s_0 and growth per cycle, while time-resolved in-situ measurements are used to resolve β_A, β_B, $\Phi_{\mathrm{ev},A}$, and $\Phi_{\mathrm{ev},B}$. In particular, the pulse-rise response constrains the corresponding reaction/sticking probability when the removal term is known, whereas the purge decay constrains the effective evaporation/removal flux through Eq. 35. These time-resolved measurements are required because the reaction and removal parameters cannot in general be separated using only the programmed recipe, the pressure trace, and a single ex-situ thickness endpoint.

For each ALD chemistry, ALDTwin therefore uses the same modeling sequence but a separately calibrated parameter set: recipe-defined timing and cycle information, sensor-derived molecular fluxes and material/chemistry-specific kinetic and growth parameters are propagated through the four stage ODE system for the prescribed number of cycles. In this paper, the formulation is instantiated for TDMAH/H_2O HfO_2 and TMA/H_2O Al_2O_3. The same workflow can be extended to a different precursor/co-reactant pair by replacing the chemistry-dependent molecular properties and recalibrating the corresponding surface-kinetic parameters, with the stage-specific reaction model revalidated when the underlying surface chemistry changes.

H Additional Model Diagnostics and Tool Drift Change

Section 5.1 reports the historical anomaly and tool drift trajectories produced by the two complementary LSTM autoencoders. This appendix provides the corresponding training diagnostics and an additional first difference view of the metric-based tool drift score. The training and validation histories are used to verify that the autoencoders reached stable reconstruction solutions before their reconstruction errors were interpreted as anomaly evidence.

H.1 Raw sensor LSTM training convergence

Figure 9 shows the training and validation mean squared error (MSE) histories for the raw sensor LSTM autoencoders. For both recipes, the loss decreases rapidly during the initial epochs and then progressively approaches a low, stable value. The training and validation curves remain closely aligned throughout training, including after the large initial reduction in error. The validation trajectory does not exhibit a sustained increase while the training loss continues to decrease which

clearly shows that our model is not overfitting. Instead, the two curves converge to comparable reconstruction-error levels, supporting the use of the frozen models for evaluating previously unseen run traces. Minor local fluctuations, such as the late-epoch variation visible for the Al_2O_3 model, do not develop into a persistent train–validation divergence.

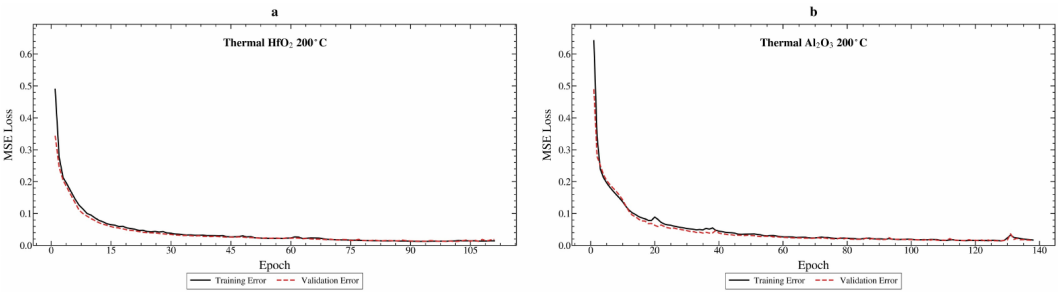


Figure 9: Training and validation reconstruction loss for the raw sensor LSTM autoencoder for (a) Thermal HfO_2 at 200 °C and (b) Thermal Al_2O_3 at 200 °C.

H.2 Metric-based LSTM training convergence and change in tool-drift score

The metric-based LSTM autoencoders show the same overall convergence behavior. As shown in Figure 10, the reconstruction loss decreases sharply over the early epochs and then decays smoothly toward a stable value. The validation curves remain close to the training curves for both recipes, with a small and stable separation after convergence rather than a growing generalization gap. Together with validation loss-based early stopping and restoration of the best weights described in Appendix E.2, this behavior indicates that the learned multirun representation is not being driven by continued reduction of training error at the expense of validation performance.

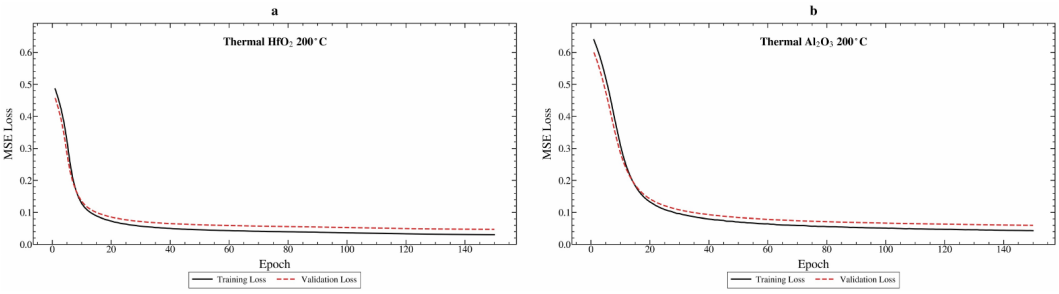


Figure 10: Training and validation reconstruction loss for the metric-based LSTM autoencoder for (a) Thermal HfO_2 at 200 °C and (b) Thermal Al_2O_3 at 200 °C.

The absolute tool drift score in Figure 4 measures how strongly a run differs from the learned same-recipe multirun behavior. To emphasize how rapidly that inferred equipment state changes from one run to the next, we additionally compute the first difference of the chronological tool drift score,

$$\Delta D_i = D_i - D_{i-1}, \tag{39}$$

where D_i is the metric-based tool drift score assigned to the ith same-recipe run. A positive value indicates an increase in the learned tool drift signal relative to the immediately preceding run, a negative value indicates a decrease and values close to zero indicate little run-to-run change in the score. Figure 11 shows that most consecutive-run changes are small, while localized excursions identify periods in which the inferred tool state changes more rapidly. These excursions complement the

slower baseline shifts and extrema visible in the absolute score. Because D_i is a reconstruction-error-derived indicator rather than a calibrated physical state variable, ΔD_i is interpreted as the magnitude and direction of change in the learned tool-drift evidence, not as a direct physical measurement of degradation or a diagnosis of a maintenance event.

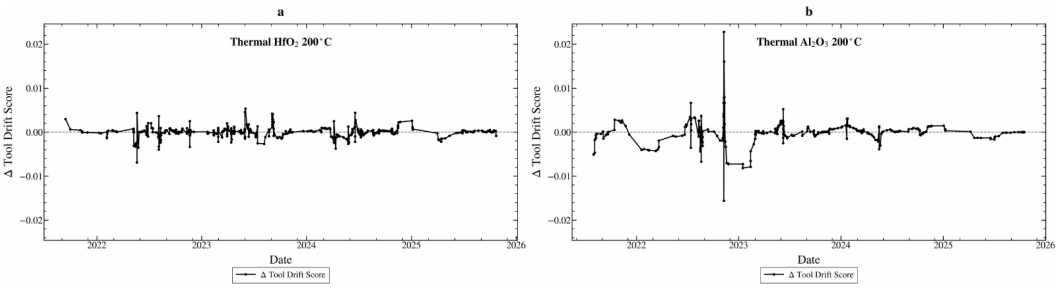


Figure 11: First difference of the metric-based LSTM tool-drift score, $\Delta D_i = D_i - D_{i-1}$, for (a) Thermal HfO_2 at 200 °C and (b) Thermal Al_2O_3 at 200 °C. Excursions away from zero highlight relatively rapid changes in the learned same-recipe tool drift signal.

H.3 Additional Kalman holdout statistics

The point-wise absolute percentage errors shown in Figure 5 are complemented by the aggregate statistics in Table 4. The rolling one-step interpretation reduces MAE and RMSE relative to the open-loop interpretation for both recipes because each newly finalized score is assimilated before the following prediction. All five held-out finalized scores for each recipe also fall within the corresponding 95% predictive intervals in this evaluation.

Table 4: Statistics of local-level Kalman forecasting for the two operating interpretations

Recipe	Mode	MAPE (%)	MAE	RMSE	95% coverage (%)
HfO_2	Open-loop	0.5171	0.00019796	0.00022886	100
HfO_2	Rolling one-step	0.3102	0.00011911	0.00012167	100
Al_2O_3	Open-loop	0.2002	0.00014052	0.00019390	100
Al_2O_3	Rolling one-step	0.1344	0.00009437	0.00011170	100

I Integrated Workflow Report Case Studies

Three incoming 2025 runs are used here to show the experimentalist-facing output of the integrated ALDTwin workflow under the three principal operational outcomes: a qualified normal run, an incomplete run and a completed run requiring engineer inspection. The cases were replayed chronologically from recorded pressure/heater pairs without modifying the underlying sensor values. These evaluation runs occur chronologically after the fixed training cutoff and were not used during model training. In every case, the authoritative run qualification, numerical model outputs, simulation permission and tool-state transition are computed before language generation. Gemini 3.8 Flash then receives this structured evidence and renders the final report in concise user-facing language for the experimentalist. Physics-based simulation output is included only when the workflow has already qualified the run for simulation. The LLM therefore formats and explains all three report types, but it does not create or override the underlying status, numerical values, interlocks or recommended operational action encoded by the monitoring workflow.

I.1 Qualified-normal report

The qualified example is a Thermal Al_2O_3 200 °C run completed on 14 November 2025 at 09:19. All authoritative completion and baseline checks pass, so the workflow classifies the run as normal

and allows the physics-based simulation. A representative experimentalist-facing report is shown below.

ALD Run Report

Status: NORMAL — completed and qualified
Recommended action: Continue routine operation.
The run completed successfully, passed the active same-recipe qualification checks and is eligible for physics-based simulation.

Run details

- **Run:** `2025_11_14-09-19_Thermal_Al2O3_200C.csv`
- **Date and time:** 14 November 2025, 09:19
- **Recipe:** Thermal Al_2O_3 at 200 °C

Current-run model evidence

- **Raw-sensor anomaly score: 0.290769 (final).** This is a contextual indicator derived from the recorded pressure, heater and MFC traces for the completed run.
- **Tool-drift score: 0.0699428 (provisional).** This provides run-to-run context for the evolving same-recipe tool state and becomes final after nine later valid same-recipe runs are available.

Next tool-drift forecast

The forecast target is the next finalized same-recipe tool-drift score rather than the current run. The open-loop forecast is **0.0704582**, and the rolling one-step forecast is also **0.0704582**. The values are identical at this origin because both modes start from the same current filter state; the rolling one-step mode updates after the next score is finalized.

Physics-based simulation output

- **Film:** Al_2O_3
- **Simulated cycles:** 50
- **Substrate temperature:** 200 °C
- **Purge time:** 10.1313 s
- **Estimated film thickness:** 5.48191 nm
- **Estimated growth per cycle:** 0.109701 nm/cycle

I.2 Incomplete-run report

The incomplete example is a Thermal HfO_2 200 °C run recorded on 7 November 2025 at 15:44. The cycle-completion check fails because `Cycles Remaining` ends at 0 whereas the expected completion condition for the monitoring workflow is 1. The report therefore suppresses the physics-based simulation but does not, from incompleteness alone, lock the tool for engineer inspection.

ALD Run Report

Status: INCOMPLETE — run stopped before the expected endpoint
Recommended action: Review the interruption. Engineer inspection is not required by this event alone, and the tool remains available for the next run.
The physics-based simulation is not executed because the process did not satisfy the run-completion requirement.

Run details

- **Run:** `2025_11_07-15-44_Thermal_HfO2_200C.csv`
- **Date and time:** 7 November 2025, 15:44
- **Recipe:** Thermal HfO_2 at 200 °C
- **Completion check:** `Cycles Remaining` ended at 0 rather than the expected value of 1; the cycle counter indicates a 100-cycle recipe.

Current-run model evidence

- **Raw-sensor anomaly score: not scored.** The run did not pass the completion requirement used by the raw-sensor model.
- **Tool-drift score: 0.0382416 (provisional).** The metric-based score is retained as longitudinal context and will finalize after nine later valid same-recipe runs.

Next tool-drift forecast

The next finalized same-recipe tool-drift score is predicted as **0.0388946** in both open-loop and rolling one-step operation at the current origin. As above, the rolling one-step mode changes only after the next finalized score is assimilated.

I.3 Inspection-required report

The inspection-required example is a completed Thermal Al_2O_3 200 °C run recorded on 14 November 2025 at 15:00. The current run exceeds the adaptive same-recipe limits for nine heater metrics, so the workflow suppresses simulation, changes the persistent tool state to require engineer inspection and pauses later queued runs until a disposition is recorded.

ALD Run Report

Status: INSPECTION REQUIRED — process checks exceeded the same-recipe baseline
Recommended action: Pause processing and have the responsible engineer inspect the heater behavior before releasing queued runs.
The physics-based simulation is not executed while the tool remains in the inspection-required state.

Run details

- **Run:** `2025_11_14-15-00_Thermal_Al2O3_200C.csv`
- **Date and time:** 14 November 2025, 15:00
- **Recipe:** Thermal Al_2O_3 at 200 °C

Why engineer inspection is required

Nine heater metrics fall outside the ranges established from the previous ten completed same-recipe runs:

- **Heater 7 median:** observed 199.43; expected 199.5502–200.4498.
- **Heater 7 final value:** observed 199.4; expected 199.5314–200.4686.
- **Heater 8 median:** observed 199.7; expected 199.7674–200.2326.
- **Heater 8 final value:** observed 199.8; expected 199.8079–200.1921.
- **Heater 9 median:** observed 199.405; expected 199.7096–200.2904.
- **Heater 9 final value:** observed 199.5; expected 199.6985–200.3015.

- **Heater 10 median:** observed 149.9; expected 149.937–150.063.
- **Heater 11 median:** observed 149.9; expected 149.937–150.063.
- **Heater 12 median:** observed 75.02; expected 75–75.

Current-run model evidence

- **Raw-sensor anomaly score: not scored.** The recorded segment began with fewer than five cycles remaining and therefore did not satisfy the raw-sensor model's eligibility rule.
- **Tool-drift score: 0.0698522 (provisional).** This remains contextual longitudinal evidence and finalizes after nine later valid same-recipe runs.

Next tool-drift forecast

The next finalized same-recipe tool-drift score is predicted as **0.0704019** in both open-loop and rolling one-step operation at the current origin.