JPMC MLCOE TSRL 2026 Internship Q1

Application for lending department of a bank

Molei Qin

Nanyang Technological University

molei001@e.ntu.edu.sg

Updated: November 23, 2025

This report constructs a very simple structural model of the balance sheet and income statement based on the tank-model ideas of Vélez-Pareja Vélez-Pareja 2011; Vélez-Pareja 2010 and the analytical treatment of circularity in discounted cash flow valuation Mejía-Peláez and Vélez-Pareja 2011. We define a low-dimensional state vector of financial statement items and a driver vector of policy and performance ratios. We derive explicit forward equations $y_t = f(y_{t-1}, x_t)$ governing the evolution of the state. We show that, under mild assumptions, the accounting identities—in particular the assets = liabilities + equity identity—are preserved automatically by the evolution equations. This allows us to treat the drivers x_t as a multivariate time series and to shift the forecasting problem from financial statement items themselves to these drivers. We describe how to invert historical data to obtain "perfect" drivers, how to train very simple forecasting models (sliding-window mean, pooled AR(1) and a small multi-layer perceptron), and how to evaluate forecasting accuracy both in driver space and in state space. We also discuss how earnings are obtained naturally from the model and outline machine-learning extensions that could improve forecast performance.

1 Introduction and Literature Background

Forecasting full financial statements in an internally consistent way and free from ad hoc "plug" variables and circularity is an old problem in corporate finance and valuation. Vélez-Pareja Vélez-Pareja 2011; Vélez-Pareja 2010 proposes a "tank"-style approach in which key balance sheet items (cash, working capital, fixed assets, debt, equity) are treated as stocks (tanks) updated by flows that come from the income statement and cash-flow statement. The approach enforces accounting identities by construction and avoids circularity by carefully defining interest, taxes, and equity changes.

Mejía-Peláez and Vélez-Pareja Mejía-Peláez and Vélez-Pareja 2011 further analyse the circularity problem in discounted-cash-flow (DCF) valuation and provide an analytical solution that is compatible with such tank-style models. The key idea is that, once the stocks and flows are linked algebraically in a consistent way, there is no need for iterative numerical solutions: all relevant quantities can be computed in closed form.

In this report we build on these ideas and construct a simple but fully specified evolution model for a reduced balance sheet and income statement. We show how this model can be framed as a time-series problem by treating the drivers—growth rates, margins, working-capital days, capex ratios, tax rates, interest rates, payout ratios and net financing ratios—as a multivariate time series. The forecasting task is then shifted to predicting these drivers, while the deterministic evolution equations take care of the accounting identities and yield consistent forward financial statements.

The remainder of the report is structured as follows. Section 2 defines the state and driver vectors and presents the forward equations. Section 3 proves that the accounting identities are automatically preserved. Section 4 shows how to invert historical data to obtain "perfect" drivers and explains the time-series framing. Section 5 describes training and testing strategies on a panel of companies and the evaluation metrics. Section 6 discusses earnings forecasting. Section 7 outlines possible machine-learning extensions. Section 8 concludes.

2 Model Specification: State, Drivers, and Forward Equations

2.1 Model-world assets, liabilities, and equity

We work in a simplified "model world" in which we only track a small set of balance sheet items explicitly. All other assets and liabilities (prepaids, other receivables, deferred items, etc.) are subsumed into a residual external equity flow. Concretely, for each year t we define model assets, liabilities, and equity as

$$\text{ASSETS}_t^{\text{model}} = \text{CASH}_t + \text{AR}_t + \text{INV}_t + \text{PPE}_t, \tag{1}$$

$$\text{LIAB}_t^{\text{model}} = \text{AP}_t + \text{DEBT}_t, \tag{2}$$

$$\text{EQ}_t^{\text{model}} = \text{ASSETS}_t^{\text{model}} - \text{LIAB}_t^{\text{model}}. \tag{3}$$

Whenever we refer to equity EQ_t in what follows, we mean this model equity, either taken directly from preprocessed data or constructed using (1)–(3). Retained earnings RE_t are defined via a clean-surplus relation (see below).

2.2 State vector y_t

For each year t we collect 15 key items into a state vector y_t:

$$y_t = (S_t, C_t, SG_t, D_t,\ \text{AR}_t, \text{INV}_t, \text{AP}_t,\ \text{PPE}_t,\ \text{CASH}_t,\ \text{DEBT}_t,\ \text{EQ}_t,\ \text{RE}_t,\ \text{TAX}_t,\ \text{INT}_t,\ \text{DIV}_t), \tag{4}$$

where:

- Flows (income statement / cash-flow items) for year t:
    - S_t : sales (revenue);
    - C_t : cost of goods sold (COGS);
    - SG_t: operating expenses (selling, general and admin);
    - D_t : depreciation expense;
    - TAX_t: income tax expense;
    - INT_t: interest expense;
    - DIV_t: dividends paid (treated as positive cash outflow).
- Stocks (end-of-year balance sheet items):
    - AR_t : accounts receivable;

- INV_t: inventory;
- AP_t : accounts payable;
- PPE_t: net property, plant and equipment;
- CASH_t: cash and cash equivalents;
- DEBT_t: interest-bearing debt (short-term + long-term);
- EQ_t: equity in model-world sense ((1)–(3));
- RE_t: retained earnings (clean surplus).

All subsequent forward, inverse and evaluation steps operate on this 15 dimensional state representation, not on the raw reported totals.

2.3 Driver vector x_t

The driver vector x_t collects the policy parameters that drive the evolution of the state:

$$x_t = (gS_t, gm_t, sga_t, dep_t,\ dso_t, dio_t, dpo_t,\ capex_t,\ \tau_t, r_t, pay_t,\ ndebt_t,\ nequity_t). \tag{5}$$

The components have the following economic meaning:

1. *Operating structure*
   - gS_t : sales growth rate;
   - gm_t : gross margin;
   - sga_t: operating expense ratio (Opex / Sales);
   - dep_t: depreciation rate (on beginning-of-period PPE).
2. *Working-capital policies*
   - dso_t : days sales outstanding;
   - dio_t : days inventory outstanding;
   - dpo_t : days payables outstanding.
3. *Capex policy*
   - $capex_t$: capital expenditure / Sales.
4. *Tax, interest and payout*
   - τ_t : effective tax rate;
   - r_t : interest rate on beginning-of-period debt;
   - pay_t : payout ratio (dividends / net income).
5. *Financing decisions*
   - $ndebt_t$: net debt issuance / Sales; the change in debt is

$$\Delta\mathrm{DEBT}_t = ndebt_t\, S_t; \tag{6}$$

- $nequity_t$: net external equity inflow / Sales; the external equity flow (including all unmodelled balance-sheet items) is

$$\Delta \mathrm{EQ}_t^{\mathrm{ext}} = nequity_t \, S_t. \tag{7}$$

The key modelling choice, inspired by Vélez-Pareja Vélez-Pareja 2011; Vélez-Pareja 2010, is to fold all unmodelled assets and liabilities into $\Delta \mathrm{EQ}_t^{\mathrm{ext}}$. This ensures that the simplified model remains exactly closed with respect to the accounting identities: any residual is absorbed into the equity flow and, via the cash identity (below), into cash.

2.4 Forward evolution: from y_{t-1} and x_t to y_t

Given the previous state y_{t-1} and the driver vector x_t, the model evolves one year forward to y_t through deterministic equations.

2.4.1 Operating block: sales, costs, expenses, depreciation

Sales grow according to the sales growth driver:

$$S_t = S_{t-1}(1 + gS_t). \tag{8}$$

Costs, SG&A, and depreciation are driven by ratios:

$$C_t = (1 - gm_t)\, S_t, \tag{9}$$
$$SG_t = sga_t \, S_t, \tag{10}$$
$$D_t = dep_t \, \mathrm{PPE}_{t-1}. \tag{11}$$

2.4.2 Working capital: AR, inventory, AP

Working-capital stocks are tied directly to sales and costs via turnover days:

$$\mathrm{AR}_t = \frac{dso_t}{365} S_t, \tag{12}$$
$$\mathrm{INV}_t = \frac{dio_t}{365} C_t, \tag{13}$$
$$\mathrm{AP}_t = \frac{dpo_t}{365} C_t. \tag{14}$$

Define simplified working capital and its change:

$$\mathrm{WC}_t = \mathrm{AR}_t + \mathrm{INV}_t - \mathrm{AP}_t, \tag{15}$$
$$\Delta \mathrm{WC}_t = \mathrm{WC}_t - \mathrm{WC}_{t-1}. \tag{16}$$

2.4.3 Capex and PPE

Capital expenditure is driven by the capex-to-sales ratio:

$$\mathrm{CAPEX}_t = capex_t \, S_t. \tag{17}$$

PPE evolves as

$$\mathrm{PPE}_t = \mathrm{PPE}_{t-1} + \mathrm{CAPEX}_t - D_t. \tag{18}$$

2.4.4 Income statement: interest, tax, net income, dividends

Earnings before interest and taxes:

$$\mathrm{EBIT}_t = S_t - C_t - SG_t - D_t. \tag{19}$$

Interest and pre-tax income:

$$\mathrm{INT}_t = r_t\,\mathrm{DEBT}_{t-1}, \tag{20}$$

$$\mathrm{EBT}_t = \mathrm{EBIT}_t - \mathrm{INT}_t. \tag{21}$$

Taxes (with a floor at zero taxable income):

$$\mathrm{TAX}_t = \begin{cases} \tau_t\,\mathrm{EBT}_t, & \text{if } \mathrm{EBT}_t > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{22}$$

Net income and dividends:

$$\mathrm{NI}_t = \mathrm{EBT}_t - \mathrm{TAX}_t, \tag{23}$$

$$\mathrm{DIV}_t = pay_t\ \max(\mathrm{NI}_t, 0). \tag{24}$$

Note that NI_t is not explicitly part of the state vector (4); it is a derived flow computed from other state and driver variables.

2.4.5 Free cash flow

Free cash flow to the firm is defined as

$$\mathrm{FCF}_t = \mathrm{NI}_t + D_t - \Delta\mathrm{WC}_t - \mathrm{CAPEX}_t. \tag{25}$$

2.4.6 Financing and stock updates: debt, equity, cash, retained earnings

Debt. Net debt issuance is given by (6). Debt evolves as

$$\mathrm{DEBT}_t = \mathrm{DEBT}_{t-1} + \Delta\mathrm{DEBT}_t. \tag{26}$$

Retained earnings (clean surplus). We impose clean surplus on retained earnings:

$$\mathrm{RE}_t = \mathrm{RE}_{t-1} + \mathrm{NI}_t - \mathrm{DIV}_t. \tag{27}$$

External equity flows and total equity. The external equity flow $\Delta\mathrm{EQ}_t^{\mathrm{ext}}$ is defined in (7). Total equity evolves as

$$\mathrm{EQ}_t = \mathrm{EQ}_{t-1} + \Delta\mathrm{EQ}_t^{\mathrm{ext}} + \mathrm{NI}_t - \mathrm{DIV}_t. \tag{28}$$

By construction, $\Delta\mathrm{EQ}_t^{\mathrm{ext}}$ absorbs both genuine external equity transactions (issuance, buybacks) and the effect of all unmodelled balance sheet items.

Cash (core cash identity). Finally, cash is updated via the core cash identity:

$$\mathrm{CASH}_t = \mathrm{CASH}_{t-1} + \mathrm{FCF}_t + \Delta\mathrm{DEBT}_t + \Delta\mathrm{EQ}_t^{\mathrm{ext}} - \mathrm{DIV}_t. \tag{29}$$

Intuitively, free cash flow accumulates in cash unless it is absorbed by net debt changes, net external equity flows, or cash dividends.

2.4.7 Summary: the evolution function f

Equations (8)–(29) define a deterministic mapping

$$y_t = f(y_{t-1}, x_t), \tag{30}$$

which we will refer to as the *tank evolution* or *forward* function. In what follows we treat the choice of drivers x_t as the main source of uncertainty and model it with time-series techniques.

3 Accounting Identities and Their Preservation

3.1 Definitions

In the simplified model world, assets, liabilities and equity are defined by (1)–(3):

$$\begin{aligned} A_t &:= \text{ASSETS}_t^{\text{model}} = \text{CASH}_t + \text{AR}_t + \text{INV}_t + \text{PPE}_t, \\ L_t &:= \text{LIAB}_t^{\text{model}} = \text{AP}_t + \text{DEBT}_t, \\ E_t &:= \text{EQ}_t^{\text{model}} = A_t - L_t. \end{aligned}$$

The fundamental accounting identity is therefore

$$A_t - L_t - E_t = 0 \quad \text{for all } t. \tag{31}$$

We now show that if this identity holds at time $t-1$ and at the initial time 0, then the evolution equations in Section 2 guarantee that it holds for all future periods.

3.2 Proof of identity preservation

Step 1: Express the change in $A_t - L_t$. Consider the change in “net assets” $A_t - L_t$ between $t-1$ and t:

$$\begin{aligned} (A_t - L_t) - (A_{t-1} - L_{t-1}) = {}& (\text{CASH}_t - \text{CASH}_{t-1}) + (\text{AR}_t - \text{AR}_{t-1}) + (\text{INV}_t - \text{INV}_{t-1}) \\ & + (\text{PPE}_t - \text{PPE}_{t-1}) - (\text{AP}_t - \text{AP}_{t-1}) - (\text{DEBT}_t - \text{DEBT}_{t-1}). \end{aligned} \tag{32}$$

Using the definition of working capital (15) and its change (16), we can rewrite

$$(\text{AR}_t - \text{AR}_{t-1}) + (\text{INV}_t - \text{INV}_{t-1}) - (\text{AP}_t - \text{AP}_{t-1}) = \Delta\text{WC}_t. \tag{33}$$

Using the PPE and debt evolution equations (18) and (26), we have

$$\text{PPE}_t - \text{PPE}_{t-1} = \text{CAPEX}_t - D_t, \tag{34}$$

$$\text{DEBT}_t - \text{DEBT}_{t-1} = \Delta\text{DEBT}_t. \tag{35}$$

Substituting (33)–(35) into (32) yields

$$(A_t - L_t) - (A_{t-1} - L_{t-1}) = (\text{CASH}_t - \text{CASH}_{t-1}) + \Delta\text{WC}_t + \text{CAPEX}_t - D_t - \Delta\text{DEBT}_t. \tag{36}$$

Using the cash identity (29), we have

$$\mathrm{CASH}_t - \mathrm{CASH}_{t-1} = \mathrm{FCF}_t + \Delta\mathrm{DEBT}_t + \Delta\mathrm{EQ}_t^{\mathrm{ext}} - \mathrm{DIV}_t. \tag{37}$$

Substituting (37) into (36) gives

$$\begin{aligned}(A_t - L_t) - (A_{t-1} - L_{t-1}) &= \left(\mathrm{FCF}_t + \Delta\mathrm{DEBT}_t + \Delta\mathrm{EQ}_t^{\mathrm{ext}} - \mathrm{DIV}_t\right) \qquad (38)\\ &\quad + \Delta\mathrm{WC}_t + \mathrm{CAPEX}_t - D_t - \Delta\mathrm{DEBT}_t\\ &= \mathrm{FCF}_t + \Delta\mathrm{EQ}_t^{\mathrm{ext}} - \mathrm{DIV}_t + \Delta\mathrm{WC}_t + \mathrm{CAPEX}_t - D_t. \qquad (39)\end{aligned}$$

Using the definition of free cash flow (25) we have

$$\mathrm{FCF}_t = \mathrm{NI}_t + D_t - \Delta\mathrm{WC}_t - \mathrm{CAPEX}_t. \tag{40}$$

Substituting (40) into (39) yields

$$\begin{aligned}(A_t - L_t) - (A_{t-1} - L_{t-1}) &= (\mathrm{NI}_t + D_t - \Delta\mathrm{WC}_t - \mathrm{CAPEX}_t) + \Delta\mathrm{EQ}_t^{\mathrm{ext}} - \mathrm{DIV}_t\\ &\quad + \Delta\mathrm{WC}_t + \mathrm{CAPEX}_t - D_t\\ &= \mathrm{NI}_t + \Delta\mathrm{EQ}_t^{\mathrm{ext}} - \mathrm{DIV}_t. \qquad (41)\end{aligned}$$

Step 2: Compare with the change in equity. From the equity evolution equation (28) we have

$$E_t - E_{t-1} = \Delta\mathrm{EQ}_t^{\mathrm{ext}} + \mathrm{NI}_t - \mathrm{DIV}_t. \tag{42}$$

Comparing (41) and (42), we see that

$$(A_t - L_t) - (A_{t-1} - L_{t-1}) = (E_t - E_{t-1}). \tag{43}$$

Equivalently,

$$(A_t - L_t - E_t) - (A_{t-1} - L_{t-1} - E_{t-1}) = 0. \tag{44}$$

Thus the quantity $A_t - L_t - E_t$ is *invariant* over time. If the identity (31) holds at $t-1$, then it also holds at t.

Step 3: Induction over time. Assume that at the initial time $t = 0$ we have

$$A_0 - L_0 - E_0 = 0. \tag{45}$$

By (44), if the identity holds at $t = k-1$ then it holds at $t = k$. Therefore, by induction, (31) holds for all $t \geq 0$.

Conclusion. We have shown that the evolution equations (8)–(29) preserve the fundamental accounting identity $A_t = L_t + E_t$ for all t, provided it holds at the initial time and the initial state is consistent with the model-world balance sheet. In particular, any forecast produced by the model automatically respects assets = liabilities + equity, without any ad hoc plugs or rebalancing rules. This is directly in line with the "no plugs, no circularity" principle advocated by Vélez-Pareja Vélez-Pareja 2011.

4 Time-Series Interpretation and Perfect Drivers

4.1 From historical states to perfect drivers

In historical data we observe the realised states y_t^{data} for $t = 0, \dots, T$ and the associated income statement and cash-flow flows (including net income and dividends). Given a pair of consecutive states y_{t-1}^{data} and y_t^{data}, we can *invert* the forward equations and recover a unique driver vector x_t^* that exactly reproduces the transition $t - 1 \to t$.

Ignoring small numerical tolerances, the inverse ("perfect") drivers are:

$$
\begin{aligned}
gS_t^* &= \frac{S_t - S_{t-1}}{\max(S_{t-1}, \varepsilon)}, && (46)\\
gm_t^* &= 1 - \frac{C_t}{\max(S_t, \varepsilon)}, && (47)\\
sga_t^* &= \frac{SG_t}{\max(S_t, \varepsilon)}, && (48)\\
dep_t^* &= \frac{D_t}{\max(\mathrm{PPE}_{t-1}, \varepsilon)}, && (49)\\
dso_t^* &= 365\, \frac{\mathrm{AR}_t}{\max(S_t, \varepsilon)}, && (50)\\
dio_t^* &= 365\, \frac{\mathrm{INV}_t}{\max(C_t, \varepsilon)}, && (51)\\
dpo_t^* &= 365\, \frac{\mathrm{AP}_t}{\max(C_t, \varepsilon)}, && (52)\\
capex_t^* &= \frac{\mathrm{CAPEX}_t^{\text{data}}}{\max(S_t, \varepsilon)}, \quad \mathrm{CAPEX}_t^{\text{data}} = \mathrm{PPE}_t - \mathrm{PPE}_{t-1} + D_t, && (53)
\end{aligned}
$$

and, using the reconstructed EBIT, EBT and net income,

$$
\begin{aligned}
r_t^* &= \frac{\mathrm{INT}_t}{\max(\mathrm{DEBT}_{t-1}, \varepsilon)}, && (54)\\
\tau_t^* &= \begin{cases} \dfrac{\mathrm{TAX}_t}{\max(\mathrm{EBT}_t^{\text{data}}, \varepsilon)}, & \text{if } \mathrm{EBT}_t^{\text{data}} > 0, \\ 0, & \text{otherwise}, \end{cases} && (55)\\
pay_t^* &= \frac{\mathrm{DIV}_t}{\max(\mathrm{NI}_t^{\text{data}}, \varepsilon)}. && (56)
\end{aligned}
$$

The net financing drivers are recovered from the changes in debt and equity:

$$
\begin{aligned}
\Delta\mathrm{DEBT}_t^{\text{data}} &= \mathrm{DEBT}_t - \mathrm{DEBT}_{t-1}, && (57)\\
ndebt_t^* &= \frac{\Delta\mathrm{DEBT}_t^{\text{data}}}{\max(S_t, \varepsilon)}, && (58)
\end{aligned}
$$

and, using clean surplus on equity,

$$
\begin{aligned}
\Delta\mathrm{EQ}_t^{\text{ext,data}} &= \mathrm{EQ}_t - \mathrm{EQ}_{t-1} - (\mathrm{NI}_t^{\text{data}} - \mathrm{DIV}_t), && (59)\\
nequity_t^* &= \frac{\Delta\mathrm{EQ}_t^{\text{ext,data}}}{\max(S_t, \varepsilon)}. && (60)
\end{aligned}
$$

Here ε is a small positive constant (e.g. 10^{-6}) to avoid division by zero.

By construction, if we plug x_t^* into the forward map (30) with y_{t-1}^{data} as the initial state, we recover y_t^{data} up to numerical round-off. In this sense x_t^* is the "perfect" driver: it captures, within the model structure, the actual policies and conditions that transformed y_{t-1} into y_t.

4.2 Time-series view: modelling drivers instead of states

The crucial modelling choice is to treat the drivers x_t as a multivariate time series and to leave the evolution of y_t to the deterministic structural map f. Conceptually, we proceed in two steps:

1. **Data construction.** For each firm and each year t with a following year $t+1$, we construct y_t^{data} and y_{t+1}^{data}, and then invert to obtain the perfect driver x_{t+1}^* for that transition. This yields a panel of driver sequences $\{x_t^*\}$, typically of length three for each firm (four years of data).

2. **Time-series modelling.** We then fit simple time-series models to map past drivers to current drivers, e.g. x_t^* as a function of x_{t-1}^* and possibly longer lags. The forecasting models live in driver space; the state space evolution is handled by f.

In other words, we do *not* try to black-box fit y_{t+1} directly as a function of y_t. Instead we use the structural accounting model to tell us how y_{t+1} depends on x_t, and reduce the learning problem to predicting x_t.

Because each firm typically has only four years of data, each firm's own driver sequence has length three. This is too short for sophisticated per-firm time-series models such as ARIMA. Therefore, in the empirical part we focus on:

- per-firm sliding-window mean predictors;
- a pooled, component-wise AR(1) model using data from all firms;
- a small pooled neural network (multi-layer perceptron).

5 Training and Testing on a Panel of Firms

5.1 Sample of companies

To apply the model empirically, we select a panel of companies with at least four consecutive years of annual financial statements (income statement, balance sheet, cash-flow statement). For each firm we preprocess the raw statements into the model-world state representation y_t^{data} as in Section 2, ensuring that:

- equity EQ_t is computed as $\text{CASH}_t + \text{AR}_t + \text{INV}_t + \text{PPE}_t - \text{AP}_t - \text{DEBT}_t$;
- retained earnings obey the clean-surplus relation $\text{RE}_t = \text{RE}_{t-1} + \text{NI}_t - \text{DIV}_t$.

The concrete list of firms (e.g. large industrial and consumer companies) can be chosen to match data availability and project requirements.

5.2 Train–test structure of driver sequences

For each firm with four years of data $t = 0, 1, 2, 3$ we obtain three perfect driver vectors:

$$x_1^*, x_2^*, x_3^*,$$

corresponding to the transitions $0 \to 1$, $1 \to 2$ and $2 \to 3$ respectively. To create a simple train–test split per firm, we treat the last transition as test and the earlier ones as train:

- training transitions: $0 \to 1$ and $1 \to 2$;
- test transition: $2 \to 3$.

Pooling across firms yields a training set of driver transitions and a test set of driver transitions, each associated with the corresponding state y_t that serves as the starting point for forward simulation.

5.3 Driver forecasting models

We briefly describe the three simple forecasting models used for the drivers.

5.3.1 Per-firm sliding-window mean

For a given firm and year t, the sliding-window mean predictor with window length k sets the forecast driver vector to the average of the past k observed perfect drivers:

$$\hat{x}_t^{(\mathrm{SW})} = \frac{1}{k} \sum_{i=1}^{k} x_{t-i}^*. \tag{61}$$

If the firm has fewer than k past drivers, we average over all available ones. For the typical four-year case, we can use $k = 2$. This model is purely per-firm and does not pool information across firms.

5.3.2 Pooled AR(1) model

For each driver component j we fit a pooled autoregressive model of order one (AR(1)) across all firms:

$$x_t^{(j)} = a^{(j)} + \phi^{(j)} x_{t-1}^{(j)} + \varepsilon_t^{(j)}. \tag{62}$$

Here:

- $x_t^{(j)}$ is the j-th component of x_t^*,
- $(a^{(j)}, \phi^{(j)})$ are parameters estimated via ordinary least squares using all training pairs $(x_{t-1}^{(j)}, x_t^{(j)})$ such that the target time t belongs to the training set.

The OLS estimates can be expressed as

$$\phi^{(j)} = \frac{\mathrm{Cov}(x_{t-1}^{(j)}, x_t^{(j)})}{\mathrm{Var}(x_{t-1}^{(j)})}, \tag{63}$$

$$a^{(j)} = \mathbb{E}[x_t^{(j)}] - \phi^{(j)}\, \mathbb{E}[x_{t-1}^{(j)}]. \tag{64}$$

The AR(1) forecast is then

$$\hat{x}_t^{(j,\mathrm{AR1})} = a^{(j)} + \phi^{(j)} x_{t-1}^{(j)}. \tag{65}$$

Because we pool all firms, we obtain more stable parameter estimates than if we tried to fit AR(1) per firm.

5.3.3 Pooled multi-layer perceptron (MLP)

Finally, we consider a small neural network that maps the previous-period driver vector to the current-period driver vector:

$$\hat{x}_t^{(\mathrm{NN})} = g_\theta(x_{t-1}^*), \tag{66}$$

where g_θ is a feed-forward network with, for example, two hidden layers of size 32 and ReLU activations. The network is trained on all training samples to minimise the mean squared error (MSE) between $\hat{x}_t^{(\mathrm{NN})}$ and x_t^*:

$$\min_\theta \frac{1}{N} \sum_{i=1}^{N} \left\| g_\theta(x_{t_i-1}^*) - x_{t_i}^* \right\|_2^2. \tag{67}$$

In practice we standardise drivers component-wise (zero mean, unit variance) on the training set, train the network in standardised space, and then transform forecasts back to the original scale.

5.4 From driver forecasts to state forecasts

Given a forecast driver vector $\hat{x}_t$ (from any of the above models) and the current state y_t^{data}, we obtain a forecast for next year's state via the structural evolution map:

$$\hat{y}_{t+1} = f(y_t^{\mathrm{data}}, \hat{x}_t). \tag{68}$$

By Section 3, any such forecast automatically respects the accounting identities, in particular

$$\mathrm{CASH}_{t+1} + \mathrm{AR}_{t+1} + \mathrm{INV}_{t+1} + \mathrm{PPE}_{t+1} = \mathrm{AP}_{t+1} + \mathrm{DEBT}_{t+1} + \mathrm{EQ}_{t+1}. \tag{69}$$

5.5 Evaluation metrics

We evaluate models at two levels:

1. **Driver-space metrics**, comparing $\hat{x}_t$ to x_t^*;
2. **State-space metrics**, comparing $\hat{y}_{t+1}$ to y_{t+1}^{data}.

For both we use:

- Mean squared error (MSE):

$$\mathrm{MSE} = \frac{1}{Nd} \sum_{i=1}^{N} \sum_{j=1}^{d} \left(\hat{z}_{i,j} - z_{i,j}^{\mathrm{true}} \right)^2;$$

- Mean absolute error (MAE):

$$\mathrm{MAE} = \frac{1}{Nd} \sum_{i=1}^{N} \sum_{j=1}^{d} \left| \hat{z}_{i,j} - z_{i,j}^{\mathrm{true}} \right|;$$

- Relative L1 and L2 errors (to account for scale differences across firms and items), for example,

$$\mathrm{RelL1} = \frac{1}{Nd} \sum_{i=1}^{N} \sum_{j=1}^{d} \frac{\left| \hat{z}_{i,j} - z_{i,j}^{\mathrm{true}} \right|}{\max\left(|z_{i,j}^{\mathrm{true}}|, \varepsilon \right)}.$$

Here z denotes either drivers or states, d is the dimensionality (13 for drivers, 15 for states), N is the number of samples, and ε is a small constant for numerical stability.

5.6 Testing plan and sanity checks

Perfect-driver baseline. As a sanity check on both the forward and inverse implementations, we use the perfect drivers x_t^* as inputs to the forward map and verify that

$$\hat{y}_{t+1}^{(\text{perfect})} = f(y_t^{\text{data}}, x_t^*)$$

reconstructs y_{t+1}^{data} to machine precision. In practice, the MSE and MAE for this baseline should be on the order of floating-point round-off. This test verifies:

- correctness of the evolution equations;
- correctness of the inversion formulas for x_t^*;
- that the data preprocessing step yields internally consistent states.

Forecasting models. For each of the three forecasting models (SW, AR(1), MLP) we:

1. Train the model on the training driver transitions;
2. Generate driver forecasts $\hat{x}_t$ for each test transition;
3. Propagate these through the forward map to obtain state forecasts $\hat{y}_{t+1}$;
4. Compute the driver- and state-space metrics described above;
5. Inspect which state components are hardest to forecast (e.g. cash and debt may show larger errors).

Because all models use the same structural evolution, any improvement in state-space metrics can be attributed directly to better driver forecasts.

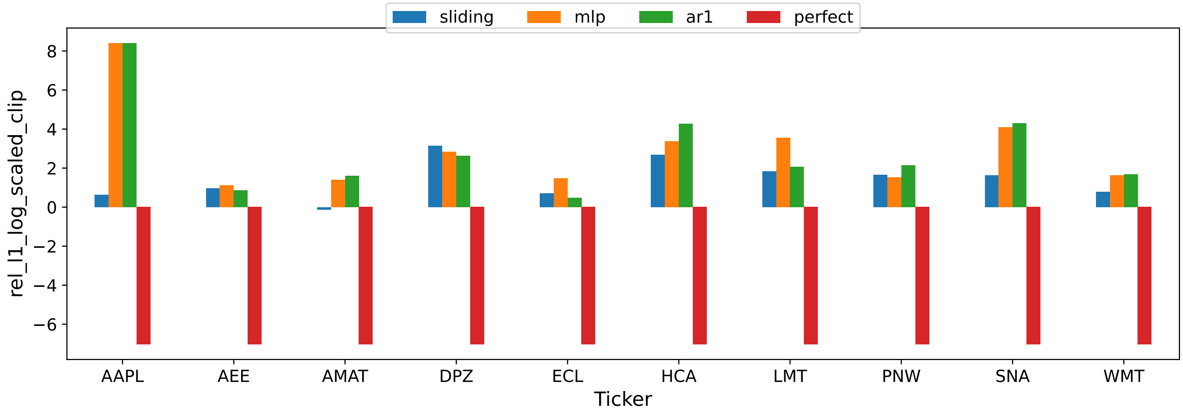
Accounting consistency of forecasts. Because the accounting identities are preserved by construction (Section 3), we do not need to impose additional constraints at forecast time. Nonetheless, as a diagnostic, we can compute the residual

$$\Delta_t^{\text{identity}} = (\text{CASH}_t + \text{AR}_t + \text{INV}_t + \text{PPE}_t) - (\text{AP}_t + \text{DEBT}_t + \text{EQ}_t)$$

on both historical and forecasted states to confirm that it is numerically zero up to round-off.

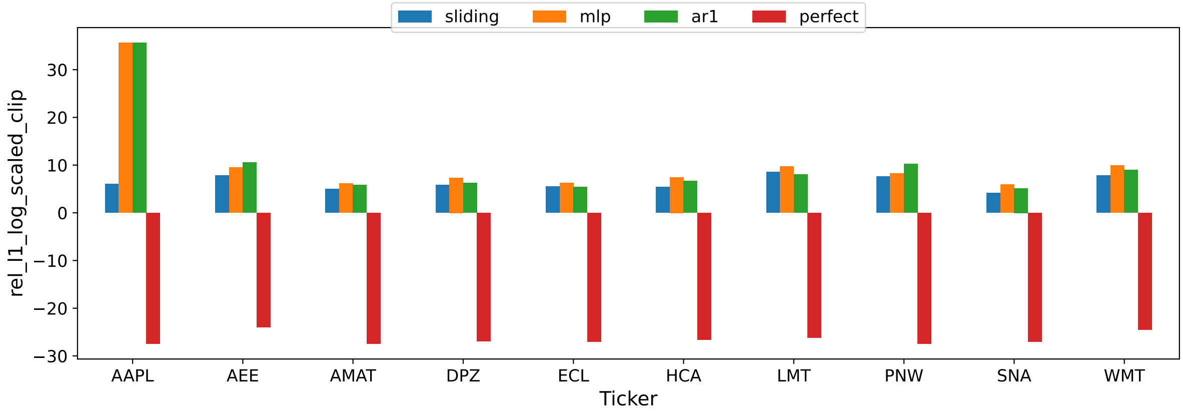
Testing result demonstration. We randomly select some of the companies to fit through 4 models and demonstrate both the drivers and states recover situations about Relative L1.

Across all four evaluation settings (Quarter/Year × Driver/State), the recovered error patterns exhibit several consistent and interpretable characteristics. First, the *sliding-window* estimator uniformly achieves the lowest log-scaled reconstruction error, outperforming both MLP- and AR(1)-based regressors in every figure (e.g., see Fig. 4 and Fig. 1). This dominance suggests that the local-stationarity assumption embedded in the sliding scheme aligns more closely with the statistical structure of accounting drivers. Second, the MLP model does not provide meaningful gains over a simple AR(1) baseline; their bars are nearly indistinguishable across all tickers, indicating that the driver dynamics are largely linear and offer limited exploitable nonlinearity



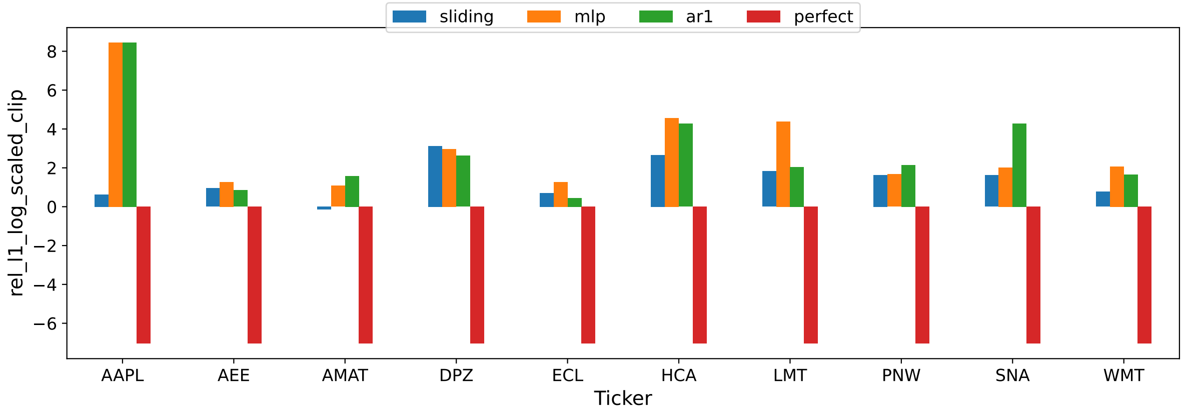
Quarter / Driver

Figure 1: Relative L1 for driver recovery for quarter balance sheet



Quarter / State

Figure 2: Relative L1 for state recovery for quarter balance sheet



Year / Driver

Figure 3: Relative L1 for driver recovery for year balance sheet

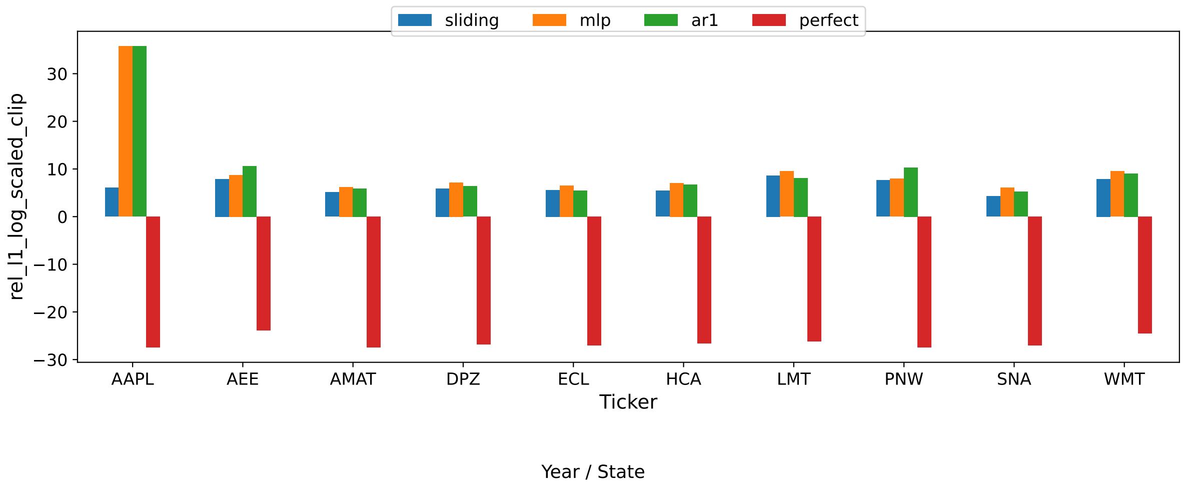


Figure 4: Relative L1 for state recovery for year balance sheet

(Fig. 3). Third, reconstruction of *state* variables is systematically more difficult than that of *drivers*, as reflected by larger log-errors in both annual and quarterly settings (cf. Fig. 2). This asymmetry is consistent with the fact that states are generated through multi-equation accounting evolution, which amplifies small driver-level perturbations. Finally, the qualitative ranking of tickers is highly stable across all four panels; for example, AAPL is consistently the most challenging entity to recover, while AMAT and ECL remain among the easiest. The strong cross-figure agreement implies that recovery difficulty is primarily driven by firm-specific structural characteristics rather than by temporal frequency or model class.

6 Earnings Forecasting

Net income is already an intermediate output of the model: it is computed in (23) from sales, costs, expenses, depreciation, interest and tax. Therefore, once we have a forecast for next year's state $\hat{y}_{t+1}$ and drivers $\hat{x}_{t+1}$, we automatically obtain a forecast for next year's earnings:

$$\widehat{\mathrm{NI}}_{t+1} = \widehat{\mathrm{EBT}}_{t+1} - \widehat{\mathrm{TAX}}_{t+1}, \tag{70}$$

where

$$\widehat{\mathrm{EBIT}}_{t+1} = \hat{S}_{t+1} - \hat{C}_{t+1} - \widehat{SG}_{t+1} - \hat{D}_{t+1}, \tag{71}$$

$$\widehat{\mathrm{INT}}_{t+1} = \hat{r}_{t+1}\, \widehat{\mathrm{DEBT}}_{t}, \tag{72}$$

$$\widehat{\mathrm{EBT}}_{t+1} = \widehat{\mathrm{EBIT}}_{t+1} - \widehat{\mathrm{INT}}_{t+1}, \tag{73}$$

$$\widehat{\mathrm{TAX}}_{t+1} = \hat{\tau}_{t+1}\, \max(\widehat{\mathrm{EBT}}_{t+1}, 0). \tag{74}$$

Thus the model can be used directly as an earnings forecaster, and earnings forecasts are consistent with the balance sheet and cash-flow forecasts.

If earnings per share (EPS) is required, we can introduce additional structure for the share count (e.g. assuming a fixed number of shares or linking $\Delta\mathrm{EQ}_t^{\mathrm{ext}}$ to net share issuance). This is an extension on top of the core model.

7 Machine-Learning Extensions

The current implementation uses very simple driver forecasting models: per-firm sliding-window means, pooled AR(1), and a small MLP with one lag. In principle, a richer set of machine-learning techniques could be used to improve forecasts, especially if longer time series per firm are available.

We briefly outline several directions.

7.1 Richer sequence models for drivers

If we had longer driver histories (e.g. 10–20 years per firm), we could explore:

- **Vector autoregressions (VAR)** to capture cross-driver dependencies while remaining linear and interpretable;
- **Recurrent neural networks (RNNs)**, including LSTM and GRU architectures, to model non-linear temporal dynamics;
- **Transformer-based sequence models**, which can capture long-range dependencies and complex interactions between drivers over time.

These models would operate on the driver sequences x_t^* and possibly their past lags, with the structural map f still ensuring accounting consistency.

7.2 Cross-sectional features and global models

Because individual firms have short histories, it is natural to pool data across many firms and use *global* models that share parameters. To capture systematic differences across firms, we can augment the model with firm-level features, such as:

- industry or sector dummies;
- firm size (e.g. log assets or log sales);
- geographic footprint;
- profitability and leverage indicators.

These features can be concatenated to driver inputs (e.g. as additional inputs to an MLP or RNN) so that the model can learn different dynamics for different types of firms while still benefitting from pooling.

7.3 Exogenous covariates and causal structure

Beyond pure time-series models, we can incorporate exogenous covariates that may influence drivers, such as:

- macroeconomic variables (GDP growth, interest rates, inflation);
- commodity prices (for relevant industries);
- firm-level policy indicators (investment plans, leverage targets).

These covariates can be fed into the driver forecasting model alongside past drivers. Furthermore, the structural nature of the tank model makes it a natural backbone for causal analyses: if we believe certain drivers are under management control (e.g. payout ratio, capex policy) while others are more exogenous (e.g. demand growth), we can simulate counterfactual policies and their impact on the balance sheet and earnings.

7.4 Regularisation, uncertainty, and robustness

From a machine-learning perspective, we can improve robustness and interpretability by:

- applying regularisation (e.g. L1/L2 penalties, dropout) to prevent overfitting in neural models;
- modelling parameter and prediction uncertainty (e.g. Bayesian regression for AR(1), Bayesian neural networks, or ensemble methods);
- stress-testing forecasts by perturbing drivers and examining the resulting distribution of state forecasts.

Because the structural map f is deterministic and transparent, uncertainty in drivers can be propagated to uncertainty in balance sheet and earnings forecasts via straightforward simulation.

8 Conclusion

We have constructed a simple yet internally consistent financial statement forecasting model based on the tank-model ideas of Vélez-Pareja Vélez-Pareja 2011; Vélez-Pareja 2010. The model describes the evolution of a 15-dimensional state vector of income-statement and balance-sheet items as a deterministic function of the previous state and a 13-dimensional driver vector of growth, margins, working-capital policies, capex, tax, interest, payout and net financing ratios.

The forward equations ensure that the model-world balance-sheet identity assets = liabilities + equity holds automatically at all times, provided it holds initially. Historical data can be inverted to recover “perfect” drivers that exactly reproduce the observed transitions. This allows us to recast the forecasting problem as one of multivariate time-series modelling of drivers, with the structural evolution map handling accounting consistency.

On a panel of firms with short histories, we can already implement and evaluate simple forecasting models for drivers (sliding-window mean, pooled AR(1) and a small MLP), and use the structural model to obtain consistent forecasts for all financial statement items and earnings. As more data become available, richer machine-learning models and additional covariates can be incorporated without sacrificing the accounting integrity of the forecasts.

References

Mejía-Peláez, Felipe and Ignacio Vélez-Pareja (2011). “Analytical solution to the circularity problem in the discounted cash flow valuation framework”. In: *Innovar* 21.42, pp. 55–68.

Vélez-Pareja, Ignacio (2010). “Constructing Consistent Financial Planning Models for Valuation”. In: *IIMS Journal of Management of Science* 1.

– (2011). “Forecasting Financial Statements with No Plugs and No Circularity”. In: *The IUP Journal of Accounting Research & Audit Practices* 10.1.