Observation-Aware Calibration of Agent-Based Epidemic Simulations from Public Aggregate Reports: A COVID-19 City-Report Benchmark Study

Mingyuan Wu*
Faculty of Science and Technology, Beijing Normal-Hong Kong Baptist University, Zhuhai, 519087, China
Corresponding author: Mingyuan Wu (email: 15066370762@163.com).

ABSTRACT City reports record confirmation, recovery or discharge, and death by publication date; agent-based simulations track latent disease events. We calibrated Covasim through an observation layer that converts cumulative reports to daily increments and aligns them with simulation outputs. A two-parameter Optuna-TPE search estimated transmission and initial infections. Shenzhen was reproduced under fixed stochastic settings, followed by independent recalibration in 12 non-Hubei first-wave cities. The Shenzhen fit reproduced the reported peak within one day (RMSE 7.2072, MAE 5.5848) but underestimated total reported cases by 20.66%. Gompertz fitted the same short curve more closely (RMSE 3.6874). Across the city panel, median RMSE was 3.3049 and peak-normalized RMSE was 0.1933. The framework links report-curve calibration to stochastic, auxiliary, and scenario outputs within a historical COVID-19 benchmark. Policy evaluation lies outside this benchmark.

INDEX TERMS agent-based model, Covasim, calibration, COVID-19

I. INTRODUCTION

Public city reports make retrospective epidemic benchmarks possible, but they record administrative events rather than infection histories. The available fields are cumulative confirmed cases, recovery or discharge counts, and deaths by report date. Infection dates and individual contacts are absent, and an early-wave window may allow different parameter combinations to fit the same curve. Our calibration begins with an observation layer between the public series and the simulator output [1-2].

OAIC-ABM uses that layer around a Covasim simulation kernel. The workflow cleans the reports, specifies their alignment with simulated events, calibrates a restricted parameter set, and reruns the fitted model across stochastic seeds. Logistic and Gompertz curves provide compact references for the same short single-peak target [3]. The agent-based fit is then retained for analyses that require run- to-run intervals, recovery and death outputs, or contact- scenario comparisons [4-6].

A reported count cannot be read directly as a simulator state. Covasim advances latent disease states on its internal clock; the public table is updated after testing, discharge decisions, data consolidation, and publication. If the two clocks are treated as identical, reporting delay can be absorbed into the fitted transmission parameters. We test confirmed-report alignment and recovery/discharge delay as observation choices while leaving the simulated disease transitions unchanged.

Repeat count, seed block, population scaling, and the fitting window all affect the average trajectory. We treat those settings as part of the experiment and apply the same protocol to the Shenzhen reproduction and every city in the formal panel.

Compared with the conference draft, this version specifies the observation mapping and experiment matrix and separates the Shenzhen reproduction from sensitivity, baseline, ablation, and multicity results. It adds a comparison of two-, three-, and four-parameter calibrations. Low-loss draws from the formal search also feed incremental scenario plots relative to a zero-additional-burden reference [7].

The empirical scope is the historical COVID-19 city- report benchmark described here. The scenario module propagates calibrated uncertainty into candidate trade-offs under an assumed burden score. Other diseases and policy evaluation need different data and a disease-specific simulation kernel [8-9].

A calibration wrapper that keeps public report dates distinct from latent infection histories.

A reproducible Shenzhen case with separate sensitivity analyses for confirmed and recovery/discharge reports [10].

Independent recalibration of a 12-city first-wave panel chosen by administrative and data-quality rules before fitting.

Identifiability checks and scenario comparisons based on the retained low-loss calibration sets.

II. RELATED WORK

A. AGGREGATE EPIDEMIC CURVE MODELS

SIR and SEIR models link incidence to transmission and removal processes [11-14]. Logistic and Gompertz curves offer a different reference: with few parameters, they can closely trace a smooth cumulative trajectory over a short single wave [3]. We fit both families to the same target series used for OAIC-ABM so that the simulation is compared with credible aggregate alternatives.

An empirical growth curve summarizes the observed sequence with a compact shape. A compartment model represents flows between epidemiological states. Both can fit a short reporting window well, so their errors establish the pointwise accuracy available without individual-based simulation. OAIC-ABM is assessed separately on outputs that depend on its stochastic kernel.

B. AGENT-BASED EPIDEMIC SIMULATIONS

Agent-based epidemic models represent people, contact layers, disease states, and stochastic transitions. Covasim provides these mechanisms for COVID-19, including disease progression, testing, and interventions [4-6]. We use its existing kernel and study the calibration of public reports rather than proposing new disease-state dynamics. Repeated runs and layer-specific contacts are the features needed for the later validation and scenario analyses.

One fitted Covasim configuration still produces different trajectories across random seeds. We summarize those runs by their mean and empirical interval. The runs also provide recovery and death series that are not included in the main confirmed-case loss. We use these channels as auxiliary checks and compare pointwise fit against the empirical curves rather than assuming that stochastic output implies better fit.

C. CALIBRATION FROM PUBLIC AGGREGATE REPORTS

Public reports identify publication dates; infection onset, recovery onset, and individual paths remain latent. Aggregation and reporting delay can shift estimates of epidemic timing and intensity [1-2]. With a short window, extra free parameters may create several near-equivalent fits. In our calibration, report alignment and recovery/discharge delay are observation assumptions, and every fit is checked against its parameter bounds.

Confirmation and recovery/discharge do not pass through the same administrative process. The former depends on testing and case publication, whereas the latter also depends on clinical discharge criteria and recording practice. We examine their delays on separate grids and judge each grid with metrics for the report field it changes.

Identifiability is most fragile when the window contains only one rise and decline. We restrict the main search to transmission and initial infections to limit compensation. The later M2P, M3P, and M4P comparison tests whether freeing intervention parameters produces a consistent gain across the sampled cities.

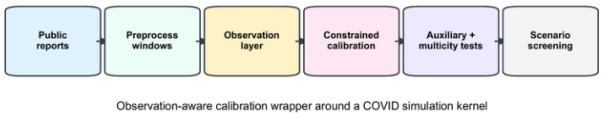

FIGURE 1. Data and Simulation Flow in OAIC-ABM

D. SCENARIO SCREENING AND MULTI-OBJECTIVE COMPARISON

The scenario module compares candidate contact reductions rather than selecting a policy. Each scenario is described by simulated total cases, peak pressure, and an assumed burden score. Non-dominated sorting and incremental plots retain alternatives when those objectives conflict [7]. Because the burden score has no empirical economic valuation, its use is restricted to comparisons inside the tested grid.

We run each scenario over several low-loss calibrations instead of only the best parameter pair. Nondominated frequency records how often a scenario remains competitive across those draws. Prevented cases per burden compares the same scenario with the zero-additional-burden reference. Both summaries preserve the trade-off without imposing a weighted utility function.

III. METHODOLOGY

A. FRAMEWORK OVERVIEW

Cumulative public reports are cleaned into daily increments before observation mapping and constrained Covasim calibration. The resulting simulation records feed the auxiliary checks and scenario screening under the same target (Figure 1).

For every experiment, we record the input file, preprocessing rule, fitting window, parameter bounds, repeat count, and seed block. These settings are shown because changing a stochastic repeat budget or seed block can alter the averages that feed later tables.

The processed city-day table stores observed increments. Each simulation record couples a parameter set with its seed- level trajectories and metrics; the low-loss subset is then passed to sensitivity and scenario analyses. Keeping these objects separate prevents later figures from mixing seed blocks, repeat counts, or calibration targets.

Algorithm 1 gives the experimental sequence for converting the public table into calibration targets, scoring repeated simulations, and passing the best and low-loss fits to auxiliary analyses.

(1) Read cumulative public city reports.

(2) Retain the daily cumulative maximum after cleaning city and date fields.

(3) Reindex the selected window and forward-fill cumulative fields.

(4) First-difference the cumulative reports to obtain daily increments.

(5) Map confirmed and recovery/discharge reports to their simulation outputs.
(6) Register free calibration parameters and the fixed protocol settings.
(7) Evaluate Optuna-TPE trials with repeated stochastic simulations.
(8) Calculate RMSE, MAE, peak-day error, total relative error, and normalized RMSE.
(9) Retain the best fit and the low-loss parameter set for validation and scenario screening.

B. OBSERVATION LAYER

The observation layer defines what is compared before any loss is calculated. For confirmed reports, the target is the daily increment obtained from cumulative confirmed counts.

$$Y_t^{\text{obs}} = C_t - C_{t-1} \tag{1}$$

A candidate parameter vector is run over repeated stochastic seeds. We score its mean simulated trajectory.

$$\overline{Y}_t^{\text{sim}}(\theta) = \frac{1}{R}\sum_{r=1}^{R} Y_{t,r}^{\text{sim}}(\theta) \tag{2}$$

Recovery and discharge form an auxiliary channel. In the Shenzhen analysis, a four-day layer shifts simulated recovery output before it is compared with public recovery/discharge counts. Only the mapping changes; the disease simulation does not.

For confirmed reports, we test shifts from zero to seven days against the same observed series. This grid measures sensitivity to alignment rather than reconstructing infection dates. E2 gives its lowest RMSE at zero days, so the packaged Shenzhen fit remains unshifted.

E3 shifts simulated recovery by 0, 2, 4, 6, or 8 days and recalculates cumulative RMSE, daily MAE, and final relative error. The four-day shift minimizes curve-level RMSE; the closest endpoint occurs at a different delay. Curve and endpoint metrics are therefore kept separate.

Death counts, where available, are checked after calibration and do not enter the confirmed-case loss. This leaves the fitting target unchanged. We report absolute errors for this low-count output, since percentage error would be dominated by its small denominator.

C. IDENTIFIABILITY-CONSTRAINED CALIBRATION

The main protocol frees the transmission coefficient and initial infections; intervention day and multiplier remain fixed. That restriction reduces compensation within the short fitting window. We use Optuna-TPE because the stochastic simulation loss does not provide a useful differentiable objective [15-16].

Each trial is evaluated over a fixed seed block, and the mean trajectory is scored against the observed daily series. The loss and parameter values are stored before TPE updates its sampling distribution. In the formal panel, the complete search is run independently for every city. Shenzhen parameters are never transferred to another city.

We also record whether a fitted value reaches a search boundary. A hit would suggest that the allowed range cuts off a lower-loss region and would make an interior optimum difficult to defend. Neither transmission nor initial infections reaches a boundary in the 12-city formal panel, so fit metrics are reported together with this search-position check. Scenario screening uses all calibrations within a stated tolerance of the best loss.

Raw Covasim isolates the gain from case-specific

$$L_\epsilon = \{\theta : L(\theta) \le (1+\epsilon)L(\theta^*)\} \tag{3}$$

calibration. Naive calibrated Covasim and OAIC-ABM share the same fitted trajectory and therefore the same RMSE. OAIC-ABM also retains explicit report mappings, auxiliary outputs, and low-loss fits used in later analyses.

D. MODEL VARIANTS AND ABLATION DEFINITIONS

Table 1 groups the comparisons by fitted quantity and analytical role: empirical curves, compartment models, raw or calibrated Covasim, and OAIC-ABM. RMSE supports the curve-fit comparison. The ablation variants also test observation mapping and downstream output coverage.

E. EVALUATION METRICS

We describe the main fit with pointwise error, peak timing, total relative error, and RMSE normalized by the observed peak.

$$\text{RMSE} = \sqrt{\frac{1}{H}\sum_{t=1}^{H}\left(\overline{Y}_t^{\text{sim}} - Y_t^{\text{obs}}\right)^2} \tag{4}$$

$$\text{MAE} = \frac{1}{H}\sum_{t=1}^{H}\left|\overline{Y}_t^{\text{sim}} - Y_t^{\text{obs}}\right| \tag{5}$$

$$\text{TRE} = \left(\frac{\sum_{t=1}^{H}\overline{Y}_t^{\text{sim}} - \sum_{t=1}^{H}Y_t^{\text{obs}}}{\sum_{t=1}^{H}Y_t^{\text{obs}}}\right)\times 100\% \tag{6}$$

$$\text{nRMSE}_{\text{peak}} = \frac{\text{RMSE}}{\max_t Y_t^{\text{obs}}} \tag{7}$$

RMSE is sensitive to large daily misses, whereas MAE describes the typical absolute miss. Peak-day error measures timing and total relative error measures cumulative scale. Dividing RMSE by the observed peak makes city windows with different report counts more comparable. We keep the unnormalized RMSE beside it, and no one metric determines the model ranking.

TABLE 1. Compared Models, Fitted Quantities, and Analytical Use

Group	Model or variant	Fitted quantities	Paper role
Empirical curve	Logistic	Curve parameters	Short-window curve-fitting reference
Empirical curve	Gompertz	Curve parameters	Strong smooth single-wave reference
Compartment	SIR	Transmission/removal parameters	Mechanistic aggregate reference
Compartment	SEIR-delay	SEIR parameters and report delay	Aggregate reference with delay
Simulation	Raw Covasim template	No case-specific calibration	Ablation baseline
Simulation	Naive calibrated Covasim	Main case parameters	Calibration-only simulation reference
Simulation	OAIC-ABM	Main case parameters plus observation workflow	Main workflow with auxiliary validation and scenario interface

F. SCENARIO-SCREENING INTERFACE

The scenario grid varies school, work, and community contact multipliers over the low-loss calibration set. A layer-specific reduction contributes to an assumed burden score, which is used only to compare scenarios in this grid.

$$B(\pi) = \sum_{l \in \{s,w,c\}} c_l (1 - m_l)(T - \tau) \tag{8}$$

$$P(\pi) = C(S0) - C(\pi) \tag{9}$$

A scenario is non-dominated when no alternative is better on every reported objective.

$$P_{\text{set}} = \{\pi \in \Pi : \nexists \pi' \in \Pi,\ \pi' \prec \pi\} \tag{10}$$

S0 preserves the calibrated contact setting and has zero additional burden. For S1-S7, we calculate prevented cases relative to S0 and, when needed, divide them by incremental burden. Because S0 must be non-dominated on the burden axis, the useful comparison concerns the scenarios that add contact reduction. Repeating the calculation over low-loss draws shows how often each alternative remains non-dominated.

IV. EXPERIMENTAL SETUP

A. DATA SOURCE AND PREPROCESSING

We use DXYArea.csv, which reports cumulative city-level confirmed, recovered/discharged, and death counts [17]. After filtering cities and normalizing dates, we retain the maximum cumulative value reported within each day. The selected date window is reindexed, cumulative fields are forward-filled, and daily increments are obtained by first differencing. Negative increments are counted during screening and clipped only when increments are constructed. The Shenzhen window runs from 2020-01-25 to 2020-02-28.

The order of preprocessing matters. Taking the within-day maximum avoids summing repeated snapshots of one cumulative total. Reindexing exposes dates with no update, and forward filling records no change until the next report. First differencing is applied last, so repeated records and missing dates do not become artificial incidence.

The first increment is left-censored by the chosen window and is not interpreted as a complete count of earlier infections. Before calibration, screening records window length, total and peak cases, nonzero-increment days, and days with negative increments. Table.2 shows the same processed schema for Shenzhen and the multicity cohort.

B. SHENZHEN REPRODUCTION PROTOCOL

For the Shenzhen reproduction, we load the archived best-parameter snapshot and rerun Covasim [10]. The fixed protocol is beta 0.01572306529144963, initial infections 60, intervention day 12, intervention multiplier 0.16, 30 stochastic repeats with seeds 0 through 29, target length 35, population scale 300, and 50,000 simulated agents.

With stochastic simulation, the same parameter values can yield different means when repeat count or seed block changes. Population scaling changes the simulated representation as well. We therefore record all three as experimental settings. We reproduce the archived configuration rather than running a new search. The resulting table contains the observed series, all 30 seed-level trajectories, their mean, and the empirical 10th to 90th percentile interval. Shenzhen sensitivity and auxiliary results are calculated from this single reproduced output.

C. BASELINES AND ABLATION PROTOCOL

The baseline set contains Logistic, Gompertz, SIR, and SEIR-delay. The ablation set contains the raw Covasim template, naive calibrated Covasim, and OAIC-ABM. Together they separate pointwise fitting from the observation and scenario functions added around the simulation.

All references use the same daily target and date window. Logistic and Gompertz fit empirical shapes; SIR supplies a compartment trajectory, and SEIR-delay adds an exposed state with report alignment. Raw Covasim has no case- specific calibration. Naive calibrated Covasim uses the archived main parameters without the observation and scenario modules.

D. MULTICITY VALIDATION COHORT

The formal panel contains 12 ordinary non-Hubei city-level windows from the 2020 first wave, selected before any model was fitted. We recalibrate each city with 100 Optuna-TPE trials, 30 repeats, seeds 0 through 29, beta from 0.005 to 0.08, and initial infections from 1 to 2000. Intervention day 12, intervention multiplier 0.16, 10,000 simulated agents, and population scale 300 remain fixed. We retain boundary and anomaly diagnostics for every city.

Administrative screening removes Hubei cities, district entries under municipalities, unclear regional aggregates, and

TABLE 2. Report Sources, Windows, and Validation Cohorts

Dataset or cohort	Role	Cities	Window	Target length	Observed total	Notes
Shenzhen main case	Reproducibility and main case	1	2020-01-25 to 2020-02-28	35	402.0	Historical city-level DXY public reports
E1R3A 2020 first-wave ordinary non-Hubei panel	Formal multicity validation	12	City-specific early-report windows	44 in current formal panel		Selected by administrative and data-quality rules

TABLE 3. Experiments, Protocol Scale, and Research Questions

Experiment	Research question	Cities	Run mode	Paper role
E0	Can Shenzhen be reproduced from packaged parameters?	Shenzhen	smoke	Main case reproduction
E2	How sensitive is fit to case-report alignment?	Shenzhen	smoke	Observation-layer sensitivity
E3	How sensitive is recovery/discharge validation to delay?	Shenzhen	smoke	Auxiliary observation-layer sensitivity
E4	What does OAIC-ABM add over raw/naive Covasim?	Shenzhen	smoke	Ablation
E6	How do empirical and compartment references compare?	Shenzhen	smoke	Baseline comparison
E1R3A	Does per-city calibration port across ordinary report windows?	12 cities	formal	Main multicity evidence
E5	Do additional free intervention parameters help?	3 cities	fast sensitivity	Identifiability check
E7	Can low-loss uncertainty be passed to scenarios?	2 cities	lightweight	Scenario-screening demonstration

imported-case aggregates. A retained series must span 30 to 45 days, contain at least 80 total cases and a daily peak of at least five, have no more than two negative-increment days, and include at least seven nonzero-increment days. We also require an interior peak outside the first and last three days and a peak-to-total ratio no greater than 0.40.

Anomaly thresholds were fixed before the formal runs. We flag normalized RMSE by peak above 0.30, absolute total relative error above 40%, absolute peak-day error above seven days, or a parameter on its search boundary. A flag is diagnostic and never removes a selected city from the table.

E. SENSITIVITY AND SCENARIO-SCREENING PROTOCOLS

E2 varies confirmed-report delay from 0 to 7 days; E3 varies recovery/discharge delay from 0 to 8 days. E5 compares M2P, M3P, and M4P in three representative formal-panel cities. E7 propagates low-loss E1R3A calibrations through eight contact-reduction scenarios for Shenzhen and Wenzhou. E5 and E7 are lightweight analyses.

These analyses reuse archived simulations or trial histories without changing the formal cohort. E2 and E3 alter only the relevant observation mapping; E5 changes the number of free intervention parameters; E7 leaves calibration unchanged and applies scenarios afterward. The experiment matrix is given in Table 3.

V. RESULTS

A. SHENZHEN REPRODUCTION

The reproduced Shenzhen series has RMSE 7.2072, MAE 5.5848, a one-day peak error, and total relative error of -20.66%. The one-day peak error shows close timing, while the negative total error shows that cumulative scale remains too low.

The simulated mean follows the observed rise and peaks one day later before declining across the 35-day window (Figure 2). Between-run spread is widest near the peak. The later daily underestimates account for the negative total error despite the close peak timing.

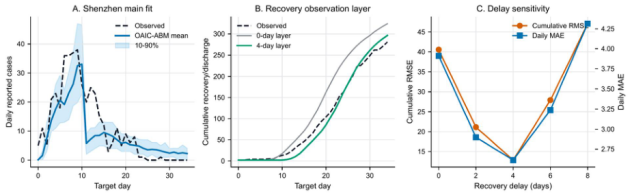

FIGURE 2. Shenzhen Main Fit and Recovery/Discharge Mapping

B. OBSERVATION-LAYER SENSITIVITY

Confirmed-report alignment and recovery/discharge alignment have different optima. Zero days minimizes E2 RMSE. In E3, four days gives recovery-curve RMSE 12.8484, daily MAE 2.6133, and final relative error 5.50%; the endpoint is closest at six days.

The E2 pattern is monotonic over the tested grid. From zero to seven days, RMSE increases from 7.2072 to 15.7605, peak-day error from one to eight days, and total relative error from -20.66% to -25.18%. For this archived trajectory, every positive post hoc shift moves the simulated peak farther from the reported peak.

Cumulative recovery RMSE falls from 40.5517 at zero days to 21.1210 at two days and 12.8484 at four days, then rises to 27.9003 and 47.0730 at six and eight days. Daily MAE follows the same U-shaped pattern. The 0.04% final relative error at six days shows why endpoint alignment selects a different delay from curve-level RMSE. Figures 3 and 4 show the delay profiles; Tables 4 and 5 give the corresponding metrics.

TABLE 4. Confirmed-Report Alignment Metrics by Delay

Delay	RMSE	MAE	Peak error	Total relative error (%)
0	7.207	5.585	1	-20.66
1	8.227	6.375	2	-21.23
2	8.815	6.713	3	-21.85
3	9.541	7.202	4	-22.42
4	10.905	7.950	5	-23.10
5	12.521	8.718	6	-23.70
6	14.431	10.111	7	-24.39
7	15.760	11.369	8	-25.18

TABLE 5. Recovery/Discharge Metrics by Delay

Delay	Cumulative RMSE	Daily MAE	Final relative error (%)
0	40.552	3.913	15.39
2	21.121	2.898	10.83
4	12.848	2.613	5.50
6	27.900	3.237	0.04
8	47.073	4.313	-7.11

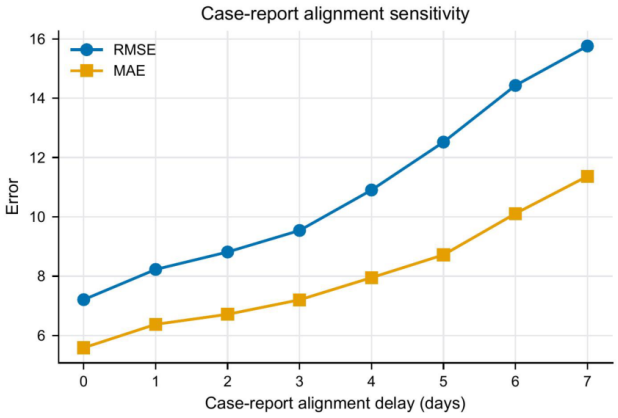

FIGURE 3. Confirmed-Report Alignment Sensitivity in Shenzhen

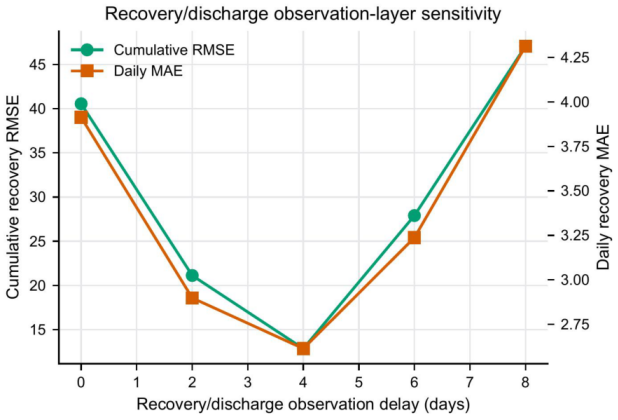

FIGURE 4. Recovery/Discharge Delay Sensitivity in Shenzhen

C. BASELINE COMPARISON AND ABLATION

Gompertz has the lowest Shenzhen RMSE in Table 6 (3.6874). OAIC-ABM reaches 7.2072, still far below raw

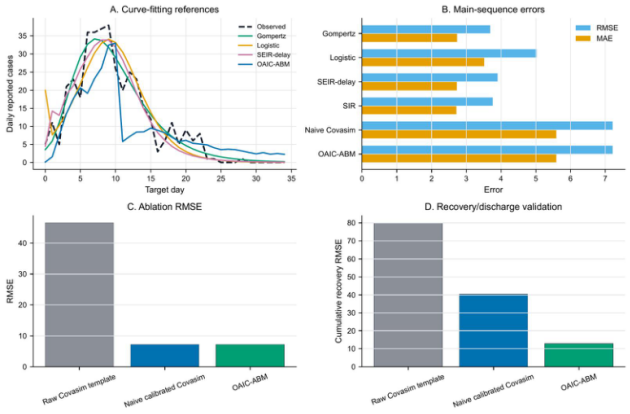

FIGURE 5. Shenzhen Baselines and Covasim Ablation

Covasim at 46.5474. Naive calibrated Covasim has the same 7.2072 value because it shares the fitted transmission trajectory. Recovery checks, stochastic intervals, the low-loss set, and contact scenarios are produced only by OAIC-ABM in this comparison.

SIR and SEIR-delay also track the daily target closely, with RMSE 3.7596 and 3.9006; Logistic gives 5.0235. Their total relative errors range from -3.29% to 0.46%, compared with -20.66% for OAIC-ABM. For reconstruction of the Shenzhen report curve alone, the aggregate models are the better choice. We retain the agent-based framework for analyses that need its simulated outputs.

The Covasim ablation confirms the effect of calibration. Raw Covasim has RMSE 46.5474, MAE 34.0381, and total relative error 159.54%; both calibrated variants reduce RMSE to 7.2072. Figure 5 and Table 6 document this numerical gain, while Sections 5.2, 5.5, and 5.6 use the observation mappings and retained simulation records.

TABLE 6. Shenzhen Fit Errors and Covasim Ablation

Model or variant	RMSE	MAE	Peak-day error	Total relative error (%)
Gompertz	3.687	2.734	-2	0.46
SIR	3.760	2.711	-1	-2.16
SEIR-delay	3.901	2.728	-1	-3.29
Logistic	5.024	3.520	0	-1.10
OAIC-ABM	7.207	5.585	1	-20.66
Raw Covasim template	46.547	34.038	25	159.54
Naive calibrated Covasim	7.207	5.585	1	-20.66

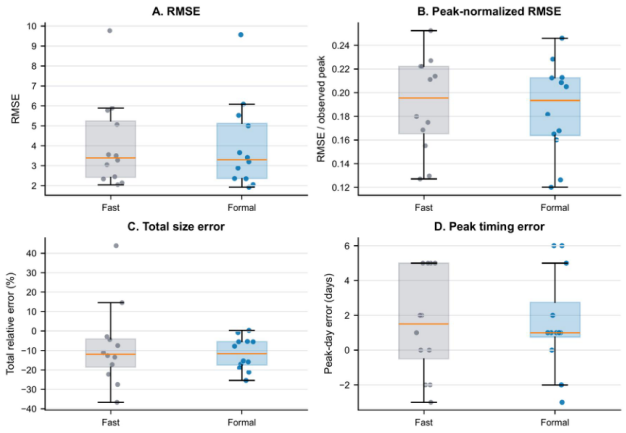

FIGURE 6. Error Distributions for the 12-City Formal Panel

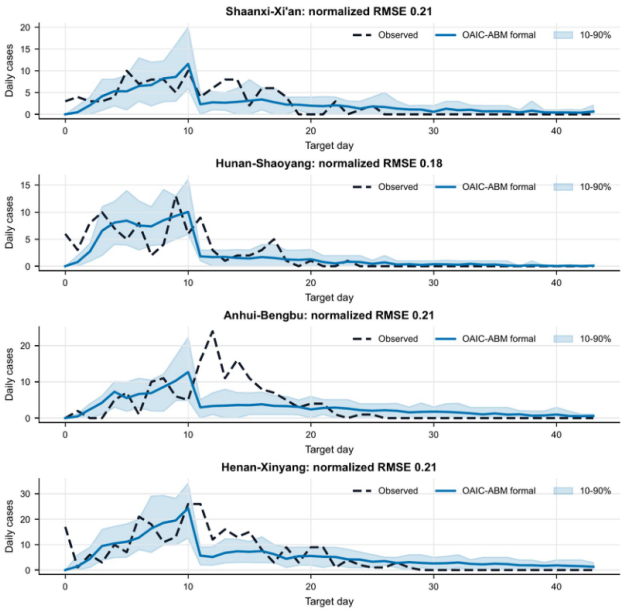

FIGURE 7. Representative Formal-Panel City Fits

D. FORMAL MULTICITY VALIDATION

The 12 independently calibrated non-Hubei city windows have median RMSE 3.3049 and median peak-normalized RMSE 0.1933. No beta or initial-infection estimate reaches a search boundary, and none of the pre-specified anomaly rules is triggered. These results support reuse of the calibration procedure within the defined first-wave benchmark.

Absolute RMSE reflects city scale. Wenzhou has the largest observed total and peak and the largest RMSE, 9.5731, yet its peak-normalized RMSE is 0.1651. Taizhou has RMSE 2.8794 and the lowest normalized value, 0.1200; Chengdu has the highest normalized value, 0.2461, still below 0.30. The paired metrics distinguish absolute miss from error relative to the city peak.

Peak timing varies more than normalized fit. Errors span - 3 to 6 days: Xinyang is exact, and Shenzhen, Ningbo, Taizhou, Shaoyang, and Shangqiu are within one day. Wenzhou and Chengdu each miss by six days, despite remaining inside the seven-day diagnostic threshold. The city-level calibrations do not reproduce peak timing uniformly.

The larger search budget changes the panel median only slightly: peak-normalized RMSE is 0.1955 in the fast panel and 0.1933 after 100 TPE trials and 30 repeats. The formal run is reported as the main result because it uses the larger city-level search and repeat budget. The distributions, representative trajectories, and city metrics appear in Figures 6 and 7 and Table 7.

E. LIGHTWEIGHT IDENTIFIABILITY ANALYSIS

Relative to M2P, median RMSE improves by 1.71% for M3P and 3.49% for M4P. Each richer variant nevertheless worsens one of the three cities. The two-parameter specification remains the main benchmark model because the added freedom gives no consistent city-level gain.

The city-level reversals affect the model choice. M3P worsens Shenzhen while improving Wenzhou and Chengdu; M4P improves Shenzhen and Chengdu but worsens Wenzhou. The median improvements therefore conceal a different ranking in each city.

E5 uses a lighter trial budget and only three cities, so M3P and M4P remain sensitivity variants. Their mixed city-level results leave the formal M2P protocol unchanged. Figure 8 and Table 8 show the comparisons.

Figure 8 plots RMSE and peak-normalized RMSE for each model variant in the three E5 cities. Shenzhen favors M4P, Wenzhou M3P, and Chengdu M3P on RMSE; no variant is best in all three cities. Table 8 gives the corresponding improvements relative to M2P.

TABLE 7. City-Level Metrics in the Formal Panel

Province	City	Total cases	Peak	RMSE	nRMSE peak	Peak error
Zhejiang	Wenzhou	494.0	58.0	9.573	0.165	6
Guangdong	Shenzhen	399.0	38.0	6.085	0.160	1
Henan	Xinyang	269.0	26.0	5.523	0.212	0
Anhui	Hefei	164.0	16.0	3.654	0.228	2
Anhui	Bengbu	159.0	24.0	5.004	0.209	-2
Zhejiang	Ningbo	151.0	27.0	3.411	0.126	1
Sichuan	Chengdu	128.0	13.0	3.199	0.246	6
Zhejiang	Taizhou	126.0	24.0	2.879	0.120	1
Shaanxi	Xi'an	115.0	10.0	2.051	0.205	5
Hunan	Shaoyang	100.0	13.0	2.361	0.182	1
Jiangsu	Suzhou	84.0	9.0	1.915	0.213	-3
Henan	Shangqiu	84.0	14.0	2.349	0.168	1

TABLE 8. E5 Fit Changes by City and Parameter Count

Province	City	Variant	Free parameters	RMSE	Improvement vs M2P (%)
Guangdong	Shenzhen	M2P	2	6.126	0.00
Guangdong	Shenzhen	M3P	3	6.774	-10.57
Guangdong	Shenzhen	M4P	4	5.769	5.83
Zhejiang	Wenzhou	M2P	2	10.229	0.00
Zhejiang	Wenzhou	M3P	3	10.054	1.71
Zhejiang	Wenzhou	M4P	4	10.368	-1.35
Sichuan	Chengdu	M2P	2	3.281	0.00
Sichuan	Chengdu	M3P	3	3.089	5.83
Sichuan	Chengdu	M4P	4	3.166	3.49

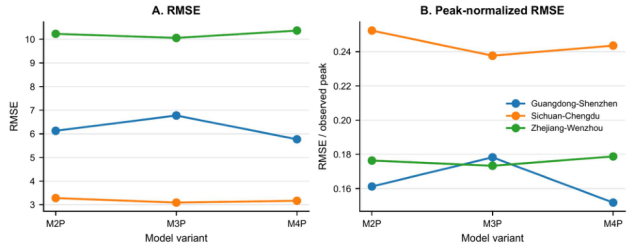

FIGURE 8. Fit Changes after Freeing Intervention Parameters

F. SCENARIO-SCREENING DEMONSTRATION

E7 retains E1R3A calibrations with RMSE no greater than 1.10 times the best city-specific value; S0 is the zero-additional-burden reference. Among the contact-reduction scenarios, S1 has the highest prevented cases per burden in Shenzhen and Wenzhou, whereas S7 prevents the most cases at higher assumed burden. The comparison is a simulation under the stated score.

For Shenzhen, mean total cases fall from 316.57 under S0 to 309.01 under S1. That difference is 7.56 cases and 90.71 prevented cases per burden unit. S7 lowers the mean to 283.22, preventing 33.35 cases at an efficiency of 60.64. Wenzhou has the same ordering: S1 prevents 8.55 cases at 102.57 per burden unit, whereas S7 prevents 37.44 at 68.08 per burden unit.

Absolute reduction and burden efficiency select different scenarios. S7 prevents more cases, whereas S1 achieves the higher reduction per burden unit. Uncertainty bands and nondominated frequency test whether this ordering persists across the low-loss calibration draws. Figure 9 reports the incremental comparison; Figure 10 gives total-case distributions.

Figure 9 relates expected prevented cases to the assumed burden score, while Figure 10 shows total-case spread across low-loss stochastic draws. The archived nondominated-frequency table records how often an alternative remains competitive without imposing a weighted score. Scenario definitions and incremental outcomes are listed in Tables 9 and 10.

VI. DISCUSSION

A. WHAT OAIC-ABM ADDS BEYOND CURVE FITTING

Table 6 changes the model choice according to the analysis goal. Gompertz, SIR, and SEIR-delay fit the Shenzhen report curve better than OAIC-ABM. Covasim becomes useful when the fitted curve must remain attached to stochastic trajectories, auxiliary disease outputs, and layer-specific contact scenarios. That requirement explains why we retained OAIC-ABM despite its larger pointwise error. Between-run intervals, recovery and death channels, and scenario-specific contact layers all come from the same calibrated simulation. The framework accepts a loss of short-window fit accuracy in exchange for those linked outputs.

The four-day recovery shift illustrates a separate source of improvement. It changes the comparison between simulated recovery and reported discharge but leaves the fitted transmission trajectory untouched. The lower recovery RMSE is an observation-mapping gain rather than a change in disease dynamics.

B. ROLE OF MULTICITY VALIDATION

Shenzhen verifies reproduction of the archived configuration. The 12-city panel goes further by recalibrating the same

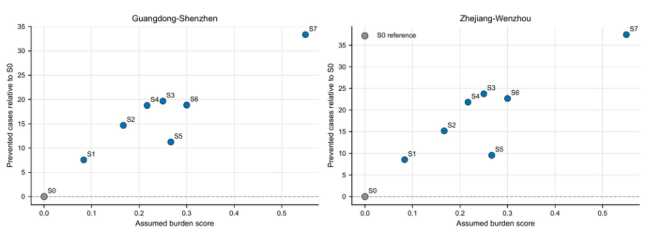

FIGURE 9. Incremental Scenario Trade-Offs Relative to S0

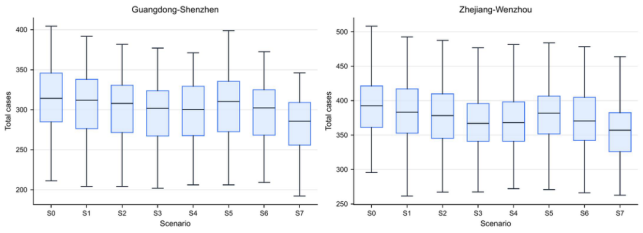

FIGURE 10. Total-Case Uncertainty across Selected Scenarios

TABLE 9. Contact Multipliers Used in the Scenario Grid

Scenario	Name	School	Work	Community
S0	Baseline calibrated intervention	1.0	1.0	1.0
S1	Mild community reduction	1.0	1.0	0.75
S2	Moderate community reduction	1.0	1.0	0.5
S3	Strong community reduction	1.0	1.0	0.25
S4	Work-focused reduction	1.0	0.5	0.85
S5	School-focused reduction	0.4	0.9	0.9
S6	Balanced moderate reduction	0.7	0.7	0.7
S7	Balanced strong reduction	0.45	0.45	0.45

TABLE 10. Incremental Scenario Outcomes Relative to S0

City	Scenario	Mean total cases	Assumed burden	Prevented cases	Cases per burden
Guangdong-Shenzhen	S0	316.57	0.00	0.00	
Guangdong-Shenzhen	S1	309.01	0.08	7.56	90.71
Guangdong-Shenzhen	S7	283.22	0.55	33.35	60.64
Zhejiang-Wenzhou	S0	394.82	0.00	0.00	
Zhejiang-Wenzhou	S1	386.27	0.08	8.55	102.57
Zhejiang-Wenzhou	S7	357.38	0.55	37.44	68.08

procedure under fixed screening rules in each reporting window. No fitted value reaches a parameter boundary, which removes one common sign that the search range is truncating the formal fits.

City selection preceded fitting and used administrative status and report shape. Each retained city was then calibrated independently. Median peak-normalized RMSE supports comparison across count scales, while Table 7 retains city-level variation. Shenzhen parameters were never transferred, and cities were selected before their errors were known.

The portability result applies to ordinary non-Hubei first-wave windows from one national reporting context. Within that cohort, the median error, boundary checks, and anomaly screen support reuse of the procedure. Later waves, other reporting systems, and other diseases remain untested.

C. IDENTIFIABILITY AND SHORT-WINDOW CALIBRATION

E5 favors the two-parameter protocol for this short aggregate window. M3P and M4P reduce median RMSE by 1.71% and 3.49%, but each worsens one city. A larger calibration space would need longer windows or dated intervention records to separate intervention effects from transmission and initial scale.

Parameter estimates are read together with their search ranges and low-loss neighborhoods. Boundary checks test whether the search limits truncate the optimum; E5 tests whether extra intervention freedom consistently lowers error. The scenario analysis carries the low-loss set forward instead of treating one best trial as exact.

The procedure is more portable than the fitted transmission coefficients. Each city is calibrated separately because reporting, population scaling, and realized contacts can all be absorbed into the coefficient. Cross-city comparisons of beta are outside the evidence provided here.

D. LIMITATIONS AND FUTURE WORK

The benchmark covers historical COVID-19 city reports. Its aggregate data contain publication dates, not individual event times, so confirmation and recovery/discharge delay remain observation assumptions. The scenario burden is an assumed comparison score. These two constraints define the scope of

the calibration and scenario results.

Fixing intervention day and multiplier keeps the main search small but compresses several behavioral and control changes into one transition [8-9]. Longer windows with dated intervention records could support piecewise or layer- specific effects. The uneven E5 gains suggest adding those parameters only when external timing information is available.

Future work could replace discrete confirmation shifts with an estimated delay distribution and allow recovery/discharge mapping to follow documented discharge practice. A hierarchical multicity model could share information while retaining city-specific parameters. These extensions require data beyond the current aggregate table.

The burden score permits relative comparison inside the scenario grid. Policy evaluation would require measured social, educational, mobility, and health-system costs with uncertainty. The present module links low-loss calibrations to multi-objective simulation outputs. Valuation of candidate actions requires a separate evidence base.

VII. CONCLUSION

We calibrated Covasim to dated public city reports and reproduced the Shenzhen result under a fixed stochastic protocol. The main empirical result is a trade-off: Gompertz and the compartment baselines fit the short Shenzhen curve more closely, while the calibrated agent-based model preserves stochastic trajectories, auxiliary outcomes, and contact-scenario outputs. Across 12 non-Hubei first-wave cities, independent recalibration produced median peak-normalized RMSE 0.1933 with no parameter-boundary hits. Added intervention freedom gave small, city-dependent gains in E5, so the two-parameter protocol remained the main specification. Scenario screening then carried low-loss fits into simulated case-burden comparisons under an assumed score. This city-report benchmark connects reproducible calibration with the simulation outputs needed for those downstream analyses. Transfer to other reporting systems, diseases, or policy evaluation requires evidence not contained in this dataset.

DECLARATION OF CONFLICTING INTERESTS

The author(s) declared no potential conflicts of interest with respect to the research, author-ship, and/or publication of this article.

DATA SHARING AGREEMENT

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

FUNDING

The author(s) received no financial support for the research, authorship, and/or publication of this article.

REFERENCES

[1] R. K. Nash, P. Nouvellet, A. Cori, et al., "Estimating epidemic dynamics from reporting-delay-adjusted incidence data," PLOS Computational Biology, vol. 19, no. 8, 2023, Art. no. e1011439.
[2] K. M. Gostic, L. McGough, E. Baskerville, et al., "Practical considerations for measuring the effective reproduction number in real time," Epidemics, vol. 38, 2022, Art. no. 100543.
[3] J. Ma, "Estimating epidemic exponential and sub-exponential growth dynamics in emerging outbreaks," Journal of Theoretical Biology, vol. 520, 2021, Art. no. 110638.
[4] C. C. Kerr, R. M. Stuart, D. Mistry, et al., "Covasim: an agent-based model of COVID-19 dynamics and interventions," PLOS Computational Biology, vol. 17, no. 7, 2021, Art. no. e1009149.
[5] K. Prem, K. Van Zandvoort, P. Klepac, et al., "Projecting social contact matrices in 152 countries using demographic data," Nature Communications, vol. 12, 2021, Art. no. 386.
[6] A. Aleta, D. Martin-Corral, A. Pastore y Piontti, et al., "Modeling the impact of testing, contact tracing and household quarantine," Nature Communications, vol. 13, 2022, Art. no. 642.
[7] K. Deb and H. Jain, "An evolutionary many-objective optimization algorithm using reference-point-based selection," IEEE Transactions on Evolutionary Computation, vol. 26, no. 2, pp. 417–430, 2022.
[8] P. Nouvellet, T. Ganyani, C. Colijn, et al., "Human mobility and intervention effects in epidemic spread modeling," Nature Reviews Physics, vol. 4, pp. 701–715, 2022.
[9] D. C. Adam, P. Wu, J. Y. Wong, et al., "Clustering and superspreading in COVID-19 transmission," Nature Reviews Microbiology, vol. 21, no. 2, pp. 103–118, 2023.
[10] Q. Bi, Y. Wu, S. Mei, et al., "Epidemiology and transmission of COVID-19 in Shenzhen, China," The Lancet Infectious Diseases, vol. 21, no. 1, pp. 1–10, 2021.
[11] W. O. Kermack and A. G. McKendrick, "A contribution to the mathematical theory of epidemics," Proceedings of the Royal Society A, vol. 115, no. 772, pp. 700–721, 1927.
[12] H. W. Hethcote, "The mathematics of infectious diseases," SIAM Review, vol. 42, no. 4, pp. 599–653, 2000.
[13] F. Brauer, "Mathematical epidemiology: Past, present, and future," Infectious Disease Modelling, vol. 6, pp. 603–615, 2021.
[14] H. J. Wearing, P. Rohani, and M. J. Keeling, "Appropriate models for infectious disease dynamics," PLOS Biology, vol. 20, no. 3, 2022, Art. no. e3001571.
[15] B. Bischl, M. Binder, and M. Lang, "Hyperparameter optimization and AutoML: A survey," Journal of Machine Learning Research, vol. 24, no. 1, pp. 1–50, 2023.
[16] T. Akiba, S. Sano, and T. Yanase, et al., "Optuna: a next-generation hyperparameter optimization framework," in Proc. KDD, 2019, pp. 2623–2631.
[17] DXY, "DXYArea.csv city-level COVID-19 public dataset (cumulative confirmed, recovered/discharged, deaths)," 2020.