Preprints of the
15th IFAC Symposium on System Identification
Saint-Malo, France, July 6-8, 2009

A System Identification Approach to PDE Modeling of a Semiconductor Manufacturing Process

Jay D. Schwartz * **Daniel E. Rivera** *

* *Control Systems Engineering Laboratory,*
Department of Chemical Engineering,
Arizona State University, Tempe, AZ 85287-6006 USA
(e-mail: jayschwartz@asu.edu, daniel.rivera@asu.edu).

Abstract: Efficient supply chain management is a crucial imperative for modern, global enterprises. Tactical decision policies based on process control principles have been developed in the literature for managing production-inventory systems and supply chain networks. To be effective these decision policies depend on accurate nominal models. With a discrete-event simulation acting as a "truth model", we employ system identification techniques to parameterize a nonlinear Partial Differential Equation (PDE) model of the semiconductor manufacturing process. A case study shows that the identified PDE model can accurately predict the output of the discrete-event simulation, but without the high computational burden.

Keywords: System Identification; Discrete Event Simulation; Semiconductor Manufacturing; Simultaneous Perturbation Stochastic Approximation

1. INTRODUCTION

The use of process control based tactical decision policies for supply chain management has been the focus of much research [Grubbström and Wikner, 1996, Dejonckheere et al., 2003]. Such a policy depends on the selection of a nominal model of the process. Previous work on the production control of semiconductor manufacturing processes has been based on a "plug-flow" fluid model abstraction, meaning material moves at a constant rate throughout the factory and is independent of load or any internal process dynamics [Schwartz et al., 2006]. While this level of abstraction provides a reasonable basis from which it is possible to develop a wide variety of decision policies, it is desirable to develop more accurate fluid representations of the manufacturing process without incurring significant computational cost. Real-life semiconductor manufacturing processes are highly stochastic and nonlinear. Therefore, the development of more sophisticated fluid models is necessary to better capture the underlying dynamics. This paper presents a data-driven approach relying on system identification to obtain a nonlinear Partial Differential Equation (PDE) based model from a discrete-event simulation of a semiconductor fab.

This paper showcases a framework for modeling discrete-event-representations of a manufacturing process as nonlinear fluid flow systems. The approach is presented in a step-by-step fashion outlined as follows.

(1) The first stage is the construction of a discrete-event simulation of the manufacturing process. A discrete-event simulation models each part as it is processed in the factory. This fine-grained approach incorporates the stochasticity and nonlinearity present in the real manufacturing system, but suffers from state explosion. Computational burden increases exponentially as the size of the system increases, this motivates the use of a nonlinear fluid model approximation. The discrete-event simulator is discussed in Section 2.
(2) The next step is to select a proper model structure and design an input signal. An appropriate model should exhibit parsimony; it should include enough terms to capture any dynamics and nonlinearities present in the real system, but not be so complicated that the model will be over-parameterized. The discrete-event simulation will be subject to an input signal that excites all the relevant modes. These topics are the focus of Section 3.
(3) The final stage is to estimate and validate the parameters in the model. An optimization scheme is used to estimate the model parameters such that any difference between estimated and actual output is minimized. A data set other than the one used for estimation will serve to test the predictive ability of the model. Section 4 presents a case study that showcases the results of this process. Conclusions are discussed in Section 5.

The end result of this research is a framework for creating simple, accurate fluid transport models from factory or validated discrete-event simulation data. These transport models have wide applicability, from being used as the internal models in predictive control algorithms, or helping to determine the effect of forecast error on a real factory.

⋆ The authors would like to acknowledge support from the National Science Foundation (CMMI-0432439) and the Intel Research Council. Corresponding author Daniel E. Rivera. Telephone (480) 965-9476. Fax (480) 965-0037.

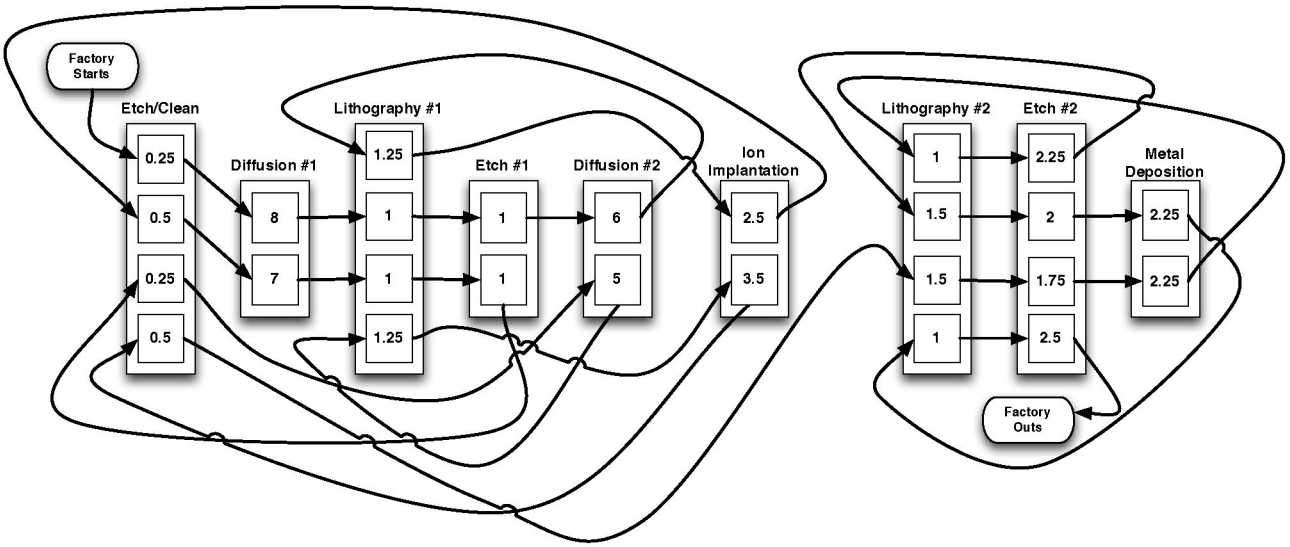


Fig. 1. Block diagram of the semiconductor manufacturing process. Numbers represent mean processing times in hours.

2. DISCRETE-EVENT SIMULATION

Discrete-event simulation of a factory can provide a true-to-life experimental testbed for parameterizing a fluid model. Discrete-event models have shown tremendous potential in the modeling of manufacturing systems. These systems track individual entities as they move through a process (such as parts moving from a buffer to a machine to an inventory). System dynamics, throughput, and cycle time are all inherent in such models. Discrete-event theory is unrivaled at the microscopic level inside a factory. The ability to track individual parts through a factory leads to extremely accurate simulations [Banks et al., 1999].

We consider a discrete-event simulation of a factory that is similar to the one shown in Armbruster et al. [2006]. The simulation, implemented in the Simulink® environment, incorporates 26 different processing steps, machine breakdown, re-entrant processing, and recycle flows. Figure 1 shows the top-level block diagram of the manufacturing process. The fab model consists of etch, diffusion, lithography, metal deposition, ion implantation, and cleaning steps. Entities (lots) must be processed in 26 different steps distributed into 9 different types of machines. Each bank of machines handles reentrant flows from other processing steps. Flow out of the last processing step is equivalent to the factory outs. A table of average throughput times for the manufacturing process is shown in Figure 2. Actual throughput times are ± 50% of the average, determined via uniform distribution.

Step	Diff 1	Diff 2	Litho 1	Etch Clean	Etch 1	Ion Imp	Met Dep	Litho 2	Etch 2	Description
1				0.25						clean wafer
2	8									grow layer
3			1							pattern
4					1					etch
5		6								grow layer
6			1.25							pattern
7						2.5				implant ions
8				0.5						remove mask
9	7									grow layer
10			1							pattern
11					1					etch
12				0.25						clean wafer
13		5								grow layer
14			1.25							pattern
15						3.5				implant ions
16				0.5						remove mask
17								1.5		pattern
18									1.75	etch contact
19							2.25			layer metal
20								1		pattern metal
21									2.25	etch metal
22								1.5		pattern contact
23									2	etch contact
24							2.25			layer metal
25								1		pattern metal
26									2.5	etch metal
# of machines	7	6	6	4	5	10	8	7	21	

Fig. 2. Table summarizing the process steps and mean throughput times shown in Figure 1.

3. MODEL DEVELOPMENT VIA SYSTEM IDENTIFICATION

It is desirable to obtain more accurate models of manufacturing processes for the purposes of analysis and control. The literature is dominated by three varieties of manufacturing line models. Historically, queuing theory has been used for the analysis of manufacturing lines [Buzacott and Shantikumar, 1993]. This approach has yielded information about the steady-state behavior of processes. However, since the presence of process dynamics is excluded from these models, their application to real manufacturing systems is limited. Discrete-event system models adequately capture the dynamics of manufacturing systems, but suffer from state explosion as problems are scaled to practical levels. Fluid models offer a reasonable trade-off between accuracy and computational scalability [Marthaler et al., 2003].

To develop a factory model that can accurately capture nonlinear dynamics, the standard fluid analogy is extended to one similar to the modeling of freeway traffic [Helbing, 2001]. The first fluid model for traffic flow dynamics, the LWR model, was developed several decades ago and named after its developers [Lighthill and Whitham, 1955, Richards, 1956]. The basis of the model is the fact that no vehicles are entering or leaving the freeway if the point of interest is far enough away from on- and off-ramps. Input into the freeway (or factory, in the manufacturing analogy) is given by the boundary conditions. The conservation of vehicles (parts) leads to the continuity equation

$$\frac{\partial \rho(x,t)}{\partial t} + \frac{\partial Q(x,t)}{\partial x} = 0 \quad (1)$$

where in manufacturing systems, t denotes the time and x denotes the position in the factory. The flow, $Q(x,t)$, is measured in parts per time and is defined as

$$Q(x,t) = \rho(x,t)v(x,t) \quad (2)$$

where the velocity, $v(x,t)$, is measured in machines per time (as opposed to distance per time). The density, $\rho(x,t)$, is in units of parts per machine. For the semiconductor manufacturing process presented in this paper, the independent variable x denotes the position within the factory. A value of $x = 0$ corresponds to the beginning of the factory, a value of $x = 26$ corresponds to the final manufacturing step. The parameter x is a continuous variable, values of x greater than 0 and less than 26 correspond to intermediate processing steps. The independent variable t denotes the time in hours, shifts, days, or weeks. Historically, traffic modeling has involved the selection of a velocity profile as a function of density. Here we model the fluid velocity as having two parts

$$v(x,t) = v_{base}(x)\left(1 - \frac{\rho(x,t)}{\rho_{max}(x)}\right) \quad (3)$$

This velocity profile specifies that the velocity of fluid in the pipe is a function of both spatial position and the local density. In a semiconductor fab each processing step will have its own throughput time, e.g. diffusion steps may take much longer than ion implantation. The inclusion of a baseline velocity term $v_{base}(x)$ allows one to model parts of the factory has being "fast" and other parts as being "slow". The density dependency is common in the traffic modeling literature; it simply states that the local velocity decreases with increasing part density. For example, parts moving through empty queues will have a faster velocity than parts moving through highly-loaded queues. At some theoretical maximum density $\rho_{max}(x)$ the fluid velocity will approach zero. Applying the velocity profile in Equation 3 yields the following model for the manufacturing problem of interest:

$$\frac{\partial \rho}{\partial t} + v_{base}\left(1 - \frac{2\rho}{\rho_{max}}\right)\frac{\partial \rho}{\partial x} + \rho\left(1 - \frac{\rho}{\rho_{max}}\right)\frac{\partial v_{base}}{\partial x} + \frac{v_{base}\rho^2}{\rho_{max}^2}\frac{\partial \rho_{max}}{\partial x} = 0 \quad (4)$$

We then seek to estimate the theoretical maximum density $\rho_{max}(x)$ and the baseline velocity profile $v_{base}(x)$ from discrete-event simulation data for the model according to Eqn. 4. This model development process is similar to that presented in Armbruster et al. [2006]. However, we will attempt to estimate the parameters of the velocity profile directly from dynamic discrete-event simulation data, as opposed to fitting a state equation that relates the Work-In-Progress to the throughput time.

Parameter Estimation A line search method based on the Nelder-Mead algorithm [Lagarias et al., 1998] is used as a parameter estimation procedure to determine values for the baseline velocity function $v_{base}(x)$ and the maximum density function $\rho_{max}(x)$. This is accomplished by discretizing the partial differential equation in Equation 4 and using a simulation-based approach for estimating a vector of discretized values for $v_{base}(x)$ and $\rho_{max}(x)$. The simulated density $\hat{\rho}_{i,k}$ is updated at each time step according to the following equation

$$\hat{\rho}_{i,k+1} = \hat{\rho}_{i,k} - \bar{v}_i\frac{\Delta t}{\Delta x}\left(1 - \frac{2\hat{\rho}_{i,k}}{\bar{\rho}_i}\right)(\hat{\rho}_{i,k} - \hat{\rho}_{i-1,k}) - \hat{\rho}_{i,k}\frac{\Delta t}{\Delta x}\left(1 - \frac{\hat{\rho}_{i,k}}{\bar{\rho}_i}\right)(\bar{v}_i - \bar{v}_{i-1}) - \frac{\bar{v}_i\hat{\rho}_{i,k}^2}{\bar{\rho}_i^2}\frac{\Delta t}{\Delta x} \quad (5)$$

where $\bar{v}_i$ is an estimate of $v_{base}(x)$, $\bar{\rho}_i$ is an estimate of $\rho_{max}(x)$, i is the index for the current manufacturing step, $i-1$ is the index for the previous manufacturing step in space, k is the index for the current time step, and $k+1$ is the index for the next time step. The ratio of the time step to the grid spacing ($\Delta t/\Delta x$) must be small to ensure numerical stability. All case studies shown in this paper utilize a time step Δt of 0.01 weeks and a grid spacing Δx of 1 processing step. Given a defined flow function at the front of the factory $Q(0,t)$ (factory starts) and estimates for the baseline velocity function $\bar{v}_i$ and the maximum density $\bar{\rho}_i$, Equation 5 can be used to determine the input and output of each manufacturing step for all time samples. The simplex-based direct search estimation method compares the 2-norm of the difference between the actual factory outs from the discrete-event simulator ($Q(26,t)$) and the simulated factory outs from the PDE model ($\hat{Q}_{26,k}$), as shown in Equation 6.

$$\|\hat{Q}_{26,k} - Q(26,t)\|_2 \leq \epsilon_{tolerance} \quad (6)$$

The Nelder-Mead search method seeks to minimize the difference between actual and simulated factory outs by iteratively updating the $\bar{v}_i$ and $\bar{\rho}_i$ functions. This process is repeated until the 2-norm of the flow error signal converges, as shown in Equation 6. The Nelder-Mead algorithm used for estimation is implemented as `fminsearch` within the MATLAB® Optimization Toolbox.

With the proposed partial differential equation model and parameter estimation procedure in place, the next step is to develop an informative input signal that will excite the modes of the manufacturing system. This is the topic of the next section.

3.1 Informative Input Signals for Identification

To generate informative data from a discrete-event simulation it will be necessary to use an input signal that excites all of the relevant factory dynamics. To this end, the experimental factory starts signal is comprised of a multi-level random waveform [Godfrey, 1993]. Multi-level random signals appear to be the most appropriate considering the nonlinearity of the problem and the need for factory starts to correspond to certain discrete values. In addition, the use of a random signal is a justifiable first approach when there is little *a priori* knowledge of the system's underlying dynamics. When subjected to the appropriate input signal the validated discrete-event simulator yields a time dependent density profile. From this data, numerical values for the parameters in the fluid models are determined.

In the next section of this paper we present a case study for the parameterization of fluid models based on partial differential equations. Figure 3(a) shows the multi-level

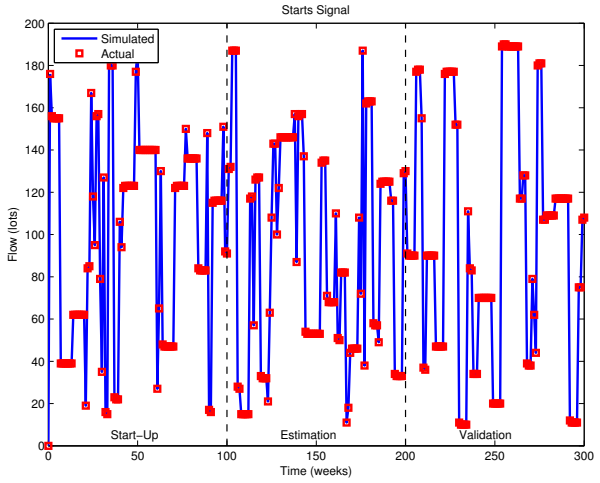
random sequence used for the estimation process. The "actual" starts obtained from the discrete-event simulation directly overlap the simulated factory starts used as a boundary condition for the partial differential equation. Factory starts vary from 10 lots/week (5% of capacity) to 190 lots/week (95% of capacity). The starts level is taken from a uniform distribution and rounded to the nearest integer; therefore the starts level may be any integer value between 10 and 190, inclusive. By selecting a wide range of factory starts values one can increase the likelihood of capturing the nonlinear dynamics. An input signal over a more narrow range would only capture the "local" dynamics in the range of the starts signal. Figure 3(b) shows the smoothed periodogram of the input signal. The periodogram indicates that the signal has a flat spectrum mimicking white noise up to a bandwidth of approximately 0.6 radians/week.

The data consist of 300 weeks of input (factory starts) and output (factory outs). The first 100 weeks of data are treated as "start-up time" for the simulator, the middle 100 weeks of data are used for estimation, and the final 100 weeks are reserved for model validation. The validation data set (shown in Figure 3 from weeks 200 to 300) is not used in the estimation process. Instead it serves as a check to ensure that the identified model is not just a good fit to the estimation data, but that it has predictive ability as well.

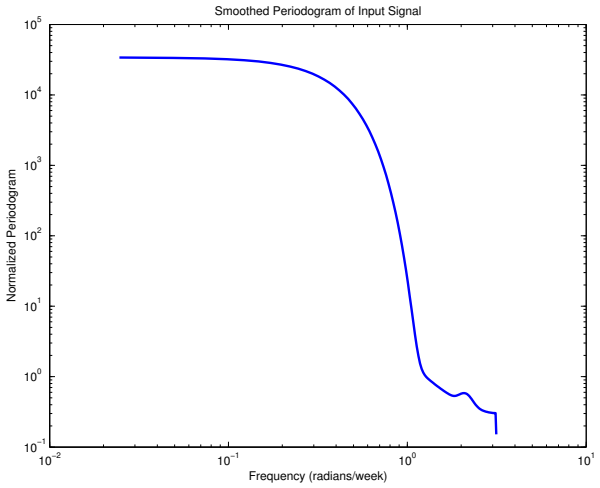
4. CASE STUDY

Figure 4 shows all of the flows within the factory as a function of space and time. These data are obtained from the discrete-event simulation discussed in the previous section and will serve as the input and output data for identification. Note that the flow at Step 0 corresponds to the factory starts level at each moment in time. Subsequent flows (Steps 1 thru 26) are the measured outflow from the relevant process step. Therefore, the outflow from Step 26 is also the Factory Outs. Our concern in this paper is with being able to estimate the parameters of a PDE model that yields a similar flow profile between weeks 200 and 300, which is the validation data set. The goodness-of-fit obtained by the parameter identification process is measured by the 2-norm of the difference between the actual factory outs obtained from the discrete-event simulation and the predicted factory outs obtained from solving the discretized PDE (Equation 5). Figure 5 shows the optimization path, the goodness-of-fit versus the number of iterations in the parameter identification process. The 2-norm is included for both estimation data (solid blue curve) and validation data (dashed red curve).

Figure 6 shows estimates of the baseline velocity field $v_{base}(x)$ and maximum density functions $\rho_{max}(x)$ determined from the parameter estimation process. While the baseline velocity is relatively constant over position (note the small change in scale), the maximum density function increases at the end of the factory. This is consistent with physical knowledge of the factory. The front half of the factory is constrained by recycle flows and batch diffusion processes with long throughput times. The low value of the maximum density profile signifies that velocity will be more negatively affected by high density levels in this



(a)



(b)

Fig. 3. The input signal used in the Case Study is shown in (a), its smoothed periodogram is shown in (b).

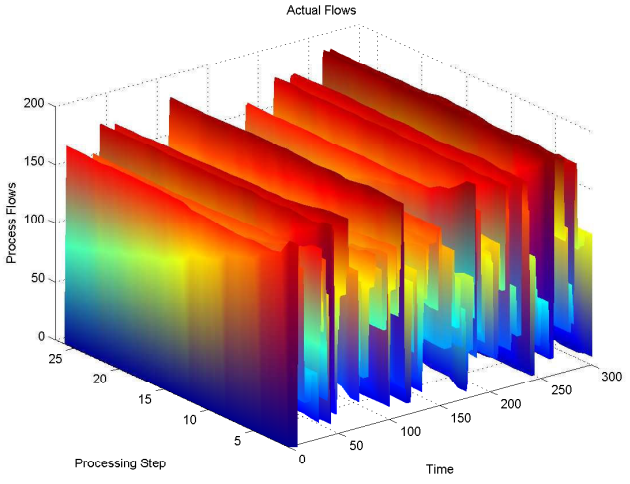


Fig. 4. Internal flows in discrete-event simulator.

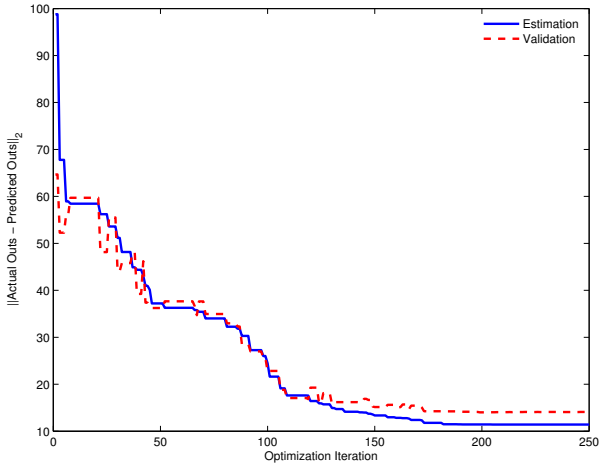


Fig. 5. Optimization path for the parameter estimation process.

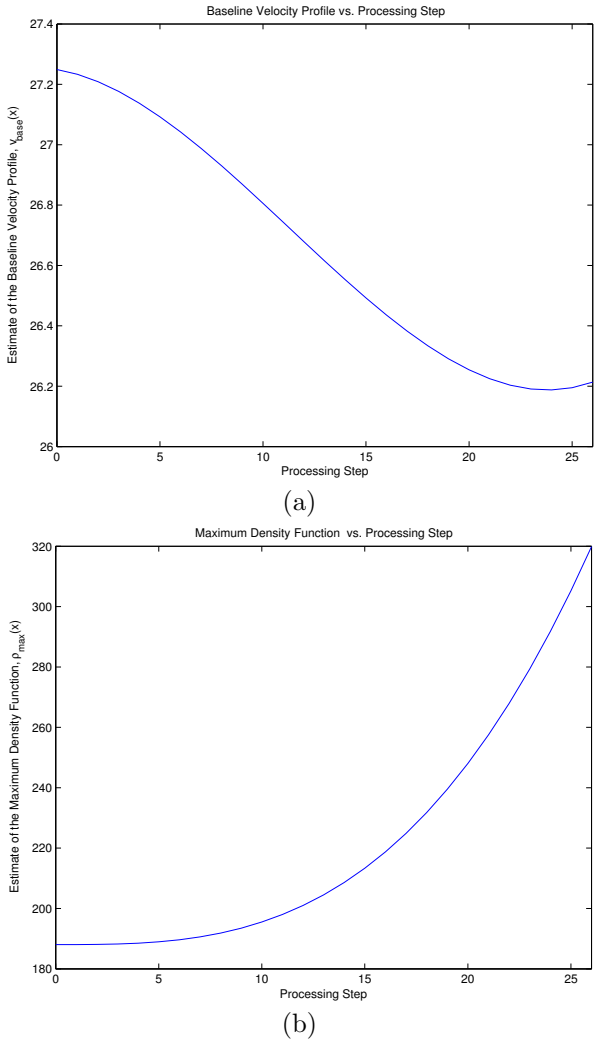


Fig. 6. (a) Estimate of the baseline velocity profile $v_{base}(x)$. (b) Estimate of the maximum density function $\rho_{max}(x)$.

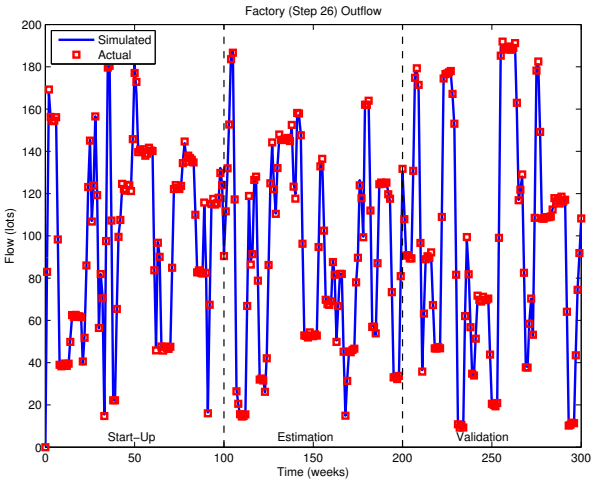


Fig. 7. Simulated and actual factory outflux.

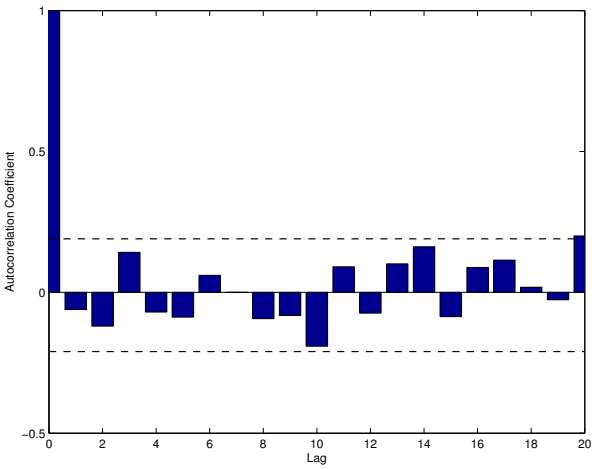


Fig. 8. Autocorrelation of residuals (from validation data). The 95% error bounds are represented by the dashed line.

half of the factory. Towards the end of the factory the maximum density profile increases, indicating that the factory is less constrained during the last few process steps. This is also consistent with physical knowledge of the factory, as there are no batch diffusion processes in the last 14 process steps.

Figure 7 overlays the output of the discrete-event simulator and the PDE-based fluid model. The actual flow signals are determined by averaging 25 stochastic discrete-event simulations. The simulator is given 100 weeks to "start-up", these data are not considered in the estimation or validation of the fluid model. The middle third of the data set (weeks 100 to 200) is for parameter estimation. The parameter identification technique optimizes the baseline velocity field $v_{base}(x)$ and the local maximum density function $\rho_{max}(x)$ to minimize the 2-norm of the difference between the simulated and actual outputs. The last third of the data (weeks 200 to 300) is set aside for model validation.

Figure 8 shows the autocorrelation of the residuals for the validation data set. The 95% standard error bounds are plotted in black dashes. Virtually all of the autocorrelation coefficients at lags greater than 0 fall inside the

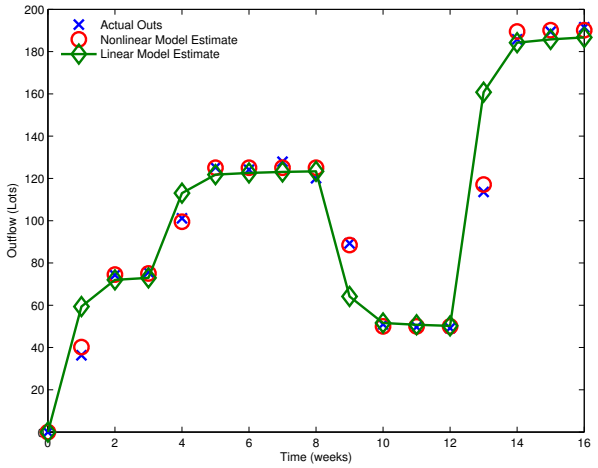


Fig. 9. The output of the nonlinear fluid model is compared against a linear flow model. The use of the nonlinear fluid model improves the accuracy of the prediction by 84.9%.

standard error bounds, indicating that the residuals are not autocorrelated.

Figure 9 contrasts actual factory outs with simulated outs from the nonlinear model when both environments are subjected to random changes in factory starts. In addition, the output of a linear flow model is included for comparison. The linear flow model is a continuous model with one zero and two poles estimated from the input and output data shown in Figure 7. To ensure a fair comparison, the same data segment (weeks 100 to 200) was used for estimating both the linear and nonlinear models. Equation 7 shows the linear model with the estimated parameters. The model was estimated using the System Identification MATLAB® toolbox. The model structure was selected as the estimated model resulted in the best fit to the validation data, relative to other continuous-time models.

$$\frac{Q_{linear}(outs,t)}{Q(0,t)} = \frac{0.99009(1.8277s+1)}{(1.9417s+1)(0.16089s+1)} \qquad (7)$$

The nonlinear model more accurately captures the time constant of the real manufacturing dynamics. Quantifying the results, the 2-norm of the residuals from the nonlinear prediction is 9.08 while the 2-norm of the residuals from the linear prediction is 60.31. The use of the nonlinear modeling process resulted in a 84.9% improvement in the prediction for this particular set of validation data.

5. CONCLUSIONS

Discrete-event simulation of manufacturing systems has been used extensively in both academia and industry. However, a PDE-based fluid model may be more amenable to analysis and better suited for end-use applications such as control system design. This paper has showcased a method for parameterizing fluid models based on partial differential equations from discrete-event system data using a system identification approach. These fluid models have been shown to provide accurate predictions of factory outs across a range of load conditions. The PDE-based fluid model accurately captures the nonlinear dynamics of the discrete-event system. Therefore, nonlinear factory models based on these fluid analogy can enable supply chain planners to develop computationally-efficient large-scale simulations without loss of accuracy.

Accurate parameter estimation for the PDE model depends on proper design of the input signal. To demonstrate the efficacy of the model as a proof-of-concept we have utilized a highly variable factory starts signal and a long identification time. Future research may involve evaluating the effectiveness of the identification approach when less informative data is available. In principle, aspects of this approach could be applied to real factory data.

With nonlinear models in place, further research could be performed on the use of nonlinear control or even nonlinear predictive control for management of starts. This would be most helpful for management of starts for the first stage of the semiconductor manufacturing process, which is the most nonlinear in practice. This approach may be similar to the design of a ramp metering algorithm for traffic flow control [Taylor et al., 204].

REFERENCES

D. Armbruster, D. E. Marthaler, C. Ringhofer, K. G. Kempf, and T. Jo. A continuum model for a re-entrant factory. *Operations Research*, 54(5):933–950, 2006.

J. Banks, J. Carson, and B. Nelson. *Discrete event system simulation*. Prentice-Hall, 1999.

J.A. Buzacott and J.G. Shantikumar. *Stochastic Models of Manufacturing Systems*. Prentice-Hall, 1993.

J. Dejonckheere, S. M. Disney, M. R. Lambrecht, and D. R. Towill. Measuring and avoiding the bullwhip effect: a control theoretic approach. *European Journal of Operational Research*, 147:567–590, 2003.

K. Godfrey, editor. *Perturbation signals for system identification*, volume 6 of *Prentice-Hall International Series in Acoustics, Speech, and Signal Processing*. Prentice-Hall, 1993.

R. W. Grubbström and J. Wikner. Inventory trigger control policies developed in terms of control theory. *International Journal of Production Economics*, 45:397–406, 1996.

D. Helbing. Traffic and related self-driven many particle systems. *Review of Modern Physics*, 73:1067–1141, 2001.

J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright. Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal of Optimization*, 9(1):112–147, 1998.

M.J. Lighthill and J.B. Whitham. On kinematic waves. ii a theory of traffic on long crowded roads. *Proceedings of the Royal Society A*, 229:281–345, 1955.

D. Marthaler, D. Armbruster, and C. Ringhofer. A mesoscopic approach to the simulation of semiconductor supply chains. *Simulation*, 79(3):157–162, 2003.

P.I. Richards. Shockwaves on the highway. *Operations Research*, 47:42–51, 1956.

J. D. Schwartz, W. Wang, and D. E. Rivera. Simulation-based optimization of model predictive control policies for inventory management in supply chains. *Automatica*, 42(8):1311–1320, 2006.

C.J. Taylor, P.G. Mckenna, P.C. Young, A. Chotai, and M. Mackinnon. Macroscopic traffic flow modelling and ramp metering control using MATLAB/Simulink. *Environmental Modelling and Software*, 19:975–988, 204.