Dynamic Costs of Resource Shocks: Evidence from Cape Town's Water Crisis

Preliminary: please do not recirculate

Emiliano Rinaldi* Pol Simpson*

March 25, 2026

Abstract

When cities face water scarcity, policymakers must decide how to allocate the burden between households and firms. We study Cape Town's 2016-2018 water crisis, during which commercial water prices quadrupled. Using administrative tax records covering all formal employment in South Africa, we document that water-intensive sectors in Cape Town experienced persistent employment and establishment declines driven by firm exit, while comparable sectors in control metros did not diverge. Within Cape Town, locations with higher pre-crisis water intensity saw larger declines that persist long after prices normalised. We develop a dynamic spatial model where temporary shocks can permanently reduce local economic activity by pushing the economy past a tipping point, and use it to evaluate how cities should allocate scarce water between households and firms during crisis episodes.

*Department of Economics, London School of Economics. Email: e.rinaldi@lse.ac.uk

1 Introduction

In early 2018, Cape Town became the first major city to announce that it might run out of water, with Day Zero projected for April. The city responded with aggressive pricing for both households and firms, consumption fell by half, and Day Zero never came; by mid-2018 the rains returned and the immediate crisis passed. But five years later, long after prices had returned to pre-crisis levels, employment in Cape Town's water-intensive sectors remained well below what pre-crisis trends would predict. A temporary shock left a permanent mark.

This paper asks why, and what it means for how cities should allocate scarce resources between households and firms. Cape Town is far from the only city to face this problem: climate change is increasing drought frequency worldwide, and in the last decade alone, São Paulo, Chennai, Melbourne, Montevideo, and London have all imposed emergency water restrictions. When a city faces a temporary shortage, a fixed quantity of water must be split between households and firms, and the textbook answer is to set price equal to social marginal cost. In practice, however, both sides of this market generate externalities, and the two externalities have very different properties.

On the household side, cutting water below a sanitation threshold creates public health costs that households do not fully internalise, and because this externality is well understood and politically salient, it tends to push policy toward protecting households. On the firm side, firms generate productivity spillovers for other firms through knowledge exchange, thick labour markets, and shared suppliers, so when a firm exits because water costs became unmanageable, it reduces productivity for the firms that remain nearby. Unlike the health externality, this cost is potentially delayed and cumulative, and it pushes policy in the opposite direction, toward protecting firms.

Across these episodes, policymakers have faced this tradeoff with limited information and political incentives that favour households. São Paulo imposed outright quantity restrictions on industrial users. Bogota used rolling supply shutoffs. Barcelona raised commercial tariffs and mandated recycling requirements that amounted to quantity constraints. In each case, cities burdened firms with little understanding of the dynamic consequences.

Cape Town is unusual in that it responded with differential prices rather than quantity restrictions. Households faced steeply increasing block tariffs while firms faced a flat price increase. This makes Cape Town the closest thing we have to a prices-versus-quantities experiment in a real city, and it means we can trace the firm-side response cleanly: from prices to costs to exit to employment, and critically, to what happened after the crisis ended.

Higher water prices raise costs for water-intensive firms. Some cannot absorb the increase and exit. Most urban equilibrium models would predict recovery when prices normalise. That prediction can fail, however, when local productivity depends on the history of economic activity in a place, because firms benefit from proximity to other firms through knowledge spillovers, thick labour markets, and shared suppliers, and these agglomeration economies do not automatically reconstitute themselves once they have been disrupted. To understand what happens in the wake of large negative shocks, we need to decompose agglomeration more carefully than is standard in the literature.

In the spirit of Allen and Donaldson (2020), we split agglomeration into two components. The first depends on who is in a location right now: if 100 manufacturing firms operate in Epping, a Cape Town industrial area, they generate knowledge spillovers, a thick pool of specialised workers, and demand for specialised suppliers, and if 50 exit tomorrow these benefits fall immediately, though they can rebuild if 50 new firms enter next year. We call this contemporaneous agglomeration.

The second component depends on what past activity built up. When those 100 firms were operating in Epping, they were not just generating today's spillovers. Over years, they built factory buildings, installed equipment, connected to water and power infrastructure, trained a local workforce, and established supplier networks. These assets persist after a firm exits, at least for a while, and they represent a legacy of past economic activity that benefits current and future firms. We call this historical agglomeration.

These two components create different dynamics after a shock. Contemporaneous agglomeration can rebuild when firms return, but historical agglomeration cannot, or at least not quickly: when a firm exits during the water crisis, it does not just reduce today's knowledge spillovers but also abandons a factory that starts to deteriorate, while its trained workers drift to other sectors and its supplier relationships dissolve. The accumulated capital that took years to build begins to erode.

If enough firms exit, this erosion can push the local economy past a tipping point below which the remaining agglomeration is too weak to make re-entry profitable: potential entrants see a thin market with depleted infrastructure and diminished prospects, workers who left face switching costs and see few jobs to return to, and the economy settles into a lower equilibrium, not because the old one is infeasible, but because the accumulated capital that sustained it has been destroyed.

In the model, the two components play distinct roles. Contemporaneous agglomeration governs whether the economy's dynamic path is unique, in the sense that at any given moment there is exactly one consistent set of wages, employment, and prices, while historical agglomeration governs whether there are multiple stable configurations the economy could settle into in the long run. When contemporaneous agglomeration is modest relative

to housing congestion the dynamic path is always unique, but when total agglomeration, combining both the contemporaneous and the historical, is strong enough, multiple long-run configurations exist and a temporary shock can permanently shift the economy from one to another. The result is path dependence without indeterminacy: the economy's trajectory is always pinned down, but which long-run outcome it reaches depends on where it has been.

We make three contributions. First, we document that a temporary resource shock caused persistent employment losses operating through firm exit. Using administrative tax records from the South African Revenue Service, we compare water-intensive to non-water-intensive sectors within Cape Town, and Cape Town to Johannesburg, Pretoria, and Durban. Before 2016, water-intensive and other sectors in Cape Town grew in parallel, but after the drought they diverged: by 2024, water-intensive employment had reached roughly 107 percent of its 2016 level while other sectors had reached 120 percent, opening a gap of about 13 percentage points that has not closed. This gap reflects firm exit, since establishment counts in water-intensive sectors fell during the crisis and have remained depressed, while wages showed no differential movement across sectors. The pattern does not appear in control metros, where water-intensive and non-water-intensive sectors evolved similarly throughout the sample period.

Second, we develop a quantitative spatial model that distinguishes contemporaneous from historical agglomeration, generating path dependence while maintaining a unique equilibrium path. The model nests both outcomes: whether the economy recovers or not depends on estimated parameters. Third, we develop a framework to evaluate whether alternative pricing policies that shift more of the conservation burden toward households lead to a reduction in welfare losses by keeping the economy above the tipping point.

Our paper connects several strands of the literature. On the empirical side, it contributes to work on the persistent economic effects of climate and resource shocks. Dell et al. (2012) show that higher temperatures reduce output growth in developing countries, and Hsiang and Jina (2014) find that national income declines from tropical cyclones persist for at least 20 years. Bilal and Känzig (2024) find similarly long-lived macroeconomic effects of temperature shocks. Cape Town offers a setting where the shock was both temporary and geographically isolated, making it possible to study the microfoundations of persistence. Two papers study resource crises closest to ours: Costa and Gerard (2021) analyse Brazil's 2001 electricity rationing and find persistent consumption reductions even after rationing ended, while Castro-Vincenzi (2024) studies how firms reorganise production geography in response to climate risk. We differ from both in studying a setting where the government mediated scarcity through differential prices, and in documenting firm-side persistence through exit and agglomeration erosion.

On the spatial economics side, Hornbeck (2012) finds that the Dust Bowl caused population declines lasting six decades in heavily eroded counties, consistent with an agglomeration mechanism even in an agricultural context. Our model builds directly on the dynamic spatial framework of Caliendo et al. (2019), adding heterogeneous firms with entry and exit and the Allen-Donaldson decomposition of agglomeration into contemporaneous and historical components. The key difference from Caliendo, Dvorkin, and Parro is that their model features a unique steady state, so temporary shocks always produce temporary effects; ours features multiple steady states, so temporary shocks can become permanent. On the empirical side of agglomeration, Greenstone et al. (2010) estimate significant productivity spillovers from large plant openings, and Kline and Moretti (2014) find that the TVA's initial investments had persistent effects on local manufacturing decades later. We study the reverse: whether destroying enough of the accumulated base causes agglomeration to unravel. For firm dynamics, we use the Melitz (2003) framework with Pareto-distributed productivity, which gives tractable aggregation and connects sunk entry costs to the accumulated capital underlying historical agglomeration.

On resource allocation, Weitzman (1974) compares prices and quantities as allocation instruments under uncertainty, but the framework is static and does not ask how the choice of instrument affects dynamic adjustment through firm exit. The applied water pricing literature, including Olmstead and Stavins (2009), Grafton et al. (2011), Ito (2014), and Jessoe and Rapson (2014), focuses on the household conservation response. We study the firm side, where the main adjustment margin is exit rather than conservation, and where the consequences extend well beyond the period of scarcity.

The rest of the paper proceeds as follows. Section 2 describes the institutional setting and the timeline of the crisis. Section 3 describes the data sources and the construction of the water intensity measure. Section 4 presents the empirical evidence on persistent divergence in employment and establishments, including event studies, triple-difference specifications, and spatial variation within Cape Town. Section 5 develops the model. Section 6 discusses identification, calibration, and the welfare framework. Section 7 concludes.

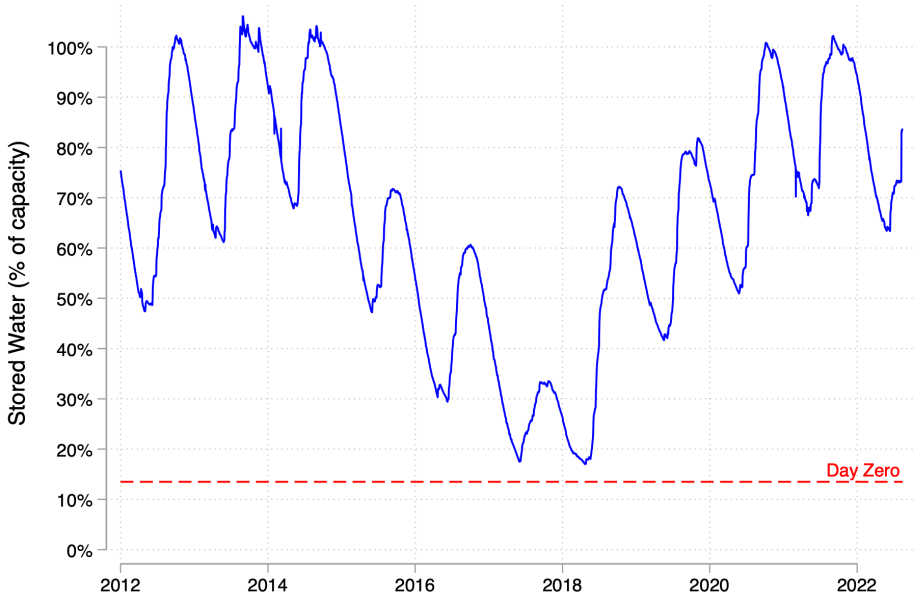
2 Institutional Background

Cape Town, with a population of five million, draws water almost entirely from surface reservoirs fed by winter rainfall. The city has limited groundwater and no major rivers nearby. A single vertically integrated municipal utility holds monopoly water rights and operates the distribution network, and water supply is effectively fixed in the short run: the city cannot increase reservoir inflows during a dry year.

2.1 Timeline of the crisis

Rainfall began declining in late 2015. Through 2016, reservoir levels fell steadily, but the city had not yet responded with emergency measures. We use 2016 as our base year: the drought had started but the policy environment had not yet changed.

Figure 1: Reservoir Storage Levels



Notes: Aggregate dam storage levels for the Cape Town metropolitan area. The horizontal line indicates the Day Zero threshold at 13.5 percent of capacity.

By early 2017, reservoir levels had dropped below 30 percent (Figure 1) and the city began raising water prices aggressively. 2017 was the driest year in Cape Town's recorded history. By January 2018, levels had fallen below 20 percent of capacity. At 13.5 percent the city would have been forced to shut off municipal supply entirely, with residents collecting water from 200 distribution points citywide. The city announced that Day Zero was weeks away.

By mid-2018, good winter rains began refilling the dams and levels recovered through the second half of the year. By 2019, emergency pricing had been largely phased out and water availability had normalised. The crisis lasted roughly two years in terms of the policy shock to firms: from early 2017, when commercial prices began rising, to mid-2018, when the rains returned.

The meteorological event behind the crisis was both severe and unusual. Three consecutive dry winters coincided with Hadley Cell expansion, a dip in rainfall intensity, and anomalous El Niño conditions. The estimated return period for a drought of this severity

is 150 to 300 years. Since the crisis, the city has invested in desalination capacity and groundwater extraction, substantially reducing its dependence on surface rainfall and its exposure to similar events in the future.

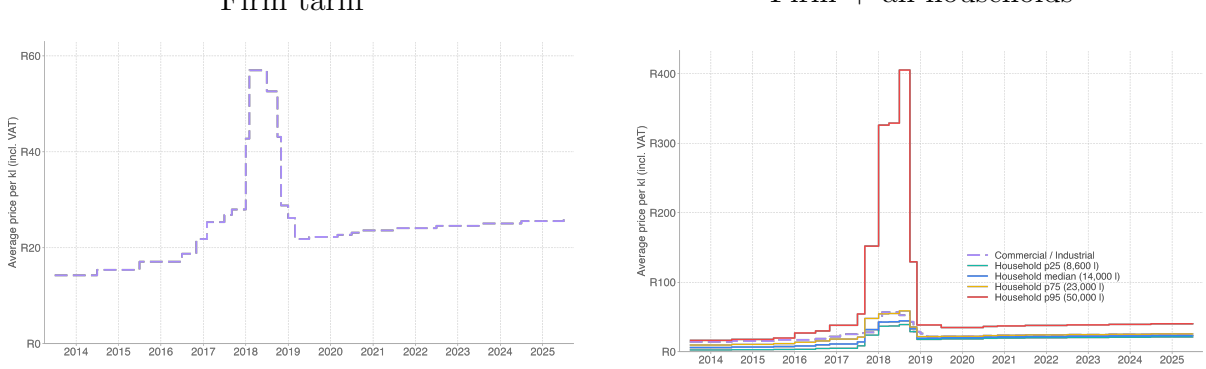
2.2 Pricing policy

Unlike most other cities that have faced recent water shortages, Cape Town's policy response was primarily price-based.

Before the drought, Cape Town used an increasing block tariff (IBT) structure for households, a standard design in South African municipalities intended to cross-subsidise low-income users. Households received six kilolitres per month free, with increasing marginal prices at higher consumption levels. Commercial users paid a flat volumetric rate. Tariffs were updated annually on July 1, and until 2015, South African regulations required utilities to recover all costs through volumetric charges.

As the crisis intensified, the city escalated both tariff schedules. On the household side, the block structure was retained but made dramatically steeper. The first six kilolitres remained free until July 2017, when the free block was removed and prices introduced for all consumption. The top household block reached over 400 rand per kilolitre at the peak of the crisis, roughly twenty times the baseline rate. The city described this as a punitive tariff designed to make discretionary use (gardens, swimming pools, long showers) prohibitively expensive while keeping basic needs affordable. Figure 2 shows the evolution of prices. The two panels illustrate, on the left, the commercial tariff alone, and on the right, the schedule of tariffs for households along the consumption distribution.

Figure 2: Water Price Schedules During the Crisis

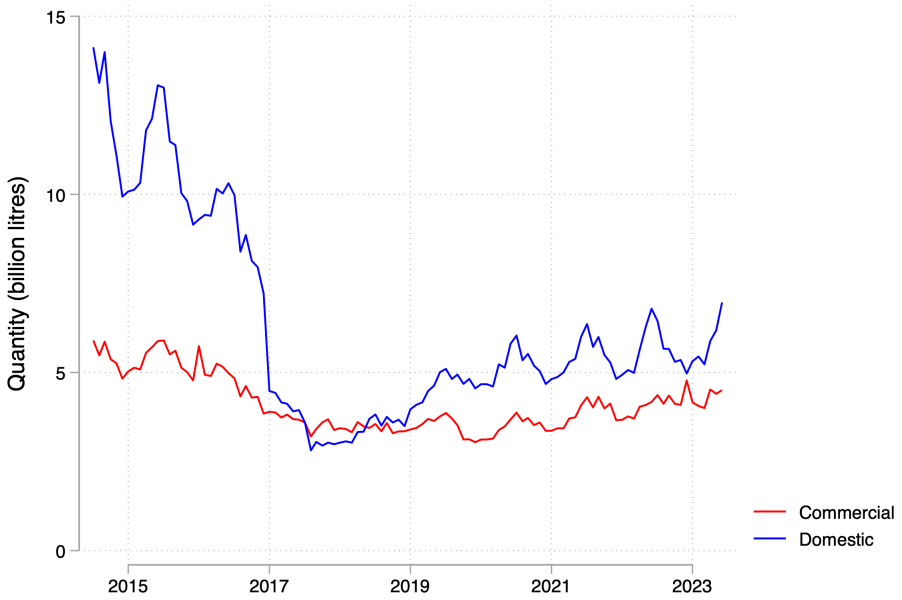


Commercial and industrial users faced a different instrument: a flat price increase that applied regardless of baseline consumption, sector, or ability to substitute away from water. The commercial rate roughly quadrupled during the drought, rising from approximately R14 to R57 per kilolitre at the peak.

This asymmetry in tariff design is what makes the empirical strategy possible. Households faced a convex price schedule that concentrated the burden on high-volume users while shielding basic use, whereas firms faced a linear price increase that was uniform across sectors and usage levels.

These price increases achieved substantial conservation, as Figure 3 shows. Domestic water use fell from approximately 14 billion litres per year to roughly 3 billion at the trough, a decline of nearly 80 percent, while commercial water use fell from about 6 billion to 3 billion litres. Consumption in both categories remained well below pre-crisis levels even after prices normalised, a pattern that echoes the household-side persistence Costa and Gerard (2021) document in the context of Brazil's 2001 electricity crisis.

Figure 3: Water Consumption by Account Type

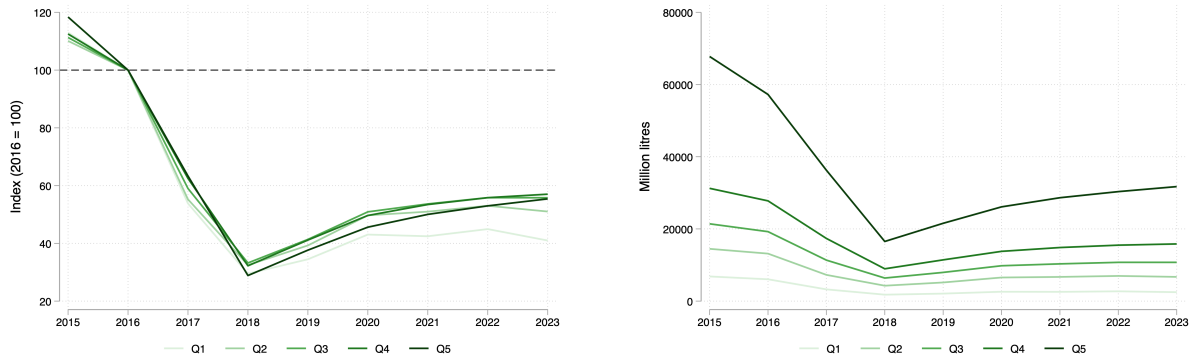


Notes: Monthly water consumption in billion litres, split by domestic and commercial accounts.

But the aggregate numbers obscure how households and firms achieved these reductions. Figure 4 splits household accounts into quintiles by pre-crisis consumption. All five quintiles cut consumption by roughly the same proportion during the crisis, falling to around 30–40 percent of their 2016 levels by 2018, with large and small consumers reducing use by similar percentages. Consumption partially recovered after the crisis but remained well below pre-crisis levels for all quintiles through 2023, with the top quintile reaching roughly 65 percent of its 2016 level.

Figure 4: Household Water Consumption by Pre-Crisis Quintile

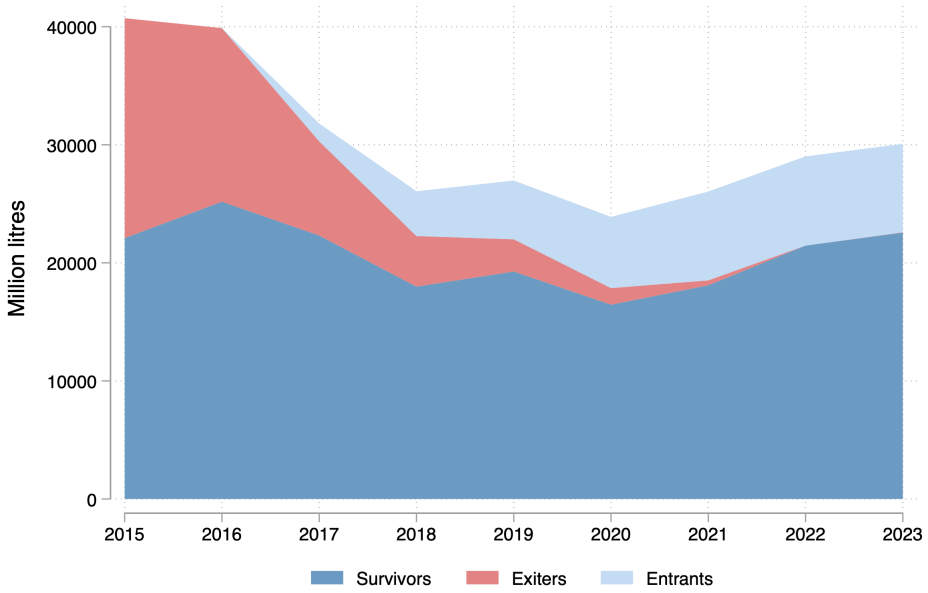
Indexed (2016 = 100) Levels (million litres)



Notes: Household water accounts split into quintiles by pre-crisis consumption. Left panel: consumption indexed to 2016. Right panel: consumption in levels.

The commercial side is different. Figure 5 decomposes total commercial water consumption into firms that survived through the full sample period, firms that exited, and firms that entered after 2015. Almost all of the decline in aggregate commercial water use comes from exiters disappearing rather than from surviving firms cutting back. Surviving firms reduced consumption modestly and largely recovered afterward, while exiters, who accounted for roughly half of commercial water use in 2015, simply vanish from the data, and entrants do not fully replace them.

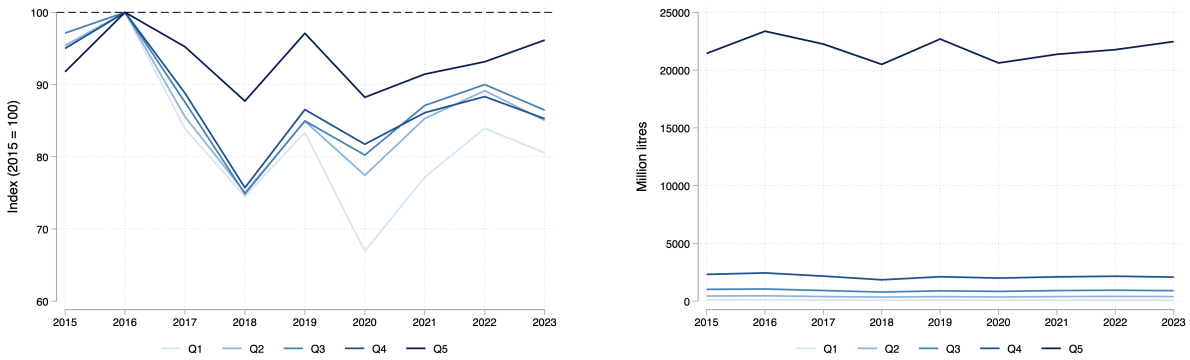
Figure 5: Commercial Water Use: Survivors, Exiters, and Entrants



Notes: Stacked area chart decomposing total commercial water consumption into firms that survive the full sample period, firms that exit, and firms that enter after 2015.

Figure 6 splits surviving commercial accounts into quintiles by pre-crisis consumption. Surviving firms in the upper quintiles reduced their water use only marginally during the crisis, with Q4 and Q5 reaching about 85–90 percent of their 2015 levels at the trough before recovering. Where household quintiles fell to 30–40 percent of baseline, surviving firm quintiles hardly moved. The conservation that shows up in aggregate commercial water use comes from firms disappearing, not from surviving firms using less water.

Figure 6: Surviving Commercial Accounts: Water Consumption by Pre-Crisis Quintile



Notes: Commercial water accounts that survive the full sample period, split into quintiles by pre-crisis consumption. Left panel: consumption indexed to 2015. Right panel: consumption in levels.

In short, households conserved by reducing consumption; firms "conserved" by ceasing to exist. The aggregate consumption figures treat these margins as equivalent, but the persistence of the commercial decline reflects the irreversibility of exit rather than a lasting change in behaviour among surviving firms.

3 Data

3.1 Administrative employment records

We use administrative tax records from the South African Revenue Service (SARS). The IRP5 certificate, filed by employers for each employee, covers the universe of formal employment. We observe earnings, employer identity, industry classification (at the 5-digit SIC level), and location at the postcode level. We distinguish firms (by tax reference number) from establishments (individual worksites), which lets us track entry, exit, and employment by postcode and 5-digit industry.

We construct an annual panel for 2014–2024 covering Cape Town and three control metros: Johannesburg, Pretoria, and Durban. These cities experienced no water scarcity

during our sample period. Our main outcomes are employment (full-time equivalent), establishment counts, firm counts, median wages, and new hires.

Additionally, we have hexagon-level aggregates for the entire country. These are constructed by mapping postcodes to H3 hexagons at resolution 7, which gives over 100 hexagons covering the South African metropolitan areas. Each hexagon has an aread of approximately $5km^2$. We use these for the spatial event study within Cape Town, and to estimate water intensity at the hexagon level.

3.2 Water consumption data

We have access to the universe of water meter readings in Cape Town, geolocated and classified by account type (commercial versus domestic), with quantities and prices. We can therefore observe consumption at the meter level and distinguish household from firm water use.

One limitation is that we cannot match individual firms in the SARS data to individual water meters. The SARS data identify firms by tax reference number and postcode; the water data identify meters by physical address and account number. Confidentiality restrictions prevent us from linking the two at the firm level.

3.3 Measuring water intensity

We classify sectors as water-intensive in two ways. For the spatial specifications, we exploit the meter-level resolution of the water data. We estimate water intensity at the hexagon level by regressing commercial water consumption per establishment on the industry composition of the hexagon, using pre-crisis data. This gives a continuous measure of water intensity for each hexagon based on the industries located there.

For the industry-level specifications, we use a regression approach to classify 2-digit SIC sectors as water-intensive or not. We regress commercial water consumption per hexagon on the number of employees in each 2-digit sector, controlling for hexagon fixed effects. The coefficients on the sector employment variables give us an estimate of the water intensity of each sector. We then take the top quintile of sectors by water intensity as water-intensive and the rest as non-water-intensive.

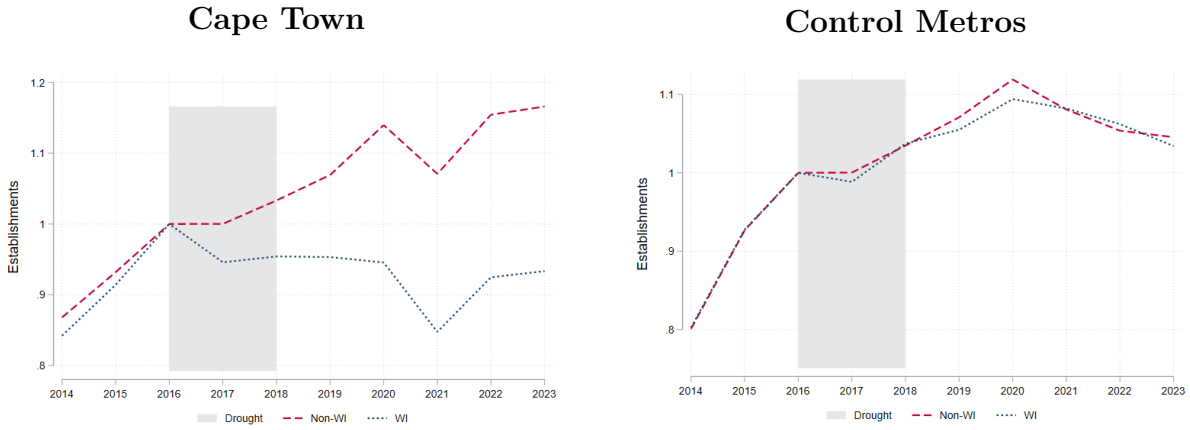
Because we cannot match firms to meters directly, the water intensity measure is estimated from spatial correlation rather than observed at the firm level. We validate this measure by verifying that the sectors classified as water-intensive correspond to those expected on economic grounds (food processing, hospitality, laundries) and by checking that the binary classification from the spatial approach aligns with the classification from the national water accounts.

4 Empirical Evidence

4.1 Descriptive Patterns

Figure 7 plots establishment counts by water intensity in Cape Town alongside the same comparison for the three control metros, both indexed to 2016. In Cape Town, water-intensive establishment counts began falling during the drought, and never recovered. Non-water-intensive establishments, by contrast, grew throughout. In the control metros, both categories of establishments evolved in near-lockstep for the entire sample period. The widening gap in Cape Town's establishment counts, and its absence in the control metros, is a strong indication that the Cape Town pattern reflects the water crisis rather than a national sectoral trend.

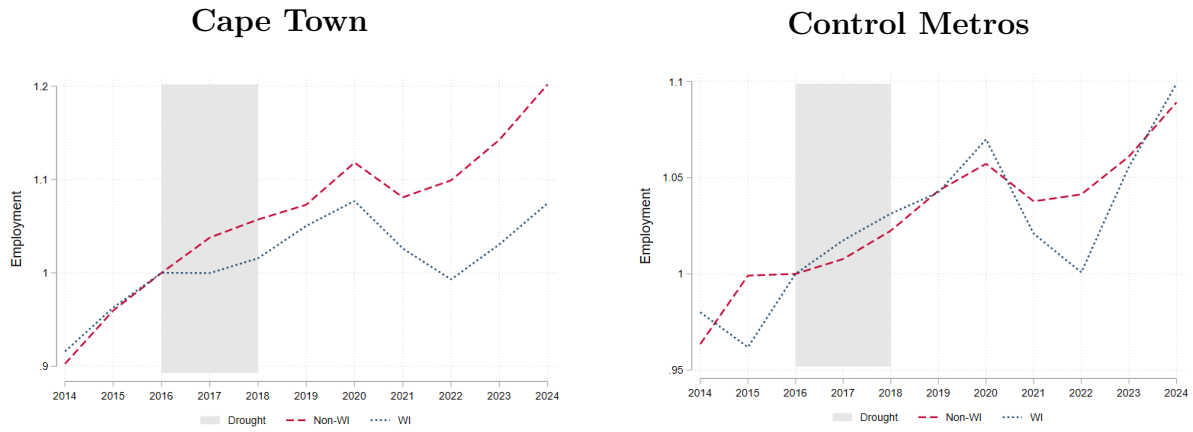
Figure 7: Establishments by Water Intensity: Cape Town vs. Control Metros



Notes: Establishment counts indexed to 2016. Grey shading marks the drought period. Control metros are Johannesburg, Pretoria, and Durban.

This firm exit also shows up in aggregate employment. Figure 8 compares employment by water intensity across Cape Town and the control metros. In Cape Town, non-water-intensive sectors grew steadily, reaching 120 percent of their 2016 level by 2024, while water-intensive sectors stagnated and reached only about 107 percent. The gap of roughly 13 percentage points opened during the crisis and has not closed. In the control metros, water-intensive and non-water-intensive employment followed broadly similar paths throughout the sample period, reaching similar levels by 2024.

Figure 8: Employment by Water Intensity: Cape Town vs. Control Metros



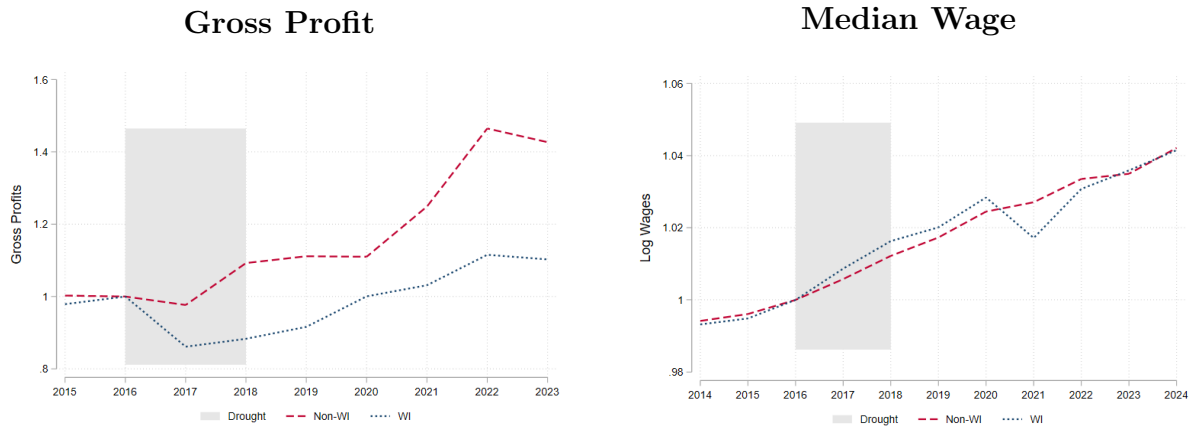
Notes: FTE employment indexed to 2016. Grey shading marks the drought period. Control metros are Johannesburg, Pretoria, and Durban.

Because the establishment decline is proportionally larger than the employment decline, the exiting establishments appear to have been predominantly small, which is what a Melitz model would predict: the least productive firms sit closest to the zero-profit cutoff and are the first to go when costs rise.

Figure 9 adds gross profit and median wage trajectories by water intensity. Profits in water-intensive sectors fell to roughly 87 percent of their 2016 level during the crisis and recovered only partially, while non-water-intensive profits grew to approximately 145 percent of their 2016 level. The pattern fits the cost shock channel: water-intensive firms faced higher input costs, which compressed margins and pushed marginal firms below break-even.

Wages, by contrast, showed no differential movement across the two groups, with water-intensive and non-water-intensive median wages tracking each other closely throughout the sample period. This is informative about the mechanism. If the employment decline were driven by negative demand shocks rather than cost shocks, we would expect wage declines among surviving firms. That wages did not diverge suggests instead that adjustment operated entirely through the extensive margin: firms exiting rather than surviving firms paying less.

Figure 9: Gross Profits and Wages by Water Intensity, Cape Town



Notes: Left panel: gross profits indexed to 2016. Right panel: median log wages indexed to 2016.

To summarise the descriptive evidence: water-intensive sectors in Cape Town experienced a persistent divergence in employment and establishments, driven by firm exit rather than wage adjustment, and this divergence is specific to Cape Town and does not appear in control metros. As we show next, the effect is also visible spatially within Cape Town, with locations that had more water-intensive activity before the crisis experiencing larger and more persistent declines.

4.2 Event Studies

We look at two event study specifications. The first is a triple-difference specification, looking at how water intensive sectors developed over time in Cape Town relative to the same sectors in the control metros:

$$Y_{cst} = \sum_{k \neq 2016} \beta_k \cdot \text{CapeTown}_c \times \text{WaterIntensive}_s \times \mathbf{1}[t = k] + \delta_{cs} + \delta_{ct} + \delta_{st} + \varepsilon_{cst} \quad (1)$$

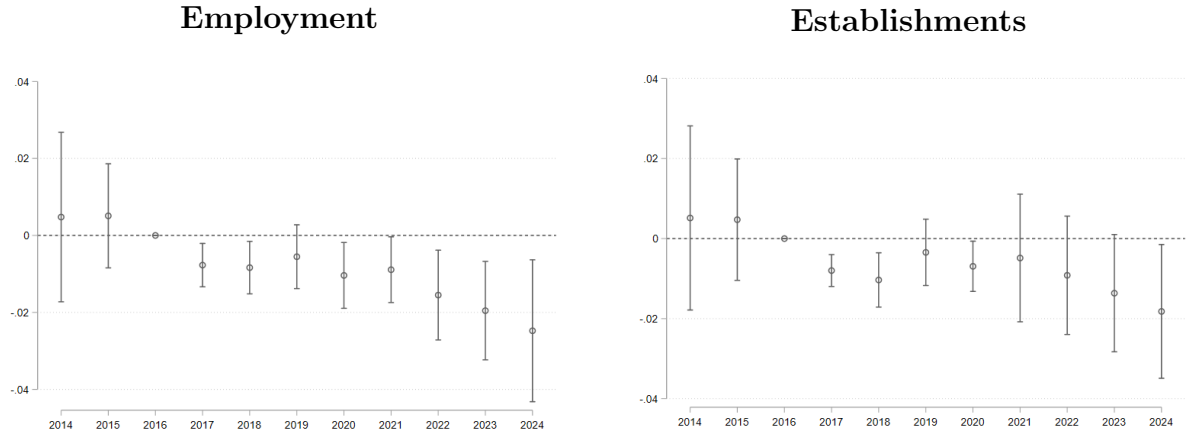
where δ_{cs}, δ_{ct}, and δ_{st} are city-sector, city-year, and sector-year fixed effects. The city-sector fixed effects absorb time-invariant differences in the level of water-intensive employment across cities, the city-year fixed effects absorb city-specific aggregate shocks, and the sector-year fixed effects absorb national trends in water-intensive sectors. Standard errors are clustered at the city-sector level.[1]

Figure 10 shows the triple-difference results for both employment and establishments. The employment specification seems to confirm what we saw with the aggregate data: flat pre-trends, then a persistent negative gap in employment between the sectors that

[1]With four cities and approximately 30 two-digit sectors, we have approximately 120 city-sector clusters. If the effective number of clusters is below 50, we also report wild cluster bootstrap confidence intervals following Cameron et al. (2008).

continues through 2024. Establishments tell a similar story, confirming that firm exit accounts for a substantial share of the employment decline and that this exit is specific to the interaction of Cape Town with water-intensive sectors.

Figure 10: Triple-Difference: Employment and Establishments



Notes: Coefficients from equation (1). Bars show 90% confidence intervals.

The post-crisis trajectory helps distinguish between two readings of the evidence. Slow convergence back to the original equilibrium should produce a concave path, with the gap narrowing over time. A permanent shift to a lower equilibrium predicts a gap that persists or keeps widening. Six years out, the coefficients show no real curvature toward zero.

4.3 Spatial Variation Within Cape Town

The triple-difference exploits sectoral variation across metros. We can also ask whether the effect shows up spatially within Cape Town, which matters for the model because agglomeration operates locally. We estimate:

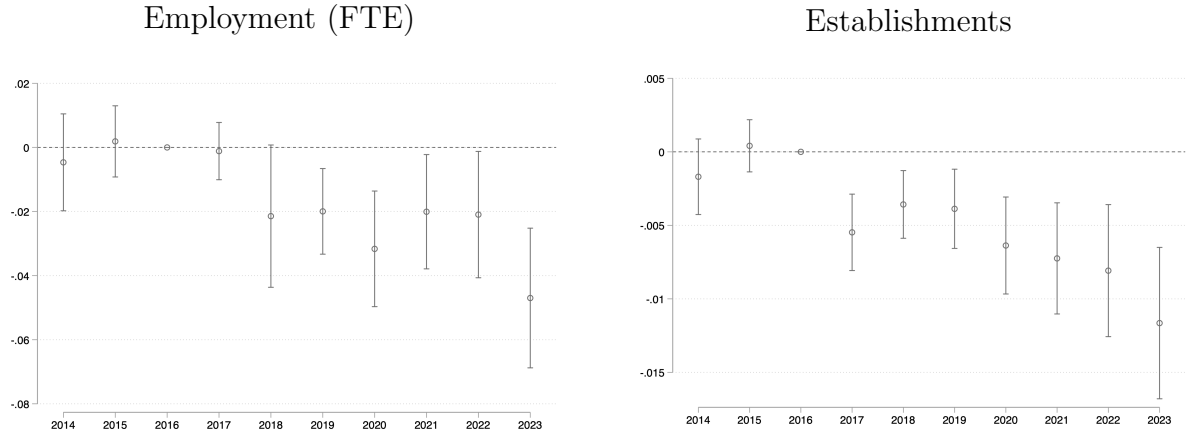
$$\ln Y_{nt} = \sum_{k \neq 2016} \beta_k \cdot \mathbf{1}[t = k] \times \ln(\text{WaterIntensity}_n) + \mu_n + \varepsilon_{nt} \tag{2}$$

where the unit of observation is hexagon $\times$ year, restricted to Cape Town. We have to restrict this analysis to Cape Town only as we do not have localised water use data for other metro areas in South Africa. Here, water intensity is continuous: average pre-crisis commercial water consumption per worker in hexagon n, computed from the meter data. Hexagon fixed effects μ_n absorb time-invariant differences across locations. Standard errors are clustered at the hexagon level.

The identifying assumption is that, absent the water crisis, hexagons with higher water intensity would have followed the same trajectory as those with lower water intensity.

The main threat is that high-water-intensity hexagons are industrial areas that might have different growth trajectories for reasons unrelated to water.

Figure 11: Spatial Event Studies: Within Cape Town



Notes: Coefficients on $\mathbf{1}[t = k] \times \ln(\text{WaterIntensity}_n)$ from equation (2). Bars show 80% confidence intervals, clustered by hexagon.

The spatial results line up with the sector-level findings. Hexagons with more water-intensive activity experienced larger declines in both employment and establishments, and the effects persist through 2023. Because agglomeration operates locally, persistence should be concentrated in exactly these locations, which is what we see.

4.4 Alternative Explanations

The persistence we document could reflect mechanisms other than agglomeration erosion. Any candidate explanation must account for the full pattern: persistent employment loss driven by firm exit, with no wage effects, concentrated in water-intensive sectors and locations, and absent in control metros. We consider four alternatives.

The most natural is that the economy is simply converging slowly back to its original equilibrium. If workers face high switching costs and firms face high sunk entry costs, recovery might take a long time. But slow convergence predicts a trajectory that is concave in time, with the gap narrowing at first quickly and then more slowly as the economy asymptotically approaches the pre-crisis level. Our event study coefficients show no such curvature; if anything, the post-crisis trajectory continues to widen through 2024, six years after prices normalised. We test this formally by estimating a linear trend on the post-crisis coefficients (2019–2024) and fail to reject a non-negative slope. Slow convergence should also be accompanied by gradually rising entry rates, and the flat establishment counts suggest this has not happened.

A second possibility is that firms avoid Cape Town because the drought could recur. If so,

the employment gap reflects increased perceived risk rather than agglomeration erosion. But the return period for a drought of this severity is estimated at 150 to 300 years, and the city has since invested in desalination and groundwater extraction, reducing future drought risk below pre-crisis levels. More to the point, risk repricing should affect all sectors in Cape Town, not specifically water-intensive ones, since a firm deciding whether to locate in the city should weigh drought risk regardless of its own water use. Our triple-difference absorbs such city-wide risk effects through the city-year fixed effects. A sector-specific version of the story, in which water-intensive firms in particular perceive higher future risk, is more plausible but should produce compensating wage increases for surviving workers in those sectors, which we do not see.

A third candidate is spatial diversification. Castro-Vincenzi (2024) shows that firms reorganise production geography in response to climate risk, shifting capacity to safer locations. If multi-establishment firms in Cape Town redirected water-intensive operations to other cities after the crisis, we would observe falling employment in Cape Town but stable employment nationally, the same pattern as agglomeration erosion but for a different reason. We can distinguish between these explanations by decomposing the employment decline by firm structure. If spatial diversification drives the result, multi-establishment firms, which can shift capacity without fully exiting, should account for most of the decline. If agglomeration erosion drives it, single-establishment firms, the marginal firms closest to exit, should dominate.

Finally, firms that survived the crisis may have adopted water-saving technologies that permanently reduced their labour demand. If so, the employment decline would reflect efficient restructuring rather than agglomeration collapse. This prediction, however, cuts in the wrong direction on two margins: it implies rising wages among surviving firms (higher productivity per worker) and employment declines concentrated within surviving firms (fewer workers per establishment) rather than in the number of establishments. We find neither. Wages show no differential movement, and the decline is driven overwhelmingly by establishment exit.

None of these alternatives can individually account for the full set of facts. Our preferred reading of the evidence, that firm exit eroded accumulated agglomeration below the threshold for re-entry, is the only interpretation we have found that fits all features of the data simultaneously. The model in the next section makes this interpretation precise and identifies the conditions under which it holds.

5 Model

A model of this episode must match three features of the data. First, a temporary cost shock caused persistent employment losses through firm exit, not wage adjustment.

Second, the decline shows no sign of reversal six years after the shock ended; if anything the trajectory widens rather than converges. Third, the effect concentrates in the locations and sectors most exposed to the price increase. Together, these facts point to a tipping point: the cost shock may have pushed certain sector-locations past a threshold toward a new, lower equilibrium.

Most quantitative spatial models fail to generate this pattern. In the workhorse urban models (Ahlfeldt et al., 2015; Redding and Rossi-Hansberg, 2017) the equilibrium is always proved to be unique. In the dynamic spatial framework of Caliendo et al. (2019), with worker mobility frictions, adjustment may be slow but the economy always converges back to its unique steady state. What is missing is a mechanism through which firm exit during the crisis damages the local economic environment in a way that is not repaired when the crisis ends, and the Allen-Donaldson decomposition of agglomeration into contemporaneous and historical components provides a helpful starting point. We develop a model that combines heterogeneous firms that enter and exit, agglomeration that depends on both current and past activity, and workers who face frictions in reallocating across sectors and locations.

5.1 Overview

The model has five components: heterogeneous firms that enter and exit (Melitz-Pareto), agglomeration externalities that depend on both current and past local activity (Allen-Donaldson), workers who choose where to work subject to switching costs and a Calvo re-optimisation friction (Caliendo-Dvorkin-Parro), a housing market that provides congestion, and households that consume water with elastic demand for discretionary use.

Firm heterogeneity generates realistic exit patterns. Agglomeration with memory creates the tipping point. Worker switching costs and the Calvo re-optimisation friction slow labour reallocation. Housing congestion keeps the equilibrium well-behaved. Household water demand provides the alternative conservation margin needed for the policy counterfactual.

5.2 Environment

The city has N locations and two sectors: water-intensive (W) and non-water-intensive ($\mathcal{N}$). Time is discrete. Firms produce differentiated varieties traded freely within the city at a common price index P_t, normalised to 1.[2] Each location has inelastic housing supply and workers live where they work.[3] We solve for perfect foresight equilibrium: all agents

[2]Total city expenditure R_t is determined by this normalisation together with the CES price index aggregating over all active varieties.

[3]Adding a separate commuting decision would roughly square the state space. Since our mechanism operates at the production level, the return from adding commuting is low relative to the cost.

correctly anticipate the path of wages, rents, and employment.[4]

The timing within each period runs as follows. Workers chose their sector and location in the previous period and arrive at the start of t; firm mass $M_{sn,t}$ is therefore predetermined. Fundamentals and agglomeration pin down local productivity. Water prices clear the conservation constraint. Wages clear the labour market; housing rents adjust given fixed supply. Firms with $\varphi < \varphi^*_{sn,t}$ exit, and potential entrants decide whether to pay F_E, determining $M_{sn,t+1}$. A fraction λ of workers then re-optimise their location-sector choice for $t+1$, determining $L_{sn,t+1}$.

5.3 Firms

A firm in sector s at location n with idiosyncratic productivity φ produces a unique variety using labour l and water x:

$$y_{sn}(\varphi) = A_{sn,t} \cdot \varphi \cdot l^{\gamma_s} \cdot x^{\theta_s}, \qquad \gamma_s + \theta_s = 1 \tag{3}$$

where $A_{sn,t}$ is local productivity, γ_s is the labour share, and θ_s is the water share. Sectors differ in θ_s: water-intensive manufacturing has a higher water share than services.

Consumers value variety. Demand for each variety follows a CES aggregator with elasticity $\sigma > 1$ across all varieties in the city.[5] Cost minimisation over labour and water gives marginal cost $\text{mc}(\varphi) = \kappa_s \cdot w_{sn,t}^{\gamma_s}(p^F_{W,t})^{\theta_s}/(A_{sn,t}\varphi)$, where $\kappa_s = \gamma_s^{-\gamma_s}\theta_s^{-\theta_s}$. Under monopolistic competition, each firm charges markup $\sigma/(\sigma-1)$ over marginal cost, and operating profit (revenue minus total variable cost) is a constant fraction $1/\sigma$ of revenue.

Because marginal cost is proportional to $1/\varphi$ and the cost index $w^{\gamma_s}(p^F_W)^{\theta_s}$ is common to all firms in (s,n), the standard Melitz results carry through: revenue ratios depend only on relative productivities, Pareto aggregation applies, and average revenue is a fixed multiple of cutoff revenue.

Each active firm also pays a per-period fixed operating cost f_s:

$$\pi_{sn}(\varphi, t) = \frac{r_{sn}(\varphi)}{\sigma} - f_s \tag{4}$$

A firm operates if $\pi \geq 0$, which defines a productivity cutoff $\varphi^*_{sn,t}$: firms below it exit. When water prices rise, marginal cost rises for every firm, revenue falls, and the cutoff

[4] This is standard in the dynamic spatial equilibrium literature. See Caliendo et al. (2019).

[5] All varieties from both sectors enter a single CES aggregator. Under the small open city assumption, Cape Town's firms are a negligible share of the national market, so the demand shifter is exogenous. When water-intensive firms exit, consumers shift spending toward other varieties, amplifying the employment decline in the water-intensive sector. An alternative with Cobb-Douglas across sectors and CES within would fix sectoral expenditure shares and generate smaller employment effects. We use the single-CES specification for tractability.

rises so that marginal firms exit. Profit for any surviving firm relative to the cutoff is:

$$\pi_{sn}(\varphi, t) = f_s \left[\left(\frac{\varphi}{\varphi^*_{sn,t}} \right)^{\sigma-1} - 1 \right] \tag{5}$$

Potential entrants draw φ from a Pareto distribution with shape $k > \sigma - 1$ and lower bound normalised to $\underline{\varphi} = 1$. Entry requires paying a sunk cost F^s_E. Free entry equates the expected value of a draw to the sunk cost when entry is positive.

Evolution of active firms

Let $M^E_{sn,t+1}$ denote new entrants. The mass of active firms evolves as:

$$M_{sn,t+1} = M_{sn,t} \cdot \min\left\{1, \left(\frac{\varphi^*_{sn,t}}{\varphi^*_{sn,t+1}} \right)^k \right\} + M^E_{sn,t+1} \cdot (\varphi^*_{sn,t+1})^{-k} \tag{6}$$

The first term counts surviving incumbents. Incumbents at t have $\varphi \geq \varphi^*_t$. Of these, the fraction with $\varphi \geq \varphi^*_{t+1}$ is $(\varphi^*_t/\varphi^*_{t+1})^k$, which is less than one when the cutoff rises and capped at one when it falls. The second term counts surviving entrants.

5.4 Agglomeration: contemporaneous and historical

Local productivity depends on both current and past activity:

$$A_{sn,t} = \bar{A}_{sn} \cdot L^{\alpha_1}_{sn,t} \cdot L^{\alpha_2}_{sn,t-1} \tag{7}$$

$\bar{A}_{sn}$ is exogenous fundamental productivity (geography, climate, proximity to the port) and is time-invariant.

$L^{\alpha_1}_{sn,t}$ captures contemporaneous agglomeration ($\alpha_1 > 0$): the productivity benefit from current activity in (s, n). Knowledge exchange between today's firms, a deep specialised labour pool, jointly supported specialised suppliers. These benefits adjust quickly. If 50 firms exit tomorrow, they are immediately halved. If 50 new firms enter next year, they are restored.

$L^{\alpha_2}_{sn,t-1}$ captures historical agglomeration ($\alpha_2 > 0$): the productivity benefit from what past activity built. Consider two locations that both have 50 manufacturing firms today. Location A had 100 firms last year and lost half during the crisis. Location B has always had 50. With only contemporaneous agglomeration, these locations would be identical, while our model hopes to distinguish them. Location A should have some remaining productivity boost from the fact that it has factory buildings still standing, water mains still

connected, laid-off workers still living nearby with sector-specific skills, and potentially surviving supplier relationships. A new firm entering Location A faces a more productive environment than one entering Location B because of what past activity left behind.

With only contemporaneous agglomeration ($\alpha_2 = 0$), the economy has no memory, and a temporary shock is always followed by recovery. With historical agglomeration ($\alpha_2 > 0$), the economy has memory, and a shock that destroys enough accumulated capital can push the economy past a tipping point after which recovery does not occur.

5.5 Workers

Workers have Stone-Geary preferences over consumption c, housing h, and water. A location-specific amenity shifter B_n scales utility multiplicatively:

$$U_{sn,t} = B_n \cdot c^{\alpha_c} \cdot h^{\alpha_h} \cdot (d^{\mathrm{w}} - \bar{w})^{\alpha_w}, \qquad \alpha_c + \alpha_h + \alpha_w = 1 \tag{8}$$

Here $\bar{w}$ is a subsistence level of water (drinking, cooking, basic hygiene). The term $(d^{\mathrm{w}} - \bar{w})$ is discretionary water consumption above subsistence. The budget constraint is $c + Q_{n,t} \cdot h + p^H_{W,t} \cdot d^{\mathrm{w}} = w_{sn,t}$, where $Q_{n,t}$ is rent and $p^H_{W,t}$ is the household water price. Define disposable income as what remains after paying the mandatory water bill:

$$\tilde{w}_{sn,t} \equiv w_{sn,t} - p^H_{W,t} \cdot \bar{w} \tag{9}$$

Workers allocate disposable income in standard Cobb-Douglas shares: fraction α_c on consumption, α_h on housing, α_w on discretionary water. Substituting the optimal allocations back into (8) and taking logs:

$$v_{sn,t} = \ln \tilde{w}_{sn,t} - \alpha_h \ln Q_{n,t} - \alpha_w \ln p^H_{W,t} + b_n \tag{10}$$

The water price term $-\alpha_w \ln p^H_{W,t}$ is city-wide, so it does not directly affect spatial comparisons. It also enters through $\ln \tilde{w}_{sn,t} = \ln(w_{sn,t} - p^H_{W,t}\bar{w})$, which is not spatially neutral when wages differ across locations. This effect is second-order when subsistence water expenditure is small relative to wages. At peak crisis prices, the ratio $p^H_{W,t}\bar{w}/w_{sn,t}$ reaches at most 0.10 and is typically below 0.03 for formal-sector workers. We treat the spatial allocation as independent of the household water price.[6]

Each period, a fraction λ of workers in each sector-location receive the opportunity to re-optimise their choice. The remaining $1 - \lambda$ stay put regardless of conditions. This

[6]The parameterisation must ensure $\tilde{w}_{sn,t} > 0$ for all sector-locations and periods, including during the crisis when $p^H_{W,t}$ is elevated.

Calvo-style friction captures the fact that most workers do not actively search for new jobs each year, whether because of contracts, inertia, or information costs. The parameter λ governs the speed at which the labour market can reallocate workers after a shock, separately from the switching costs τ that govern the cost of moving conditional on having the opportunity.

Workers who do re-optimise draw idiosyncratic preference shocks from a Type-I Extreme Value distribution with scale $1/\nu$. The parameter ν governs how responsive these active movers are to systematic utility differences: high ν means small wage gaps induce large flows; low ν means idiosyncratic preferences dominate.[7]

Workers are forward-looking. The value of being in (s, n) at t satisfies the Bellman equation:

$$\mathcal{V}_{sn,t} = v_{sn,t} + \frac{1}{\nu} \ln \sum_{(j,h)} \exp\left[\nu\left(\beta \mathcal{V}_{jh,t+1} - \tau_{(s,n)\to(j,h)}\right)\right] \tag{11}$$

Switching costs are additive: $\tau_{(s,n)\to(j,h)} = \tau^S \cdot \mathbf{1}[\text{change sector}] + \tau^N \cdot \mathbf{1}[\text{change location}]$. A worker who changes both sector and location pays both. The logit choice probabilities, applied only to the fraction λ who re-optimise, give a law of motion for employment:

$$L_{sn,t+1} = (1-\lambda) L_{sn,t} + \lambda \sum_{(i,r)} \mu_{(i,r)\to(s,n),t} \cdot L_{ir,t} \tag{12}$$

where $\mu_{(i,r)\to(s,n),t}$ is the logit probability that a re-optimising worker in (i, r) chooses (s, n). The first term captures workers who stay; the second captures the active movers. Both λ and τ slow adjustment, but they do so through different channels. Switching costs make moving expensive conditional on searching. The Calvo friction limits how many workers search at all. Empirically, the distinction matters: τ is identified from the relative transition rates across destinations among movers, while λ is identified from the overall share of workers who change sector-location in a given year.

5.6 Household Water Demand

The optimal allocation implies that each worker in (s, n) consumes:

$$d^{\mathrm{w}}_{sn,t} = \bar{w} + \frac{\alpha_w \cdot \tilde{w}_{sn,t}}{p^H_{W,t}} \tag{13}$$

The first term is subsistence, consumed regardless of price. The second is discretionary: it falls when $p^H_{W,t}$ rises and increases with disposable income. Total household water

[7] Our notation follows Caliendo et al. (2019).

consumption in the city is:

$$W_t^H = \sum_{s,n} L_{sn,t} \cdot \left[\bar{w} + \frac{\alpha_w \cdot \tilde{w}_{sn,t}}{p_{W,t}^H} \right] \tag{14}$$

5.7 Firm Water Demand

From cost minimisation, each firm spends fraction θ_s of variable cost on water. Variable cost is $(\sigma - 1)/\sigma$ of revenue. Firm φ's water consumption is:

$$x(\varphi) = \frac{\theta_s(\sigma - 1) \cdot r(\varphi)}{\sigma \cdot p_{W,t}^F} \tag{15}$$

Larger, more productive firms use more water. Total firm water demand aggregates over active firms using the Pareto average:

$$W_t^F = \sum_{s,n} M_{sn,t} \cdot \frac{\theta_s(\sigma - 1) \cdot \bar{r}_{sn,t}}{\sigma \cdot p_{W,t}^F} \tag{16}$$

where $\bar{r}_{sn,t}$ is average revenue among active firms. Higher water prices reduce firm water demand through two channels: exit (lower M) and substitution (lower x per surviving firm).

Combined with household demand, total city water use must satisfy the conservation constraint imposed by reservoir capacity $\bar{W}_t$:

$$\underbrace{\sum_{s,n} M_{sn,t} \cdot \frac{\theta_s(\sigma - 1)\bar{r}_{sn,t}}{\sigma \cdot p_{W,t}^F}}_{\text{firm water}} + \underbrace{\sum_{s,n} L_{sn,t} \cdot \left[\bar{w} + \frac{\alpha_w \cdot \tilde{w}_{sn,t}}{p_{W,t}^H} \right]}_{\text{household water}} \leq \bar{W}_t \tag{17}$$

Three margins of adjustment satisfy the constraint: firm exit (lower M), substitution toward labour within surviving firms, and reduced household discretionary consumption. In practice the substitution margin is small because θ_s is small, so firm adjustment operates mainly through exit. The policy instrument $\rho_t \equiv p_{W,t}^F / p_{W,t}^H$ determines which margin bears more of the burden.

The constraint has a hard floor. Even if all discretionary household water is eliminated and all firms exit, subsistence water demand $\bar{L}\bar{w}$ remains. We restrict attention to crises where price-based allocation remains feasible, that is, where water supply exceeds subsistence needs.

5.8 Housing

Each location n has inelastic housing supply $\bar{H}_n$. As more workers move in, rents rise and each worker consumes less housing. Workers spend fraction α_h of disposable income on housing, and market clearing gives:

$$Q_{n,t} = \frac{\alpha_h}{\bar{H}_n} \sum_s \tilde{w}_{sn,t} \cdot L_{sn,t} \tag{18}$$

More workers means higher total income bidding for the same housing stock, which raises rents. In equilibrium, the agglomeration benefit of being in a thick market is offset at the margin by the congestion cost of expensive housing.

5.9 Uniqueness of the dynamic equilibrium path

Because employment affects productivity (through α_1), which affects wages and profits, which through worker choices affect future employment, the model features a feedback loop. Whether the equilibrium path $\{L_{sn,t}\}_{t=0}^{\infty}$ is unique, given initial conditions, depends on whether this feedback is amplifying or dampening.

The net contemporaneous effect on a location's attractiveness combines the agglomeration benefit (α_1) with the congestion cost from housing ($-\alpha_h$), and with standard parameter values ($\alpha_1 \approx 0.04$, $\alpha_h \approx 0.27$) the net effect is negative: congestion outweighs agglomeration within a single period.[8] Agglomeration is not absent, but the dampening force is stronger than the amplifying force, so perturbations shrink rather than grow.

Proposition 1 (Unique Dynamic Equilibrium Path)**.** *If $\nu\gamma_1 < 1$, the dynamic equilibrium path is unique: given initial conditions (L_0, L_{-1}), there is a unique sequence $\{L_t, w_t, Q_t, M_t, \mathcal{V}_t\}_{t=0}^{\infty}$ satisfying all equilibrium conditions. When $\gamma_1 < 0$ (congestion exceeds contemporaneous agglomeration), this holds for all $\nu > 0$.*

(This is a work-in-progress proof sketch:) Both employment $L_{sn,t}$ and firm mass $M_{sn,t}$ are predetermined within each period, so wages, rents, and all within-period objects are determined given the state. The only feedback loop runs through the logit choice into future employment.

Take log-ratios of employment shares at two locations n and m. In the reduced-form representation where flow utility depends on $\ln L_n$ with net contemporaneous elasticity

[8] The precise net effect depends on the reduced-form elasticity of flow utility with respect to own employment. Substituting $\ln w \propto \alpha_1 \ln L + \alpha_2 \ln L_{t-1}$ and $\ln Q \propto \ln w + \ln L$ into indirect utility $v = \ln w - \alpha_h \ln Q + b_n$ gives the net contemporaneous elasticity $\gamma_1 = (1-\alpha_h)\alpha_1 - \alpha_h \approx -0.24$ and the historical elasticity $\gamma_2 = (1-\alpha_h)\alpha_2$. We use the exact expressions in all computations.

γ_1 and on $\ln L_{n,t-1}$ with historical elasticity γ_2 (a fixed state), the log-ratio satisfies:

$$(1 - \nu\gamma_1)(x_n - x_m) = c_{nm}$$

where c_{nm} collects predetermined terms. When $|\nu\gamma_1| < 1$, this system has a unique solution. This is a special case of the uniqueness result in Allen and Donaldson (2020).

The extension to the forward-looking case follows by backward induction. Fix terminal values. At each date, the within-period equilibrium is unique given continuation values, and continuation values are uniquely determined by the (unique) future path. Switching costs shift the attractiveness of destinations but do not create feedback between L_t and v_t, so they do not affect the contraction property.

In our simulations, $\gamma_1 \approx -0.24$, so $\nu\gamma_1 < 0$ for all $\nu > 0$. The within-period mapping is not merely a contraction but a dampening map: more workers at a location make it less attractive, so perturbations are self-correcting.

The myopic exit rule ensures the firm side does not introduce additional multiplicity. With myopic exit, firm mass is a mechanical function of current conditions, adding no independent source of indeterminacy.[9]

5.10 Multiple steady states and the tipping point

At a steady state, $L_{sn,t} = L_{sn,t-1}$. Both α_1 and α_2 are fully operative:

$$A_{sn}^{SS} = \bar{A}_{sn} \cdot (L_{sn}^{SS})^{\alpha_1+\alpha_2}$$

The net spillover at a steady state combines both agglomeration components against congestion. In terms of the reduced-form elasticities, the steady-state net elasticity is $\gamma_1+\gamma_2 = (1-\alpha_h)(\alpha_1+\alpha_2)-\alpha_h$, which exceeds the contemporaneous γ_1 by $\gamma_2 = (1-\alpha_h)\alpha_2$. Multiple stable steady states exist when this combined elasticity is strong enough:

$$\nu\big[(1 - \alpha_h)(\alpha_1 + \alpha_2) - \alpha_h\big] > 1 \tag{19}$$

A necessary condition is $\alpha_1 + \alpha_2 > \alpha_h/(1 - \alpha_h)$. With $\alpha_1 = 0.04$ and $\alpha_h = 0.27$, this requires $\alpha_2 > 0.33$. The precise threshold also depends on ν: more responsive workers (higher ν) make the condition easier to satisfy, because small utility differences translate into large employment shifts that amplify the agglomeration feedback. Whether the condition holds is an empirical question, and we will estimate α_2 and ν from the data and

[9] We verify uniqueness numerically in all simulations by initialising the inner loop from multiple starting points and confirming convergence to the same solution.

test it directly.

When condition (19) holds, two stable configurations coexist: one with high activity and strong agglomeration, another with low activity and weak agglomeration. The unstable fixed point separating them is the tipping point, the employment level at which agglomeration and congestion exactly balance.

5.11 Why the shock can be permanent

During the crisis, higher water prices raise the cost index for water-intensive firms, the cutoff rises, marginal firms exit, contemporaneous agglomeration falls immediately, and historical agglomeration begins eroding as accumulated capital depreciates.

When water prices normalise, a potential entrant evaluates the post-crisis environment. She sees fewer firms, depleted infrastructure, and a thin labour market, all of which translate into lower expected profits because expected productivity has fallen. If expected profits net of the sunk cost F_E are negative, entry does not occur. Workers considering returning see few firms and low wages and stay put. The economy settles into a self-reinforcing low-activity trap.

Expectations amplify this mechanism. Entry is a forward-looking decision: a firm pays F_E today and earns profits over many future periods, so what matters is not just current conditions but the expected path of local productivity. If potential entrants believe the crisis could recur, or simply that too many other firms have left for conditions to improve, the expected return to entry falls further. Even a small probability of future water scarcity can depress entry enough to prevent recovery, because the option value of waiting rises relative to the sunk cost of entering a weakened market. The model assumes perfect foresight, so beliefs are correct in equilibrium, but the qualitative point extends to settings with uncertainty: pessimistic expectations about re-entry by others are self-fulfilling when agglomeration depends on the mass of firms present.

Whether the model generates persistence or recovery depends on parameters. If α_2 is small there is no tipping point and the economy recovers. If F_E is small firms re-enter despite weaker agglomeration. If switching costs are low and λ is high, workers flow quickly and the labour market adjusts before agglomeration erodes too far.

6 Identification, Calibration, and Welfare

The central results hinge on the agglomeration parameters α_1 and α_2, so we spend the most space on their identification.

6.1 Parameters Set Externally

We set the discount factor $\beta = 0.95$. The CES elasticity σ comes from the trade and firm dynamics literature; we verify consistency with the Pareto shape k using the firm size distribution in the SARS data.

6.2 Technology Parameters from Pre-Crisis Cross-Section

The Pareto shape k governs how many firms are near the zero-profit cutoff and therefore how many exit when costs rise. We estimate it from the slope of the log rank-log size relationship for establishments in the SARS data. The administrative records cover the universe of formal firms, avoiding the truncation problems that affect survey-based estimates.

In the Melitz-Pareto framework, average firm employment is a function of σ, k, γ_s, and the per-period fixed cost f. We observe average employment per establishment in the pre-crisis data, and given σ, k, and γ_s, this pins down f. The sunk entry cost F_E is then identified from the free entry condition: in steady state, the expected value of a productivity draw equals F_E. The entry rate (the fraction of establishments newly registered each year) is directly observable and, given the other technology parameters, pins down F_E.

The water cost share θ_s measures the fraction of variable cost spent on water. We identify it from water consumption microdata: total commercial water expenditure per establishment divided by total variable cost (approximated by total wage bill times $\sigma/(\sigma - 1)$, the markup-adjusted variable cost) at pre-crisis prices. The water share is small for most sectors ($\theta_{\mathcal{N}} \approx 0.01$–$0.02$) and larger for water-intensive manufacturing ($\theta_W \approx 0.05$–0.10). The labour share is $\gamma_s = 1 - \theta_s$. Housing supply $\bar{H}_n$ per location comes from building stock data at the hexagon level.

Location-sector productivities $\bar{A}_{sn}$ and location amenities b_n are calibrated by inverting the model: given all other parameters and pre-crisis employment as a steady state, the unique $\bar{A}_{sn}$ and b_n that rationalise observed allocations are recovered. This follows the standard approach in quantitative spatial models after Ahlfeldt et al. (2015).

6.3 Worker Mobility Parameters

The responsiveness parameter ν, the Calvo re-optimisation rate λ, and switching costs τ are identified jointly from worker transitions in the SARS matched employer-employee panel. We observe each worker's sector-location in year t and year $t + 1$, giving a full transition matrix.

The logit structure implies that log ratios of transition probabilities among movers identify

switching costs relative to ν:

$$\ln \mu_{(s,n)\to(j,h)} - \ln \mu_{(s,n)\to(s,n)} = \nu\big[\beta(\mathcal{V}_{jh} - \mathcal{V}_{sn}) - \tau_{(s,n)\to(j,h)}\big] \tag{20}$$

Relative transition rates across destinations with the same type of switch (for example, same-sector moves to different locations) identify τ^N relative to ν. Relative rates for same-location moves across versus within sectors identify τ^S relative to ν. The overall fraction of workers who change sector-location in a given year pins down λ, while the distribution of moves across destinations, conditional on moving, pins down ν. We estimate these from pre-crisis transitions (2014–2016), when value functions can be computed from the pre-crisis steady state.

6.4 Preference Parameters

The housing share α_h comes from the South African Income and Expenditure Survey for Cape Town, where we expect $\alpha_h \approx 0.30$. The water share α_w, which governs spending on discretionary water, can be identified in two ways. The direct approach uses total household water expenditure from Cape Town billing data, minus the subsistence component $p_W^H \cdot \bar{w}$, as a fraction of household income. The indirect approach inverts the price elasticity of household water demand, which in the Stone-Geary framework equals the ratio of discretionary to total water consumption. Subsistence water $\bar{w}$ is calibrated from Cape Town's crisis guidelines, which set 50 litres per person per day as the basic needs target. This is also in line with the World Health Organization's recommendation of 50 litres per person per day for basic needs. The amenity shifters B_n are calibrated together with fundamental productivities $\bar{A}_{sn}$ by inverting the model to match pre-crisis employment patterns.

6.5 Agglomeration Parameters: Primary Specification

α_1 and α_2 enter both the uniqueness condition and the condition for multiple steady states (Section 5.9), so getting them right is essential. We estimate them from the spatial panel regression:

$$\ln w_{sn,t} = \alpha_1 \ln L_{sn,t} + \alpha_2 \ln L_{sn,t-1} + \delta_{sn} + \gamma_t + \varepsilon_{sn,t} \tag{21}$$

where δ_{sn} are sector-location fixed effects and γ_t are time fixed effects.[10] Both regressors are endogenous: positive productivity shocks simultaneously raise wages and attract workers, biasing OLS estimates of α_1 upward, and serial correlation in $\varepsilon_{sn,t}$ contaminates

[10] With Cobb-Douglas production and free entry, the cutoff condition gives $w^{\gamma_s}(p_W^F)^{\theta_s} \propto A$. Taking logs: $\gamma_s \ln w_{sn,t} = \alpha_1 \ln L_{sn,t} + \alpha_2 \ln L_{sn,t-1} + \ln \bar{A}_{sn} - \theta_s \ln p_{W,t}^F + \text{const}$. The water price is city-wide and absorbed by time fixed effects. The regression coefficients identify α_1/γ_s and α_2/γ_s. Since $\gamma_s \approx 0.95$ for water-intensive and $\gamma_s \approx 0.99$ for non-water-intensive, the rescaling is small.

α_2 through lagged employment.

We use two instruments. The first, which primarily identifies α_2, is based on the water crisis. The second, which primarily identifies α_1, is a shift-share instrument. We estimate (21) by 2SLS.

Instrument for α_2: the crisis legacy

In the post-crisis period (2019–2024), water prices have normalised and firms in water-intensive sectors no longer face higher costs. But the employment legacy persists: sector-locations that were more water-intensive experienced larger employment declines during the crisis, and these declines have not reversed. Our instrument interacts water intensity with the post-crisis period:

$$Z_{sn,t}^{\text{crisis}} = \text{WaterIntensity}_s \times \text{CapeTown}_n \times \mathbf{1}[t \geq 2019]$$

This predicts lagged employment in the post-crisis period: water-intensive sectors in Cape Town have exogenously depressed $L_{sn,t-1}$ because of crisis-era firm exit, and the time interaction ensures the instrument only operates after prices have normalised.

The exclusion restriction requires that, after prices normalise, water intensity affects productivity only through the employment channel, that is, through agglomeration. During the crisis the exclusion restriction fails because water intensity directly affects firm costs. The time interaction restricts the instrument to the post-crisis years where the direct cost channel is shut off.

Several threats to the exclusion restriction deserve discussion. First, firms that survived the crisis may have permanently adopted water-saving technologies, altering their production function and therefore their labour demand. Second, consumer preferences may have permanently shifted away from water-intensive products. The sector-year fixed effects absorb national demand shifts; the concern is Cape Town-specific demand changes, which we cannot directly control for. Third, the crisis may have permanently altered the risk profile of water-intensive locations, affecting entry decisions through a channel other than agglomeration.

Identification of α_2 comes from the following comparison: conditional on current employment, do sector-locations whose lagged employment was depressed by the crisis exhibit lower productivity? If so, it is due to the reduction in path dependent productivity captured by α_2.

Instrument for α_1: shift-share

We construct a Bartik instrument:

$$B_{sn,t} = \sum_j \text{share}_{jn,2015} \times \Delta \ln L_{j,-\text{CPT},t}$$

where $\text{share}_{jn,2015}$ is the pre-crisis employment share of industry j in hexagon n and $\Delta \ln L_{j,-\text{CPT},t}$ is the national (excluding Cape Town) growth rate of industry j at time t. This predicts current employment through demand shocks originating outside Cape Town. Following Borusyak et al. (2022), identification relies on the exogeneity of national industry shocks rather than on initial shares.

Both $\ln L_{sn,t}$ and $\ln L_{sn,t-1}$ are instrumented with $\{Z^{\text{crisis}}_{sn,t}, B_{sn,t}, B_{sn,t-1}\}$ plus sector-location and time fixed effects. Three instruments for two endogenous variables gives one overidentifying restriction.

6.6 Agglomeration Parameters: Structural Estimation

As a complement to the reduced-form approach, we estimate α_1 and α_2 using simulated method of moments (SMM). This exploits the full model structure and uses moments the regression cannot target.

For each candidate (α_1, α_2), we solve for the pre-crisis steady state, feed in the observed water price sequence, solve the full transition path, extract model-predicted moments, and compute the distance to data moments.

Identification comes from the fact that different moments load on different parameters. How much employment falls during the crisis (2016–2018) pins down α_1, because contemporaneous agglomeration governs how much firm exit amplifies. Whether the gap persists after the crisis (2019–2024) pins down α_2: with $\alpha_2 = 0$ the model predicts recovery, with large α_2 it predicts a flat or widening trajectory. Establishment counts discipline the interaction of α_2 with F_E, and the absence of wage effects jointly tests α_1 and α_h.

6.7 Validation

TBC

6.8 Welfare Measurement

Aggregate welfare at $t = 0$ is the sum of all workers' lifetime value:[11]

$$\mathcal{W} = \sum_{s,n} L_{sn,0} \cdot \mathcal{V}_{sn,0} \tag{22}$$

We assume firm profits accrue to absentee owners and are not included in worker welfare.[12]

We decompose the welfare cost by comparing $\mathcal{V}_{sn,0}^{\text{crisis}}$ to $\mathcal{V}_{sn,0}^{\text{no crisis}}$. The difference breaks into three pieces. Direct costs are temporary: higher water bills and foregone discretionary consumption during the crisis itself. Transition costs are more extended: depressed wages and employment as firms exit and workers slowly reallocate. Permanent costs arise only if the economy crosses the tipping point, in which case the discounted gap between the high and low steady states accumulates over an infinite horizon.

The "persistence multiplier" is the ratio of total costs to direct costs alone.

The policy instrument is the ratio of firm to household water prices, $\rho_t \equiv p_{W,t}^F / p_{W,t}^H$, subject to the conservation constraint (17). Lowering ρ (protecting firms) means fewer firm exits but higher household water bills and lower discretionary water consumption. Raising ρ (protecting households) means more firm exits and greater risk of crossing the tipping point.

For each value of ρ, we solve the full transition path and compute welfare. The central question is whether there exists a critical $\bar{\rho}$ below which the economy remains in the high steady state. If so, the welfare gain from setting $\rho < \bar{\rho}$ relative to the actual policy is the value of preventing the tipping point crossing.

7 Conclusion

Cape Town's water crisis provides evidence on what happens when temporary resource shocks interact with firm dynamics and agglomeration. The standard policy of protecting households and burdening firms achieved its conservation goal, but employment in water-intensive sectors fell and did not recover.

Firm exit, we argue, does more than reduce current activity. Exiting firms destroy capital that took years to build and that other firms depended on: infrastructure, supplier networks, sector-specific human capital. If enough of this capital is destroyed, the remaining agglomeration is too weak to attract re-entry. Potential entrants, evaluating a depleted

[11] The value function $\mathcal{V}$ from equation (11) captures not only current flow utility but also the option value of future mobility and the expected value of idiosyncratic taste shocks.

[12] This could be changed to a lump sum redistribution of profits to residents.

market, find that expected profits don't justify the sunk cost. The local economy settles into a lower equilibrium.

On the policy side, conservation can be achieved through two margins, and their social cost may differ. Household water demand is elastic; discretionary consumption can be cut without permanent damage. Firms can substitute toward labour when water prices rise, but the water share of costs is small, so the dominant firm response is exit. If the tipping point mechanism is operative, shifting more of the conservation burden to households could prevent enough exits to keep the economy above it, and whatever households lose to higher water bills and less garden watering during the crisis would be temporary. Crossing the tipping point is not.

If the mechanism generalises, standard damage estimates for climate shocks may understate true costs whenever shocks trigger firm exits, potentially shifting the economy to a permanent, lower, equilibrium. Temporary shocks can have permanent effects, not because of wage rigidity or search frictions, but because accumulated productive capital, once destroyed, does not spontaneously reassemble.

References

Ahlfeldt, G. M., Redding, S. J., Sturm, D. M., and Wolf, N. (2015). The economics of density: Evidence from the berlin wall. *Econometrica*, 83(6):2127–2189.

Allen, T. and Donaldson, D. (2020). Persistence and path dependence in the spatial economy. Revision requested by the Journal of Political Economy.

Bilal, A. and Känzig, D. R. (2024). The macroeconomic impact of climate change: Global vs. local temperature. *Quarterly Journal of Economics*. Forthcoming. NBER Working Paper No. 32450.

Borusyak, K., Hull, P., and Jaravel, X. (2022). Quasi-experimental shift-share research designs. *Review of Economic Studies*, 89(1):181–213.

Caliendo, L., Dvorkin, M., and Parro, F. (2019). Trade and labor market dynamics: General equilibrium analysis of the china trade shock. *Econometrica*, 87(3):741–835.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3):414–427.

Castro-Vincenzi, J. (2024). Climate hazards and resilience in the global car industry. Working Paper.

Costa, F. and Gerard, F. (2021). Hysteresis and the welfare effect of corrective policies: Theory and evidence from an energy-saving program. *Journal of Political Economy*, 129(6):1705–1743.

Dell, M., Jones, B. F., and Olken, B. A. (2012). Temperature shocks and economic growth: Evidence from the last half century. *American Economic Journal: Macroeconomics*, 4(3):66–95.

Grafton, R. Q., Ward, M. B., To, H., and Kompas, T. (2011). Determinants of residential water consumption: Evidence and analysis from a 10-country household survey. *Water Resources Research*, 47(8):W08537.

Greenstone, M., Hornbeck, R., and Moretti, E. (2010). Identifying agglomeration spillovers: Evidence from winners and losers of large plant openings. *Journal of Political Economy*, 118(3):536–598.

Hornbeck, R. (2012). The enduring impact of the american dust bowl: Short- and long-run adjustments to environmental catastrophe. *American Economic Review*, 102(4):1477–1507.

Hsiang, S. M. and Jina, A. S. (2014). The causal effect of environmental catastrophe on long-run economic growth: Evidence from 6,700 cyclones. *NBER Working Paper*, (20352). National Bureau of Economic Research.

Ito, K. (2014). Do consumers respond to marginal or average price? Evidence from nonlinear electricity pricing. *American Economic Review*, 104(2):537–563.

Jessoe, K. and Rapson, D. (2014). Knowledge is (less) power: Experimental evidence from residential energy use. *American Economic Review*, 104(4):1417–1438.

Kline, P. and Moretti, E. (2014). Local economic development, agglomeration economies, and the big push: 100 years of evidence from the tennessee valley authority. *Quarterly Journal of Economics*, 129(1):275–331.

Melitz, M. J. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica*, 71(6):1695–1725.

Olmstead, S. M. and Stavins, R. N. (2009). Comparing price and nonprice approaches to urban water conservation. *Water Resources Research*, 45(4):W04301.

Redding, S. J. and Rossi-Hansberg, E. (2017). Quantitative spatial economics. *Annual Review of Economics*, 9:21–58.

Weitzman, M. L. (1974). Prices vs. quantities. *Review of Economic Studies*, 41(4):477–491.