Institut für Verfahrenstechnik
des industriellen Umweltschutzes
Montanuniversität Leoben
Vorstand: O.Univ.Prof.Dr.mont. W.L. Kepplinger

Peter-Tunner-Straße 15, A-8700 Leoben
Tel: (43) 3842-402-5001, Fax: (43) 3842-402-5002
E-Mail: vtiu@unileoben.ac.at
http://www.vtiu.com

Diplomarbeit

Entwicklung und Simulation einer „Run to Run“ Regelungsschleife für Ofenprozesse in der Halbleiterindustrie

erstellt für

Infineon Technologies Austria AG

Vorgelegt von:
Christian Fischer, 9835172

Betreuer/Gutachter:
O.Univ.Prof.Dr.mont. W.L. Kepplinger
Dipl.-Ing. Kurt Sorschag,Infineon

Leoben, 20 10 2004

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Diplomarbeit selbständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche erkenntlich gemacht habe.

DANKSAGUNG

DIESE ARBEIT WIDME ICH ALL JENEN, DIE MIR WÄHREND MEINES STUDIUMS UND BEI DER ERSTELLUNG DIESER ARBEIT HILFSBEREIT ZUR SEITE STANDEN, MICH TATKRÄFTIG UNTERSTÜTZT HABEN UND MIR ERMÖGLICHTEN, DIESES VORHABEN SO RASCH ZU BEWERKSTELLIGEN.

ICH DANKE EUCH ALLEN!

Kurzfassung

Entwicklung und Simulation einer „Run to Run“ Regelungsschleife für Ofenprozesse in der Halbleiterindustrie

Der Output von LPCVD Ofenprozessen in der Halbleiterindustrie wird von Zustandsveränderungen des Equipments beeinträchtigt. Prozessstörungen wie Drift, Sprünge und zufällige Störungen beeinflussen das Prozessergebnis. Eine Run to Run Regelung ist eine diskrete Prozesskontrolle, die zwischen den Runs die Prozessparameter modifiziert, um Prozessstörungen und mit ihnen die Kosten durch Verwurf zu minimieren. Diese Arbeit zeigt an Hand des DHOBE-Kontrollalgorithmus eine Möglichkeit wie die Störungen bei der Siliziumnitridabscheidung verringert werden und wie Prozesse mit einem nichtlinearen Input/Output Verhalten durch diesen Algorithmus beherrschbar werden.

In Simulationen wird der Kontrollalgorithmus auf sein Regelverhalten, seine Effektivität und seine Fähigkeit Prozesse zu optimieren hin untersucht. Experimente sollen zeigen, dass die Simulationsergebnisse auch im praktischen Einsatz bestätigt werden.

Darüber hinaus wird eine Methode zur empirischen Modellfindung bei LPCVD-Prozessen erläutert, die es ermöglicht von der Planung der Experimente über die Auswertung bis zur Erstellung eines Regelmodells standardisiert vorzugehen. Dieses Vorgehen ermöglicht es, Modelle für andere LPCVD-Prozesse zeitsparend zu erstellen und sie in einen Run to Run Kontroller zu implementieren.

Abstract

Development and simulation of a run to run controller for semiconductor furnace processes

The output of LPCVD furnace processes in the semiconductor manufacturing is affected by drifts, shifts and variability and these disturbances cause equipment changes. A run to run control is a discrete process controller which modifies the recipes between runs in order to eliminate these disturbances and thereby minimize the costs. This work will show a possibility how to decrease the disturbances with the DHOBE- Controller in case of silicon nitride deposition. It shows that the algorithm is able to control nonlinear Input/Output dependencies.

Simulations of the control algorithm are used to interpret its control characteristics, its effectiveness and its ability to optimize furnace processes. In experiments the results of the simulations are verified.

In addition, a method for empirical model building is mentioned and explained. This method gives one the possibility to plan, evaluate and build a control model for LPCVD-processes in a standardized way. It enables to build models for other LPCVD-processes and implement these models into a Run-to-Run-controller.

Inhaltsverzeichnis

Seite

1 Einleitung

1.1 Einführung in das Run to Run Thema

Die Thematik die diese Arbeit behandelt, ist ein grundlegendes Problem in der Halbleiterindustrie. Die Fähigkeit, bei der Fertigung von Halbleiterprodukten die Prozesse zu kontrollieren und dabei die gestellten Anforderungen zu erreichen, ist essentiell für eine gleich bleibende Qualität der Mikrochips. Durch die zunehmende Verkleinerung der Strukturen (Gate-Längen von unter 100 nm beim CMOS-Transistor) und der gleichzeitigen Größenzunahme der Wafer[1] auf 300 mm sind bessere und konstantere Prozesse notwendig.

Abbildung 1: 300 mm Wafer

Das Ziel, die Fertigungskosten so niedrig wie möglich zu halten, setzt gleich bleibend gute Prozesse voraus, da der prozentuelle Verwurf und damit auch die Kosten direkt mit der Stabilität der Einzelprozesse zusammenhängen. Auch das Einsetzen von Fertigungsprozessen für neue Technologien auf bestehenden Anlagen erfordert immer konstantere Prozesse mit gleich bleibendem Output. Verschiedenste Lösungen, die den härteren Anforderungen genügen, dazu gehören besseres Anlagendesign, Prozess Innovationen und Prozesstuning durch Prozessingeneure sind jedoch nicht länger ausreichend. Der Einsatz von kontinuierlichen Messsystemen und Regelkreisen verbessert die Produktqualität immer mehr und ist vielfach schon Stand der Technik. Das gemeinsame Ziel all dieser Verbesserungen ist die Verringerung der Prozessstreuung des Outputs.

[1] Bezeichnung der Siliziumträgerscheiben im prozessierten als auch unprozessierten Zustand mit derzeit üblichen Durchmessern von 150 – 300 mm

In der Halbleiterindustrie hat sich in den letzten Jahren ein reges Interesse für R2R-Regelungen gebildet. Ursprünglich wurden die ersten R2R-Regelungen in anderen technischen Bereichen (z.B. chemische Industrie) entwickelt. Der Einsatzbereich dort unterscheidet sich jedoch meist deutlich von dem in der Halbleiterindustrie, da dort über einen längeren Zeitraum auf einer Anlage nur ein Produkt erzeugt wird. Bei der Herstellung von Halbleiterprodukten ist es jedoch oftmals der Fall, dass auf einer Anlage mehrere Produkte bei unterschiedlichen Prozessbedingungen erzeugt werden. Diese Vorraussetzungen erfordern ganz andere R2R-Kontroller, die den jeweiligen Umständen in den einzelnen Fertigungsfabriken angepasst werden müssen.

Um die Änderung der Prozessbedingungen, hervorgerufen durch Verschleiß und Alterung des Equipments oder durch regelmäßige Wartungsarbeiten, zu erkennen und die dadurch auftretenden Schwankungen im Prozessoutput zu beheben, wird der Einsatz von Regelungen notwendig. Dieses Ziel kann durch mannigfaltige Ansätze erreicht werden. Wo möglich werden in situ Messungen, d.h. Messungen, die während eines Prozesses stattfinden, verwendet. Ein Beispiel dazu wäre eine optische Endpunkterkennung bei einem Chemical Mechanical Planarization (CMP) [1] Prozess. Dort wird ein prozessierter Wafer plan geschliffen mit dem Ziel eine möglichst ebene, über den ganzen Wafer gleich bleibende Restschichtdicke zu erhalten. Nicht überall können derartige Regelungen eingesetzt werden weil entweder aggressive Prozessbedingungen oder die Kosten für den Einbau eines solchen Messsystems dies verhindern. Hier können Run to Run Regelungen, die kostengünstig und effizient sind, eingesetzt werden.

R2R-Regelungen sind eine Form von diskreter Prozess- und Anlagenkontrolle, wobei das Rezept ex-situ zwischen den Runs verändert wird, um die Prozessschwankungen (Drift, Sprünge) zu minimieren. Prinzipiell werden die Messergebnisse vom vorigen Run oder von nach der vorigen Anlage dazu verwendet das Rezept für den nächsten Run zu erstellen. Bildlich gesprochen, lässt sich eine R2R Regelung mit dem Backen von Keksen vergleichen. Das erste Backergebnis ist vielleicht etwas zu dunkel geraten und vom Geschmack etwas zu süß. Darauf lässt man beim nächsten Versuch etwas Zucker weg und verringert die Backzeit oder die Temperatur. Es stellt sich heraus, dass zwar der Geschmack in Ordnung ist aber die Kekse diesmal nicht richtig goldbraun sind, was für den nächsten Versuch wieder zu einer Verlängerung der Backzeit führen könnte. Der Mensch agiert hier intuitiv wie eine R2R-Regelung, was letztendlich auf einem bestimmten Ofen zu einem guten Ergebnis führt.

Das gleiche Rezept auf einem anderen Ofen führt jedoch möglicherweise zu einem nicht so guten Ergebnis. Auf gleiche Art und Weise werden momentan auch in der Halbleiterindustrie die Rezepte bei einigen Anlagen von Einzelprozessingenieuren eingestellt. Was in vielen Fällen das Ergebnis sehr stark vom Erfahrungswert des Ingenieurs und von seinem Gefühl für den Prozess abhängig macht. Grundsätzlich unterscheidet man zwei Arten von R2R-Regelungen; die Feedback Regelung (Abbildung 2) und die Feedforward Regelung (Abbildung 3).

Diese beiden Abbildungen [2] zeigen in vereinfachter Form die grundsätzliche Kontrollstruktur.

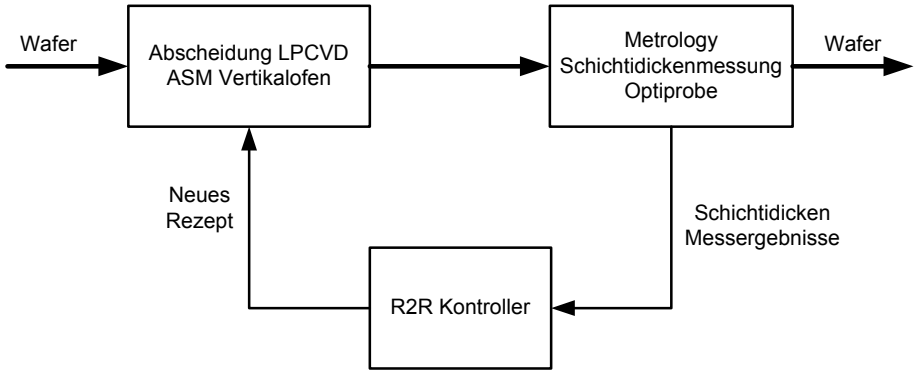

Abbildung 2: Feedback R2R Kontrollschema

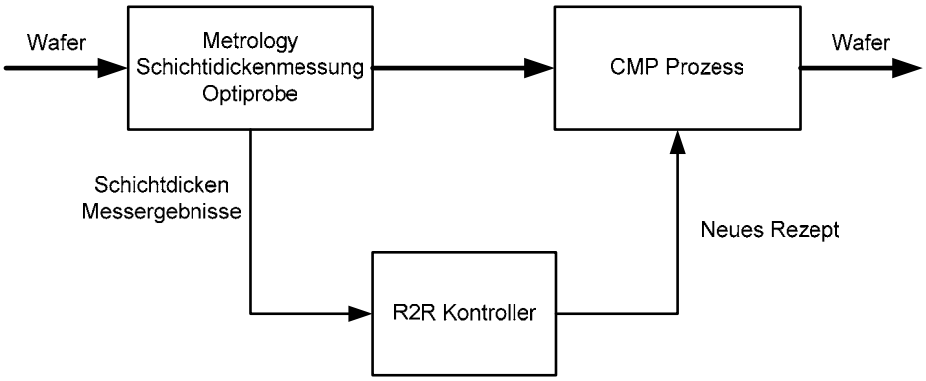

Abbildung 3: Feedforward R2R Kontrollschema

Der einfachere Fall ist die Feedforward Regelung. Hier wird aus den Werten einer vorangehenden Messung ein Rezept für die Weiterverarbeitung eines Batches[1]/Loses[2]/Wafer auf einer bestimmten Anlage erstellt. Bei der Feedback Regelung ist es notwendig, eine Vorhersage darüber zu machen, wie sich ein Prozess beim nächsten Run verhält. Die mathematischen Methoden zur Erstellung des Rezeptes für den nächsten Run aus den Daten der vorangegangenen Runs sind ebenso zahlreich wie komplex. Abgesehen von den in-situ real time Messungen (gehören nicht zu den R2R-Regelungen) verwendet man Techniken wie Expertensysteme, Fuzzy Logic, Neural Networks, Statistische Analyse,

[1] Bezeichnung für alle Wafer die gleichzeitig einen Prozess durchlaufen (Bei Ofenprozessen sind dies maximal 150 produktive Wafer)

[2] Bezeichnung für maximal 50 Wafer des gleichen Typs

Parameterschätzung und Modelbased Control. In Abhängigkeit von den Umständen der Applikation sind die einzelnen Methoden mehr oder weniger passend. Jede R2R-Methode die letztendlich verwendet wird, muss imstande sein mit Modellfehlern als auch mit unvorhersehbaren Störungen zurechtkommen.

1.2 Problemstellung

Bei der Erzeugung von Halbleiterprodukten ist es notwendig, leitende bzw. isolierende Schichten planar auf das Siliziumträgermaterial aufzubringen, um integrierte Schaltkreise zu fertigen. Dies kann durch verschiedenste Verfahren geschehen. Zwei dieser Verfahren zählt man zur Ofentechnik. Einerseits die thermische Oxidation, bei der Siliziumoxid in einer oxidierenden Atmospähre (Sauerstoff, Wasserdampf) bei Temperaturen > 800°C defektfrei hergestellt wird. Andererseits die Chemical Vapor Deposition (CVD) die als die Abscheidung eines Films durch die Zersetzung oder Reaktion von gasförmigen chemischen Stoffen an der heißen Siliziumscheibe definiert ist. Letztere kann bei atmosphärischen (APCVD) oder Niederdruck (LPCVD) Bedingungen stattfinden und ist Gegenstand dieser Arbeit.

Drei unterschiedliche LPCVD-Prozesse werden momentan in Villach auf ASM-Öfen betrieben.

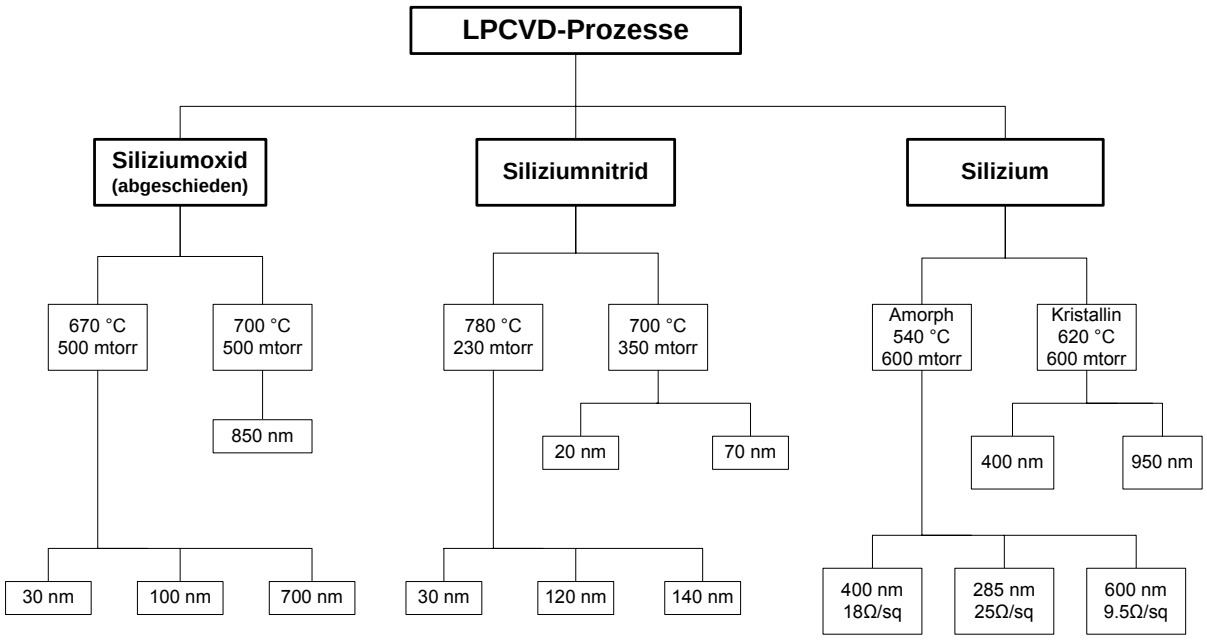

Abbildung 4: Unterteilung LPCVD-Prozesse für 8 Zoll Wafer

In Abbildung 4 sind diese Prozesse, die sich hinsichtlich Reaktionsgasen, Druck und Temperaturprofilen grundsätzlich voneinander unterscheiden, dargestellt. Jeder dieser drei Prozesse wird wiederum je nach gewünschter Schichteigenschaft, Uniformität und

Schichtdicke bei unterschiedlichen Gasflüssen, Temperaturen und Abscheidezeiten, die zu Rezepten zusammengefasst sind,auf einem oder mehreren Öfen gefahren. Die Siliziumoxidabscheidung wird auch TEOS-Abscheidung genannt, weil Tetraethylorthosilikat das Reaktionsgas ist. Die Siliziumnitridabscheidung wird zur Nitrid-Abscheidung, und die Siliziumabscheidung wird in der Halbleiterbranche Poly-Abscheidung genannt, weil früher polykristallines Silizium abgeschieden wurde (Namensgebung ist historisch bedingt). Heute wird mit dem LPCVD-Verfahren nur mehr amorphes Silizium abgeschieden.

Diese LPCVD-Prozesse werden unter anderem auf ASM-Öfen (Abbildung 5) prozessiert und erfahren während einer Rohrlebensdauer (ca. 500 Runs) starke Veränderungen, die durch Änderungen des Rohrzustandes hervorgerufen werden und Auswirkungen auf die Uniformität des Outputs haben. Schwankungen in der Boat-Down-Uniformity (Schwankungen in Strömungsrichtung über das Boot[1]) und in der R2R-Uniformität stellen ein erhebliches Problem dar, weil das manuelle Nachregeln der Öfen zeitaufwendig ist und die Ausbeute bei schlechter Uniformität sinkt.

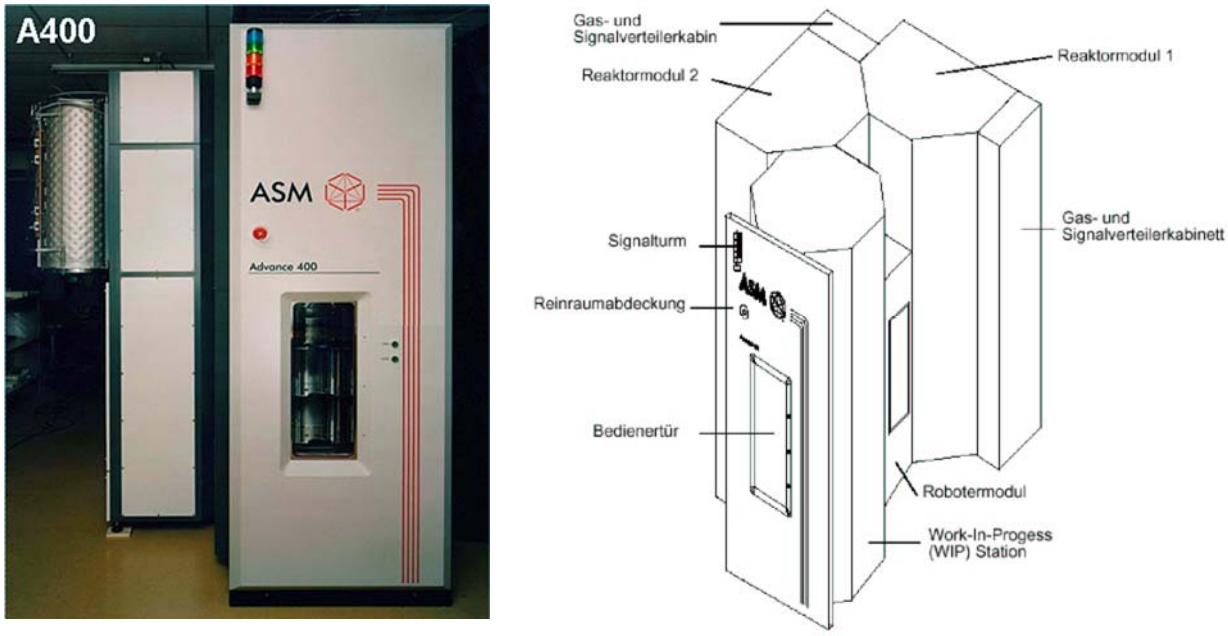

Abbildung 5: ASM A400 Ofen

Zusätzlich treten bei den einzelnen Runs nicht beeinflussbare Streuungen, so genanntes Rauschen, über den Wafer über das Boot und auch von Run zu Run auf. Auch diese

[1] Als Boot bezeichnet man einen Quarzgestell in dem bis zu 150 produktive Wafer platz haben. Es dient zum Ein- und Ausfahren der Wafer ins bzw. aus dem Rohr.

zufälligen Schwankungen müssen bei der Wahl eines R2R-Kontrollers berücksichtigt werden.

1.3 Zielsetzung und Durchführung

Aufgabe dieser Arbeit ist es eine R2R-Methode zu finden, mit der sich eine messbare Performanceverbesserung der Ofenprozesse erzielen lässt. Dies ist für einen Prozess mit Hilfe einer Simulation des R2R-Algorithmus und anschließenden Testfahrten zu beweisen. Es soll gezeigt werden, dass eine Verbesserung der R2R-Uniformity und der Boat-Down-Uniformity mit einer R2R-Regelung möglich ist. Der Algorithmus muss auf die unterschiedlichen LPCVD-Prozesse (TEOS, Nitrid, Poly) anwendbar sein und einen stabilen Betrieb gewährleisten. Des Weiteren soll er flexibel genug sein, um einerseits bei unterschiedlichen Messhäufigkeiten (d.h. die Anzahl der Scheiben die pro Ofenfahrt gemessen werden) und andererseits nach wiederkehrenden Wartungsarbeiten (Rohrwechsel, Bootwechsel,…) betrieben werden zu können. Nach Wartungsarbeiten soll der Algorithmus rasch eine mögliche Zustandsänderung des Ofens erkennen und beheben. Grundsätzliche Anforderungen, wie Stabilität und rasche Konvergenz müssen von der R2R-Regelung erfüllt werden.

In der Experimentalphase soll unter Anwendung eines DoE (Design of Experiment) ein Input/Output-Modell, das die Grundsätzlichen Abhängigkeiten

$$Schichtdicken = SD_z = f\left(T_x, t, p, \dot{V}, z(Position)\right) \tag{1}$$

eines Prozesses beschreibt, erstellt werden. Hierin sind die Schichtdicken abhängig von den Zonentemperaturen, der Abscheidezeit, dem Prozessdruck, den Volumenströmen der Reaktionsgase und der Position im Boot. Die Versuchsreihen sind mit dem hauseigenen Statistikpaket CEDA/Cornerstone auszuwerten und in passender Form als Grundlage für die Simulation zu verwenden. An Hand dieses Modells soll ein R2R-Algorithmus entwickelt, simuliert, verifiziert und beurteilt werden. Für die Simulationserstellung und Programmierung soll Matlab verwendet werden, damit in späterer Folge daraus eine Stand-Alone-Application, die modular aufgebaut ist, generiert werden kann. Eine Implementierung des Algorithmus in einen R2R-Kontroller muss möglich sein.

Der Vielfalt der Prozesse auf einem Vertikalofen und die Tatsache, dass diese Prozesse bei unterschiedlichen Bedingungen gefahren werden ist bei der Wahl des Kontrollalgorithmus Rechnung zu tragen. Darüber hinaus soll eine standardisierte Methode bei der Erstellung des Modells verwendet werden, um eine Adaptierung des Algorithmus auf andere LPCVD-Prozesse mit vertretbarem Aufwand zu ermöglichen.

2 Grundlagenteil und Prozessbeschreibung

2.1 Chemical Vapour Deposition (CVD)

Bei der chemischen Gasphasenabscheidung [3] werden gasförmige Komponenten in einem Reaktionsraum über die heiße Substratoberfläche geleitet und reagieren dort zu einem festen Film. Die Zusammensetzung der Schichten – und damit ihre Eigenschaften – werden durch die gezielte Steuerung der chemischen Reaktionen erzielt. Die Steuerung dieser Reaktionen kann über relativ wenige Parameter (Temperatur, Druck, Gasflüsse) erfolgen.

Vorteile dieses Verfahrens sind

- die vielseitigen Möglichkeiten bei der Art von dünnen Filmen die abgeschieden werden können (Siliziumoxide, Siliziumnitride, amorphes Silizium,…),
- der kostengünstige Einsatz bei hoher Reinheit und Qualität und
- die Einfachheit bei der Erzeugung von Schichten mit variabler und genau kontrollierbarer Zusammensetzung.

Die bei dieser heterogenen Reaktion an der heißen Substratoberfläche abgeschiedenen amorphen oder kristallnen Schichten können

- Dielektrika (Oxide, Nitride,….)
- Halbleiter (Poly-Silizium) oder
- Metalle (Al, AlSiCu, W, Ta,…..)

sein. Aktiviert werden diese Reaktionen durch thermische Anregung. Die Parameter, die wesentlichen Einfluss auf den Ablauf der Reaktion haben sind

- Konzentration und Menge der Gasflüsse
- Druck
- Temperatur

Zusätzlich wird die abgeschiedene Schichtdicke noch von

- Der Zeit (Abscheidedauer)
- Der Geometrie der Reaktionskammer (Strömungsverhältnisse die von der Ofen- und Bootgeometrie Abbildung 10 abhängen)

Je genauer und konstanter diese Parameter während einer Abscheidung gehalten werden können, desto einheitlicher sind die Schichten und desto besser ist deren Uniformität. Sensible Messgeräte (Thermoelemente, Druckmessdosen, Massenflussmesser) die eine regelmäßige Kalibrierung erfordern garantieren stabile Prozessbedingungen.

2.1.1 Chemische u. physikalische Grundlagen von APCVD und LPCVD

Die Prozesse, die bei einer APCVD/LPCVD bei Temperaturen > 500°C ablaufen, werden durch die Grenzschichttheorie näherungsweise beschrieben. Für den Stofftransport in axialer Richtung über das Boot ist die Konvektion, von und zur Waferoberfläche die Diffusion verantwortlich. Die chemische Reaktion an der Oberfläche wird von der Reaktionskinetik der gasförmigen Substanzen bestimmt. Der Mechanismus welcher die Vorgänge bei der APCVD beschreibt lässt sich wie folgt einteilen:

- Schritt 1: Transport der Reaktanden an die Oberfläche (Diffusion)
- Schritt 2: Adsorption der Reaktanden an der Oberfläche
- Schritt 3: Chemische Reaktion der adsorbierten Stoffen und Bildung einer Schicht
- Schritt 4: Desorption der nicht verbrauchten Reaktanden von der Oberfläche
- Schritt 5: Transport der Reaktionsprodukte in die Gasphase (durch die Grenzschicht)

Welcher dieser Schritte geschwindigkeitsbestimmend ist hängt von der Konzentration der Reaktanden in der Gasphase, von der Temperatur und vom Druck (LPCVD) ab. Bei einigen Prozessen sieht man auch, dass die Abscheiderate, bei einer Betrachtung über den gesamten Reaktor (bis zu 150 Wafer), auch von der Position des Wafers im Reaktor abhängt was an der Verarmung der Reaktanden in axialer Reaktorrichtung liegt

Die Strömungen in der Gasphase sind im allgemeinem laminar und werden über die Reynoldszahl beschrieben. In Abbildung 6 sieht man eine schematische Darstellung der Grenzschicht und der Profile über einen Wafer.

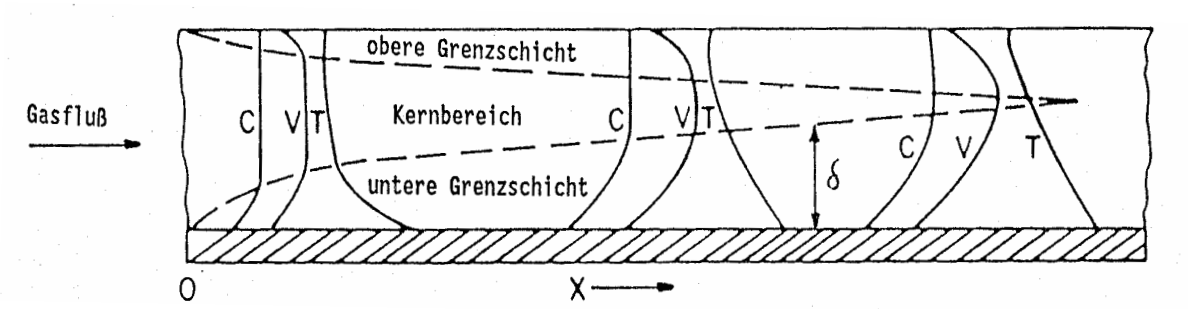

Abbildung 6: Schematische Schnitt durch einen CVD-Reaktor

Wobei C das Konzentrationsprofil, V das Geschwindigkeitsprofil, T das Temperaturprofil, x die Distanz in Strömungsrichtung und δ die Grenzschichtdicke darstellt.

Das Schema zeigt, dass innerhalb der Grenzschicht die Diffusion für den Stofftransport verantwortlich ist und außerhalb dieses Bereiches die Konvektion entscheidend ist. Die Grenzschichtdicke variiert über den Reaktorquerschnitt (Waferdurchmesser) und die Konzentration der Reaktanden nimmt indirekt proportional zur Grenzschichtdicke ab. Dadurch verändert sich die Abscheiderate über die Waferoberfläche was Schwierigkeiten bei der Within-Wafer-Uniformity bereitet. Eine Verbesserung kann durch Optimierung der Prozessparameter ($T, p, \dot{V}$) erreicht werden.

Die Abscheiderate wird bei der atmosphärischen CVD über die Temperatur bestimmt. Bei hohen Temperaturen ist die Abscheidung hauptsächlich diffusionskontrolliert – wenn man davon ausgeht, dass immer genügend Reaktanden an der Grenzschicht vorhanden sind – und zeigt eine nur leichte Temperaturabhängigkeit. Bei niedrigen Temperaturen wird der Prozess von der Geschwindigkeit der chemischen Oberflächenreaktion kontrolliert und ist stark temperaturabhängig.

2.1.2 Unterschied zwischen APCVD und LPCVD

Grundsätzlich unterscheidet sich die APCVD von der LPCVD durch wesentlich höheren Prozessdruck und viel geringer Strömungsgeschwindigkeit der Reaktionsgase. In der Halbleiterindustrie werden zum Abscheiden von dünnen Schichten (< 1000 nm) hauptsächlich LPCVD-Reaktoren verwendet. Begründet wird dies einerseits durch die viel geringere Partikelgeneration (weniger Defekte) der Wafer und andererseits dadurch, dass die Abscheiderate hauptsächlich durch die Reaktionskinetik bestimmt wird. Letzteres ist für die Prozessentwicklung und für die Prozessregelung von Vorteil. Dadurch kann der Prozess über die Einstellung weniger Parametern kontrolliert werden und ist wesentlich einfacher in seiner Handhabung.

Tabelle 1: Gegenüberstellung APCVD/LPCVD

Prozess	**Vorteile**	**Nachteile**	**Anwendung**
APCVD	einfacher Reaktor hohe Abscheiderate niedrige Temperatur	schlechte Kantenbedeckung hohe Partikelanzahl geringer Durchsatz	Niedertemperatur-Oxide (dotiert und undotiert)
LPCVD	hohe Reinheit hohe Gleichförmigkeit gute Kantenbedeckung hoher Durchsatz	hohe Temperatur niedrige Abscheiderate	Oxide(dotiert und undotiert) Siliziumnitrid Poly-Silizium

Die Diffusionskonstante beim LPCVD-Verfahren ist um den Faktor 1000 größer als bei atmosphärischen Bedingungen. Es lassen sich daher bei diesem Verfahren viel bessere Schichteigenschaften wie Homogenität und Kantenbedeckung erreichen weil die schwierig zu bestimmenden Parameter der Diffusionskonstanten und der Grenzschichtdicke keinen Einfluss haben.

2.2 Der ASM A400 Vertikalofen

Der ASM A400 Vertikalofen (Abbildung 5) ist ein Tool, das sowohl für Oxidations- als auch Abscheideprozesse eingesetzt wird. Je nach Anwendungsgebiet können sich die Öfen in ihrem Versorgungs- und Zusatzequipment deutlich voneinander unterscheiden. Jeder Ofen hat zwei Reaktormodule (Rohre) die für unterschiedliche Prozesse (z.B. ein Rohr TEOS, das Andere Nitrid) ausgelegt sein können. Der gesamte Reaktor inkl. des Beladekarussells hat eine Stickstoffatmosphäre. Die Wafer werden von einem Beladeroboter aus der WIP in ein Boot (Quarz oder Siliziumcarbid) mit Platz für max. 150 produktive Wafer transportiert und anschließend das beladene Boot in die vertikale Prozesskammer von unten eingefahren. Nach dem Schließen der Prozesskammer, sind einige Prozessschritte vor und auch nach der eigentlichen Abscheidung zu durchlaufen. Dazu gehören unter anderem, die Überprüfung von Pumpen und Ventilen, das Hochfahren der Temperaturen auf die gewünschten Abscheidetemperaturen, das Stabilisieren der Temperaturen und das Evakuieren auf Abscheidedruck. Beim eigentlichen Abscheideschritt strömen nun die aufbereiteten Reaktionsgase durch den Einlassflansch von unten in die Prozesskammer und die Abscheidereaktion beginnt zu.

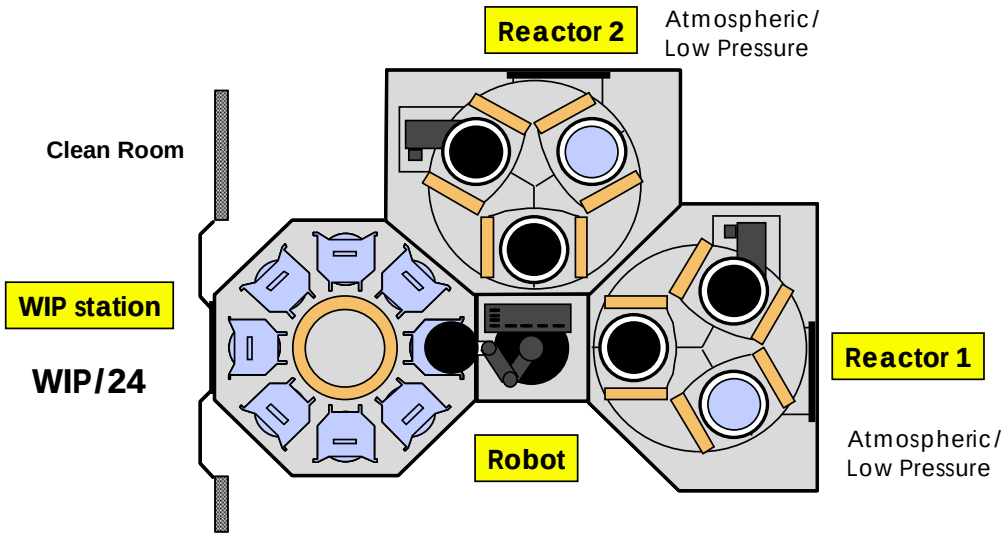

Abbildung 7: Hauptkomponenten ASM A400 Vertikalofen

Die Wafer werden für die Dauer der Abscheidezeit prozessiert und nach Abschluss der Folgeschritte aus dem Reaktor ausgefahren. Nach dem Abkühlvorgang werden die Wafer wiederum vom Roboter entladen und zurück in die Horden[1], welche sich im Karussell befinden, gehoben.

In Abbildung 7 sieht man schematisch den ASM-Ofen mit seinen wichtigsten Hauptkomponenten. Der Operator hat die Aufgabe die einzelnen Lose in ihren Horden, nachdem er das entsprechende Programm gestartet hat, in das Beladekarussell (WIP Station) zu stellen und anschließend die fertig prozessierten Wafer wieder zu entladen. Die restlichen Schritte bis zum Einfahren in die Prozesskammer werden vom ASM-Ofen automatisiert durch die Prozesssteuerung durchgeführt. In Abbildung 8 lässt sich die Zweiboot-Strategie gut erkennen. Während ein Boot in die Prozesskammer eingefahren und prozessiert wird, wird gleichzeitig das zweite Boot vom Beladeroboter befüllt. Dieser Prozessablauf bringt eine erhebliche Zeitersparnis mit sich, da unmittelbar nach Ausfahren des einen Bootes der Ofen bereits mit dem Zweiten beladen werden kann.

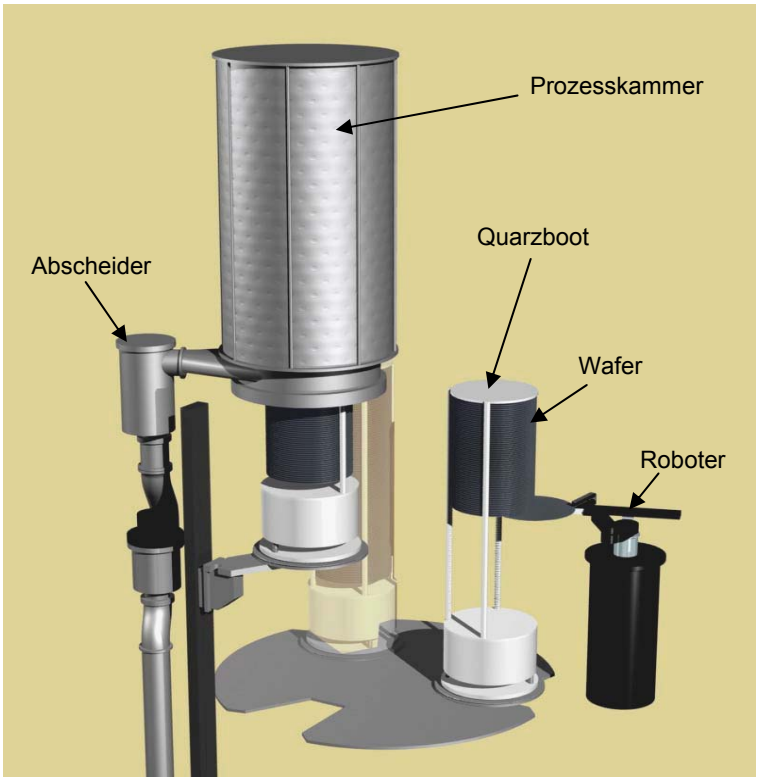

Abbildung 8: Beladevorgang

[1] Behälter zum Transport der Wafer mit Platz für 25 Wafer

2.3 Der Nitridprozess

Die Verbesserung der Prozessuniformität der LPCVD-Prozesse, im Speziellen des Nitridprozesses, soll durch eine R2R-Regelung verwirklicht werden. In diesem Abschnitt wird der Nitridprozess genauer beschrieben und erklärt. In Abbildung 9 ist ein vereinfachtes Verfahrensfließbild mit den wichtigsten Armaturen, Apparaten und Instrumenten dargestellt.

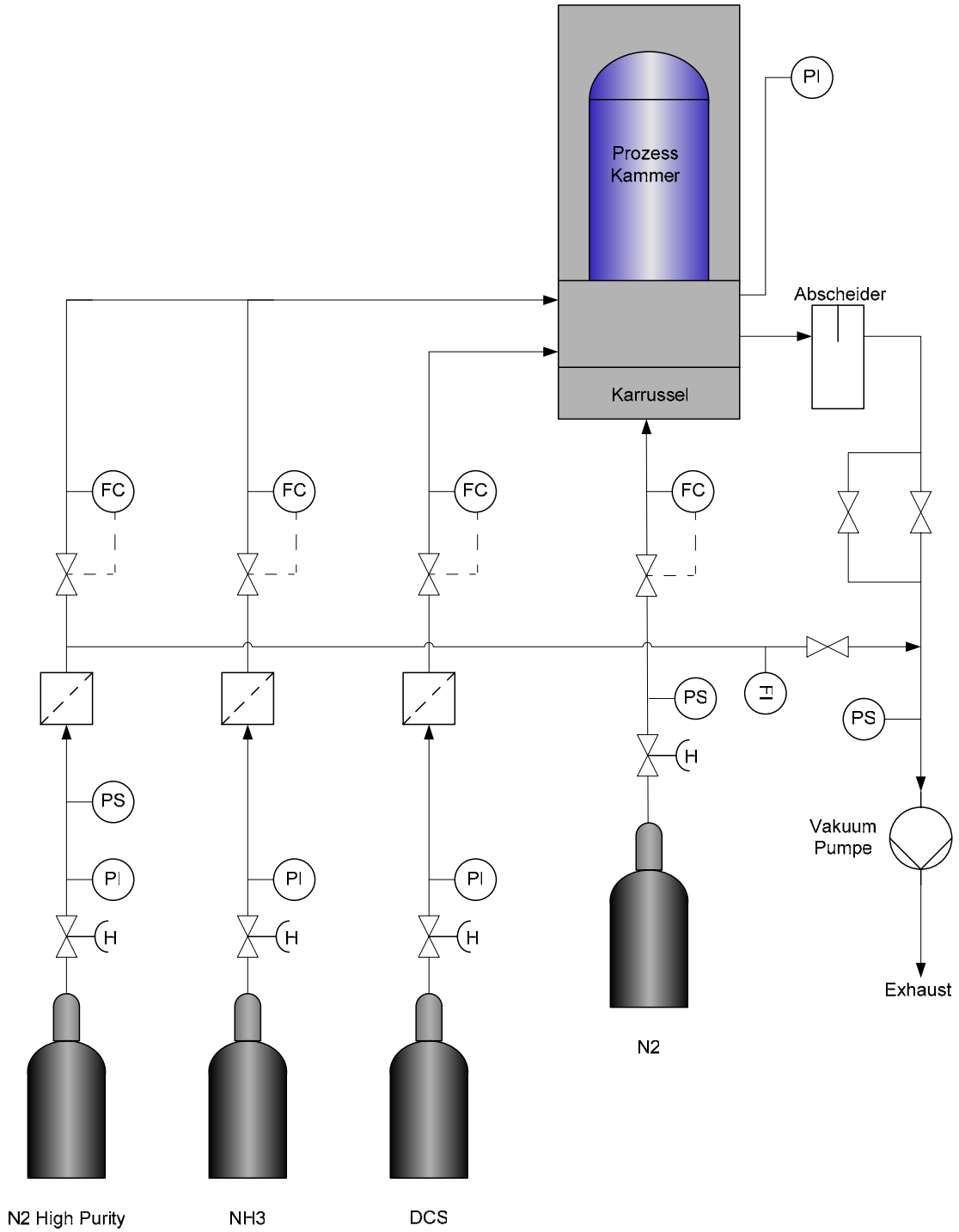

Abbildung 9: Verfahrensfließbild Nitridprozess

Beim Nitridprozess wird ein Quarzboot mit Platz für 165 Wafer verwendet das mit maximal 150 produktiven Wafer beladen wird. Die restlichen 15 Wafer sind Füllscheiben bzw. Testscheiben (für Schichtdickenmessung und Defektdichtemessungen). Grundsätzlich muss das Boot immer voll beladen sein, um die Oberfläche für die Abscheidung konstant zu halten. Dies ist notwendig, weil die Abscheiderate (nm/min) direkt von der Gesamtoberfläche (Wafer und Prozesskammer) abhängt. Die Oberflächenreaktionen beim Nitridprozess sehen wie folgt aus:

$$3\ SiH_2Cl_2 + 4\ NH_3 \rightarrow Si_3N_4 + 6\ HCl + 6\ H_2$$

$$3\ SiH_2Cl_2 + 10\ NH_3 \rightarrow Si_3N_4 + 6\ NH_4Cl + 6\ H_2$$

Diese Reaktionen finden bei Temperaturen von 700 – 800°C, bei Drücken von 230 – 350 mTorr und bei einem Verhältnis Dichlorsilan (DCS) zu Ammoniak von 1:4 statt. Über das Boot muss bei diesen Gasflüssen ein Temperaturgradient eingestellt werden, um trotzt der Verarmung der Reaktanden gleiche Schichtdicken zur produzieren. Die genauen Prozessbedingungen sind von den Faktoren

- Produktanforderungen (Schichtdicke, Schichteigenschaften wie Spannungen Brechungsindex…)
- Abscheidegeschwindigkeit (Durchsatz)
- Uniformität über den Wafer
- Position des Wafers

abhängig und sind in produktspezifischen Rezepten hinterlegt. In diesen Rezepten sind die Prozessschritte vom Einfahren in das Ofenrohr über einige Zwischenschritte wie den Abscheideschritt bis zum Ausfahren der Wafer aus dem Ofenrohr bestimmt. Beim Abscheidevorgang strömen die Reaktionsgase durch einen Einlassflansch von unten in den Reaktor vorbei am Pedestal (Wärmeaustausch) zu den Wafern. Siliziumnitrid wird auf den heißen Wafern abgeschieden und die nicht verbrauchten Edukte sowie die Reaktionsnebenprodukte strömen oben über den Liner[1] in Richtung Gasauslass. Ein Teil der Abgase kristallisiert im Abscheider. Der Rest gelangt über die Leitung für toxische Abgase in einen basischen Wäscher.

Die Schichtdicken der einzelnen Wafer werden durch die Abscheidezeit und durch eine fünf Zonen Temperaturregelung kontrolliert wobei der Druck und die Gasflüsse konstant gehalten werden. Grundsätzlich könnte der Prozess auch über das Regeln von Druck bzw.

[1] Strömungsleitrohr zwischen den Wafern und dem Ofenrohr

Gasflüssen durchgeführt werden. Die Praxis zeigt aber, dass eine derartige Regelung nur die Ergebnisse verschlechtert. Im Reaktor wird an insgesamt zehn Positionen die Temperatur gemessen. In jeder Heizzone befindet sich ein Thermoelement (Paddle Thermoelemente) innerhalb des Reaktors und eines direkt bei den Heizkassetten (Spike Thermoelemente). Die Paddle Thermoelemente werden bei einer Kalibrierungsfahrt auf die Spike Thermoelemente eingestellt und ermöglichen dadurch eine genaue Temperatureinstellung in den jeweiligen Zonen.

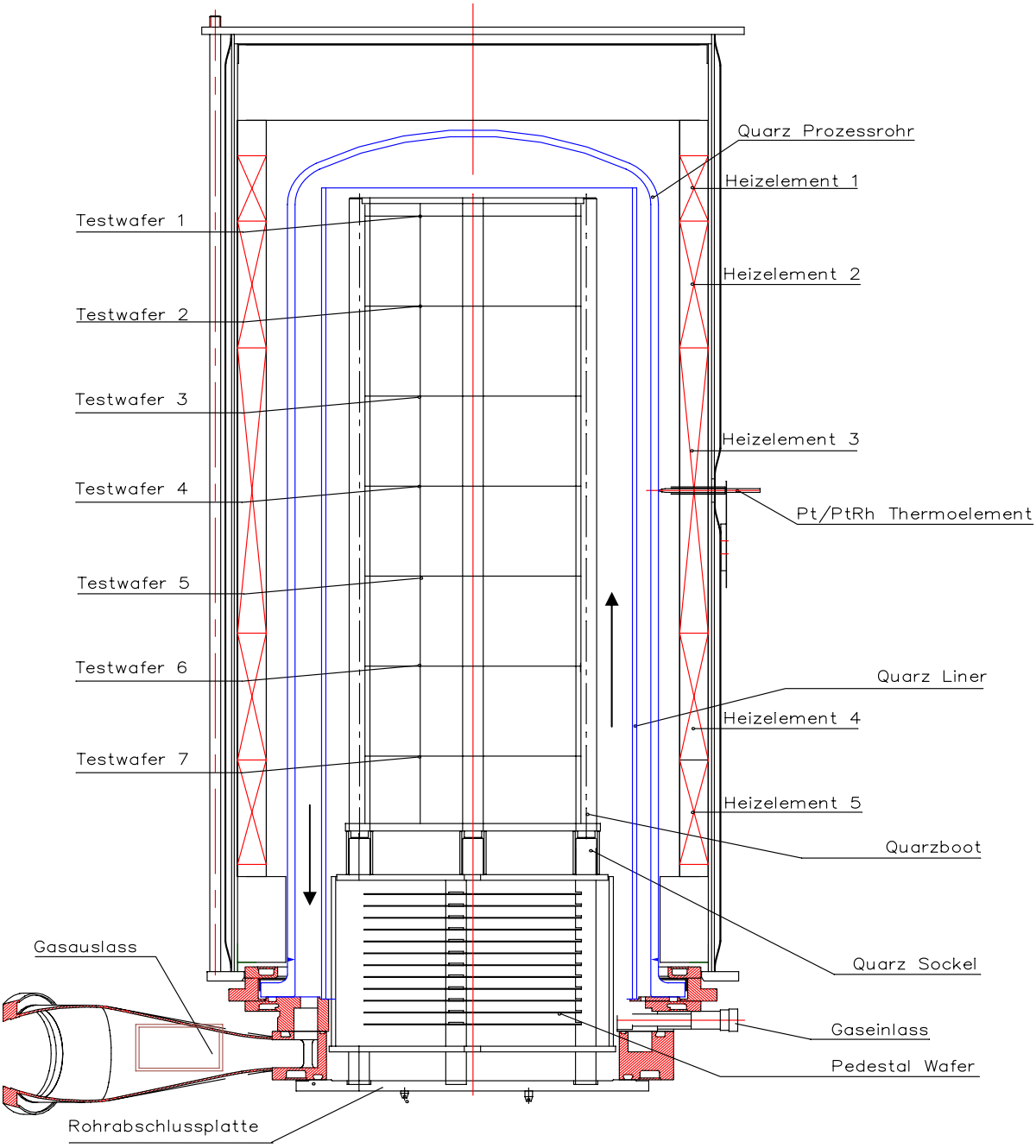

Abbildung 10: ASM A400 Reaktor mit Quarzboot

In obiger Abbildung sehen sie einen Reaktor im Querschnitt wobei jedoch nur das Spike Thermoelement in Zone 3 eingezeichnet ist. Die fünf Paddle Thermoelemente befinden sich in einem Quarzrohr an der Innenseite des Liners. Zusätzlich wurden die bei den Experimenten verwendeten 7 Testscheiben lagerichtig eingezeichnet. Im Produktionsprozess werden derzeit nur 3 Testscheiben pro Fahrt an den Positonen 1, 4 und 7 eingesetzt da wirtschaftliche Gründe gegen eine höhere Anzahl von Testscheiben sprechen und das Prozessresultat innerhalb der gewünschten Spezifikationen liegt. Um jedoch ein aussagekräftiges Prozessmodell zu erstellen war ein Mehr an Testscheiben pro Fahrt erforderlich.

2.4 Optiprobe Schichtdickenmessgerät

Das Schichtdickenmessgerät vom Typ Optiprobe [22] arbeitet mit verschiedenen Messmethoden. Eine davon ist die „Beam-Profile-Ellipsometrie“ die zur Messung der Schichtdicken von Siliziumnitriden verwendet wird. Es soll hier kurz auf die Messmethodik eingegangen werden.

Ein Ellipsometer misst die Änderung der Polarisation des Lichtes das von der Probenoberfläche reflektiert wird. Das Licht eines Lasers mit einer bestimmten Wellenlänge, das in einem bestimmten Winkel auf die Oberfläche des Wafers auftrifft wird reflektiert und von einem speziellen Detektor gemessen. Durch Analyse dieser Polarisationsänderungen kann die Schichtdicke bestimmt werden. Dabei können Schichtdicken von 1 nm bis zu 1000 nm mit einer Genauigkeit von 0,1 nm gemessen werden.

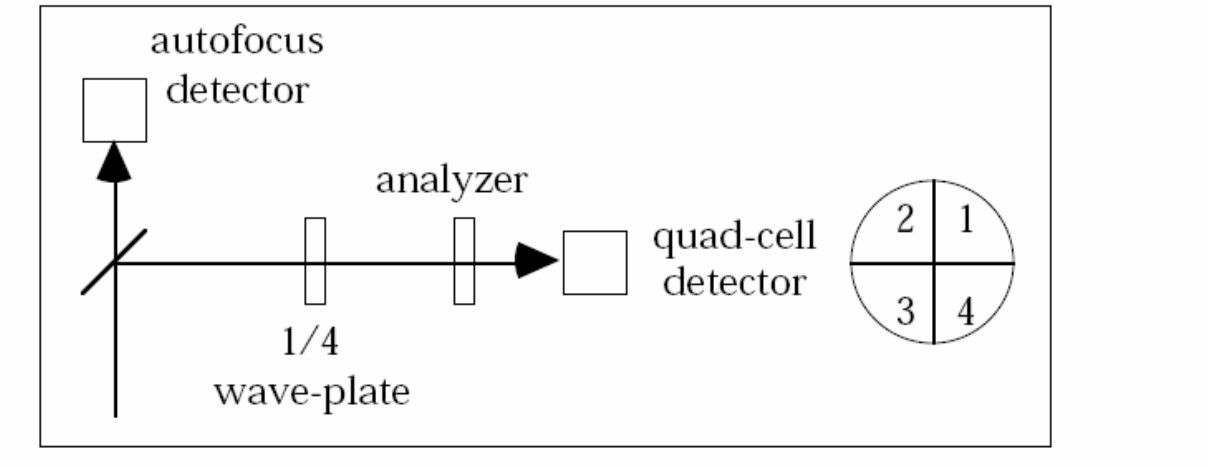

Abbildung 11: Strahlengang Ellipsometer

Der Nachteil der Ellipsometrie ist der relativ große Messpunkt, der aber bei der Schichtdickenmessung auf unstrukturierten Testscheiben, wie im vorliegenden Fall, keine Rolle spielt.

Der Fehler der Messungen spielt in Verhältnis zum Prozessrauschen eine untergeordnete Rolle. Er liegt im Bereich von ± 10% des Prozessrauschens und wurde bei den Experimenten über Vergleichsmessungen an zwei verschiedenen Messgeräten ermittelt. Der Messfehler wurde bei den Simulationen nicht berücksichtigt da er vernachlässigbar klein ist.

2.5 Modellbildung

Strategien die der Modellfindung dienen lassen sich in drei unterschiedliche Kategorien einordnen, die sich grundlegend voneinander unterscheiden [4]. Nutzung theoretischer Kenntnisse wie physikalisch bzw. chemische Zusammenhänge führen zur theoretischen Modellbildung. Bei der experimentellen Modellbildung werden am System Versuche durchgeführt um die Systemgrößen zu finden. Die dritte Methode heißt maschinelles Lernen und beruht auf der Erfahrung des Menschen.

Darüber hinaus werden oftmals noch Mischformen wie z.B. semiempirische Methoden – eine Mischung aus theoretischer und experimenteller Modellbildung – angewendet. Diese Methoden beruhen jedoch auf den Grundlagen der drei Erstgenannten.

2.5.1 Theoretische Modellbildung

Bei dieser ursprünglichsten Methode werden physikalische und chemische Zusammenhänge eines Systems meist in Form von Differenzialgleichungen abgebildet. Grundlage sind Bilanzgleichungen von Massen, Energien, Kräften, Impuls mit deren Hilfe sich physikalisch parametrische Modelle in Form von Zustandsmodellen aufstellen lassen.

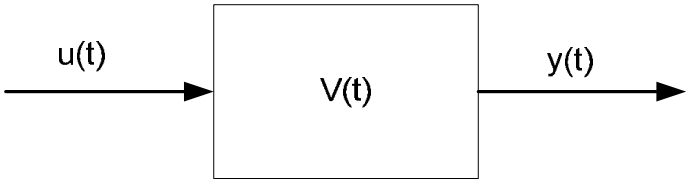

Abbildung 12: Schematische Darstellung eines Systems

$$\frac{dV(t)}{dt} = u(t) - y(t) \tag{1}$$

Vorteile dieser Vorgangsweise sind die gute Abbildung der Systemstrukturen und dass die Parameter den Systemkomponenten zugeordnet werden können. Schwierigkeiten ergeben sich bei der Bestimmung der Parameter und bei der Erfassung von zeitvariantem Verhalten insbesondere wenn das System nicht beeinflussbaren Störungen unterliegt. Diese zwei Punkte sind auch der Grund warum in der Halbleiterindustrie theoretische Modellbildung selten zum Einsatz kommt. In [5] wird ein theoretisches Modell für den TEOS-Prozess auf einem ASM-A 412 300 mm Equipment beschrieben. Hierin wird über Transportgleichungen

(Kontinuitätsgleichung, Navier-Stokes-Gleichung….) ein Modell für die Abscheiderate in Abhängigkeit vom TEOS-Fluss, Temperatur und Druck hergeleitet. Es mussten dabei zahlreiche Vereinfachungen hinsichtlich Reaktorgeometrie, Gaseinströmöffnungen, Waferposition und Strömungsverhältnisse vorgenommen werden um die Parameter für ein zweidimensionales Modell zu erlangen. Dieses Modell ist für einen bestimmten Ofentyp geeignet und nur mit viel Aufwand auf andere Öfen anwendbar. Ein halbempirisches Modell, das versucht die Within-Wafer-Nonuniformity bei einer Poly-Abscheidung zu modellieren wird in [6] beschrieben. Dort werden die tatsächlichen Oberflächentemperaturen über insitu Thermoelemente auf mehreren Positionen über den Wafer gemessen. Beide Modelle zeigen im Vergleich mit Experimentaldaten gute Ergebnisse in der Übereinstimmung Modell – Experiment. Sie sind jedoch auf rein zu Testzwecken abgestelltem Equipment durchgeführt worden. Diese Abstellung eines Ofens nur für Experimente lässt sich bei Infineon Villach bei der momentanen Auftragslage nicht verwirklichen.

2.5.2 Experimentelle Modellbildung

Ein anderer Weg, wie man zu einem Modell gelangt wird durch Abfahren einer Versuchsreihe gezeigt. Es werden Experimente direkt am Equipment durchgeführt um die Abhängigkeiten der Regelgrößen von den Stellgrößen zu eruieren. Als Ergebnis erhält man Input/Output-Modelle. Bei diesem Ansatz ist die Bestimmung bzw. Schätzung der Modellparameter durch statistische Analysemethoden gut möglich.

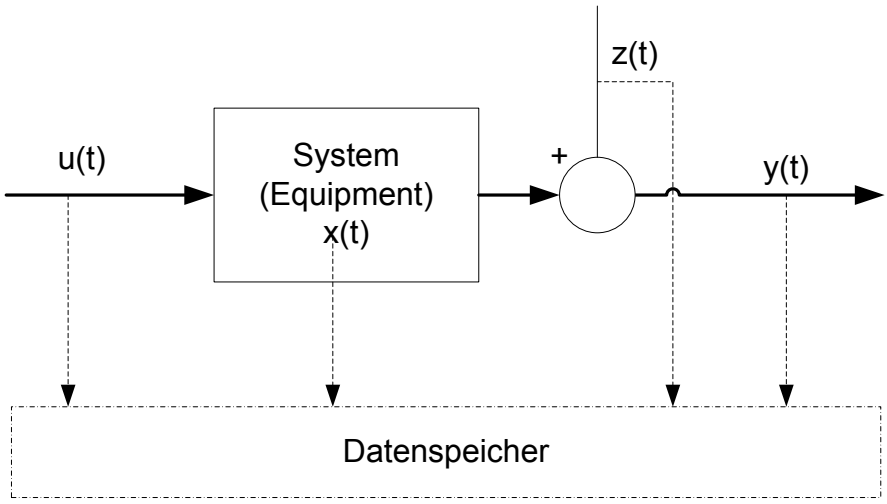

Abbildung 13: Input/Output Modell mit Störgrößen und Systemzustand

$$y(t) = f\left(u(t), z(t), x(t)\right) \tag{2}$$

Das System wird in der realen Umwelt abgebildet, was eine gute Übereinstimmung Modell mit realem Verhalten garantiert. Nachteilig wirken sich hingegen eine nur schwer mögliche

Strukturerkennung und eine schlechte Parameterzuordnung zu den Systemkomponenten aus. Überdies steigt mit zunehmender Komplexität der Zusammenhänge die Anzahl der Versuche, sodass der Aufwand für Planung, Durchführung und Auswertung erheblich zunimmt.

Trotzdem wird diese Methode in der Halbleiterbranche oftmals angewendet. Eine standardisierte Vorgehensweise – das Design of Experiment (DoE) - und der Einsatz von elektronischer Datenverarbeitung in einer CIM-Umgebung erleichtern die Durchführung erheblich.

2.5.3 Maschinelles Lernen

Als maschinelles Lernen bezeichnet man die Erstellung eines Modells durch Befragung von Experten und die Analyse der Handlungen von Experten bei konkreten Problemen. Diese Methode ist mit geringem Aufwand durchzuführen und spiegelt das reale Equipmentverhalten wider.

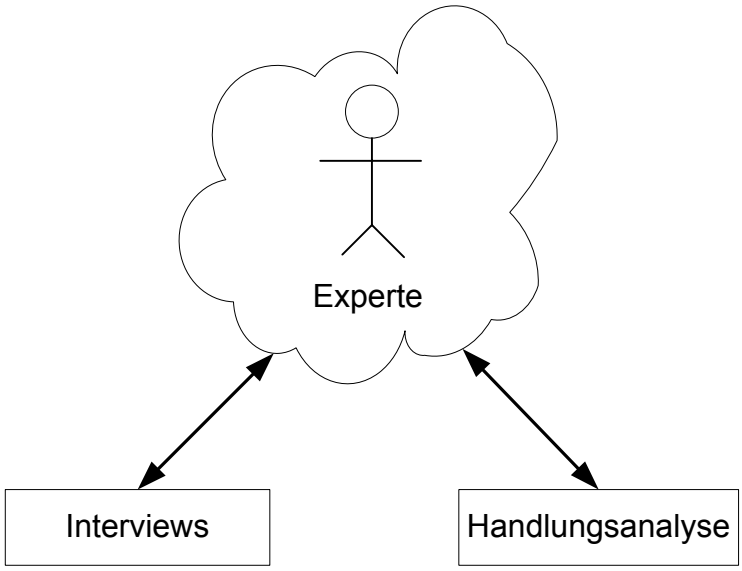

Abbildung 14: Maschinelles Lernen (schematisch)

Die Befragungen führen zu linguistischen qualitativen Modellen deren Güte stark vom Wissen und der Erfahrung der Experten abhängt. Die Modelle haben einen subjektiven Charakter und deren Einbindung in eine automatisierte Regelung gestaltet sich als schwieriges Unterfangen, da mit zunehmender Komplexität die Überführung dieser linguistischen Modelle in eine mathematische Beschreibung nur schwer möglich ist. Überdies ist es notwendig, dass die Technologien und Prozesse bereits existieren müssen. Neue Verfahren bei denen noch kein Expertenwissen vorhanden ist lassen sich mit dieser Methode nicht bearbeiten. Folglich, ist dieser Ansatz zwar die einfachste und zeitsparendste Vorgehensweise welche aber in der Praxis nur beschränkt einsetzbar ist.

2.6 Design of Experiment

Als Design of Experiment (DoE) bezeichnet man eine systematische und organisierte Methode um die Abhängigkeiten von prozessbeeinflussenden Faktoren auf den Output eines Prozesses festzustellen [7]. Es ist eine Methode, wie ein Experiment auf eine möglichst effektive Weise geplant, durchgeführt und ausgewertet wird, um die notwendigen und erwünschten Informationen zu erhalten. Ziel eines DoE ist das Erlangen eines Modells, welches das Verhalten von Equipment, eines Prozesses oder von statistischen Daten erklärt [8].

Die Vorteile die ein DoE bringt, sind:

- Bessere Effektivität durch das Variieren von mehr als einer Variable pro Experiment (Informationsdichte ist wesentlich höher als wenn man nur eine Stellgröße variiert)
- Reduzierung der Anzahl der Versuche, die notwendig sind um ein bestimmtes Experiment durchzuführen
- Erlangen von Grunddaten, die mit zukünftigen Experimenten verglichen werden können
- Standardisiertes Vorgehen, dass die Reproduzierbarkeit, Aussagekraft und Vergleichbarkeit erhöht

Der Grundsätzliche Ablauf eines DoE lässt sich in 4 Bereiche aufteilen:

- Planung des Experiments
- Durchführung des Experiments und Datensammlung
- Auswertung und Analyse der Daten
- Ziehen von Schlussfolgerungen (Erstellung eines Modells)

Bei einem DoE werden die Stellgrößen kontrolliert verändert um Auswirkungen auf den Output des Prozesses zu untersuchen. Dabei soll bei der Planung aus der Vielzahl von möglichen Versuchsanordnungen (voll-faktoriell, teil-faktoriell, Box-Behken, Central Composite, D-Optimal…) jene Anordnung gefunden werden, die für das vorliegende Problem am Besten geeignet ist. Vorab ist es notwendig Faktoren die man kontrollieren kann von jenen die man nicht kontrollieren kann zu trennen, um die Anzahl der veränderbaren Größen so gering als möglich zu halten. Je mehr veränderbare Größen, deren Änderung sich auf den Output auswirkt, in einem Experiment vorhanden sind, desto höher wird die Anzahl der notwendigen Versuche. Will man z.B. lineare Abhängigkeiten und auch Abhängigkeiten durch Interaktion einzelner Größen bei einem Prozess mit drei Stellgrößen herausfinden sind

bei einer voll - faktoriellen Versuchsanordnung 2^3 = 8 Versuche notwendig. Quadratische Abhängigkeiten können bei dieser Versuchsanordnung nur abgeschätzt werden wenn zusätzlich Fahrten bei normalen Prozessbedingungen (Center-Fahrten) durchgeführt werden. Deswegen ist es oft sinnvoll einige regelbare Größen konstant zu halten bzw. sie zu einer Größe zusammenzufassen. Dies wird oft bei Vorab-Experimenten gemacht, um die grundsätzlichen Abhängigkeiten (linear, quadratisch,…) abschätzen zu können und um bei den Folgeexperimenten jene Versuchsanordnung zu finden, die das jeweilige Problem bestmöglich löst. Mit der Response-Surface-Method (RSM), einer Technik die das Input/Output Verhalten eines Prozesses vollständig erklärt, kann ein exaktes Prozessmodell erstellt werden, welches auch quadratische Abhängigkeiten berücksichtigt. Meist wird dafür ein D-Optimales[1] Design [21] (bei mehr als 4 regelbarer Größen), welches computergeneriert ist verwendet, um die Anzahl der Versuche so gering als möglich zu halten. Will man eine Aussage über das zufällige Prozessrauschen machen sind auch hier wiederholte Centerfahrten notwendig. Grundsätzliche Überlegungen bezüglich der Versuchsreihenfolge, dem momentanen Zustand des Equipment und der zeitlichen Versuchsabfolge müssen wie bei jedem Experiment berücksichtigt werden um aussagekräftige und reproduzierbare Ergebnisse zu erhalten.

In [8] wird beispielhaft eine voll–faktorielle Versuchsanordnung mit drei veränderbaren Größen A, B, C gezeigt. Jeder der Faktoren kann zwei unterschiedliche Zustände einnehmen. Die Faktoren sind auf den Bereich [-1,1], wie in Gleichung (3) beschrieben, normiert. Dies verhindert eine Gewichtung auf Grund der unterschiedlichen Absolutwerte und ist für die Auswertung erforderlich. Außerdem eliminiert durch diese Normierung eine Korrelation zwischen den Schätzungen Haupteffekte (A, B, C) und den Interaktionseffekten (AB,…).

Tabelle 2:Voll-faktorielles Design mit 2^3 Versuchen

	Design Faktoren		
Runs	A	B	C
1 (1)	-1	-1	-1
2 (A)	1	-1	-1
3 (B)	-1	1	-1
4 (C)	-1	-1	1
5 (AB)	1	1	-1
6 (AC)	1	-1	1
7 (BC)	-1	1	1
8 (ABC)	1	1	1

[1]Kriterium für D-Optimalität: max |X´X| - maximiere die Determinante der Informationsmatrix

$$X_{normiert} = \frac{X - \frac{(X_{max} - X_{min})}{2}}{\frac{(X_{max} - X_{min})}{2}} \quad (3)$$

Mit dieser Anordnung wird neben linearen Abhängigkeiten der einzelnen Faktoren auch die Interaktion von zwei bzw. drei Faktoren herausgefunden. Erhöht sich jedoch die Anzahl der Parameter auf 5 und höher, so geht man zu anderen Versuchsanordnungen (D-Optimal,…) über. Entscheidend für die Wahl der Versuchsanordnung ist die Tatsache, dass bereits vor der Planung festgelegt werden muss welche Abhängigkeiten (linear, linear und quadratisch) man untersuchen will.

Nach der Durchführung der Experimente werden die Daten statistisch ausgewertet und daraus ein Input/Output Modell generiert. In allgemeiner Form sieht so ein Multi-Input Single-Output (MISO) Modell mit zwei Stellgrößen wie folgt aus:

$$Y = f(X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_4 X_1^2 + \beta_5 X_2^2 + ... \quad (4)$$

Worin Y die Regelgröße (Response), X_i die Stellgrößen (Predictors) und β_i die Regressionsparameter bzw. Modellparameter sind. Ein SISO Modell ersten Grades (linear) hat die Form einer einfachen Geradengleichung und wird mit der Methode der kleinsten Quadrate (LSM) aus den Experimentaldaten gewonnen:

$$Y = f(X_1) = \beta_0 + \beta_1 X_1 \quad (5)$$

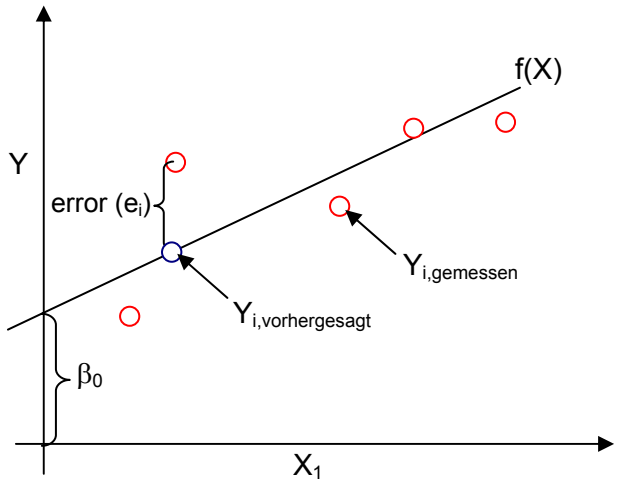

Abbildung 15: Lineares Modell durch LSM

Bei komplexeren Problemen, wie in dieser Arbeit, können die Abhängigkeiten nicht mehr so einfach dargestellt werden. Bei diesen Multi-Input-Multi-Output (MIMO) Modellen erhält man aus den Experimenten Systeme von linearen (quadratischen) Gleichungen. Statistische Analyse und Bewertungsmethoden, wie ANOVA-Test (Analysis of Variance), Lack of fit Test, Residuenanalyse und noch einige Andere [16] machen es möglich die Qualität dieser MIMO Modelle zu beurteilen.

Letztendlich gewinnt man mit dieser Methode ein Modell das die Abhängigkeiten des Prozesses für den Zeitpunkt der Experimente beschreibt. Dieses Modell kann als Initialmodell für einen R2R-Kontroller dienen.

2.7 Run to Run Methoden

2.7.1 Überblick Run to Run Algorithmen

In den letzten 30 Jahren wurde eine Vielzahl von R2R Kontrollalgorithmen für Prozesse in der Halbleiterindustrie entwickelt bzw. adaptiert. Es ist Ziel all dieser Kontroller, Prozesse deren Output einer Drift oder Sprüngen unterliegen besser zu beherrschen und sie automatisch nachzuregeln. Anfänglich wurden hauptsächlich lineare Kontrollalgorithmen (EWMA) zur Regelung der Prozesse verwendet und auch eingesetzt. Unglücklicherweise hat man in der Halbleiterindustrie Prozesse, wie z.B. CMP und Abscheidung, die leichte bis starke Nichtlinearitäten aufweisen. Diese Prozesse erfordern dynamische Regelmodelle und Algorithmen die auch nichtlineare Probleme lösen.

Eine weit verbreitete Möglichkeit ist die Verwendung von Algorithmen [9], die versuchen eine Vorhersage zu machen wo sich die Regelgröße beim nächsten Run befindet. Wenn der Vergleich der Vorhersage mit dem gemessenen Ergebnis eine zu große Abweichung zeigt, versuchen diese Kontroller durch eine Änderung der Modelparameter entgegenzusteuern und dadurch dem Zielwert näher zu kommen.

Zu den möglichen R2R Kontrollalgorithmen die nach obigem Schema arbeiten zählen

- EWMA Controller (Exponential Weighted Moving Average)
- OAQC Controller (Optimizing Adaptive Quality Controller)
- OVE Controller (Optimal Volume Ellipsoid)
- DHOBE Controller (Dasgupta – Huang Optimal Bounded Ellipsoid)

um nur die bekanntesten zu nennen.

Der am längsten bekannte und auch einfachste Controller ist der EWMA Kontroller [10], der von E. Sachs vom MIT entwickelt wurde. Dieser Controller ist jedoch nur gut geeignet für

Prozesse die mit linearen Modellen beschrieben werden können. Er ist der am besten untersuchte Controller und gewichtet die Daten – im konkreten Fall die Schichtdicken Messergebnisse – exponentiell, umso geringer je weiter sie in der Vergangenheit liegen. Vom EWMA Kontroller gibt es einige Modifikationen, wie z.B. die double EWMA Methode [13] oder die ANN EWMA Methode [14].

Der OAQC Controller [11], [12] ist eine „Least Square Recursive“ Methode(LSR) und wird sowohl zur Modelloptimierung als auch als reiner Controller eingesetzt. Im Optimierungsmodus erneuert er das zugrunde liegende Modell bei jedem Run. Im Kontrollermodus wird eine quadratische Kostenfunktion dazu verwendet um die Zielwerte zu berechnen. Es wird dabei auf die Variation der einstellbaren Parameter Rücksicht genommen.

Der OVE Algorithmus zählt zu den „set-valued“ R2R Kontrollmethoden. Er wurde zuerst publiziert von J.S. Baras und N.S. Patel. Entscheidend bei dieser Methode ist es, dass die Unsicherheit bei der Modellidentifizierung mit berücksichtigt wird. Diese Unsicherheit existiert, da bei vielen Halbleiterprozessen ein Rauschen vorhanden ist und dadurch bei der Erstellung von empirischen Modellen ein Modellfehler nicht zu vermeiden ist. Deswegen wird bei dieser Methode nicht ein exaktes Regelmodell, sondern eine Gruppe von möglichen Modellen vorhergesagt in dem sich das reale Prozessmodell mit hoher Wahrscheinlichkeit befindet. Aus den möglichen Regelmodellen wird jenes mit der größten Kostenfunktion (worst case approach) herausgesucht und damit die Regelgrößen für den nächsten Run vorhergesagt.

Der DHOBE Controller, ist wie der zuvor genannte auch ein Ellipsoid Algorithmus. Er ist jedoch etwas allgemeiner gehalten und wurde von Dasgupta und Huang entwickelt. Sie führten einen „forgetting factor“ ein der den OBE Algorithmus von Fogel und Huang modifizierte. Von Fogel und Huang wurde 1993 auch noch eine Notfallsprozedur hinzugefügt. Er gehört zu der Gruppe der „bounded error estimation“ Algorithmen, was ihn auch vom OVE Algorithmus unterscheidet.

Tabelle 3: Zusammenfassung R2R Methoden

R2R Control Method	Linear Process	Light non-linear Process	Non-Linear Process	Complexity
EWMA (double, enhanced)	Y	Y	N	Low
Machine Learning (KIRC)	Y	N	N	High
LSR (OAQC)	Y	Y	Y	Medium
Probability Method	Y	?	?	Medium
Neural Network	Y	Y	?	High
Model - predictive (DHOBE, OVE)	Y	Y	Y	Medium

Andere Algorithmen, wie der Knowledge Based Interactive Controller (KIRC), das Artifical Neural Network (ANN) oder der Probability Approach unterscheiden sich in ihrer Funktionsweise deutlich von den bisher genannten. Detaillierte Informationen dazu findet man in [15]. Dort werden die einzelnen Algorithmen noch einmal zusammengefasst und ihr mögliches Einsatzgebiet erläutert.

2.7.2 Der DHOBE R2R Kontroller

Dieser Arbeit liegt der DHOBE Algorithmus zugrunde. Bei dieser Methode werden die Modellparameter nicht bei jedem Run erneuert, sondern nur wenn neue Informationen vorhanden sind. Mit anderen Worten es wird nur dann das Regelmodel erneuert, wenn die Differenz von Messwert zu vorhergesagtem Wert deutlich voneinander abweicht. Innerhalb eines bestimmten freiwählbaren Bereiches um den Zielwert wird für den nächsten Run das vorherige Modell zur Bestimmung der Stellgrößen verwendet. Ein weiterer Vorteil des Algorithmus ist es, dass man nur die Grenzen des Rauschens kennen muss und nicht dessen exakte Verteilung (Normal, Logonormal….). Dies ist die einzige und auch hinreichende Bedingung dafür dass der Algorithmus konvergiert.

Durch die Einführung der Notfallsprozedur können auch plötzlich auftretende größere Abweichungen und große Fehler zwischen Regelmodell und Prozessmodell durch diesen Kontroller geregelt werden, was beim Equipment in der Halbleiterbranche insbesondere nach Instandhaltungsarbeiten häufiger vorkommt. Bei diesen plötzlich auftretenden größeren Zustandsveränderungen können die meisten anderen Algorithmen keine neuen Parameter mehr berechnen. Zusätzlich ist dieser R2R Kontroller auch imstande quadratische Modelle die linear in ihren Parametern sind zu regeln. Die letzten beiden Eigenschaften sind die Hauptgründe dafür, dass der DHOBE Kontroller bestens für den Einsatz bei Ofenprozessen geeignet ist.

Die Berechnung der neuen Modelparameter erfolgt dadurch, dass eine Gruppe möglicher Modellparameter, welche außen durch ein Ellipsoid begrenzt sind, vorhergesagt wird. Aus all diesen möglichen Parametern werden im vorliegenden Fall die Parameter, die durch Mittelpunkt des Ellipsoids beschrieben werden dazu verwendet um die regelbaren Größen des nächsten Runs zu bestimmen.

2.7.3 Der DHOBE Algorithmus

In diesem Abschnitt wird der DHOBE Kontrollalgorithmus im Einzelnen erläutert und beschrieben [9].

Die grundsätzliche Idee des Algorithmus ist es die äußeren Grenzen des Ellipsoids, welcher die möglichen Modelle enthält zu finden. Diese Methode kann auf Modelle bis zur zweiten Ordnung, die linear in den Parametern sind, angewendet werden. Damit wird sichergestellt,

dass die Parameterschätzung für Modelle mit einem quadratischen Input/Output Verhalten angewendet werden kann. Die folgende Abbildung zeigt eine schematische Darstellung wie der Algorithmus arbeitet und wie er die Modellparameter von Run zu Run erneuert.

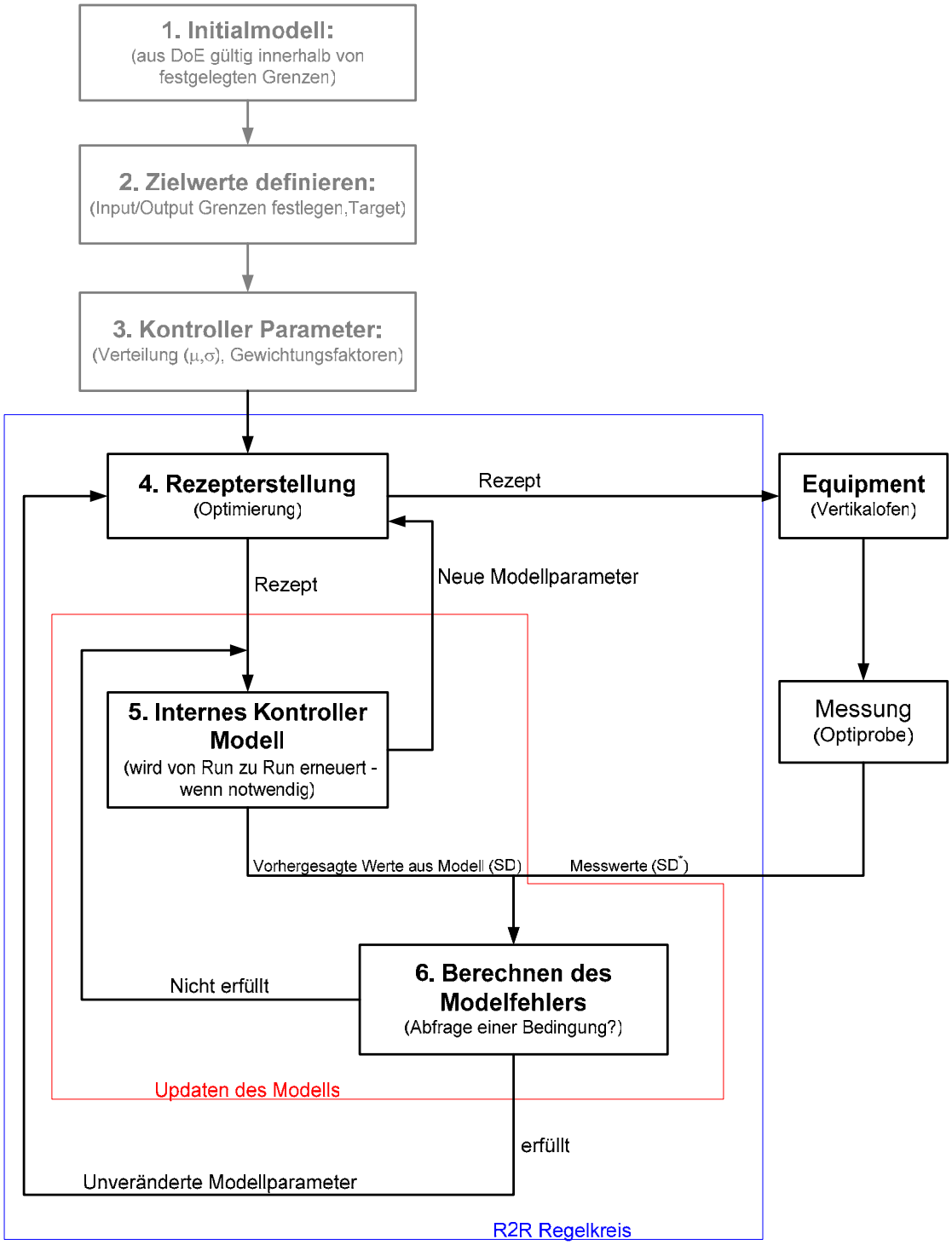

Abbildung 16:Blockdiagramm DHOBE-Algorihmus

Um die Verständlichkeit zu erhalten wird der Algorithmus an Hand eines MISO (Multi Input Single Output) System beschrieben. Er kann natürlich auch auf MIMO Systeme (wie im untersuchten Fall) angewendet werden.

Gehen wir davon aus, dass das Prozessmodell in folgender Form vorliegt:

$$y_t = \theta^{*T} u_t + e_t \tag{6}$$

Wobei y_t die Regelgröße, θ^{*T} der wahre Parametervektor, u_t der Vektor der Stellgrößen und e_t das Rauschen des Modells darstellt. Der Term e_t wird begrenzt von γ:

$$|e_t| \leq \gamma \tag{7}$$

Nehmen wir an, dass für den Zeitpunkt t-1, die mögliche Gruppe der Modellparameter vom Ellipsoid E_{t-1} begrenzt wird. Dieses Ellipsoid kann durch seinen Mittelpunkt θ_{t-1}, seine positiv definite Orientierungsmatrix P^{-1}_{t-1} [19] [20] und seine Größe über den Unsicherheitsfaktor σ^2_{t-1} definiert werden:

$$E_{t-1} = \left\{ \theta_t \in \Re^n : (\theta_t - \theta_{t-1})^T P_{t-1}^{-1} (\theta_t - \theta_{t-1}) \leq \sigma_t^2 \right\} \tag{8}$$

Wenn man dann zum Zeitpunkt t eine Beobachtung des Prozesses y_t hat, kann eine Ebene wie folgt definiert werden:

$$S_t = \left\{ \theta \in \Re^n : (y_t - \theta^T u_t)^2 \leq \gamma^2 \right\}. \tag{9}$$

Diese Ebene S_t schneidet E_{t-1}, was uns ermöglicht ein neues Ellipsoid E_t rekursiv zu berechnen:

$$E_t = \left\{ \theta \in \Re^n : (1-\lambda_t)(\theta - \theta_{t-1})^T P_{t-1}^{-1} (\theta - \theta_{t-1}) + \lambda_t (y_t - \theta^T u_t)^2 \leq (1-\lambda_t)\sigma_{t-1}^2 + \lambda_t \gamma^2 \right\} \tag{10}$$

Hier wird der „Updatefaktor“ λ_t und der „Forgettingfactor“ $1-\lambda_t$ eingeführt. λ_t wird dazu verwendet um die Größe des Ellipsoids von Run zu Run zu verkleinern. Dies wird erreicht indem man die Gleichung

$$E_t = \left\{ \theta \in \Re^n : \left(\theta - \theta_{t-1}\right)^T P_{t-1}^{-1} \left(\theta - \theta_{t-1}\right) \le \sigma_{t-1}^2 \right\} \tag{11}$$

transformiert und daraus folgenden rekursiven Ausdruck erhält.

$$P_t^{-1} = \left(1 - \lambda_t\right) P_{t-1}^{-1} + \lambda_t u_t u_t^T \tag{12}$$

$$\sigma_t^2 = \left(1 - \lambda_t\right) \sigma_{t-1}^2 + \lambda_t \gamma^2 - \frac{\lambda_t \left(1 - \lambda_t\right)\left(y_t - u_t^T \theta_{t-1}\right)}{1 - \lambda_t + \lambda_t u_t^T P_{t-1} u_t} \tag{13}$$

$$\theta_t = \theta_{t-1} + \lambda_t P_t u_t \left(y_t - u_t^T \theta_{t-1}\right) \tag{14}$$

Invertiert man Gleichung (12) so erhält man

$$P_t = \left(\frac{1}{1 - \lambda_t}\right)\left(P_{t-1} - \frac{\lambda_t P_{t-1} u_t u_t^T P_{t-1}}{1 - \lambda_t + \lambda_t u_t^T P_{t-1} u_t}\right). \tag{15}$$

Nachdem man den neuen Mittelpunkt des Ellipsoids θ_t berechnet hat, dient er als Schätzung der neuen Parameter für das Modell.

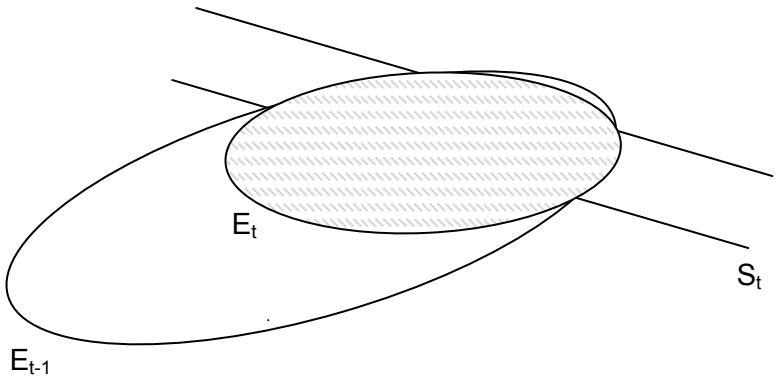

Abbildung 17: Rekursive Bildung des neuen Ellipsoids E_t

Diese Updateprozedur wird nicht bei jedem Run durchgeführt, sondern der Kontroller entscheidet wann eine Erneuerung der Modellparameter notwendig ist. Die Bedingung für ein Update ist folgende:

$$\sigma_{t-1}^2 + \delta_t^2 \leq \gamma^2 \tag{16}$$

Dabei ist σ_{t-1} der Parameterunsicherheitsfaktor d.h. die Ungenauigkeit mit der die neuen Parameter vorhergesagt werden und δ_t der Vorhersagefehler d.h. der Fehler zwischen vorhergesagter Schichtdicke und gemessener Schichtdicke. Der Vorhersagefehler ergibt sich aus

$$\delta_t = y_t - u_t^T \theta_{t-1} \tag{17}$$

All die Modellparameter und daraus folgend auch die regelbaren Größen (T_1,…) bleiben bei ihren alten Werten.

Ist Bedingung (16) nicht erfüllt wird λ_t folgendermaßen berechnet:

$$\lambda_t = \min\left(\lambda_{max}, \nu_t\right),$$

wobei

$$\nu_t = \begin{cases} \lambda_{max} & \text{wenn } \delta_t = 0 \\ \dfrac{(1-\beta_t)}{2} & \text{wenn } G_t = 1 \\ \dfrac{\left(\left(1 - \sqrt{\dfrac{G_t}{(1+\beta_t(G_t-1))}}\right)\right)}{(1-G_t)} & \text{wenn } \beta_t(G_t-1)+1 > 0 \\ \lambda_{max} & \text{wenn } \beta_t(G_t-1)+1 \leq 0 \end{cases} \tag{18}$$

Der Faktor λ_{max} kann eingestellt werden und liegt im Bereich von [0,1]. Er sagt etwas darüber aus wie schnell man Richtung Zielwert konvergiert. Jedoch muss man beachten dass bei zu

hohen Werten von λ_{max} der Algorithmus instabil wird. Der maximale Wert, für die jeweilige Problemstellung, kann jedoch leicht in Offline-Simulationen herausgefunden werden.

G_t und β_t sind skalare Hilfsvariablen:

$$G_t = u_t^T P_{t-1} u_t$$
$$\beta_t = \frac{\left(\gamma^2 - \sigma_{t-1}^2\right)}{\delta_t^2} \qquad (19)$$

Eine wesentliche Erweiterung des Algorithmus ist die Rettungsprozedur. Im Falle von großen Störungen bzw. Sprüngen der Regelgrößen, die größer ± 3% der Zielwerte sind, gibt der Algorithmus keine Lösung bei der Schätzung der Parameter mehr zurück. Das lässt sich darauf Rückführen, dass E_{t-1} und S_t, erläutert in Abbildung 18, keinen Schnittpunkt mehr haben.

Die Rettungsprozedur versucht nun das Ellipsoid E_{t-1} so lange zu vergrößern bis ein Schnittpunkt vorhanden ist d.h. Der Parameterunsicherheitsfaktor σ_{t-1} wird vergrößert. Dabei kommt es auch zu einer Verschiebung des Mittelpunktes, der eine Schätzung der Modellparameter ist, bis die realen Parameter θ^* wieder innerhalb des Ellipsoids liegen.

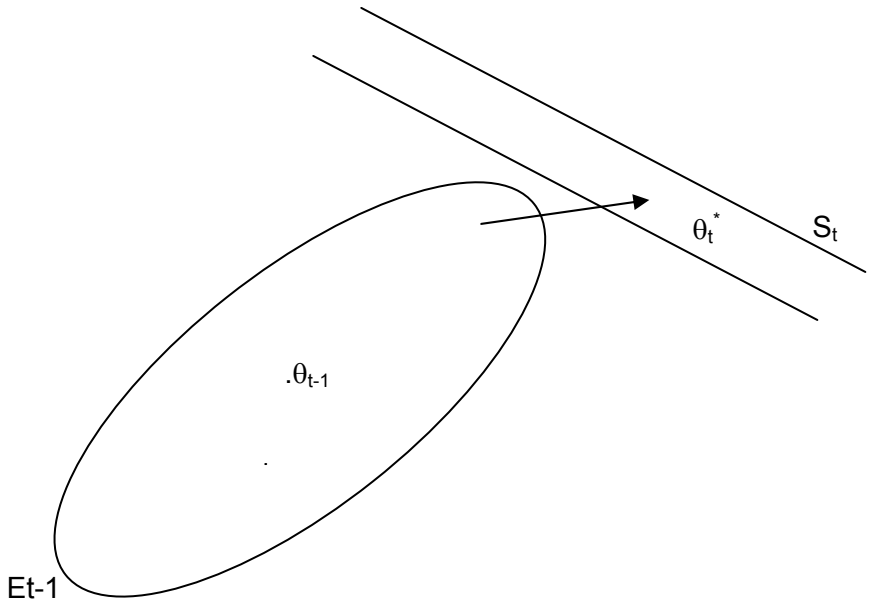

Abbildung 18: Notfallsprozedur

In der Notfallsprozedur wird überprüft ob der neue Unsicherheitsfaktor σt folgende Bedingung erfüllt:

$$\sigma_t^2 > 0 \tag{20}$$

Wenn diese Bedingung nicht erfüllt ist, wird die Notfallsprozedur wie folgt durchgeführt:

Berechnung einer skalaren Hilfsvariablen κ

$$\kappa = \begin{cases} \delta_t^2 + \gamma^2 - 2\gamma\left|\delta_t\right| & \text{wenn } \lambda_t \neq \lambda_{\max} \\ \lambda_{\max}\left(\dfrac{\delta_t^2}{1-\lambda_{\max}+\lambda_{\max}G_t} - \dfrac{\gamma^2}{1-\lambda_{\max}}\right) & \text{wenn } \lambda_t = \lambda_{\max} \end{cases} \tag{21}$$

Zurücksetzen des Unsicherheitsfaktors für den vorigen Run

$$\sigma_{t-1}^2 = \kappa + \zeta \tag{22}$$

Wobei ζ ein frei wählbarer Inflationsparameter ist. Er sagt etwas darüber aus wie schnell das Ellipsoid vergrößert wird um neue Modellparameter berechnen zu können.

Um nun eine Vorhersage der regelbaren Größe y_t (SD_1) zu erhalten muss man eine definierte Kostenfunktion minimieren. Diese Funktion hat normalerweise die Form des quadratischen Fehlers zwischen Zielwert und Vorhersage des Models. Sollen alle Regelgrößen gleich gewichtet werden, hat sie die Form:

$$\min_u \left(T - \theta^T u_t\right) \tag{23}$$

Wobei hier T der Zielwert ist.

Bei mehren Regelgrößen kann die Kostenfunktion wie folgt aussehen:

$$\min_u \left[\left(T - \theta^T u\right)^T W \left(T - \theta^T u\right)\right] \tag{24}$$

Hier ist W eine diagonale Matrix mit der Gewichtung für die einzelnen Regelgrößen. Je höher der Wert eines einzelnen Elementes desto wichtiger ist es, diese vorhergesagte Regelgröße nahe dem Zielwert zu halten.

3 Erstellung eines empirischen Modells für den Nitrid-prozess

Viele R2R-Algorithmen benötigen ein Initialmodel um für den ersten Run eine Vorhersage treffen zu können. Dieses Initialmodel, welches das Input/Output Verhalten des Ofens beschreibt, soll für den in Kapitel 2.3 beschriebenen Nitridprozess erstellt werden. Um das Verhalten des Prozesses näher zu untersuchen ist es sinnvoll zuerst die Abhängigkeiten an Hand der Nominalrezepte bzw. der historischen Datenaufzeichnungen (SPC) zu bestimmen. Bei der Analyse der SPC-Daten von vergangenen Ofenfahrten ist insbesondere die Plausibilität der Schichtdickenmessergebnisse zu überprüfen. Untersuchungen zeigten, dass mit diesen Datensätzen weder die Zeit- noch die Temperaturabhängigkeit der Schichtdicken in den einzelnen Zonen bestimmbar ist. Begründen lässt sich dies durch die händische Eingabe der Messergebnisse durch den Operator bzw. dass die Messergebnisse nicht mehr eindeutig einer bestimmten Ofenposition zuordenbar sind. Überdies sind die Auslenkungen der Stellgrößen durch den Prozessingenieur beim Nachregeln von produktiven Fahrten nur sehr gering (± 1°C bzw. ± 1 min) sodass eine eindeutige Zuordnung der Regelgrößen zu einer Stellgrößenänderung oftmals im Prozessrauschen untergeht. Diese Fakten schließen das Erstellen eines empirischen Modells aus historischen Daten aus.

Betrachtet man hingegen die Abscheidezeiten der Nominalrezepte von Prozessen die mit den gleichen Temperatureinstellungen verschiedene Schichtdicken produzieren, so lässt sich zumindest die Ordnung der Abhängigkeit der Schichtdicken von der Abscheidezeit bestimmen.

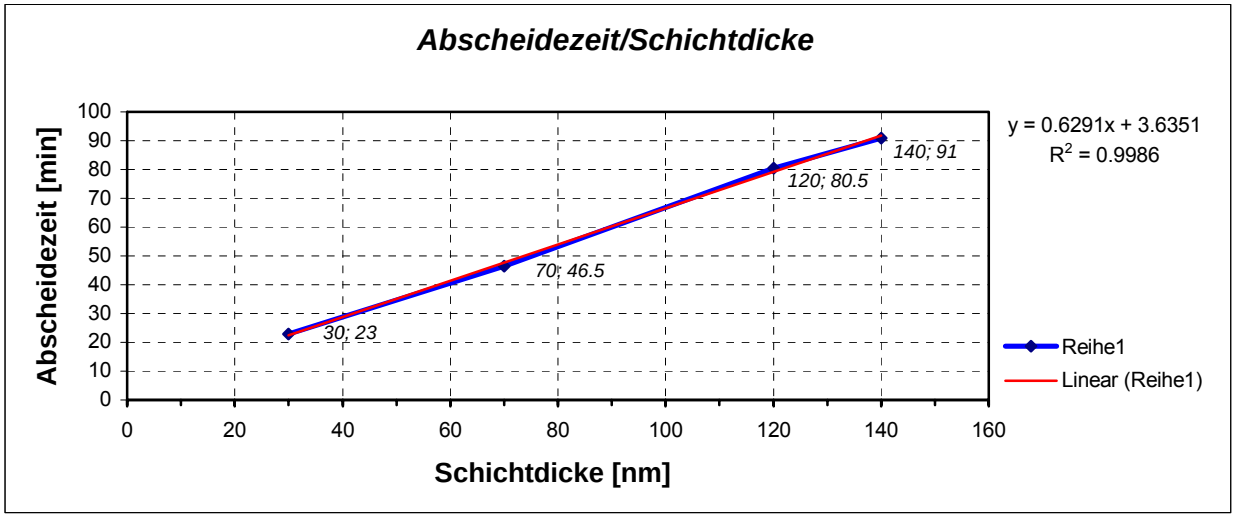

Abbildung 19: Nominale Abhängigkeit Schichtdicke/Abscheidezeit

In guter Näherung gilt für den untersuchten Nitridprozess eine lineare Zeitabhängigkeit der Schichtdicken über den gesamten Ofen und deckt sich mit praktischen Erfahrungen. Diese

Annahme verringert für die folgenden Versuchsreihen die Anzahl der Experimente, weil die Abscheidezeit nicht mehr in Linearkombination von Faktoren oder in quadratischer Form in die Modellbildung eingeht.

Grundsätzlich soll der Nitridprozess über das Einstellen der fünf Stellgrößen Abscheidezeit t und der Temperaturen T_1,T_2 ,T_4 ,T_5 geregelt werden. Die Temperatur der mittleren Heizzone T_3 soll konstant gehalten werden, da in diesem Bereich die Schichtdicken (SD) über die Abscheidezeit zu regeln sind. Ein Nachregeln von T_3 kann unter Umständen zu veränderten Schichteigenschaften führen. Außerdem muss bei der Prozessqualifikation für ein bestimmtes Produkt in der Entwicklungsphase die Nominaltemperatur für T_3 festgelegt werden.

Für die Experimente wurde jener ASM A400 Vertikalofen mit den stabilsten Ergebnissen verwendet. Von den auf diesem Ofen gefahrenen Prozessen wurde der Durchsatzstärkste herangezogen. Die Experimente wurden auf Ofen V03/2D für den Nitridprozess mit der Prozessbezeichnung VEH 10008 unter folgenden Bedingungen durchgeführt:

Zielschichtdicke:	Target = 140 nm
Temperatur:	T_3 = 780 °C (konstant gehalten)
Stellgrößenbereich:	T_1 = [792,802]°C
	T_2 = [790,800]°C
	T_4 = [753,763]°C
	T_5 = [750,760]°C
	t = [86,96] min
Regelgrößen:	SD_1, SD_2,…..,SD_7 [nm]

Der Ofen auf dem die Versuche durchgeführt wurden stand die ganze Zeit im Produktionsprozess. Das heißt, dass zwischen den einzelnen Experimenten produktive Fahrten durchgeführt wurden. Dies kann bei mehreren Fahrten zwischen zwei Experimenten zu einer Zustandsveränderung des Ofens führen. Daraus resultierend ergibt sich ein weniger gutes Prozessmodell, was allerdings in Hinblick auf den verwendeten R2R-Algorithmus kein allzu großes Problem darstellt, weil dieser in der Lage ist innerhalb von wenigen Runs ein geeignetes Modell zu finden.

3.1 Versuchsreihe mit zusammengefassten Temperaturen

Bei der ersten Versuchsreihe – bestehend aus 7 Versuchen – wurden die Temperaturen T_1 und T_2 zur Gruppe $T_{1,2}$ bzw. T_4 und T_5 zur Gruppe $T_{4,5}$ zusammengefasst um die Anzahl der nötigen Versuche niedrig zu halten und um das prinzipielle Verhalten des Ofenprozesses zu evaluieren. Dies ist möglich weil diese Zonen im Vergleich zur mittleren Temperaturzone relativ geringe Abmessungen haben. Die Differenzen zwischen T_1 und T_2 bzw. T_4 und T_5 wurden bei allen Experimenten konstant gehalten. Zusätzlich wurden noch 7 weitere

Datensätze (Run 8 -14) rechnerisch aus den ersten 4 Runs ermittelt. Diese Ermittlung setzt ein lineares Schichtdicken/Zeitverhalten voraus. Run 5 – 7 sind so genannte Centerfahrten. Sie sind notwendig um eine Aussage über das Prozessrauschen machen zu können. Die dort eingestellten Stellgrößen entsprechen den bei produktiven Fahrten verwendeten. Um bei diesem Prozess signifikante Input/Output Abhängigkeiten zu erzielen wurden die Stellgrößen um ±5°C bzw. ±5 min von Run zu Run variiert. Die Ergebnisse dieser Versuchsreihe befinden sich in Tabelle 4. Dort werden die Temperaturen T_2 bzw. T_5 stellvertretend für die Temperaturgruppen $T_{1,2}$ bzw. $T_{4,5}$ angeführt.

Tabelle 4: Datensätze der ersten Versuchsreihe

Run	$T_{1\ 2}$	$T_{4\ 5}$	t	SD_1	SD_2	SD_3	SD_4	SD_5	SD_6	SD_7
1	792	760	91	126.1	123.6	127.9	133.9	136.4	137.5	144.6
2	802	750	91	149.5	146.2	143.8	140.5	131.4	121.1	119.4
3	792	750	91	129.3	128.8	138.2	142.7	138.1	125.1	119.9
4	802	760	91	143.3	139.2	134.3	134	134.2	137.2	144.9
5	797	755	91	142.3	140.4	141.6	142.6	138.8	133.4	131.3
6	797	755	91	142.5	140.6	141.8	142.5	138.7	133.5	132
7	797	755	91	141.3	139.4	140.8	142.5	138.5	133	131.3
8	792	760	96	132.6	130.9	134.4	141.0	143.2	144.2	152.7
9	802	750	86	142.0	137.7	135.1	133.7	125.0	114.6	113.4
10	792	750	86	122.8	122.5	129.9	135.3	130.9	117.7	114.2
11	792	750	96	135.6	136.2	144.9	151.1	145.6	131.2	126.1
12	802	760	96	152.3	145.9	141.7	142.1	140.6	144.7	153.3
13	792	760	86	119.3	117.7	121.9	126.2	128.4	129.6	135.6
14	802	760	86	135.8	131.2	127.8	127.0	126.5	130.4	137.6

Zur Planung und Auswertung der Experimente wurde das hauseigene Statistikpaket CEDA/Cornerstone verwendet wobei nach den Grundsätzen des in 2.6 beschriebenen DoE vorgegangen wurde. Als Ergebnis erlangt man ein Prozessmodell für die 7 Schichtdicken an unterschiedlichen Ofenpositionen mit folgenden Modellparametern:

Coefficient/Std Error Table

Term	SD1	SD2	SD3	SD4	SD5	SD6	SD7
Constant	142,033 0,487557	140,133 0,437511	141,400 0,523104	142,533 0,170720	138,667 0,180514	133,300 0,380030	132,221 0,217399
T1_2	9,00444 0,259522	8,14889 0,232883	2,85111 0,278443	-0,360897 0,0916913	-2,28269 0,0969518	-0,989103 0,204109	Out
T4_5	-2,22111 0,259522	-2,86556 0,232883	-5,01778 0,278443	-3,77244 0,0916913	0,499359 0,0969518	6,63910 0,204109	12,0898 0,245604
t	7,60667 0,327063	6,92333 0,293491	7,67667 0,350909	7,64615 0,115981	8,03846 0,122635	7,45385 0,258180	7,39023 0,308458
T1_2 * T4_5	Out	Out	Out	0,377564 0,0916913	0,999359 0,0969518	0,655769 0,204109	Out
T4_5^2	-5,01222 0,552326	-5,88444 0,495631	-5,43222 0,592595	-4,74423 0,193785	-3,61603 0,204903	-3,03910 0,431374	Out
R-Squared	**0,9947**	**0,9952**	**0,9898**	**0,9988**	**0,9987**	**0,9966**	**0,9967**
Adj. R-Squared	**0,992408**	**0,992998**	**0,985287**	**0,998090**	**0,997928**	**0,994444**	**0,996118**
RMS Error	**0,844473684**	**0,757790811**	**0,906042814**	**0,295695616**	**0,312660215**	**0,658231897**	**0,805750406**
Residual DF	**9**	**9**	**9**	**8**	**8**	**8**	**11**

Abbildung 20: Modellparameter bei 3 Stellgrößen

Die Gleichung für Schichtdicke 1 (SD_1) die das Prozessverhalten beschreibt sieht für das erstellte Modell wie folgt aus:

$$SD_1 = 142,03 + 9,00 \cdot T_{1,2} - 2,22 \cdot T_{4,5} + 7,61 \cdot t - 5,01 \cdot T_{4,5}^2 \qquad (25)$$

Der R-Squared-Wert sagt etwas darüber aus wie gut das Modell die Variabilität der Regelgrößen erklärt. Je näher dieser Wert bei 1 ist, desto besser passt das angenommene Prozessmodell zu den Daten aus den Experimenten. Allerdings bedeutet ein hoher Wert nicht zwangsläufig auch ein qualitativ gutes Modell. Gleiches gilt für den Ajusted-R-Squared-Wert. Er sagt etwas über die Güte des Modelles aus,wenn man Terme hinzu bzw. weg gibt, wobei seine verhältnismäßige Veränderung gegenüber dem R-Squared-Wert zu betrachten ist. Der RMS-Error ist eine Schätzung der Standardabweichung des Modellfehlers.

Am Beispiel von Schichtdicke 1 – gleiches gilt für die anderen Regelgrößen - werden zwei Analysemethoden die bei der Auswertung der Datensätze durchgeführt wurden näher beschrieben.

Analysis of Variance (ANOVA)

Die Analyse der Varianz des Modells zeigt die Summe der Variabilität, die durch das Modell erklärt werden kann. Die erste Spalte zeigt die Quelle der Regelgrößenvariation und unterscheidet wie viel von dieser durch die Regression erklärt werden kann und wie viel nicht.

Tabelle 5: Analyse der Varianz

Source	Df	Sum of Squares	Mean Square	F-Value	Significance
Total	13	1221.16			
Regression	4	1214.74	303.68	425.84	0.0000
Linear	3	1108.16	369.39	517.97	0.0000
Non-linear	1	58.73	58.73	82.35	0.0000
Residual	9	6.42	0.71		
Lack of Fit	7	5.59	0.80	1.93	0.3828
Pure Error	2	0.83	0.41		

Df steht für die Freiheitsgrade und gibt uns Auskunft darüber, wie viele unabhängige Parameter notwendig sind um eine Modellkomponente (linear, nichtlinear,…) zu beschreiben. Sum of Squares ist ein Maß für die Variation der Regelgrößen. Die Mean Square Werte sind die Sum of Squares/Df. Werte von signifikanten Komponenten tendieren dazu größer als die Mean Square Werte der Residuen zu sein. Die F-Values sind die Werte aus einem statistischen F-Test. In der letzten Spalte stehen die Signifikanzen der einzelnen Komponenten: Je kleiner die Werte, desto signifikanter sind die Komponenten.

Residuenanalyse

Die Residuen sind die Differenzen zwischen den aktuellen Regelgrößen (Schichtdicken aus den Versuchen) und der Vorhersage der Schichtdicken durch das verwendete Modell. Eine guter Fit des Modells bedingt gleichmäßig verteilte Residuen die entlang einer Geraden liegen.

In Abbildung 21 sieht man den Residuenplot für Schichtdicke 1 (Gasauslass) der einen relativ guten Fit zeigt.

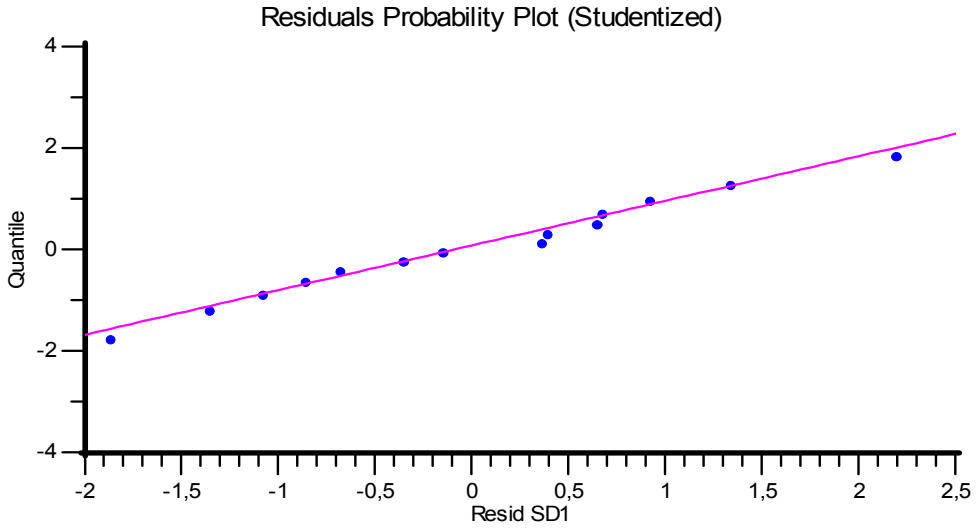

Abbildung 21: Residuenplot für SD_1

Über diese zwei Analysemethoden hinaus gibt es noch einige andere. Genaueres dazu steht in der Onlinehilfe zu CEDA/Cornerstone bzw. in [16]. Daraus ging hervor, dass die Abhängigkeiten zwischen Input ($T_{1,2}$, $T_{4,5}$, t) und Output (SD_1 – SD_7) nicht bzw. nur sehr schlecht durch das Bilden eines linearen Models beschrieben werden können.

Um das erstellte quadratische Prozessmodell praktisch zu überprüfen wurde eine weitere Testfahrt mit den vom Modell vorhergesagten Stellgrößen durchgeführt und die vorhergesagten Schichtdicken mit den tatsächlich erreichten verglichen.

Tabelle 6: Verifikation des Temperaturgruppenmodells

	$T_{1,2}$	$T_{4,5}$	t	SD_1	SD_2	SD_3	SD_4	SD_5	SD_6	SD_7
Vorhersage Modell	795	757	92,35	142,3	139,8	140,0	141,8	140,0	137,8	140,0
Testfahrt	795	757	92,35	138,9	135,4	136,7	138,4	136,9	134,8	140,3

Obige Werte lassen erkennen, dass die Vorhersage nur sehr schlecht mit den Werten der Testfahrt übereinstimmt. Insbesondere die Positionen 2, 3, 5, 6 zeigen eine erhebliche Abweichung vom Zielwert. Dadurch wird ersichtlich, dass eine Regelung mit zusammengefassten Stellgrößen nicht das gewünschte Resultat liefert.

Das quadratische Verhalten und die Anzahl der Stellgrößen machten beim Design of Experiment (DoE) für die zweite Versuchreihe eine D-Optimale Versuchsanordnung notwendig.

3.2 D-Optimale Vesuchsanordnung

Bei dieser zweiten Versuchsreihe kann durch diese spezielle Anordnung die Anzahl der Experimente auf 24 beschränkt werden. Dies war bei einer Dauer von ca. 5 Stunden pro Experiment auch unbedingt notwendig. Durch eine geschickte Veränderung der Stellgrößen können auch nichtlineare Abhängigkeiten mit hoher Wahrscheinlichkeit erkannt werden. Es gelten die gleichen Bedingungen wie bei der ersten Versuchsreihe. Jedoch wurden diesmal keine Temperaturgruppen gebildet. Es soll ein Modell in der Form

$$SD_i = f(T_1, T_2, T_4, T_5, t) \tag{26}$$

erstellt werden.

Ergebnis der Auswertung und Analyse mit CEDA/Cornerstone waren zwei Modelle, die in ihrer allgemeinen Form unten angeführt sind. Dabei wurden auch die Datensätze der vorangegangenen Versuchsreihe mitberücksichtigt. Um aus diesen Datensätzen zwei aussagekräftige Modelle erstellen zu können wurden einzelne Datensätze nicht berücksichtigt. Möglich war dies durch eine sogenannte „Outlieranalyse“. Dort werden die

Daten statistisch darauf untersucht ob sie in das erstellte Modell passen. Die Datensätze der Experimente die zu diesen Modellen führten liegen im Anhang.

3.2.1 Initialmodell

Das erste Modell, welches als Initialmodell für die Simulationen dient und gleichzeitig die Ordnung des Regelmodells festlegt hat folgende Form:

$$y = \beta_{n,m} x_n + e_m \quad \text{mit:} \quad n = 10; \quad m = 7 \tag{27}$$

$$\beta = \begin{bmatrix} 142 & 10 & -1.2 & 7.3 & -2 & 0 & -0.7 & 0 & -0.9 & -4 \\ 140 & -2.4 & 10.6 & 7.7 & -3 & 0 & 0 & 0 & -2.2 & -3 \\ 141 & -2.3 & 5 & 7 & -5 & 0.7 & 0 & 0.3 & -2.7 & -2.3 \\ 142 & 0 & -0.4 & 7.3 & -4 & 0.7 & 0.4 & 0 & -2.4 & -2.1 \\ 138 & 0 & -2.2 & 6.5 & 0.7 & -0.7 & -0.6 & 1.5 & -3.2 & 0 \\ 133 & 0 & -1.3 & 6.3 & 9 & -2.3 & 0 & 0.8 & -2.5 & 0 \\ 132 & 0 & 0 & 6.7 & 2.9 & 9 & 0 & 0.3 & 0 & 0 \end{bmatrix}$$

Wobei y der Vektor der Regelgrößen (SD_1, SD_2…, SD_7) ist und x_n der Vektor der regelbaren Größen (T_1, T_2, t, T_4, T_5) mit bilinearen (T_2T_4,…) und quadratischen (T_2^2,…) Termen. Bei diesem Modell wurden die Parameter β bereits abgeändert um bei den Simulationen einen Anfangsmodellfehler zu erzeugen. Die exakten Werte die aus den Analysen hervorgehen findet man in Abbildung 22.

In skalarer Form für nur eine Regelgröße sieht im speziellen Fall das Initialmodell wie folgt aus:

$$SD_1 = \beta_{1,1} + \beta_{1,2}T_1 + \beta_{1,3}T_2 + \beta_{1,4}t + \beta_{1,5}T_4 + \beta_{1,6}T_5 + \beta_{1,7}T_2T_4 + \beta_{1,8}T_2T_5 + \beta_{1,9}T_2^2 + \beta_{1,10}T_4^2 \tag{28}$$

Dieses skalare Modell beschreibt mit genügender Genauigkeit den untersuchten Nitrid Prozess um für den Algorithmus als Initialmodell verwendet werden zu können.

Bei den einzelnen Simulationen wurde das Initialmodell unterschiedlich modifiziert. Im ersten Fall wurden nur die linearen Terme verwendet:

$$SD_1 = \beta_{1,1} + \beta_{1,2}T_1 + \beta_{1,3}T + \beta_{1,4}t + \beta_{1,5}T_4 + \beta_{1,6}T_5 \tag{29}$$

In skalarer Form für nur eine Regelgröße sieht im speziellen Fall das Initialmodell wie folgt aus:

$$SD_1 = \beta_{1,1} + \beta_{1,2}T_1 + \beta_{1,3}T_2 + \beta_{1,4}t + \beta_{1,5}T_4 + \beta_{1,6}T_5 + \beta_{1,7}T_2T_4 + \beta_{1,8}T_2T_5 + \beta_{1,9}T_2^2 + \beta_{1,10}T_4^2 \tag{30}$$

Coefficient/Std Error Table

Term	SD1	SD2	SD3	SD4	SD5	SD6	SD7
Constant	142,238 0,232739	140,080 0,228199	141,330 0,205859	142,477 0,187475	138,481 0,253722	133,260 0,301485	132,395 0,229910
T1	10,7522 0,351714	-2,47429 0,344512	-2,30180 0,357812	Out	Out	Out	Out
T2	-1,26753 0,312214	10,6681 0,305784	4,97883 0,335737	-0,410507 0,113814	-2,29255 0,154061	-1,32668 0,182289	Out
T4	-2,22950 0,142000	-3,02224 0,138305	-5,63472 0,280452	-4,43143 0,221143	0,808935 0,291723	9,50739 0,346063	2,68421 0,497365
T5	Out	Out	0,733210 0,271673	0,722197 0,207794	-0,712854 0,275817	-2,30466 0,328158	9,37942 0,471276
t	7,39434 0,163380	7,74692 0,160154	7,58700 0,143202	7,35396 0,134082	6,98859 0,170952	6,78039 0,194781	6,95545 0,297734
T2 * T4	-0,705431 0,148093	Out	Out	0,665737 0,115517	-0,407041 0,351888	Out	Out
T2 * T5	Out	Out	0,260060 0,122935	Out	1,54295 0,337521	0,826272 0,180775	0,536106 0,277905
T2^2	-0,964722 0,417835	-2,23767 0,409395	-2,73577 0,401502	-2,46366 0,347399	-3,27121 0,306733	-2,86814 0,363724	Out
T4^2	-4,43716 0,408685	-3,34878 0,400046	-2,38325 0,381846	-2,48104 0,329006	Out	Out	Out
R-Squared	**0,9978**	**0,9982**	**0,9982**	**0,9974**	**0,9932**	**0,9949**	**0,9934**
Adj. R-Squared	**0,996741**	**0,997470**	**0,997175**	**0,996244**	**0,990082**	**0,992978**	**0,991889**
RMS Error	**0,540179477**	**0,529909159**	**0,447499949**	**0,421969172**	**0,579961273**	**0,690223064**	**1,062619645**
Residual DF	**15**	**16**	**14**	**15**	**15**	**16**	**18**

Abbildung 22: Modellparameter bei 5 Stellgrößen

Dieses skalare Modell beschreibt mit genügender Genauigkeit den untersuchten Nitrid Prozess um für den Algorithmus als Initialmodell verwendet werden zu können.

Bei den einzelnen Simulationen wurde das Initialmodell unterschiedlich modifiziert. Im ersten Fall wurden nur die linearen Terme verwendet:

$$SD_1 = \beta_{1,1} + \beta_{1,2}T_1 + \beta_{1,3}T + \beta_{1,4}t + \beta_{1,5}T_4 + \beta_{1,6}T_5 \tag{31}$$

Im zweiten Fall wurden auch die bilinearen Terme hinzugefügt:

$$SD_1 = \beta_{1,1} + \beta_{1,2}T_1 + \beta_{1,3}T + \beta_{1,4}t + \beta_{1,5}T_4 + \beta_{1,6}T_5 + \beta_{1,7}T_2T_4 + \beta_{1,8}T_2T_5 \tag{32}$$

Im dritten Fall diente das komplette Modell inkl. quadratischer Terme (30) als Initialmodell.

3.2.2 Prozessmodell

Das zweite Modell dient in dieser Simulation als reales Prozessmodell. Das heißt, dass mit diesem Modell die im realen Fall gemessenen Schichtdicken berechnet werden. Es hat dieselbe Form wie obiges Modell, jedoch mit dem Unterschied, dass x_n ein Vektor mit 16 Faktoren und β eine [16 x 7] Matrix ist.

Tabelle 7: Parameter β Prozessmodell

	SD_1	SD_2	SD_3	SD_4	SD_5	SD_6	SD_7
Konstante	141.79	139.43	141.10	142.11	138.37	133.11	132.21
T_1	14.52	-2.52	-1.81	-0.45	-0.25	0.00	1.74
T_2	-5.37	10.66	4.44	-0.15	-2.03	-1.26	-1.79
T_4	-2.49	-3.09	-5.69	-4.29	1.89	10.59	3.18
T_5	0.00	0.01	0.88	0.65	-1.79	-3.41	9.03
t	8.14	7.86	7.70	7.65	7.54	7.34	7.84
$T_1 * T_2$	0.00	0.00	4.96	5.22	3.12	0.00	0.00
$T_2 * T_4$	-0.90	0.00	-1.43	-1.15	-0.87	0.00	0.00
$T_2 * T_5$	0.00	0.00	1.69	1.76	1.90	0.82	0.00
$T_4 * T_5$	0.00	-1.84	0.00	0.00	0.00	0.00	0.00
$t * T_1$	3.86	0.62	1.14	0.00	1.31	0.00	2.26
$t * T_2$	-3.62	0.00	-1.12	0.00	-1.63	-0.62	-2.37
$t * T_4$	0.00	0.00	0.00	0.00	0.40	0.74	0.00
T_1^2	-4.61	0.00	-4.87	-4.70	-3.05	0.00	-2.89
T_2^2	0.00	-3.14	-4.91	-4.94	-3.28	-2.67	0.00
T_5^2	0.00	0.00	0.00	0.00	0.00	0.00	2.82

Erstellt wurde dieses Modell wie in Kapitel 3.1. Stellvertretend ist unten das Input/Output Verhalten für SD_1 angeführt

$$SD_1 = 141.79 + 14.52T_1 - 5.37T_2 - 2.49T_4 + 8.14t - 0.9T_2T_4 + 3.86tT_1 - 3.62tT_2 - 4.61T_1^2 \quad (33)$$

4 Implementierung des DHOBE Algorithmus in eine MatLab Simulation

4.1 Abbildung des Algorithmus in Matlab

Das Modell, welches den Zustand einer Anlage beschreibt, ist für viele Prozesse in der Halbleiterindustrie linear in den Parametern. Diese Form beinhaltet Interaktionsterme und Terme zweiter Ordnung wobei jedoch die Parameter (Koeffizienten) Linear sind. Die Implementierung des Algorithmus in eine Simulation wurde mit Hilfe von Matlab 6.5 durchgeführt. Dies war notwendig, um das System nichtlinearer Gleichungen (welches bei Anwendung des Algorithmus auf ein MIMO System entsteht) zu lösen.

Für ein MISO System sieht das Modell wie folgt aus:

$$y_t = \theta^{*T} u_t + e_t \tag{34}$$

Hierin sind

y_t die Regelgröße (Schichtdicke),

θ Parametervektor,

u_t Vektor der Stellgrößen (Temperaturen, Zeit)

e_t der zufällige Fehler von Run zu Run (Rauschen)

Der Algorithmus kann auf solche Modelle angewendet werden, indem er einen Satz von Parameterschätzungen produziert. Diese Schätzungen sind begrenzt durch einen Ellipsoid:

$$E_t = \left\{ \theta_t \in \Re^n : \left(\theta_t - \theta_t^+\right)^T P_t^{-1} \left(\theta_t - \theta_t^+\right) \le \sigma_t^2 \right\} \tag{35}$$

wobei

P_{t-1} eine positiv definite Matrix ist, die die gleiche Dimension wie der Parametervektor hat. Sie beschreibt die Form des Ellipsoids.

θ_t^+ der Mittelpunkt des Ellipsoids ist und als aktuelle Parameterschätzung verwendet wird.

σ_t^2 die Unsicherheit der Parameterschätzung definiert durch die Größe des Ellipsoids ist.

All die eben erwähnten Parameter werden von Run zu Run neu berechnet und dazu verwendet die Modellparameter für den nächsten Run vorherzusagen. Die einzig hinreichende Bedingung, um die Konvergenz des Algorithmus sicherzustellen,

ist Kenntnis über die strikten Grenzen des Prozessrauschens e_t zu haben. Die Bedingung

$$|e_t| < \gamma \tag{36}$$

muss erfüllt sein.

Der Berechnungsablauf des DHOBE Algorithmus sieht wie folgt aus:

Schritt 1)

Vorhersage der Regelgröße (SD) mit Initialmodell:

$$y_t = u_t^T \theta_{t-1} + e_t \tag{37}$$

Dabei müssen die Stellgrößen u_t über die Minimierung der Kostenfunktion

$$\min_u \left(T - \theta_{t-1}^T u_t \right) \tag{38}$$

berechnet werden.

Schritt 2)

Bei jedem Run muss der Vorhersagefehler berechnet werden:

$$\delta_t = y_{t,Messung} - y_t \tag{39}$$

Wobei

$y_{t,Messung}$ die gemessene Schichtdicke[1],

y_t die vorhergesagte Schichtdicke und

δ_t der Vorhersagefehler

ist.

[1] In der Simulation wird die gemessene Schichtdicke dadurch erhalten, dass man die vorhergesagten Stellgrößen in das Prozessmodell einsetzt. Zusätzlich ist es notwendig Fehler (Rauschen, Shift, Drift) einzubauen, um das Regelverhalten des Algorithmus zu untersuchen.

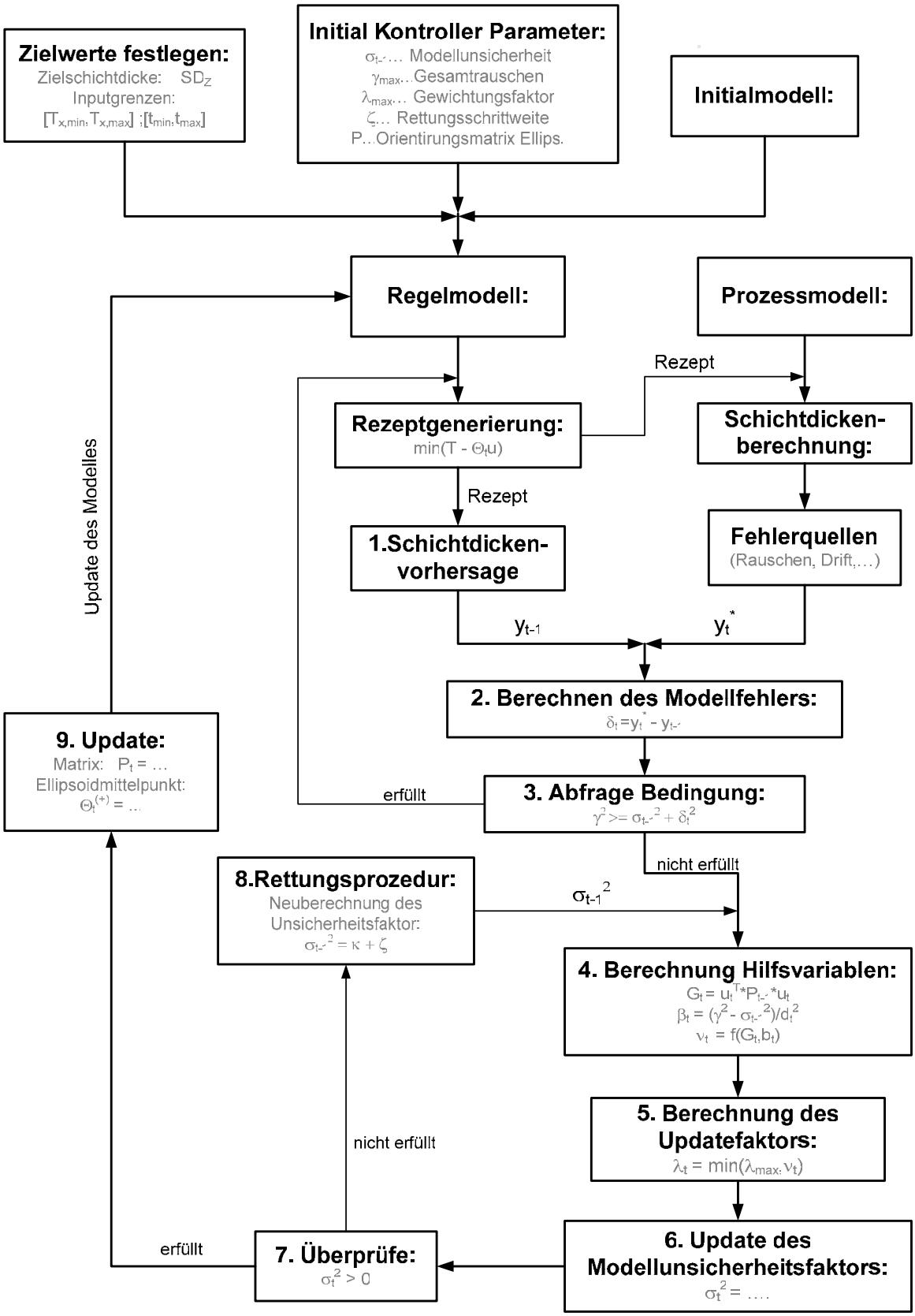

Abbildung 23: Blockdiagramm des DHOBE-Algorithmus in Matlab

Schritt 3)

Überprüfung der Ungleichung:

$$\sigma_{t-1}^2 + \delta_t^2 \leq \gamma^2 \tag{40}$$

Wenn diese Ungleichung erfüllt ist, liegt der Prozess innerhalb von akzeptablen Grenzen. Es ist nicht notwendig die Modellkoeffizienten zu erneuern. Das Rezept (Stellgrößen) bleibt für den nächsten Run gleich. Man beginnt wieder bei Schritt 1. Wenn die Bedingung nicht erfüllt ist, erneuert der Algorithmus in den weiteren Schritten die Modellkoeffizienten.

Schritt 4)

Berechnung von zwei Hilfsvariablen:

$$\begin{aligned} G_t &= u_t^T P_{t-1} u_t \\ \beta_t &= \left(\frac{\gamma^2 - \sigma_{t-1}^2}{\delta_t^2} \right) \end{aligned} \tag{41}$$

Schritt 5)

Berechnung der Hilfsvariable ν_t und des „Updatefaktors“ λ_t:

$$\lambda_t = \min\left(\lambda_{\max}, \nu_t\right),$$

wobei

$$\nu_t = \begin{cases} \lambda_{\max} & \text{wenn } \delta_t = 0 \\ \dfrac{(1-\beta_t)}{2} & \text{wenn } G_t = 1 \\ \dfrac{\left(\left(1 - \sqrt{\dfrac{G_t}{(1+\beta_t (G_t - 1))}}\right)\right)}{(1-G_t)} & \text{wenn } \beta_t (G_t - 1) + 1 > 0 \\ \lambda_{\max} & \text{wenn } \beta_t (G_t - 1) + 1 \leq 0 \end{cases} \tag{42}$$

Der Updatefaktor λ_t bestimmt die Geschwindigkeit der Konvergenz in Richtung Zielwert. Die Wahl von λ_{max} [0,1], im Zusammenspiel mit der Wahl von P_t und dem

tatsächlichen Prozessrauschen ist ein entscheidendes Kriterium für die Stabilität des Algorithmus.

Bei zu hohem λ_{max} kommt es zu einem Aufschwingen. Wählt man hingegen λ_{max} zu klein konvergiert die Regelgröße zu langsam.

Schritt 6)

Berechnen des Modellunsicherheitsfaktors:

$$\sigma_t^2 = (1-\lambda_t)\,\sigma_{t-1}^2 + \lambda_t \gamma^2 - \frac{\lambda_t (1-\lambda_t)\,\delta_t^2}{1-\lambda_t + \lambda_t G_t} \tag{43}$$

Schritt 7)

Überprüfe ob Bedingung

$$\sigma_t^2 > 0 \tag{44}$$

erfüllt ist.

Wenn ja, überspringt man die Rettungsprozedur und fährt mit **Schritt 9** fort.

Schritt 8)

Diesen Schritt nennt man die Rettungsprozedur. Er ist erforderlich wenn σ_t^2 negativ wird und es keinen Schnittpunkt zwischen Ellipsoid und der Fläche, die die Messergebnisse widerspiegelt, gibt.

Berechnung einer Hilfsvariablen κ und Rückstellen des Parameterunsicherheitsfaktors σ_{t-1}:

$$\kappa = \begin{cases} \delta_t^2 + \gamma^2 - 2\gamma\left|\delta_t\right| & \text{wenn } \lambda_t \neq \lambda_{max} \\ \lambda_{max}\left(\dfrac{\delta_t^2}{1-\lambda_{max}+\lambda_{max}G_t} - \dfrac{\gamma^2}{1-\lambda_{max}}\right) & \text{wenn } \lambda_t = \lambda_{max} \end{cases}$$

$$\sigma_{t-1}^2 = \kappa + \zeta \tag{45}$$

ζ bezeichnet den Inflationsparameter, der das zusammengebrochene Ellipsoid vergrößert um, für den nächsten Run wieder Modellkoeffizienten zu erhalten. Anschließend kehrt man zurück zu **Schritt 4** und berechnet σ_t^2 erneut.

Schritt 9)

Berechnung der rekursiven Matrix der kleinsten Quadrate P_t und des neuen Mittelpunktes des Ellipsoids θ_t^+.

$$P_t = \left(\frac{1}{1-\lambda_t}\right)\left(P_{t-1} - \frac{\lambda_t P_{t-1} u_t u_t^T P_{t-1}}{1-\lambda_t + \lambda_t u_t^T P_{t-1} u_t}\right) \tag{46}$$

$$\theta_t^+ = \theta_{t-1}^+ + \lambda_t P_t u_t \delta_t$$

Die Koordinaten des neuen Mittelpunktes θ_t^+ sind die Modellkoeffizienten für den nächsten Run. Danach beginnt man für den nächsten Run wieder bei **Schritt 1**.

4.2 Parametereinstellung:

Um den DHOBE Kontrollalgorithmus zu implementieren und ihn an die Prozessbedingungen und das Equipment anzupassen, ist es notwendig die einstellbaren Parameter entsprechend zu konfigurieren. Um einen stabilen Einsatz gewährleisten zu können, ist dieser Punkt, ebenso wie die Wahl eines guten Initialmodells entscheidend. Im Folgenden wird die Wahl dieser Parameter und deren Einfluss auf das Kontrollverhalten diskutiert. Zusätzlich wird darauf eingegangen welche Daten der Algorithmus für einen ordentlichen Betrieb benötigt.

- Erstellen eines Initialmodells, welches man aus einem DoE erhält
- Definition der Zielwerte (gewünschte Schichtdicke)
- Festlegen des Stellgrößenbereiches d.h. festlegen jenes Bereiches in welchem das Initialmodell gültig ist und der Prozess physikalisch und auch technisch betrieben werden kann. Damit wird sichergestellt, dass der Algorithmus nur Stellgrößenwerte innerhalb dieses Bereiches vorschlägt.
- Angabe eines Initialrezeptes (Stellgrößen) mit welchem das Equipment (ASM-Ofen) im Normalfall betrieben wird.

Die genannten Informationen lassen sich einerseits aus Experimenten und andererseits aus Rezepten von bereits betriebenen Öfen gewinnen.

Der nächste Schritt, ist die Einstellung der algorithmusrelevanten Parameter. Das Finden dieser Werte ist für einen reibungslosen Betrieb nahe am Optimum nicht ganz einfach. Es lässt sich jedoch mit einigen Offline-Simulationen bewerkstelligen. Zu diesen Parametern gehören:

- die Grenzen des zufälligen R2R Fehlers (γ), die man aus Herstellerangaben, aus historischen Daten oder aus den vorangegangenen Offline-Experimenten erhält.
- der maximale Update-faktor λ_{max}
- die Rettungsschrittweite κ welche im Normalfall auf 1 gesetzt werden kann
- die Orientierungsmatrix P welche beim ersten Run meist die Einheitsmatrix I ist
- der Parameterunsicherheitsfaktor σ

Die Größe des zufälligen Fehlers γ ist für das R2R Regelverhalten sehr wichtig. Wählt man γ um einiges größer als die nicht beeinflussbare R2R Variabilität des Prozesses, so erhöht man zwar die Stabilität des Algorithmus.Man nimmt dabei jedoch in Kauf, dass nur bei signifikanten Sprüngen eine Erneuerung des Modells durchgeführt wird. Kleinere Abweichungen, die eine konstante Drift hervorrufen würde, bleiben so einige Zeit unerkannt und die Sensibilität des Algorithmus geht verloren. Simulationen für den konkreten Fall haben gezeigt, dass für γ im Bereich von 0,8 bis 1,2 gute Ergebnisse erzielt werden. Diese Werte entsprechen in etwa der zwei- bis dreifachen der realen R2R Standardabweichung.

Für die Konvergenzeigenschaften des Algorithmus ist der Update Faktor λ_{max} gemeinsam mit der Wahl der Orientierungsmatrix P verantwortlich. Wählt man λ_{max} zu klein, so konvergieren die Regelgrößen nur sehr langsam in Richtung Zielwert. Ist λ_{max} zu groß ändert sich die Größe des Ellipsoids so schnell, dass die realen Modellparameter nicht mehr innerhalb des Ellipsoids liegen. Dies hat im schlimmsten Fall zur Folge, dass die Rettungsprozedur immer und immer wieder aufgerufen wird und kein Ergebnis mehr zustande kommt bzw. dass der Algorithmus instabil wird. Simulationen haben gezeigt, dass λ_{max} bei Messverzögerungen kleiner gewählt werden muss. Der Bereich von λ_{max} liegt im Allgemeinen zwischen 0,1 und 0,6.

Grundsätzlich muss die Parametereinstellung an das Equipment und an die darauf laufenden Prozesse angepasst werden. Dies kann teilweise in Offline-Simulationen erfolgen. Die Feineinstellung, die für einen optimalen Betrieb notwendig ist, sollte jedoch an realen Ofenfahrten durchgeführt werden um sicherzustellen, dass der Kontroller möglichst nahe dem optimalen Arbeitspunkt operiert.

5 Simulation des DHOBE Kontrollers für den Nitridprozess

5.1 Performance und Bewertungskriterien

Grundsätzlich muss jeder R2R-Algorithmus folgende Anforderungen erfüllen, um im praktischen Betrieb einsetzbar zu sein. Er soll

- die Regelgrößen möglichst nahe am gewünschten Zielwert zu halten
- auf Prozessstörungen, die den Prozessoutput beeinflussen passend reagieren
- auf nicht beeinflussbare zufällige Störungen, keine Reglertätigkeit setzen d.h. dass innerhalb der Grenzen des Prozessrauschen nicht geregelt werden darf
- die Regelgrößen möglichst rasch wieder in Richtung Zielwert korrigieren d.h. er soll möglichst wenige Runs benötigen, damit die Schichtdicken wieder nahe am Zielwert sind.

Um die Fähigkeiten eines R2R-Algorithmus beurteilen zu können müssen allgemeingültige Kriterien herangezogen werden. In der Halbleiterindustrie werden statistische Kennzahlen verwendet deren Aufgabe es ist, die Stabilität und die Güte einzelner Prozesse zu beschreiben. Diese Kennzahlen werden in diversen Programmen wie Advanced Process Control (APC) oder Statistic Process Control (SPC) automatisiert berechnet. Um die Simulationen mit den aktuellen SPC-Daten zu vergleichen und um eine Aussage über die Fähigkeit des Algorithmus zu machen, werden diese Kennzahlen auch in den Simulationen berechnet und dadurch eine Vergleichsbasis geschaffen. Im Folgenden werden einige häufig verwendete Kennzahlen erläutert.

Dabei ist allerdings zu unterschieden, dass in SPC die Varianz aller Einzelmessungen in einem betrachteten Zeitraum zur Berechnung der Kennzahlen herangezogen wird. In der Simulation hingegen wird die Varianz der Mittelwerte über die simulierten Runs verwendet. Erklären lässt sich diese unterschiedliche Berechnung dadurch, dass durch eine R2R-Regelung die Schichtdickenverteilung über den Wafer nicht beeinflussbar ist und somit für die Beurteilung nicht heran gezogen werden kann.

Varianz (σ^2) in Space (SPC-Software bei Infineon):

$$\sigma^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_{,i,j} - \overline{\overline{X}}\right)^2}{n \cdot m - 1} \quad \text{mit} \quad \overline{\overline{X}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} x_i}{n \cdot m} \tag{47}$$

Hier sind die $x_{i,j}$ die Einzelwerte der m Punkte Schichtdicken Messung auf allen Testscheiben im betrachteten Zeitraum und $\overline{\overline{X}}$ der Mittelwert all dieser Messungen.

Varianz (σ^2) für R2R-Simulation:

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 \tag{48}$$

Hier sind die x_i die Mittelwerte der neun Punkte Schichtdicken Messung auf einer Testscheibe und $\overline{x}$ der Mittelwert der Schichtdicken über alle simulierten Runs (n).

Erklärung:

Da bei einem R2R-Kontroller nur die Boat-Down-Uniformity und die R2R-Uniformity betrachtet und verbessert wird, macht es keinen Sinn die Within-Wafer-Uniformity und damit die einzelnen Messwerte pro Scheibe zu betrachten. σ^2 ist bei der Simulation also die Varianz der Mittelwerte. In Space hingegen, ist σ^2 die Varianz aller Messung im betrachteten Zeitraum.

Prozessfähigkeitsindex (cp):

$$cp = \frac{\text{obere Spezifikationsgrenze - untere Spezifikationsgrenze}}{6\sigma} = \frac{OSG - USG}{6\sigma} \tag{49}$$

Mit Hilfe des cp-Wertes wird eine Aussage darüber gemacht wie stark der Prozess innerhalb der Spezifikationsgrenzen streut.

Prozessfähigkeitsindex (cpk):

$$cpk = \min\left(\left|\frac{OSG - \overline{\overline{x}}}{3\sigma}\right|, \left|\frac{USG - \overline{\overline{x}}}{3\sigma}\right|\right) \tag{50}$$

Beim cpk-Index wird auch die Lage der Verteilung zum Mittelwert miteinbezogen, d.h. wie gut zentriert ist der Prozess. Wobei $\overline{\overline{x}}$ der Mittelwert der Schichtdicken über alle simulierten Runs ist. Er sagt jedoch nichts darüber aus wie nahe man dem gewünschten Zielwert ist.

Erklärung zu cp und cpk:

Die unterschiedlichen Varianzberechnungen in der Simulation und in SPC machen die Ergebnisse nicht direkt Vergleichbar. Darüber hinaus kommt erschwerend hinzu, dass die Messwerte momentan noch vom Operator per Hand ins Netzwerk eingegeben werden und dadurch keiner exakten Ofenposition zuverlässig zuordenbar sind. Um trotzdem näherungsweise einen Vergleich durchzuführen habe ich bei der cp- bzw. cpk-Berechnung unterschiedliche Spezifikationsgrenzen verwendet. Für die Simulation wurden um die mittlere Streuung über den Wafer (für das 140 nm Rezept sind das 8 nm) verringerte bzw. erhöhte Spezifikationsgrenzen verwendet. Das Problem der nicht zuverlässigen Messungen und der unterschiedlichen Messhäufigkeiten (nur 1 – 3 Wafer pro Fahrt) lassen sich dadurch jedoch nicht beheben. Ein Vergleich der Ergebnisse ist dadurch nur bedingt möglich.

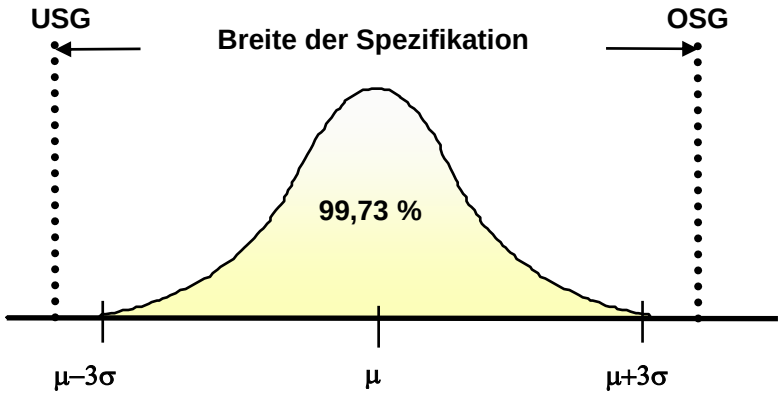

Abbildung 24: Schichtdickenverteilung

Root mean squared error (RMSE):

Dieser Index gibt Auskunft darüber, wie weit die Summe aller Werte vom Zielwert entfernt ist. Dafür wird zuerst der „Sum of squared error“ (SSE) berechnet.

$$SSE = \sum_{j=1}^{m} \sum_{i=1}^{n} \left(x_{i,j} - T \right)^2 \tag{51}$$

$$RMSE = \sqrt{\frac{SSE}{n \cdot m}} \tag{52}$$

Hier sind die $x_{i,j}$ die berechneten Schichtdicken pro Run und Position, T (Target) die Zielschichtdicke, n die Anzahl der Runs und m die Anzahl der Testpositonen.

Mit Hilfe dieser Kennzahlen ist es nun möglich eine objektive Aussage über die Fähigkeit des Algorithmus bzw. einen bedingten Vergleich mit dem aktuellen Ist-Zustand der ASM-Ofenprozesse zu machen.

Eine Weitere Analysemöglichkeit ist das Auswerten der Diagramme. Dort werden einerseits die mittleren Schichtdicken für jede Testposition und andererseits die normierten Stellgrößen über die einzelnen Runs aufgetragen. In diesen Diagrammen lassen sich sehr gut das Konvergenzverhalten, das Regelverhalten und die Stärke des zufälligen Rauschens ablesen.

Kennzahlen und Diagramme zusammen bieten eine hervorragende Möglichkeit die einzelnen Simulationen zu vergleichen und Schlussfolgerungen zu ziehen.

5.2 Allgemeines zu den Simulationen

Alle Simulationen wurden in Matlab 6.5 durchgeführt. Dies war notwendig da bei der Minimierung der Kostenfunktion (23) in Abhängigkeit von der Anzahl der Regelgrößen (die einzelnen Schichtdicken) und von der Komplexität des Modells ein System von multivariablen quadratischen Gleichungen, das unterbestimmt, bestimmt oder überbestimmt sein kann, zu lösen ist.

Zu diesem Zwecke wurde ein Matlab Programm erstellt, welches einerseits den Algorithmus abbildet und mit einer speziellen Funktion aus der „Optimization Toolbox“ das Minimierungsproblem löst, und andererseits eine Funktion welche die Messergebnisse mit bewusst eingebauten Störungen (Drift, Shift, Rauschen) simuliert.

Im konkreten Fall wurde ein Nitridprozess (VEH 10008) unter folgenden Bedingungen untersucht:

Zielschichtdicke:	Target = 140 nm
Temperatur:	T_3 = 780 °C (konstant gehalten)
Stellgrößen:	T_1 = [792,802]°C
	T_2 = [790,800]°C
	T_4 = [753,763]°C
	T_5 = [750,760]°C
	t = [86,96] min
Regelgrößen:	$SD_1, SD_2, \ldots, SD_7$

Bei den Simulationen kann die Anzahl der Regelgrößen (3,5 oder 7) und die Anzahl der Modelparameter (6,8 oder 10) frei gewählt werden. Dadurch kann man verschiedene Reglermodelle (linear, quadratisch,...) gegeneinander evaluieren und durch die Wahl der Anzahl der Regelgrößen die Auflösung der Simulation – d.h. wie genau betrachtet man den Ofen (je mehr Positionen desto höher die Auflösung) – bestimmen. In den Simulationen werden die Schichtdicken immer fortlaufend nummeriert. Im Falle von drei simulierten Schichtdicken werden diese für die Position 1, 4, 7 berechnet. Bei fünf simulierten Schichtdicken entsprechen diese den Berechnungen für die Positionen 1, 2, 4, 6, 7.

In den einzelnen Simulationen gilt es herauszufinden ob dieser Algorithmus im Prinzip funktioniert und ob er für das vorliegende Problem einsetzbar ist. Vorweg kann gesagt werden, dass diese Methode auf Grund ihrer Fähigkeit mit quadratischen Abhängigkeiten umgehen zu können prädestiniert für Ofenprozesse aber auch andere Prozesse in der Halbleiterindustrie ist. Der Algorithmus hat die Fähigkeit binnen weniger Runs (1-3) von einer Zielwertabweichung (±5 % der Zielschichtdicke) wieder in Richtung Ziel zu regeln.

Die Simulationen können außerdem zur Bestimmung der optimalen Algorithmusparameter verwendet werden. Diese liegen im Bereich von

P_{t-1}	Strukturmatrix des Ellipsoids (Startwert = Einheitsmatrix * [0.1;1])
λ_t	Updateparameter [0,1;0.6]
σ_{t-1}	Parameterunsicherheitsfaktor [0.3;0,8]
ζ	Rettungsschrittweite [1;5]
γ	Gesamtrauschen [0,8;2]

und hängen vom Prozess und von den simulierten Fehlern (Drift, Rauschen, Shift, Messverzögerung) ab.

Das Initialmodell, welches die Anzahl der Parameter festlegt und als Reglermodell für den ersten Run eingesetzt wird, ist für alle Simulationen das in Gleichung (27) beschriebene.

Um den Modellfehler zu simulieren wurde ein sehr genaues quadratisches Prozessmodell mit 16 Parametern verwendet, um die Messwerte die man sonst aus der Schichtdickenmessung erhält, zu ersetzen.

Allgemeine Erläuterung der Simulationsdiagramme:

Jede Simulation wird an Hand zweier Diagramme und den statistischen Kennzahlen beschrieben und analysiert. Das obere dieser beiden Diagramme zeigt den Verlauf der einzelnen Schichtdicken (SD_1, SD_2,..., SD_7) die sich an unterschiedlichen Positionen im Ofen befinden, verteilt von Gaseinlass (SD_7) bis Gasauslass (SD_1) mit äquidistanten Abständen (Abbildung 10). SD n.c. zeigt stellvertretend den Verlauf für eine Schichtdicke (SD_3) wenn sie nicht kontrolliert wird.

Das untere Diagramm zeigt wie sich die Stellgrößen von Run zu Run verändern. Sie sind normiert von [-1; 1] d.h. ein Wert von 0 für die Zeit t bedeutet, dass beim untersuchten Prozess t mit 91 min eingestellt wird (Bereiche der einzelnen Stellgrößen - siehe oben). Bei einigen Simulationen musste der Stellgrößenbereich auf [-1,5; 1,5] erweitert werden damit die Regelgrößen auf Ziel gehalten werden können.

Die Genauigkeit mit der die Stellgrößen ins Prozessmodell eingesetzt wurden, wurde der Einstellgenauigkeit des ASM-A400 angepasst. Eine Rundung der Temperaturen auf 0,1 °C bzw. der Zeit auf Sekunden war erforderlich

Um die Ergebnisse in Zahlen zu fassen wurden folgende Parameter berechnet und ins Diagramm eingetragen

cpk-Werte
cp-Werte
RMSE (Root mean squared error) in nm und
die Standardabweichung σ in nm

5.3 Angenommene Störungen in den Simulationen

Alle Simulationen wurden für eine Anzahl von 60 Runs durchgeführt. Bei einigen Simulationen wurde eine Messverzögerung von einem Run eingebaut d.h. die Messergebnisse vom aktuellen Run stehen erst für die Generierung des Rezeptes für den übernächsten Run zur Verfügung. Diese Annahme macht Sinn, da in der Praxis solche Verzögerungen häufiger vorkommen.

Die R2R Standardabweichung beträgt

$$\sigma_{Ofen} = \pm 0{,}3\% * \text{Target [nm]}$$

Einmaliger Sprung des gesamten Ofens nach oben bei Run 10 um 5 nm, der z.B. durch eine Fehlmessung oder eine Leckage hervorgerufen werden könnte. Dabei wird der Algorithmus auf seine Impulsantwort untersucht.

Sprung nach oben für SD1 um 4 nm bzw. für SD2 um 3nm ab Run 20 bzw. Sprung nach unten für gesamten Ofen ab Run 40 von 4 nm. Dieses Verhalten kann in der Praxis durch Nichterreichen der Temperaturen auftreten und der Algorithmus wird dabei auf seine Sprungantwort analysiert.

Der gesamte Ofen (alle Schichtdicken) unterliegt einer Drift von - 0.1 nm/Run. Dieser Effekt kann durch die Zunahme der Schichtdicke am Prozessrohr erklärt werden. Das Prozessrohr wird ca. alle 500 Runs gewechselt.

Ein Fehler des Initialmodells liegt grundsätzlich immer vor, da das Regelmodell mit dem Prozessmodell verglichen wird.

5.4 Simulationen in Matlab

5.4.1 Simulation ohne Prozessrauschen:

Diese erste Simulation soll das grundsätzliche Regelverhalten des Algorithmus zeigen. Deshalb wird hier ohne das zufällige Rauschen und ohne Messverzögerung simuliert.

Hier wird ein quadratischen Regelmodell mit 10 Parametern für fünf Positionen (1,2,4,6,7) verwendet.

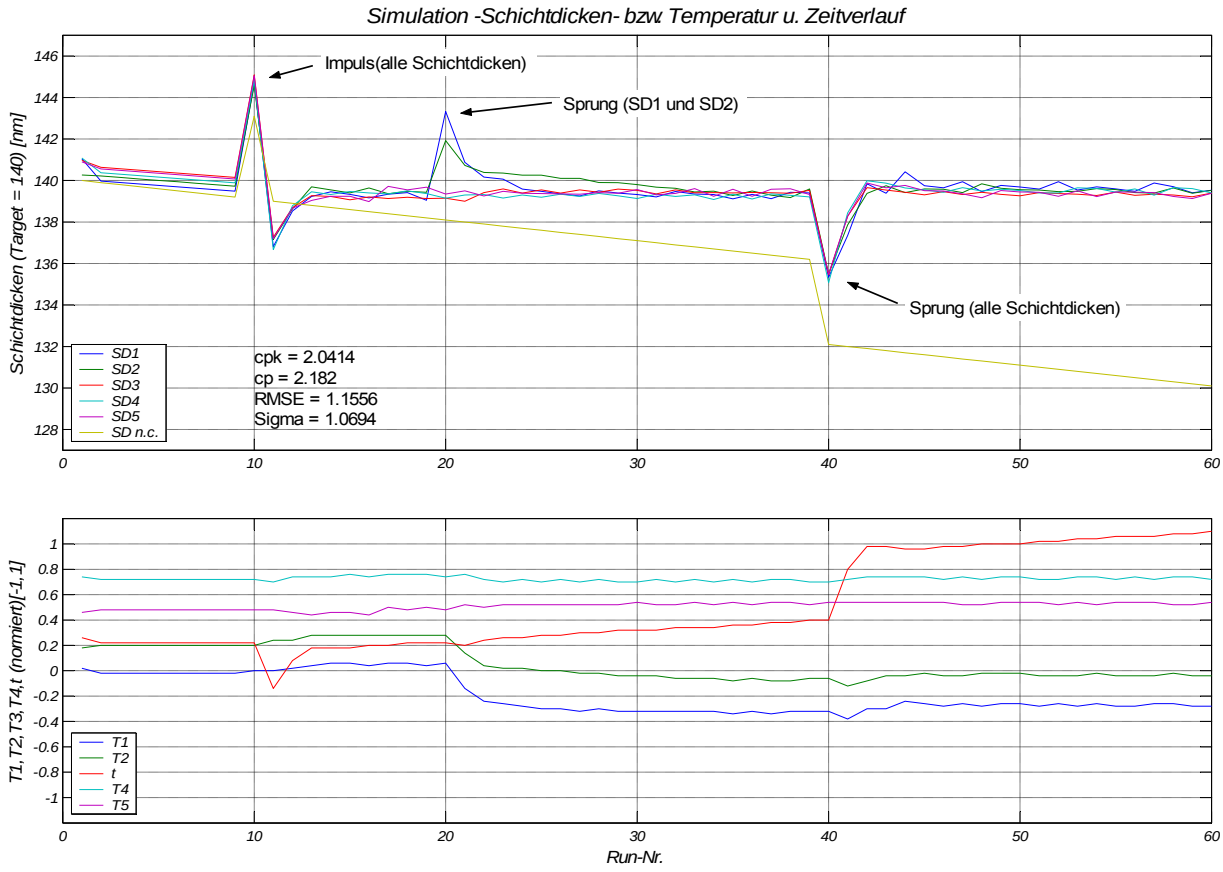

Abbildung 25: Simulation für fünf Schichtdicken ohne Messverzögerung und Rauschen

Man kann am Verlauf der Schichtdicken sehr gut erkennen wie schnell dieser Algorithmus bei Störungen wieder Richtung Ziel konvergiert (1-3 Runs). Der Stellgrößenverlauf (Abbildung 25, unteres Diagramm) zeigt, dass die Korrekturen, die der Kontroller setzt der Praxis sehr gut entsprechen. Zwischen Run 20 und 40 reagiert der Kontroller auf die reine Drift nur durch Nachregeln der Zeit, was wiederum mit den praktischen Erfahrungen übereinstimmt.

5.4.2 Simulation bei reiner Prozess-Drift:

Hier wird ohne Sprünge (RunNr. 10, 20, 40) und ohne Messverzögerung bei reiner Drift mit zufälligem Rauschen simuliert. Es wurde das gleiche Regelmodell wie in Simulation 1 verwendet.

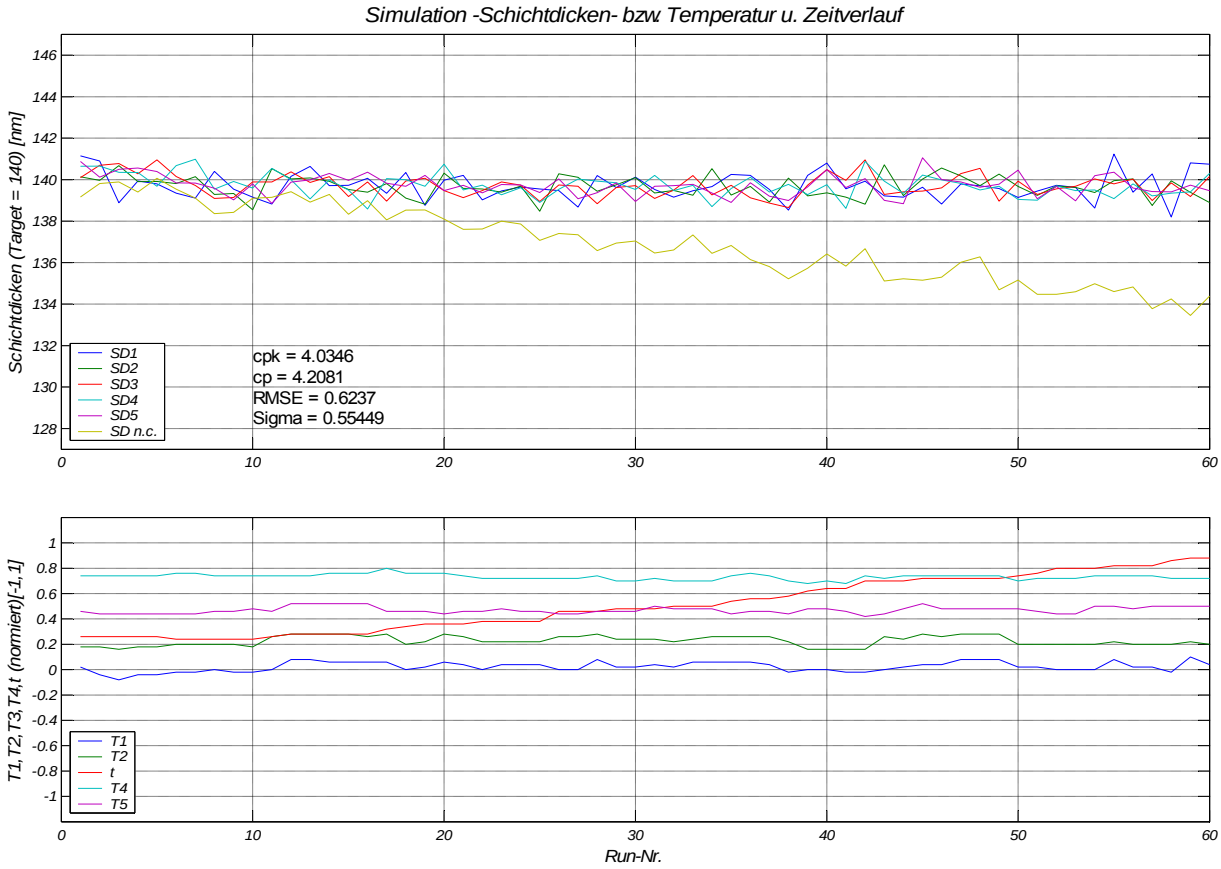

Abbildung 26: Simulation für fünf Schichtdicken bei reiner Drift

An Hand der Kennzahlen kann man die erstklassige Performance des Algorithmus erkennen und dass keine Schwierigkeiten beim Umgang mit Drift und Rauschen auftreten.

Wie gut der Kontroller ist, hängt von der Wahl der einstellbaren Parameter ab (λ_{max}, P_{t-1}, ζ….). Diese können für den realen Prozess sehr gut in Offline-Simulationen bestimmt werden. Letztendlich wird es aber notwendig sein das „Finetuning“ der Parameter an produktiven Fahrten durchzuführen, da das exakte Rauschen, prozess- und equipmentspezifisch ist.

5.4.3 Simulationen mit 3 Regelgrößen:

Bei den nächsten drei Simulationen wurden nur die drei Regelgrößen SD_1 (Testwafer 1 am Gasauslass), SD_2 (Testwafer 4 in der Ofenmitte), SD_3 (Testwafer 7 am Gaseinlass) verwendet. Alle Störungen - Rauschen, Drift, Impuls und Sprung - mit Ausnahme der Messverzögerung sind integriert.

Simulation 1:

Für die erste Simulation wurde ein lineares Regelmodell wie in Gleichung (31) beschrieben verwendet

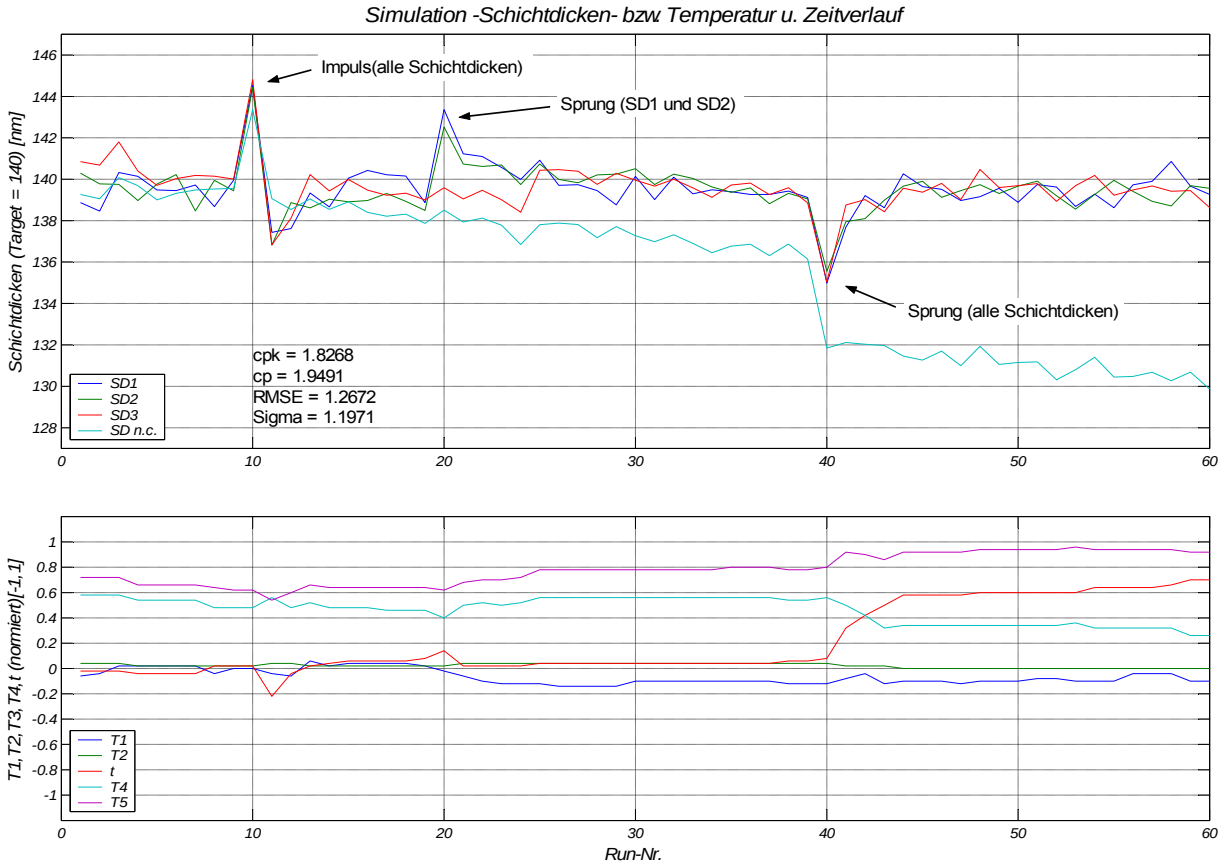

Abbildung 27: Simulation lineares Regelmodell bei 3 Schichtdicken

Bei Run 1 ist der Anfangs relativ große Modellfehler erkennbar. Außerdem sieht man bei Run 11 - hervorgerufen durch die einmalige Fehlmessung bei Run 10 - eine sehr starke Korrektur nach unten. Erkennbar ist auch das bei nur 3 Regelgrößen die Stellgrößenvorhersage nicht mit den Vorhersagen bei 5 oder 7 Regelgrößen übereinstimmt.

Simulation 2:

In dieser Simulation gelten die gleichen Annahmen wie bei Simulation 3, jedoch wurde diesmal ein bilineares Regelmodell mit 8 Parametern (32) verwendet.

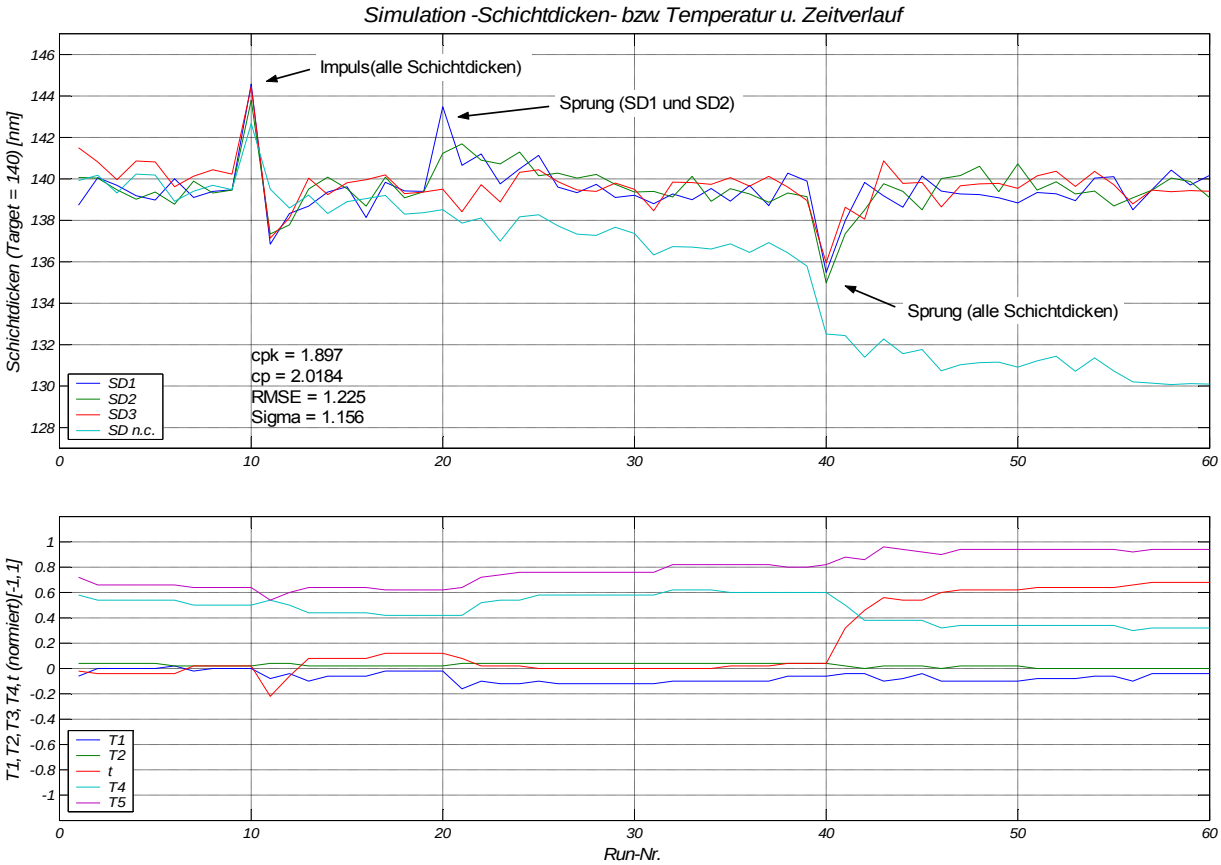

Abbildung 28: Simulation bilineares Regelmodell bei 3 Schichtdicken

Hier zeigt sich ein ähnliches Regelverhalten wie in der vorangegangenen Simulation.

Simulation 3:

Hier dient ein quadratisches Modell wie in Gleichung (30), welches dort für nur eine Schichtdicke beschrieben ist, als Regelmodell

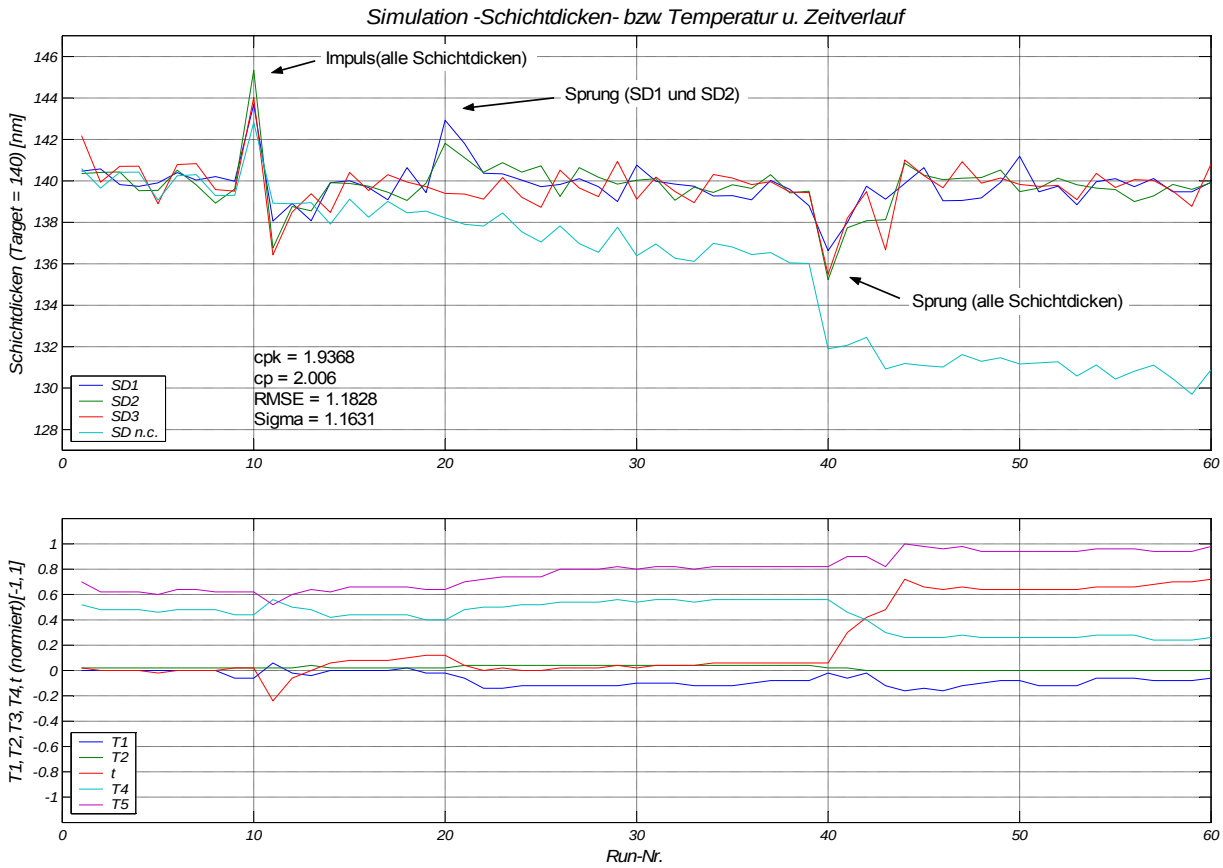

Abbildung 29: Simulation quadratisches Regelmodell bei 3 Schichtdicken

Auch diese Simulation zeigt gleiche Regeleigenschaften wie Simulation 3 und 4. Beim Sprung nach unten (Run 40) dauert es 4 Runs bis die Regelgrößen wieder auf Ziel sind.

5.4.4 Simulationen für fünf Regelgrößen:

Bei den nächsten zwei Simulationen wurde ein Modell mit fünf Regelgrößen verwendet. Das Regelverhalten bei allen Störungen mit Ausnahme der Messverzögerung soll untersucht werden.

Simulation 1:

Hier wird für die Regelung ein lineares Modell mit 6 Parameter verwendet.

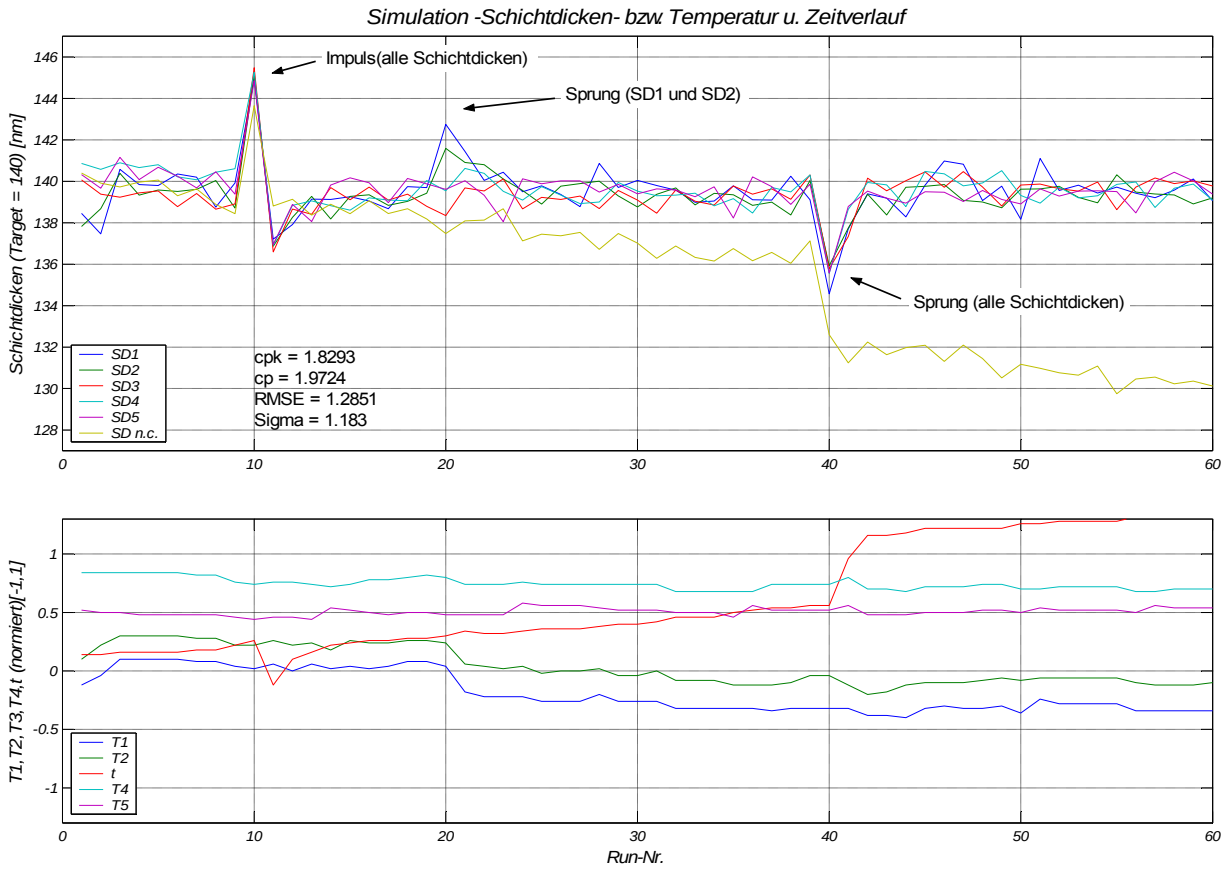

Abbildung 30: Simulation lineares Regelmodell bei 5 Schichtdicken

Da bei dieser und der nächsten Simulation die Schichtdicken an 5 Ofenpositionen simuliert werden, hat man eine höhere Auflösung. Dadurch unterscheiden sich die vorgeschlagenen Stellgrößen wesentlich von denen in den Simulationen 3 -5. Vor allem der Drift wird hier fast ausschließlich durch Änderung der Zeit kompensiert. Dieses Verhalten entspricht den praktischen Erfahrungen und zeigt, dass für eine vernünftige Regelung des Nitridprozesses die Schichtdicken an mindestens 5 Positionen gemessen werden müssen. Diese Messhäufigkeit garantiert, dass auch zwischen den gemessenen Positionen die Schichtdicken den gewünschten Werten entsprechen.

Simulation 2:

Hier gelten die gleichen Annahmen wie bei der vorangegangenen Simulation. Allerdings wird hier ein quadratisches Regelmodell mit 10 Parametern verwendet.

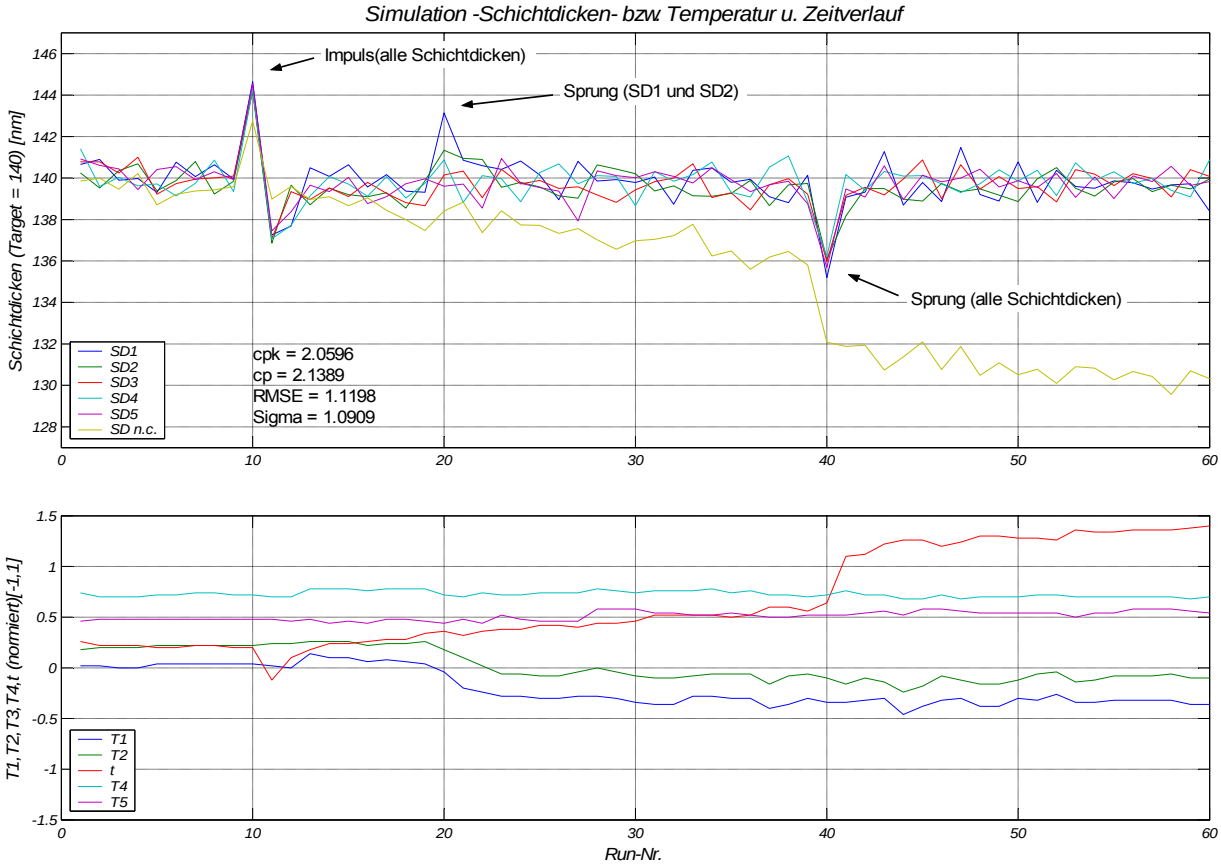

Abbildung 31: Simulation quadratisches Regelmodell bei 5 Schichtdicken

Sehr geringer Initialmodellfehler und gute Regeleigenschaften (cpk>2). Das Nachregeln bei Störungen (Sprung, Impuls) geht sehr rasch (1-2 Runs). Obwohl man mit einem Linearen Regelmodell den Prozess zufrieden stellend regeln kann, erreicht man bei Verwendung eines quadratischen Regelmodells eine bessere Performance. Wenn man den Regelgrößenverlauf von Run 40 auf 41 betrachtet, lässt sich sofort das Potential dieser Methode erkennen. In nur einem Regelschritt schafft es der Algorithmus die Zielwerte zu erreichen.

5.4.5 Simulationen für 7 Regelgrößen

Die sieben Regelgrößen bei dieser Simualtion waren SD_1 (Gasauslass) bis SD_7 (Gaseinlass). Es wird dadurch die beste Auflösung erreicht.

Simulation 1:

Bei dieser Simulation wurde für die Regelung ein lineares Modell mit 6 Parametern verwendet.

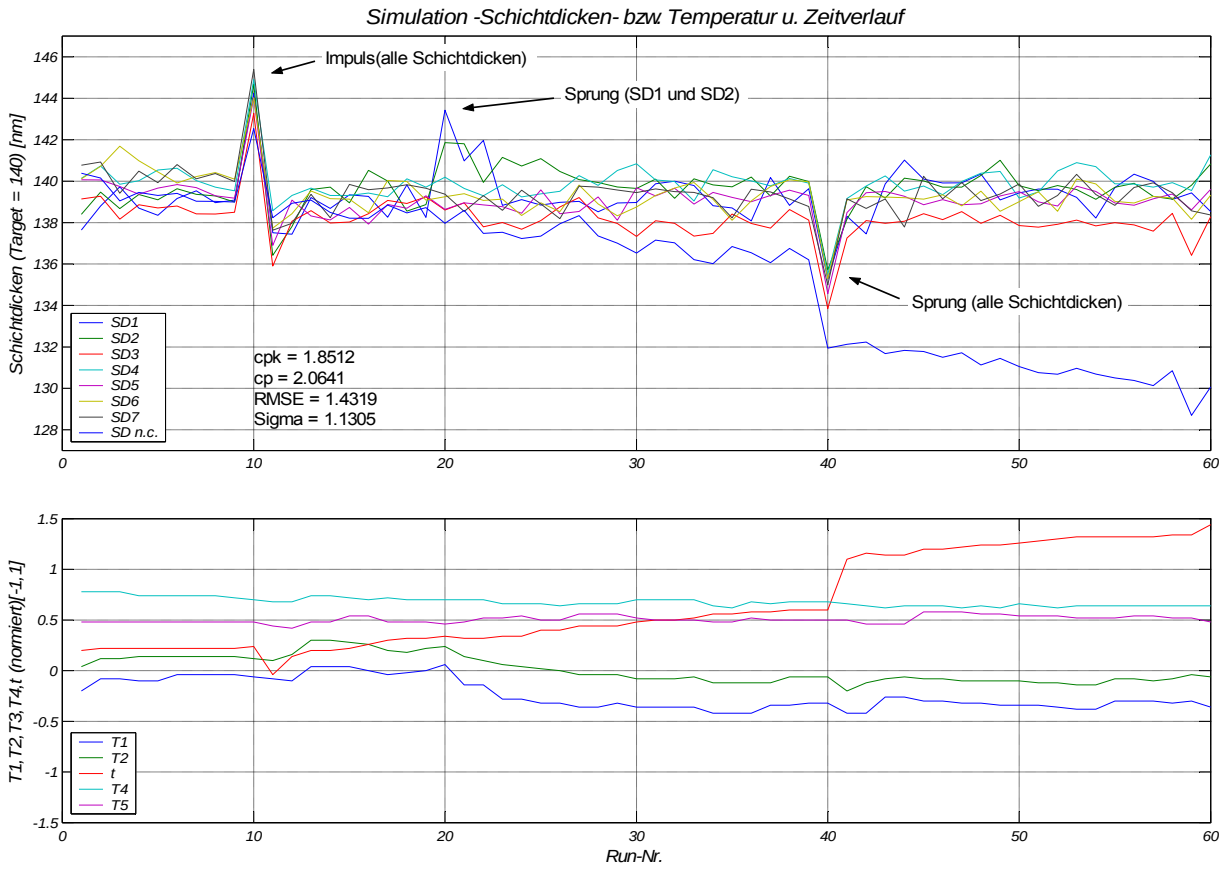

Abbildung 32: Simulation lineares Regelmodell bei 7 Schichtdicken

Auch hier zeigt sich wie bei allen Versuchen mit einem linearen Berechnungsmodell, dass der Algorithmus einige Runs benötigt um den Initalmodellfehler zu beheben. Die Störungen werden mit ansprechender Geschwindigkeit behoben.

Simulation 2:

Gleich wie obige Simulation nur mit dem Unterschied, dass ein quadratisches Modell verwendet wurde.

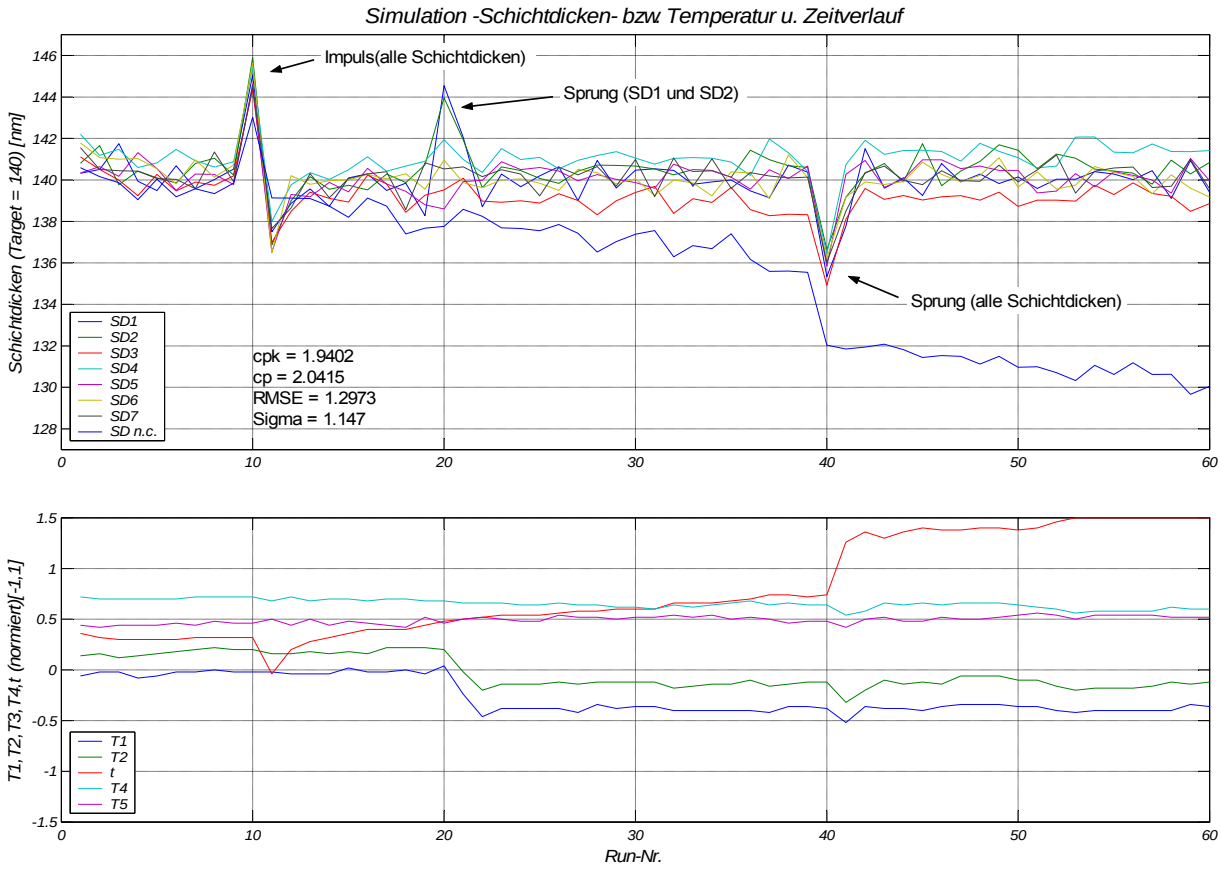

Abbildung 33: Simulation quadratisches Regelmodell bei 7 Schichtdicken

Vergleicht man Simulationen mit 7 Regelgrößen mit jenen mit 5 Regelgrößen zeigen sich ähnliche Stellgrößenveränderungen. Dadurch wird klar, dass es zum Regeln des Ofens ausreicht, wenn man an 5 Positionen misst. Eine Verbesserung der Performance ist durch eine höhere Anzahl an Messungen nicht zu erreichen und würde nur unnötigen Mehraufwand bedeuten. Auf Grund dieser Tatsache empfiehlt sich eine Regelung über fünf Messpositionen mit einem quadratischen Regelmodell.

In den weiteren Simulationen wird deshalb nur mehr ein quadrtisches Regelmodell für fünf Regelgrößen (Schichtdicken) verwendet.

5.4.6 Simulation mit Messverzögerung ohne Rauschen:

Bei dieser Simulation wurde das Verhalten des quadratischen Modells für 5 Messpositionen. Es wurde ein „delay“ von einem Run angenommen, d. h. dass für die Erstellung des Rezeptes für die nächste Ofenfahrt das Messergebnis der letzten Fahrt nicht zur Verfügung steht, sondern nur das der Vorletzten. Zusätzlich wurde ohne zufälligen Fehler (Rauschen) simuliert um das Regelverhalten anschaulicher zu machen.

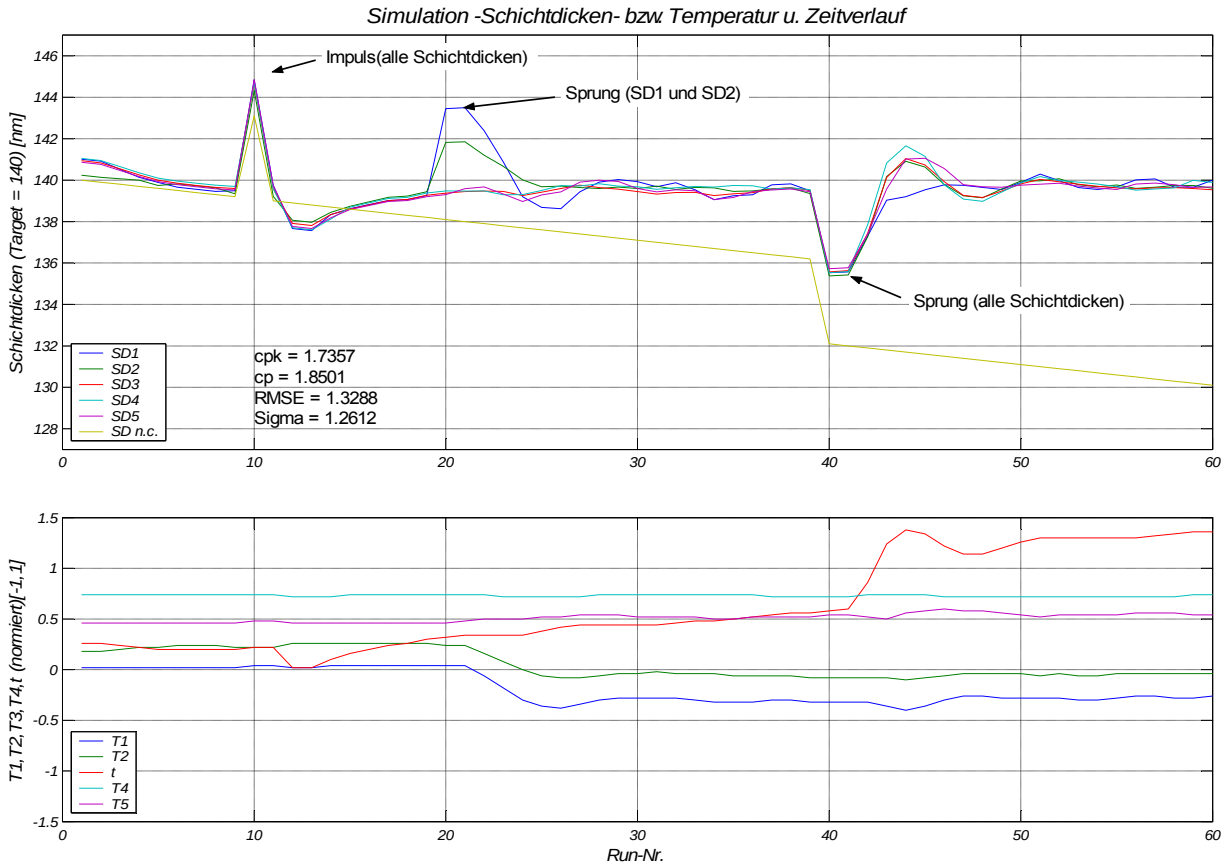

Abbildung 34: Simulation mit Messverzögerung von einem Run

Es ist auch hier deutlich zu erkennen, dass der Algorithmus mit ansprechender Geschwindigkeit in Richtung Zielwert konvergiert. Man muss allerdings einen Performanceverlust (cpk) hinnehmen, der sich einerseits aus der Messverzögerung selbst, und andererseits aus einer weniger aggressiven Wahl der einstellbaren Algorithmusparameter (λ_{max}) ableiten lässt. Die Wahl von λ_{max} ist nahe dem maximalen Wert der noch eine stabile Regelung produziert. Erkennbar ist dies am leichten Überschwingen beim Beheben der Störungen.

5.4.7 Simulation mit Messverzögerung:

Hier gelten dieselben Bedingungen wie bei obiger Simulation, außer dass diesmal auch das Prozessrauschen mit eingebaut wurde.

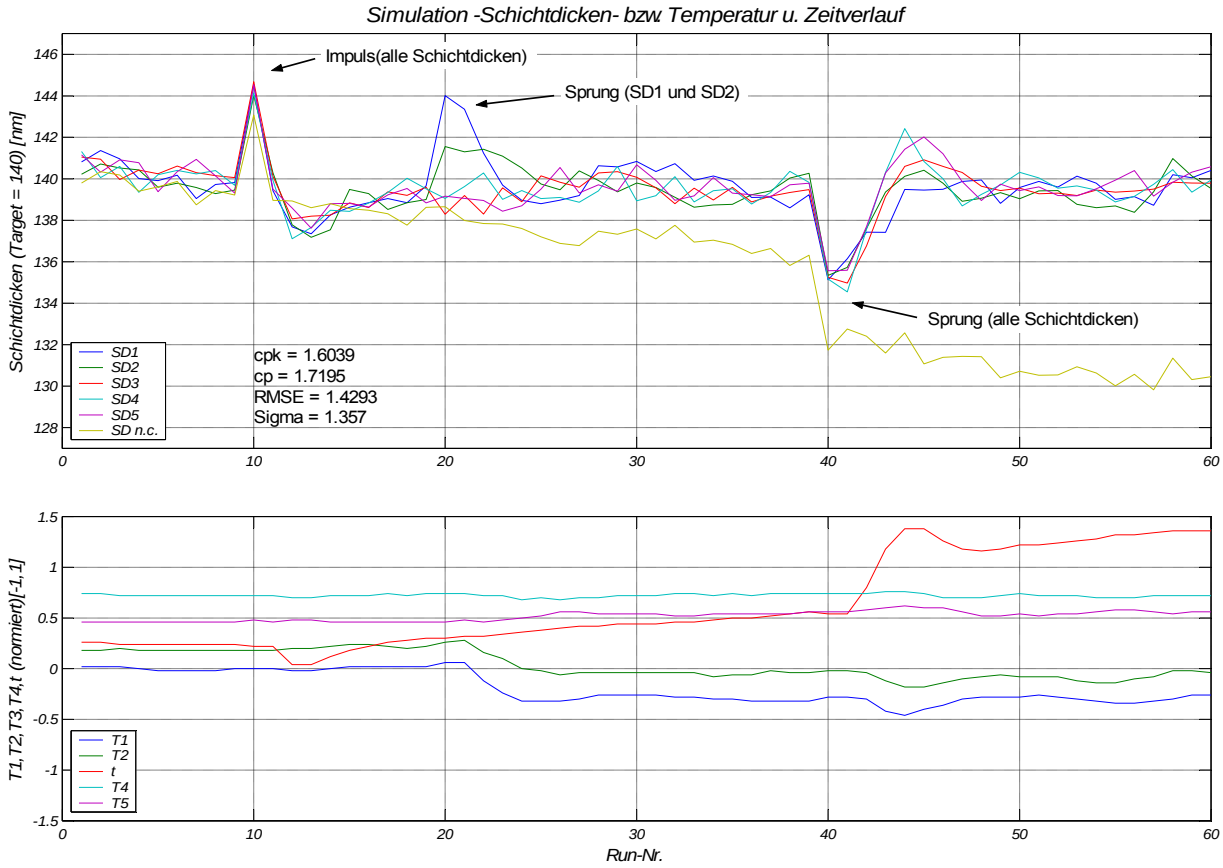

Abbildung 35: Simulation mit Messverzögerung und Rauschen

Man sieht, dass für den Algorithmus eine Messverzögerung kein Problem darstellt. Allerdings muss ein Performanceverlust in Kauf genommen werden. Geht man davon aus, dass in der Praxis diese Messverzögerung nicht bei jedem Run – wie hier simuliert – vorkommt, lassen sich sicherlich bessere cpk-Werte erreichen.

5.4.8 Simulation bei reiner Drift und Messverzögerung:

Bei diesem Test gelten dieselben Rahmenbedingungen wie bei den beiden vorhergehenden Simulationen. Hier wurde diesmal bei reiner Drift simuliert.

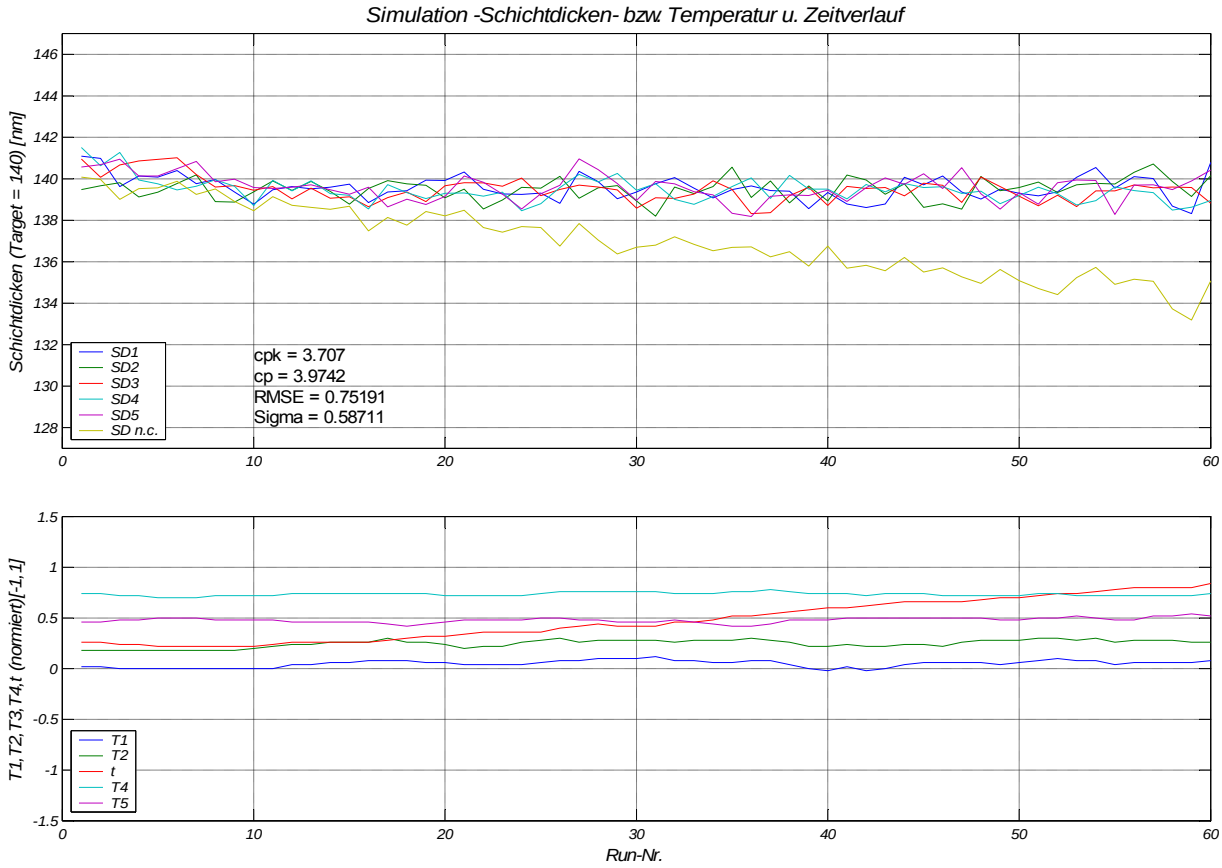

Abbildung 36: Simulation mit Messverzögerung bei reiner Drift

Der Algorithmus regelt reine Drift wie erwartet durch Neuberechnen der Abscheidzeit. Die Performancekennzahlen dürfen allerdings nicht überbewertet werden, da in der Praxis wesentlich häufiger Störungen auftreten.

5.4.9 Simulation ohne exaktes Regelmodell

Bei dieser Simulation wird davon ausgegangen, dass kein exaktes Regelmodell aus einem DoE vorhanden ist. Bekannt sind allerdings die Stellgrößen, die einen Einfluss auf das Input/Output Verhalten haben. Das bedeutet, dass man Informationen darüber haben muss welche Terme (Lineare, Interaktion,…) im Modell auftauchen sollen. Es wurden wiederum sämtliche Störungen mit Ausnahme der Messverzögerung in die Simulation eingebaut. Als

Regelmodell dient das in Gleichung (27) beschriebene mit dem Unterschied, dass diesmal der Parameter β wie folgt angenommen wird.

$$\beta = \begin{pmatrix} 140 & 1 & 1 & 5 & 0.1 & 0.1 & 0 & 0 & 0 & 0 \\ 140 & 1 & 1 & 5 & 0.1 & 0.1 & 0 & 0 & 0 & 0 \\ 140 & 1 & 1 & 5 & 0.1 & 0.1 & 0 & 0 & 0 & 0 \\ 140 & 0.1 & 0.1 & 5 & 1 & 1 & 0 & 0 & 0 & 0 \\ 140 & 0.1 & 0.1 & 5 & 1 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{53}$$

Hier werden zwar quadratische und Interaktionsterme zur Regelung herangezogen, jedoch ist im Vorhinein kein Wissen über den Einfluss dieser Terme vorhanden. Den Einfluss der Modellkoeffizienten zu finden und aufeinander abzustimmen erledigt der Algorithmus von selbst.

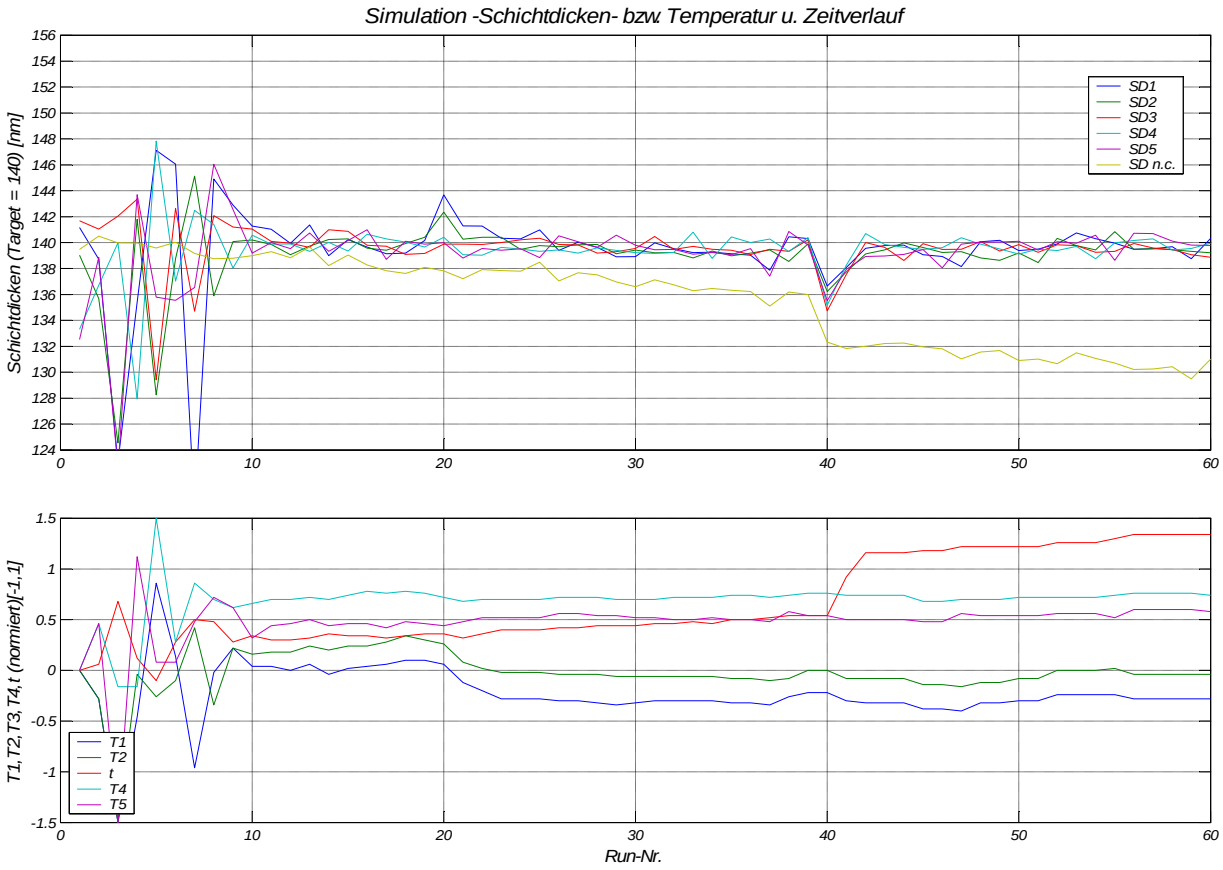

Abbildung 37: Simulation ohne exaktes Regelmodell

Der Algorithmus kann innerhalb von nur 10 Runs, ein Prozessmodell finden mit dem er imstande ist den Prozess zu regeln. Diese besondere Fähigkeit der Modellfindung für einen

Prozess bei dem kein detailliertes und in Zahlen fassbares Prozessverständnis gegeben ist, zeichnet diese R2R-Methode aus. Hier kann man das wahre Potenzial des Algorithmus erkennen. Trotzdem gilt es zu beachten, dass ganz ohne Wissen über den Prozess ein Optimierungsversuch auch misslingen kann wie die nächste Simulation zeigt.

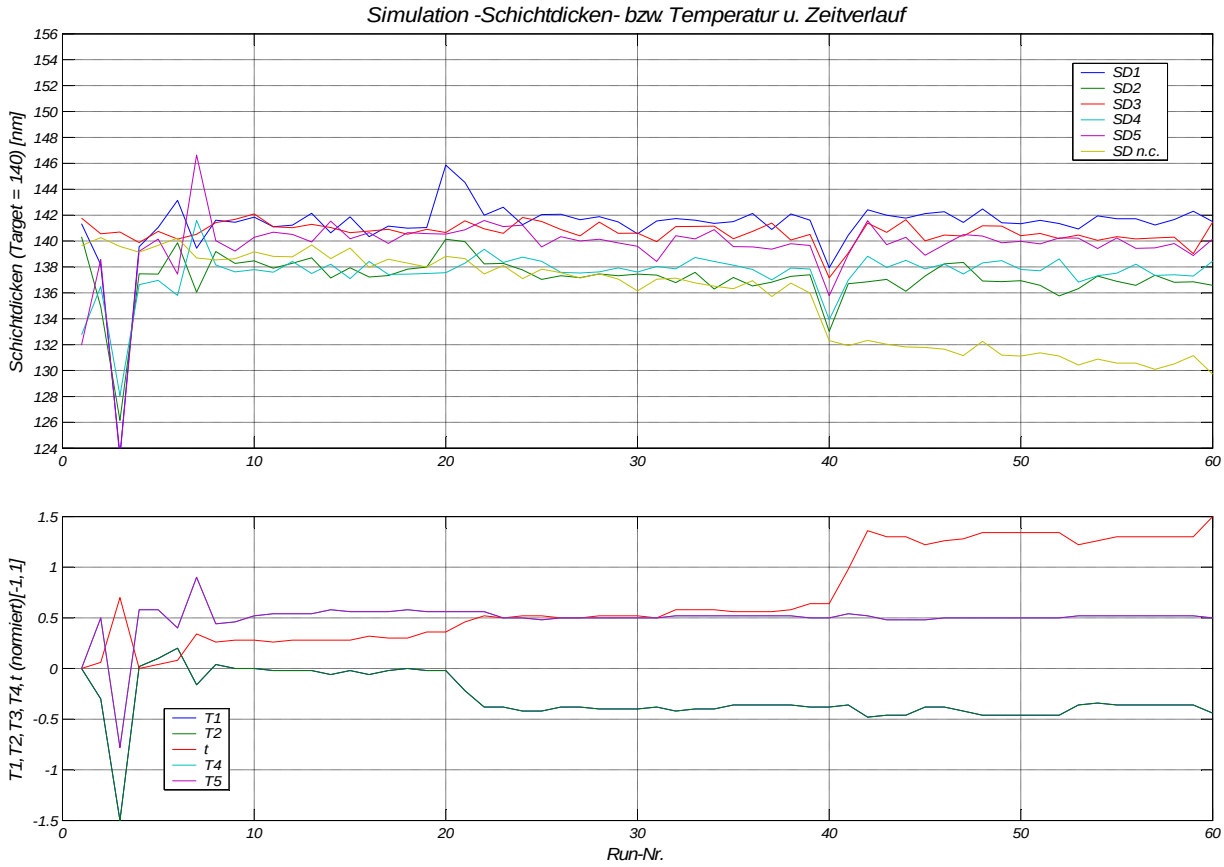

Abbildung 38: Simulation ohne exaktes Regelmodell – nur lineare Terme

Hier ist der Kontroller nicht mehr fähig die Stellgrößen so zu wählen, dass die Zielwerte (140 nm) erreicht werden. T_1 und T_2 bzw. T_4 und T_5 haben während allen Runs die gleichen Werte (T_1 und T_4 sind deshalb in der Abbildung nicht sichtbar).

5.4.10 Simulation bei rollierender Messung

Diese Simulation soll zeigen, dass auch bei rollierender Messung der Algorithmus in der Lage ist den Prozess zu kontrollieren. Als rollierende Messung bezeichnet man eine Messung bei der pro Run nur eine Schichtdickenmessung durchgeführt wird. Die Position an der gemessen wird variiert von Run zu Run so dass eine bestimmte Position nur bei jedem fünften Run gemessen wird. Dies verringert den Arbeitsaufwand der zur Messung notwendig ist deutlich und bringt eine Kostenersparung mit sich.

Als Regelmodell dient wieder das quadratische Modell für fünf Positionen. Einmal soll ohne Rauschen und einmal mit Rauschen simuliert werden. In beiden Fällen gab es eine Messverzögerung von einem Run.

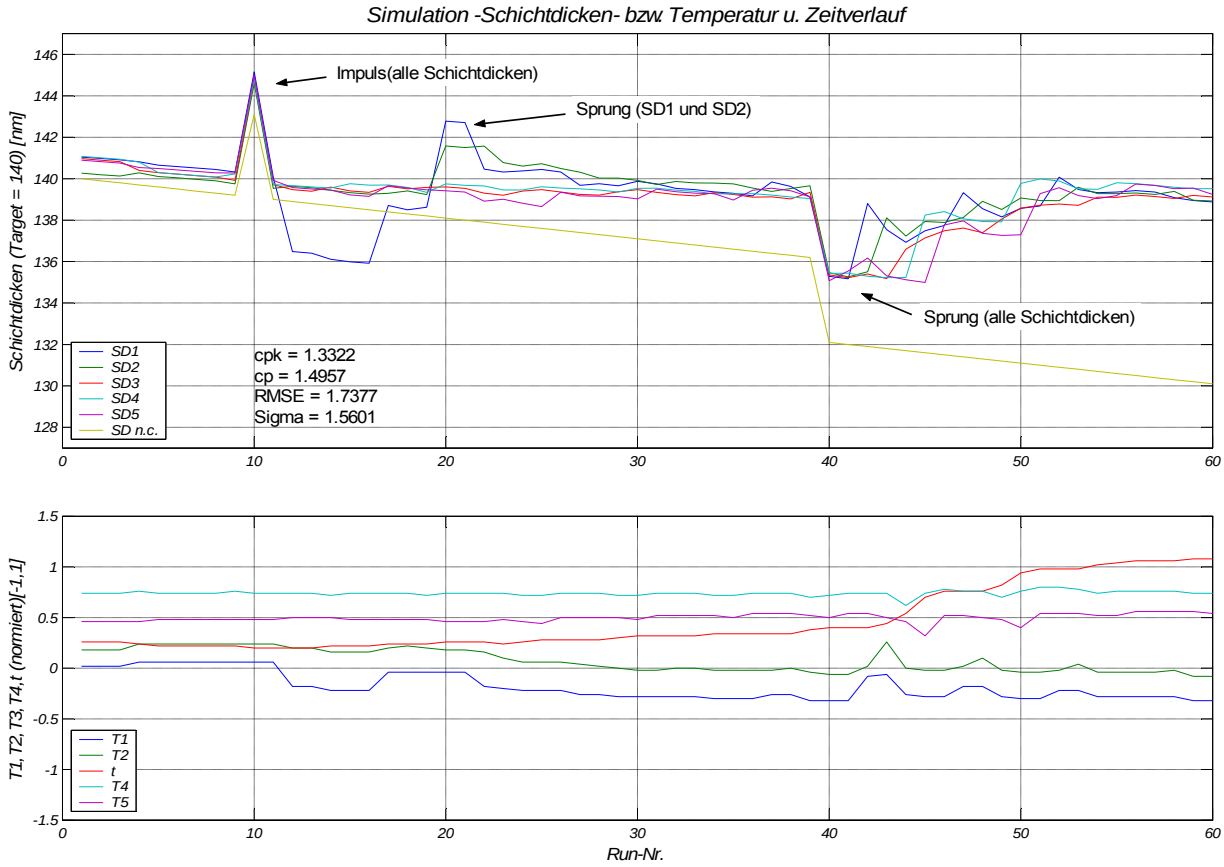

Abbildung 39: Simulation mit rollierender Messung ohne Rauschen

Diese Simulation zeigt, dass der Algorithmus auch bei rollierender Messung auf Störungen, entsprechend schnell reagiert.

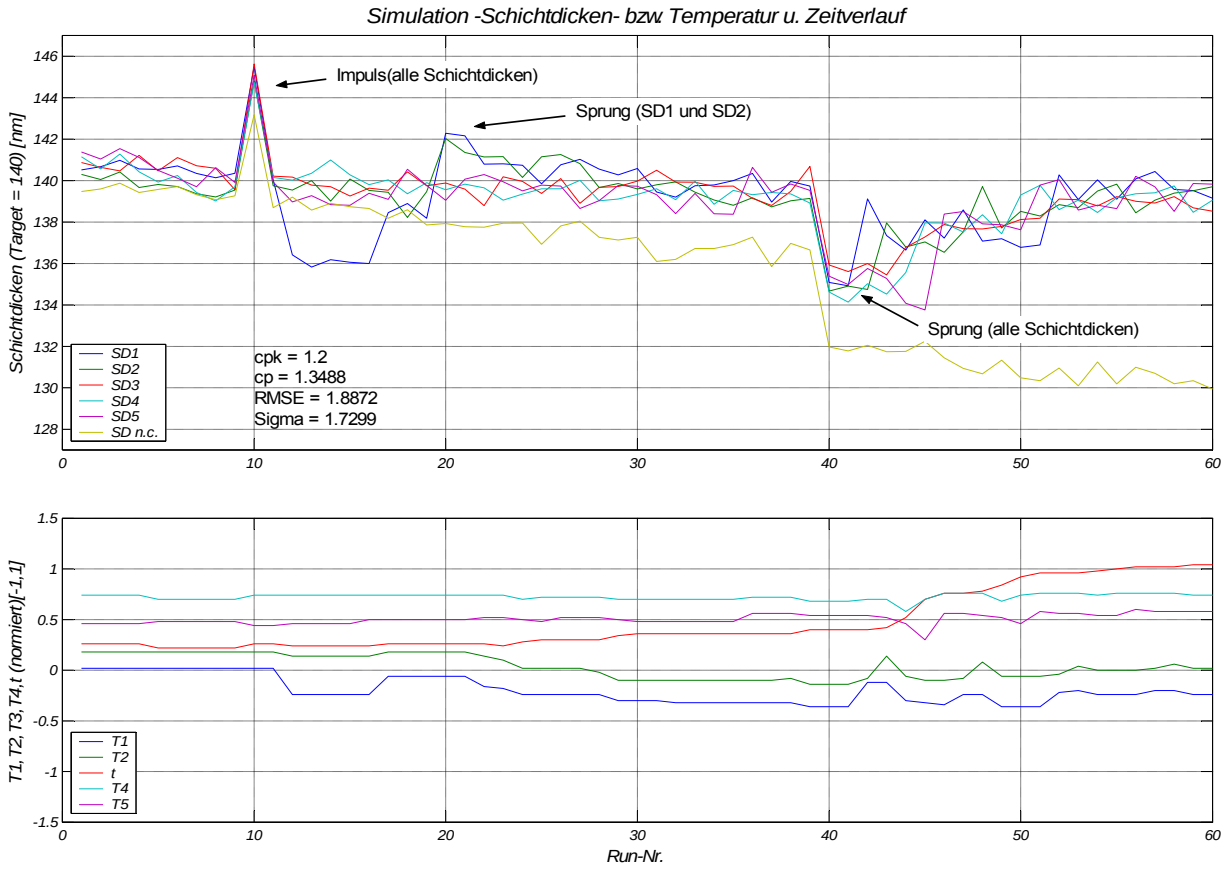

Abbildung 40: Simulation mit rollierender Messung

Abbildung 40 zeigt, dass es grundsätzlich möglich ist die Messhäufigkeit zu verringern. Eine schlechtere Performance ist dabei aber unvermeidbar. Obwohl man festhalten muss, dass hier auch jene Schichtdicken in die Kennzahlberechnung einfließen bei denen in der Praxis keine Informationen vorhanden wären. Das heißt, dass man in der Simulation die Werte aller fünf Schichtdicken in jedem Run berücksichtigt. Hingegen werden In der Praxis nur bei jedem fünften Run die Informationen über eine bestimmte Schichtdicke zur Kennzahlberechnung berücksichtigt.

6 Experimentelle Verifikation des DHOBE Kontrollers

Um die Funktion des Algorithmus zu überprüfen, sind erste Testfahrten unter Anwendung der vom Kontroller vorhergesagten Stellgrößen durchgeführt worden. Die Versuchsfahrten wurden auf verschiedenen Öfen durchgeführt um die Fähigkeit dieser Methode in der Praxis zu evaluieren und um zu belegen, dass die in den Simulationen gefundenen Algorithmusparameter (γ, P_{t-1}…) auch unter realen Bedingungen verwendet werden können. Durch die Versuche auf jenen Öfen für die kein eigenes Modell erstellt wurde, erkennt man das Potenzial des Algorithmus auch mit großen Modellfehlern zurechtzukommen. Diese Eigenschaft ermöglicht es auch bestehende Rezepte mit dem R2R-Kontroller zu optimieren.

6.1 Verifikation auf Ofen V03/2D

Die Versuche die für die Erstellung eines Initialmodells notwendig waren wurden auf diesem Ofen durchgeführt.

Es wurde das in Gleichung (27) beschriebene quadratische Regelmodell mit 10 Faktoren für 5 Schichtdicken verwendet.

In der untenstehenden Tabelle sieht man anhand den Ergebnissen von vier Testfahrten wie der Algorithmus reagiert. Man kann erkennen, dass bei der 2.Fahrt ein durch die Einstellungen nicht begründbarer Ausreißer nach unten (über den gesamten Ofen) auftrat, der aber vom Kontroller problemlos behoben wurde.

Um die Ergebnisse besser deuten zu können und um sicherzustellen, dass auch zwischen den Testscheibenpositionen die gewünschte Schichtdicke erreicht wird, wurden wie bei den Experimenten 7 Testscheiben über den Ofen verteilt und deren mittlere Schichtdicke gemessen.

Tabelle 8:Vorhersage des Algorithmus V03/2D

Run.Nr.:	T1[°C]	T2[°C]	t [min, sec]	T4[°C]	T5[°C]
Bisher	797	795	91´00´´	760	758
1	797,1	795,9	92´18´´	761,7	757,3
2	797	796	92´02´´	761,6	756,7
3	797	795,8	92´21´´	761,4	757
4	797,1	795,8	92´12´´	761,5	756,7

Tabelle 9: Schichtdicken [nm] V03/2D

Run.Nr.:	SD1	SD2	SD3	SD4	SD5	SD6	SD7
1	141,3	140,8	139,9	141,3	140,2	141,3	143,2
2	138,4	138,8	137,4	138,4	138,0	139,3	138,2
3	140,5	139,9	139,1	140,7	139,5	139,5	141,7
4	141,2	140,6	139,3	140,1	140,4	141,0	139,9

Es zeigt sich, dass der Algorithmus ausgezeichnet funktioniert. Die Vorhersage für Run 4 wurde bereits für produktive Fahrten angewendet.

Wenn das bisherige Rezept mit dem vom Algorithmus Vorgeschlagenen verglichen wird, erkennt man sofort das Potenzial der Methode. Der Algorithmus verbessert nicht nur die R2R-Uniformity sondern erzielt auch eine wesentlich verbesserte Boat-down-Uniformity.

Man sieht auch, dass einmalige Sprünge, wie von mir simuliert, auch in der Praxis vorkommen. Beim zweiten Run gab es so einen Sprung um ca. 2 – 3 nm nach unten, der nicht auf die Wahl der Stellgrößen zurückzuführen ist. Ohne diesen Sprung wäre das Rezept bereits bei der zweiten Fahrt produktiv einsetzbar.

6.2 Verifikation auf Ofen V14/2D

Bei dieser Versuchsreihe wurde der Algorithmus auf einem anderen ASM-Vertikalofen für das gleiche Rezept (140 nm) überprüft. Für Fahrt 1 wurde dasselbe Regelmodell wie für Ofen V03/2D verwendet. Damit soll überprüft werden, wie gut der Algorithmus mit großen Modelfehlern zurechtkommt und wie lange er für die Erstellung eines für die Regelung ausreichenden Modells braucht. Die unterschiedlichen Rezepte zeigen, dass die gewünschte Zielschichtdicke auf verschiedenen Öfen nur durch unterschiedliche Wahl der Stellgrößen erreicht werden kann.

Tabelle 10:Vorhersage des Algorithmus V14/2D

Run.Nr.:	T1[°C]	T2[°C]	t [min, sec]	T4[°C]	T5[°C]
Bisher	803,5	802	88´30´´	754,5	754
1	797,1	795,9	92´18´´	761,7	757,3
2	799,2	797,7	91´48´´	760,4	755,5
3	802,1	800,4	94´41´´	757,9	753,1
4	804,5	801,2	91´14´´	757,1	752,5
5	804,5	802	91´25´´	757,7	753,9
6	804,1	801,2	91´02´´	758,2	754,2

Tabelle 11: Schichtdicken [nm] V14/2D

Run.Nr.:	SD1	SD2	SD3	SD4	SD5	SD6	SD7
1	128,2	131,9	138,4	140,1	143,8	148,4	151,8
2	125,9	128,5	130,6	132,9	140,2	144,8	144,2
3	141,6	145,6	146,2	146,0	146,0	145,7	145,5
4	143,1	140,8	141,9	141,8	140,1	138,8	134,3
5	141,2	142,5	141,4	140,9	140,2	139,6	140,2
6	139,3	139,3	139,9	140,3	139,6	140,3	139,6

Hier zeigt die Methode ihre Stärken. Sie ist imstande bei großen Modellfehlern (d.h. dass das Initialmodell den realen Prozess sehr schlecht beschreibt), innerhalb von 5 Ofenfahrten ein brauchbares Regelmodell zu finden das den Zustand des Ofens beschreibt. Dabei wird die Boat-Down-Uniformity von anfangs ca. 24 nm auf 1 nm reduziert. Sehr gut erkennbar ist, dass sich das bisherige Rezept – aus Messungen an nur 3 Ofenpositionen entstanden - deutlich vom vorhergesagten unterscheidet. Daraus lässt sich ableiten, dass für eine vernünftige (± 1,5 nm) Boat-Down-Uniformity beim Nitridprozess mindestens an fünf Positionen gemessen werden muss. Diese Erkenntnis ist für die Messhäufigkeiten (Anzahl der Testscheiben pro Fahrt) bei Einsatz des Kontrollers in der Praxis von entscheidender Bedeutung.

Überdies wurden an diesem Ofen während der Versuchsreihe Wartungsarbeiten (Abscheiderwechsel) durchgeführt. Dies zeigt, dass der Algorithmus mit den im praktischen Betrieb vorkommenden Störungen ohne Probleme zurechtkommt.

6.3 Verifikation auf Ofen V13/2D

Bei der dritten Versuchsreihe zur Verifikation des Algorithmus wurde wieder ein anderer ASM-Vertikalofen verwendet. Die Ausgangsbedingungen sind die gleichen wie bei den Versuchen auf Ofen V14/2D. Auch hier wurde das Initialmodell, welches für Ofen V03/2D gefunden wurde verwendet.

Im Speziellen soll hier untersucht werden, wie sich der Algorithmus verhält wenn seine Stellgrößen an die vorgegebenen Grenzwerte stoßen (T_1 bei Fahrt 3 bis 8). Außerdem soll geklärt werden, wie sich der Algorithmus verhält wenn man während des Betriebes die Stellgrößengrenzen (von Fahrt 4 auf 5) verändert d.h. wenn man den Kontroller außerhalb des Stellgrößenbereiches der aus dem DoE hervorgeht betreibt.

Tabelle 12:Vorhersage des Algorithmus V13/2D

Run.Nr.:	T1[°C]	T2[°C]	t [min, sec]	T4[°C]	T5[°C]
Bisher	800	797	89´00´´	754	754
1	797,1	795,9	92´18´´	761,7	757,3
2	799,4	797,8	91´40´´	760,3	755,2
3	802	799,6	91´27´´	758,1	753,7
4	802	799,1	90´21´´	757,4	753,6
5	804,5	800,8	91´08´´	756,1	751
6	804,5	801,5	89´29´´	757,1	752,7
7(produktiv)	804,5	801,3	89´59	757,5	753

Tabelle 13: Schichtdicken [nm] V13/2D

Run.Nr.:	SD1	SD2	SD3	SD4	SD5	SD6	SD7
1	128	133	137,4	141,6	146,6	150,9	155,9
2	129,6	134,3	137,1	141,6	144,4	147,7	146,8
3	138	140,1	141,9	144,9	144,3	143,3	143,6
4	135,7	136,4	139,0	141,9	141,7	139,8	141,6
5	145,5	144,4	144,9	145,9	143,8	140,5	134,8
6	138,8	139,7	140,3	141,5	140,2	138,2	137,4
7	141,7	141,1	141,6	142,6	141,4	140,0	141,6

Hier zeigt sich ein ähnliches Ergebnis wie bei der vorangegangenen Versuchsreihe auf Ofen V14/2D. Ein Überschreiten der Stellgrößengrenzen (Fahrt 4 auf 5), stellt für den Algorithmus ebenfalls kein Problem dar. Dies eröffnet dem Kontroller einen viel weiteren Einsatzbereich. Er kann ohne gesondertes DoE ein Regelmodell erstellen selbst wenn ein vollkommen anderes Standardrezept z.B. mit T3 = 700°C vorgegeben wird. Vorraussetzung dafür ist allerdings die Kenntnis der grundsätzlichen Abhängigkeiten zwischen Stellgrößen und Regelgrößen. Auch ein Anstoßen einer Stellgröße (T1 Run 3 – 8) an seinen maximalen Wert, wird vom Algorithmus, sofern es physikalisch möglich ist, durch Änderung der anderen Stellgrößen kompensiert.

7 Schlussfolgerungen und Ausblick

Der in dieser Arbeit untersuchte Algorithmus zeigt, dass auch Probleme die nicht linearisierbar sind und ein quadratisches Input/Output Verhalten haben, durch den DHOBE-Kontroller in einer R2R-Regelung gelöst werden können. Diese Fähigkeit vergrößert das Einsatzgebiet eines R2R-Kontrollers in der Halbleiterindustrie auf Prozesse die momentan noch nicht oder nur durch starke Vereinfachungen und Annahmen automatisch regelbar sind.

Darüber hinaus ist der Algorithmus in der Lage Prozessergebnisse zu optimieren. Er kann jenes Rezept finden bei dem eine ausgezeichnete Schichtdickenuniformität erreicht wird. Für diese Aufgabe kann er bereits jetzt eingesetzt werden, ohne das begleitende Maßnahmen, wie Messgerätekopplung, Anbindung an das Firmennetzwerk, Anbindung an den ASM-Ofen, die für eine erfolgreiche Integration des Algorithmus in eine R2R-Regelung notwendig sind, durchgeführt wurden.

Die Verbesserung der Boat-Down-Uniformity und der R2R-Uniformity bei der Abscheidung von Siliziumnitrid beim LPCVD-Verfahren war Ziel dieser Arbeit. Simulationen zeigen, dass der Algorithmus mit allen auch in der Praxis vorkommenden Störungen zurechtkommt. Bei geeigneter Wahl der Algorithmusparameter zeigt sich ein ausgezeichnetes Regelverhalten. Die Geschwindigkeit mit der die Zielwerte nach einer Störung wieder erreicht werden beeindruckt und ist für den Einsatz bei Ofenprozessen erforderlich, da es bei jedem Rohrwechsel unter Umständen zu großen Zustandsänderungen kommen kann. Dort muss der Kontroller innerhalb einer Testfahrt (Fahrt ohne produktive Scheiben) passende Stellgrößen finden. Auch diese Anforderung kann von einem DHOBE-Kontroller erfüllt werden, wenn ein gutes Regelmodell vorhanden ist. Messverzögerungen wie sie in der Praxis vorkommen stellen ebenso, wie der Wunsch nach einer Verringerung der Messhäufigkeit kein großes Problem dar. Der Algorithmus kann auch damit zufrieden stellend umgehen wie die Simulationen mit rollierender Messung zeigen.

Es gilt jedoch klar zu stellen, dass zur Erstellung des Regelmodells für einen Ofenprozess oder für die Adaptierung eines Modells auf einen anderen Ofen eine bestimmte Anzahl an Testfahrten notwendig ist. Jedoch kann aus Abbildung 37 entnommen werden, dass mit maximal 10 – 12 Runs ein ausreichend präzises Regelmodell gefunden werden kann. Versucht man über Experimente ein solches Modell zu finden sind mindestens 20 Testfahrten notwendig um über eine statistische Auswertung ein brauchbares Modell zu erstellen. Diese Eigenschaft ermöglicht eine rasche Modellbildung für einen neuen Prozess. Nur der OAQC-Kontroller [17] besitzt ähnliche Eigenschaften.

Will man jedoch in Simulationen den Algorithmus auf seine Fähigkeiten hin untersuchen und die Algorithmusparameter für einen anderen Prozess (TEOS, Poly, doped-Poly) in Offline Simulationen bestimmen, empfiehlt es sich ein Prozessmodell über ein DoE zu erstellen. So

ist es möglich den Algorithmus auf sein Verhalten bei Modellfehlern zu untersuchen. Denn jedes Modell eines Prozesses bei dem Prozessrauschen vorhanden ist, und sei es auch noch so exakt, ist fehlerhaft und unterscheidet sich vom realen Prozessmodell.

Der Umgang mit diesem Modelfehler bei der Vorhersage der Stellgrößen für den nächsten Run zeichnet den DHOBE-Algorithmus aus. Dadurch, dass er immer eine Gruppe von möglichen Modellen vorhersagt wird dieser Fehler berücksichtigt.

Betrachtet man nun die Performance des Kontrollers so wird schnell klar, dass diese Methode zur Prozessregelung für Ofenprozesse ausgezeichnet geeignet ist. Die cpk-Werte liegen bei allen Simulationen höher als die momentan in der Praxis (cpk = 0,85 -1,4) erreichten. Trotzdem muss festgehalten werden, dass die Werte aus den Simulationen nur bedingt mit den Werten aus der Praxis verglichen werden können. Das liegt daran, dass die Störungen die bei Ofenprozessen auftreten zwar in ihrer Art in einer Simulation berücksichtigt werden können jedoch deren Anzahl und Häufigkeit nicht exakt bekannt sind.

Vergleicht man die Simulationen untereinander so hat sich gezeigt, dass für eine Regelung des Nitridprozesses mindestens an fünf Ofenpositionen gemessen werden sollte. Dies gewährleistet eine Ausreichende Kontrolle der Boat-Down-Uniformity. Nimmt man für den Nitridprozess ein quadratisches Regelmodell so erhält man die besten R2R-Ergebnisse. Um die Stabilität des Algorithmus zu gewährleisten ist alleine die Kenntnis über die Höhe des Prozessrauschens ausreichend. Dieser Fakt, der über einen einzigen Parameter in das Modell eingeht, erleichtert die praktische Anwendung erheblich.

Dieser R2R-Kontroller kann ohne großen Aufwand auf andere LPCVD-Prozesse oder auch auf andere Prozesse der Halbleiterindustrie wie z.B. CMP angewandt werden, sofern ein Prozessmodell vorhanden ist. Dieses Prozessmodell muss nicht sehr exakt sein, da mit einigen Testfahrten der Algorithmus ein brauchbares Regelmodell findet. Schwieriger ist die Verknüpfung einzelner Kontroller die unterschiedliche Rezepte des gleichen Prozesses am selben Equipment regeln. Jene Rezepte die nur unterschiedliche Schichtdicken produzieren aber die gleiche Centertemperatur haben, lassen sich durch die lineare Abhängigkeit der Zeit auf die Schichtdicken mit nur einem Regelmodell beschreiben. Ist jedoch die Centertemperatur eine andere braucht der Kontroller ein anderes Regelmodell.

Um eine Zustandsveränderung die bei einer Ofenfahrt durch ein bestimmtes Regelmodell erkannt wird, auf ein anderes Regelmodell übergeben zu können bedarf es noch an weiterer Entwicklungsarbeit. In [18] erhält man einen weiterführenden Einblick über die Vorgehensweise und Probleme die bei der Implementierung in eine CIM Produktionsumbebung auftreten. Dort wird in einzelnen Aufsätzen die R2R-Thematik über die reine Algorithmuserstellung hinaus beleuchtet und ein Einblick in die Verknüpfung mehrerer R2R-Kontroller in einer gegeben.

Zusätzlich werden in Fallbeispielen andere R2R-Kontroller beschrieben und auf ihre Fähigkeiten Prozesse zu regeln analysiert. Man gewährt dort einen allgemeinen Einblick warum der Einsatz von R2R-Kontrollern in der Halbleiterbranche, die Qualität und die Stabilität der Prozesse deutlich verbessern kann, und skizziert dabei einige mögliche Vorgehensweisen bei der Umsetzung der Implementierung in eine Produktion.

8 Verzeichnisse

8.1 Literatur

[1] James Moyne: Advancements in Chemical Mechanical Planarization Process Automation and Control: Aufsatz aus Run to Run Control in Semiconductor Manufacturing, S 279 – 288: 2001

[2] Michael L. Miller: Impact of multi-product and -process manufacturing on run-to-run control: Aufsatz aus Proceedings of the SPIE Vol. 3213 S138 – 146, 1997.

[3] Dr. Johannes Baumgartl: Infineon Technologies Austria AG: Foliensatz CVD Prozesse Kurzfassung, zur internen Verwendung, 2002

[4] Prof. Dr.-Ing. habil. J. Wernstedt: Versuchsplanung und Modellierung Skript: Fakulät für Informatik und Automatisierung – Institut für Automatisierungs- und Systemtechnik: www.systemtechnik.tu-ilmenau.de, 2002

[5] G.J. Schoof, C.R. Kleijn H.E.A. Van der Akker, T.G.M. Oosterlaken, H.J.C.M. Terhorst, F. Huussen: Simulation and validation of SiO_2 LPCVD from TEOS in a vertical 300 mm multi-wafer reactor: Aufsatz, 2002.

[6] Thomas A. Badgwell, Thomas F. Edgar, Isaac Trachtenberg: Experimental Verification of a Fondamental Model for Multiwafer Low-pressure Chemical Vapor Deposition of Poysilicon: J. Electrochem. Soc., Vol. 139, No. 2, February 1992

[7] De Luca Cristina: Statistical Concepts and Methods for Run-to-Run Control in Smiconductor Lithography: September 2003

[8] Sematech Statistical Methods Group: Introduction to Design of Experiments Using JMP: Schulungsunterlagen zu JMP: März 2003

[9] Hao Deng, Chang Zhang andJohn S. Baras: Run-to-Run Control Methods based on the DHOBE Algorithm: www.isr.umd.edu/CSHCN: November 1999

[10] Enrique Del Castillo, Arnon M. Hurwitz: Process Control and Optimization Methods for Run-to-Run Applcation: Aufsatz aus Run to Run Control in Semiconductor Manufacturing, S 47 – 50: 2001

[11] Enrique Del Castillo: Learning and Optimization Algorithm for an Optimizing Adaptive Quality Controller: Aufsatz aus Run to Run Control in Semiconductor Manufacturing, S 81 - 90: 2001

[12] Enrique Del Castillo: An Adaptive Run-to-Run Optimization Controller for Linear and Nonlinear Semiconductor Processes: IEEE Transaction on Semiconductor Manufaturing. Vol. 11: Mai 1998

[13] S.W. Butler, J.A. Stefani: Supervisory run-to-run control of polysilicon gate etch using in situ ellipsometry: IEEE Trans. Semicondut. Manufact., Vol. 7, S 193 -201: 1994

[14] T. H. Smith, D.S. Boning: Artifical neural network exponentially weighted moving average controller for semiconductor processes: J. Vacuum Science Tech. A, Vol. 15, no. 3, S 236 -239: 1997

[15] Chang Zhang, Hao Deng, John S. Baras: Comparision of Run-to-Run Control Methods in Semiconductor Manufacturing Processes: August 2000

[16] Carroll Croarkin, Paul Tobias,…:NIST/SEMATECH e-Handbook of Statistical Methods: www.itl.nist.gov/div898/handbook/index.html : 2004

[17] Arnon Max Hurwitz, Enrique Del Castillo: An Adaptive Run-to-Run Optimizing Controller for Linear and Nonlinear Processes:Paper out of Run to Run Control in Semiconductor Manufacturing, S 91 -100: 2001

[18] James Moyne, Enrique Del Castillo, Arnon Max Hurwitz: Run-to-Run Control in Semiconductor Manufactoring: CRC Press LLC: 2001

[19] Erwin Kreyszig: Advanced Engineering Mathematics, Third Edition: John Wiley and Sons, Inc.:ISBN 0-471-50728-8: 1972

[20] G. W. Stewart: Matrix Algorithms, Volume I: Basic Decompositions: Society for Industrial and Applied Mathematics: ISBN 0-89871-414-1: 1998

[21] Enrique Del Castillo: Learning and Optimization Algorithms for an Adaptive Quality Controller:Paper out or Run to Run Control in Semiconductor Manufacturing, S 81 – 89: 2001

[22] Thermavave: Opti-Probe Theory of Operation: PN 76-016554: 1997

8.2 Abkürzungsverzeichnis

R2R	Run to Run
CMP	Chemical Mechanical Planarization
TEOS	Tetraethylorthosilikat
CVD	Chemical Vapor Deposition
LPCVD	Low Pressure Chemical Vapor Deposition
APCVD	Atmospheric Pressure Chemical Vapor Deposition
DoE	Design of Experiment
DCS	Dichlorsilan
CIM	Computer Integrated Manufacturing
RSM	Response Surface Method
ANOVA	Analysis of Variance
SISO	Single Input Single Output
MISO	Multi Input Single Output
MIMO	Multi Input Multi Output
EWMA	Exponential Weighted Moving Average
OAQC	Optimizing Adaptive Quality Controller
DHOBE	Dasgupta – Huang Optimal Bounded Ellipsoid
ANN	Artificial Neural Network
LSR	Least Square Recursive
SSE	Sum of Squared Error
SPC	Statistical Process Control
RMS-Error	Root Mean Squared Error
DoE	Design of Experiment
APC	Advanced Process Control
T_x [°C]	Zonentemperaturen
t [min,sec]	Abscheidezeit
p [mtorr]	Abscheidedruck
SD_i [nm]	Schichtdicken
cp,cpk	Prozessfähigkeitsindizes
$x_{i,j}$	Schichtdickeneinzelmessung
OSG [nm]	Obere Spezifikationsgrenze
USG [nm]	Untere Spezifikationsgrenze
y_t	Regelgröße (Regelgrößenvektor)
u_t	Stellgrößenvektor
θ^*	wahrer Parametervektor
e_t	Prozessrauschen
γ	Algorithmusparameter
θ_t	Vektor der Modellparameter (Mittelpunkt des Ellipsoids)

σ_t^2	Unsicherheitsfaktor
E_t	Ellipsoid
S_t	Ebene
λ_{max}	maximaler Updatefaktor
λ_t	Updatefaktor
$(1-\lambda_t)$	Forgettingfaktor
P_t-1	Strukturmatrix des Ellipsoids
δ_t [nm]	Vorhersagefehler
ν_t	Hilfsvariable
G_t	Hilfsvariable
β_t	Hilfsvariable
κ	Hilfsvariable
ζ	Inflationsparameter
T [nm]	Zielwert (Target)
W	Gewichtungsmatrix

8.3 Tabellen

8.4 Abbildungen

Anhang

Tabelle 14: Datensätze für Initial- und Prozessmodell (T_3 = 780 °C)

Run	T1	T2	T4	T5	Time	SD1	SD2	SD3	SD4	SD5	SD6	SD7
	[°C]	[°C]	[°C]	[°C]	[min]	[nm]	[nm]	[nm]	[nm]	[nm]	[nm]	[nm]
1	**797.0**	**795.0**	**760.0**	**756.0**	**91.5**	138.9	135.9	137	138.8	137.5	135	134.7
2	**792.0**	**790.0**	**753.0**	**750.0**	**86.0**	121.3	122.1	131.3	134.6	131.6	118.0	113.9
3	**802.0**	**800.0**	**753.0**	**750.0**	**86.0**	141.3	137.1	135.8	132.1	124.8	115.3	113.7
4	**794.8**	**790.0**	**758.0**	**750.0**	**86.0**	129.8	120.7	126.5	132.5	133.5	129.0	118.3
5	**797.4**	**792.0**	**761.0**	**750.0**	**86.0**	134.0	122.0	124.7	130.1	133.3	134.4	118.4
6	**792.0**	**795.0**	**760.8**	**750.0**	**86.0**	119.3	141.1	129.4	130.5	132.5	134.1	117.1
7	**797.0**	**795.0**	**760.0**	**756.0**	**91.5**	141.5	138.8	139.2	140.6	139.0	136.9	135.0
8	**792.0**	**790.0**	**758.0**	**755.0**	**86.0**	121.7	120.2	125.5	130.3	128.6	122.2	120.4
9	**797.0**	**798.5**	**762.6**	**755.0**	**86.0**	127.2	133.2	128.9	127.7	129.7	133.3	126.8
10	**795.0**	**790.0**	**753.0**	**756.0**	**86.0**	131.4	122.4	129.5	134.2	126.2	113.3	122.6
11	**797.0**	**800.0**	**757.0**	**760.0**	**86.0**	133.5	141.1	138	133.7	127.1	119.5	136.1
12	**802.0**	**797.0**	**758.0**	**760.0**	**86.0**	138.8	131.2	130.8	130.8	125.8	120.6	132.0
13	**792.0**	**790.0**	**763.0**	**760.0**	**86.0**	118.8	115.9	121.5	126.3	128.6	131.0	135.9
14	**802.0**	**800.0**	**763.0**	**760.0**	**86.0**	135.6	131.8	126.1	127.2	126.6	128.9	137.1
15	**801.3**	**797.0**	**753.0**	**750.0**	**96.0**	162.0	150.3	151.0	151.3	142.4	130.7	129.4
16	**797.0**	**798.8**	**753.0**	**750.0**	**96.0**	145.8	155.0	153.2	149.6	140.4	129.1	127.7
17	**797.0**	**795.0**	**760.0**	**756.0**	**91.0**	141.0	138.3	139.1	140.5	138.9	136.8	137.4
18	**792.0**	**790.0**	**762.8**	**755.0**	**96.0**	135.3	133.2	137.9	144.4	148.8	151.8	141.9
19	**802.0**	**800.0**	**762.7**	**755.0**	**96.0**	155.2	150.8	145.4	144.7	145.9	150.8	144.4
20	**797.0**	**795.0**	**753.0**	**755.0**	**96.0**	153.6	150.7	154.2	153.8	143.2	129.8	138.6
21	**792.0**	**790.0**	**757.0**	**759.0**	**96.0**	137.7	135.2	141.9	146.8	142.2	133.9	144.6
22	**802.0**	**800.0**	**757.0**	**758.5**	**96.0**	163.6	159.4	155.3	151.3	143.1	136.4	148.7
23	**801.3**	**797.0**	**763.0**	**760.0**	**96.0**	155.7	144.8	142.6	144.3	145.1	148.6	156.3
24	**797.0**	**798.5**	**763.0**	**760.0**	**96.0**	141.2	147.4	143.9	142.7	142.6	146.0	152.4
25	**792.0**	**790.0**	**763.0**	**760.0**	**91.0**	126.1	123.6	127.9	133.9	136.4	137.5	144.6
26	**802.0**	**800.0**	**753.0**	**750.0**	**91.0**	149.5	146.2	143.8	140.5	131.4	121.1	119.4
27	**792.0**	**790.0**	**753.0**	**750.0**	**91.0**	129.3	128.8	138.2	142.7	138.1	125.1	119.9
28	**797.0**	**795.0**	**758.0**	**755.0**	**91.0**	142.3	140.4	141.6	142.6	138.8	133.4	131.3
29	**802.0**	**800.0**	**763.0**	**760.0**	**91.0**	143.3	139.2	134.3	134.0	134.2	137.2	144.9
30	**797.0**	**795.0**	**758.0**	**755.0**	**91.0**	142.5	140.6	141.8	142.5	138.7	133.5	132.0
31	**797.0**	**795.0**	**758.0**	**755.0**	**91.0**	141.3	139.4	140.8	142.5	138.5	133.0	131.3
32	**802.0**	**800.0**	**753.0**	**750.0**	**96.0**	157.3	153.7	151.1	147.2	138.9	127.0	126.3
33	**792.0**	**790.0**	**763.0**	**760.0**	**96.0**	132.5	130.7	135.7	140.5	144.3	145.7	151.7
34	**792.0**	**790.0**	**753.0**	**750.0**	**96.0**	136.4	136.5	146.7	149.5	145.6	132.3	127.3
35	**802.0**	**800.0**	**763.0**	**760.0**	**96.0**	150.9	148.0	142.3	141.2	141.2	144.0	151.9