Calibrating a Hypertension Model Across 20 Cities Using Open Data: Workflow and Parameter Identifiability

Submitted to ISDC 2026, Delft, Netherlands, 20-24 Jul 2026

Abstract. Adapting System Dynamics (SD) models from a single city to many cities remains a practical challenge: each city demands its own parameter set, data sources are fragmented across databases, and manual calibration does not scale. Full Bayesian methods offer rigorous uncertainty quantification but require specialized infrastructure and substantial computational cost. We present a standardized calibration workflow that strikes a balance between these extremes. The workflow uses multi-start optimization with exclusively publicly available data to calibrate a sex-stratified hypertension (HTN) cascade-of-care and cardiovascular disease (CVD) model across 20 cities in 15 countries. The multi-start design not only improves calibration quality but also reveals structural non-identifiability. The calibration contributes to a workflow that produces an initial, data-informed simulator for each city - sufficient to begin stakeholder engagement without requiring local survey data - and scales to additional cities with minimal adaptation.

Keywords: model calibration, parameter identifiability, multi-city deployment, hypertension, uncertainty quantification

1 Introduction

Hypertension affects an estimated 1.3 billion adults worldwide (WHO, 2023), yet fewer than one in four achieve blood pressure control (Zhou et al., 2021). Multi-city initiatives such as CARDIO4Cities (Reiker et al., 2023; Saric et al., 2026) are deploying data-driven, simulation-supported approaches to improve urban cardiovascular health across diverse settings. Computational simulation models, including System Dynamics (SD) models of the hypertension (HTN) cascade of care (Loo et al., 2026a), are increasingly used within such initiatives to project the population-level impact of intervention strategies and inform health policy. By representing the full cascade from undiagnosed prevalence through diagnosis, treatment, and control, these models allow decision-makers to compare intervention scenarios (e.g., expanding screening programs vs. improving treatment adherence) and project their downstream effects on cardiovascular outcomes, supporting policy making on resource allocation that cross-sectional data alone often cannot offer. When such models must be deployed across many cities, each with its own epidemiological profile, calibration becomes the primary practical bottleneck: each city demands a distinct parameter set, the number of observable indicators is small relative to the number of free parameters, and manual tuning does not scale.

Existing approaches to SD model calibration span a wide spectrum. At one end, standard practice calibrates a model using a mix of manual and optimization methods from one or a small number of starting points (Sterman, 2000, Ch. 21; Oliva, 2003), yielding a single “best” parameter set. Sensitivity analysis around this point estimate is recommended (isee systems, 2020) and provides local information about the objective function landscape, but does not reveal whether other, equally good parameter sets exist elsewhere in the space. For multi-site (e.g., multiple cities) calibration, this process needs to be repeated multiple times. At the other end, full Bayesian estimation samples the entire posterior distribution (Rahmandad et al., 2021; Andrade & Duggan, 2021), providing rigorous uncertainty quantification

and the ability to share information across comparable sites through hierarchical priors, but at the cost of specialized software infrastructure and substantial computational investment.

In this paper, we take a middle path. We calibrate a hypertension cascade-of-care and CVD model across 20 cities using 500 independent Latin Hypercube starting points per city (McKay et al., 1979), each optimized through a staged Powell procedure. This multi-start design serves two purposes: first, it increases the probability of finding the global optimum in a rugged objective-function landscape (Raue et al., 2009); second, it produces an ensemble of optimized solutions that, taken together, characterize the shape of the solution space, revealing which parameters are well-constrained by the data and which are not (Rahmandad & Sterman, 2012).

The 20 cities were selected in coordination with the CARDIO4Cities initiative, which is progressively engaging cities worldwide in simulation-supported cardiovascular health planning. The cities span 15 countries across five continents, representing diverse epidemiological profiles: from high-diagnosis, high-control settings to low-diagnosis, low-control settings. All calibrations use exclusively publicly available data: population demographics from the United Nations World Population Prospects 2024 (United Nations, 2024), hypertension cascade estimates from NCD-RisC (Zhou et al., 2021), and CVD burden from the Global Burden of Disease study (IHME, 2021), making the approach reproducible without access to local survey data.

2 Methods

2.1 Calibration workflow

The work presented in this paper is situated within a broader four-step workflow for arriving at a validated, city-specific parameterization (Figure 1):

- **Literature-informed parameter bounds.** The feasible range for each parameter is defined based on published epidemiological evidence and physiological plausibility. These bounds constrain the sampling and the optimizer's search space.
- **Data-based multi-start calibration.** Multiple starting points are optimized against historical data. All optimized solutions are retained and analyzed for parameter identifiability and compensation patterns.
- **Expert review.** Domain experts examine the compensating parameter pairs flagged by the ensemble diagnostics and resolve ambiguities using knowledge that aggregate time-series data cannot provide, for example, whether a high diagnosis rate / low treatment uptake combination or the reverse is more consistent with the local healthcare system.
- **Stakeholder engagement.** During engagement with local cities, health authorities and experts contribute city-specific data (e.g., screening program coverage, treatment protocols) and local estimates that are not available in global databases, further constraining the parameter space.

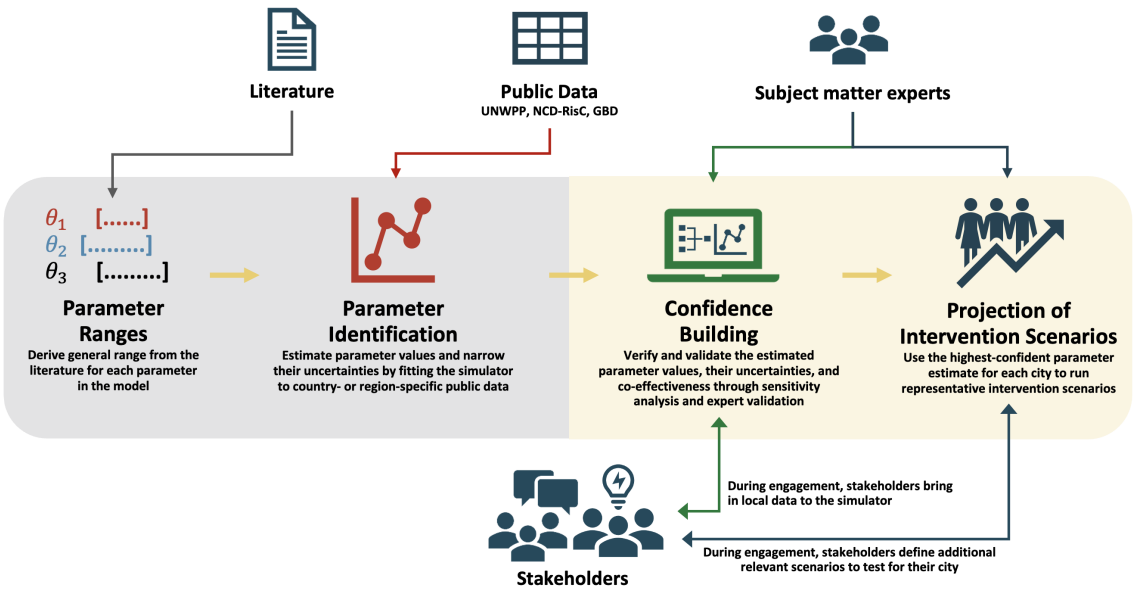

Figure 1: Four-step calibration workflow — from literature-informed bounds through data calibration, expert review, and stakeholder engagement.

This paper focuses on steps 1 and 2. Steps 3 and 4 are being carried out as part of the CARDIO4Cities initiative, where the calibrated simulators are used to support city-level intervention planning in collaboration with local health authorities (Saric et al., 2026). The identifiability diagnostics produced by step 2 directly inform what step 3 should focus on, making the computational analysis a prerequisite for efficient expert review rather than an end in itself.

2.2 Model and data

The model used in this study is a continuous-time, sex-stratified stock-and-flow representation of the adult (30+) hypertension cascade of care (Loo et al., 2026b). The cascade tracks individuals through normotension, pre-hypertension (undiagnosed, diagnosed, treated), and hypertension stages (undiagnosed, diagnosed, treated, controlled, and treatment dropout). Downstream, the model represents CVD events (both first-time and recurrent) and CVD-attributable mortality. The population module tracks 5-year age cohorts (30–94) by sex, with fertility, aging, migration, and non-CVD mortality flows driven by UN demographic projections.

38 parameters per city characterize the model's behavior. These include pre-hypertension onset, progression, and regression rates; hypertension diagnosis, treatment uptake, control, and dropout rates; CVD risk modifiers; and initial-condition ratios that partition the population across cascade stages at the simulation start (2011). Parameter bounds are informed by published epidemiological evidence: hypertension incidence and pre-hypertension progression rates draw on cohort studies (Pereira et al., 2012; Kim et al., 2011; Vasan et al., 2001), CVD risk parameters on the Framingham risk framework (D'Agostino et al., 2008), and treatment adherence rates on clinical follow-up data (Park et al., 2011).

Three publicly available, global data sources provide the calibration targets and model inputs:

- **UN World Population Prospects 2024** (United Nations, 2024): age- and sex-specific population counts, fertility rates, and mortality rates by country, used as exogenous time-series inputs to the population module.

- **NCD-RisC** (NCD Risk Factor Collaboration; Zhou et al., 2021): age- and sex-specific estimates of hypertension prevalence and the proportions diagnosed, treated, and controlled, for 200 countries from 1990 to 2019.
- **Global Burden of Disease** (IHME, 2021): age- and sex-specific estimates of CVD incidence and CVD-attributable mortality by country.

Country-level estimates are scaled to city populations using the ratio of the city's total population to the country's total population, obtained from national census or municipal boundary reports. Rates and proportions (e.g., hypertension prevalence, cascade ratios) are applied directly at country level, while absolute counts (population, CVD cases, deaths) are scaled by this ratio.

2.3 Multi-stage calibration design

For each of the 20 cities, we generate 500 starting points via Latin Hypercube Sampling (LHS) with a fixed random seed (seed = 99) across the feasible parameter region defined by parameter-specific bounds. The same LHS design (same seed, same bounds) is used for all 20 cities, so that differences in calibration outcomes reflect genuine city-specific epidemiological differences rather than differences in starting-point coverage. The payoff function is the log-scale sum of squared errors (log-SSE) computed across 16 indicators: 12 HTN cascade stocks (total, undiagnosed, diagnosed, treated, controlled, and dropout populations, each for Male and Female) and 4 CVD indicators (CVD incidence and CVD death rate, each for Male and Female).

Each starting point is then optimized using Powell's conjugate direction method through five sequential stages aligned to the model's causal structure:

Table 1. Calibration design by stages.

Stage	Parameter group	Parameters	Max evaluations
1	Pre-HTN prevalence and onset rates	16	150
2	HTN cascade flow rates (diagnosis, treatment, control, dropout)	12	100
3	CVD risk and death rates	14	100
4	Initial cascade ratios	4	50
5	Joint tuning (all parameters unfrozen)	38	1,600
	Total		2,000

Note: Stages share some parameters; unique total is 38.

The motivation for staged calibration is twofold. First, Powell's conjugate direction method cycles through the full parameter set sequentially, so that within any given evaluation budget, each parameter receives a limited number of update opportunities. When all 38 parameters are optimized simultaneously under a constrained budget – as is typical in practice – there is a risk that the optimizer does not explore each parameter's dimension sufficiently. Partitioning parameters into smaller groups allows more iterations per parameter within each stage, improving exploration of the local landscape and reaching good fit earlier in the evaluation budget.

Second, the model's cascade structure lends itself naturally to sequential estimation. The hypertension cascade is largely feedforward: the upper flow (prevalence, onset) determines the population entering the downstream stocks (diagnosed, treated, controlled, and CVD), but the downstream stocks do not strongly feed back to the upper flow — for example, CVD deaths do not substantially affect the total population. This allows us to first match the prevalence of hypertension (Stage 1), then distribute that population across cascade stages (Stage 2), then fit CVD outcomes conditional on the cascade populations (Stage

3) and finally adjust initial stock ratios (Stage 4). Each later stage builds on a well-fitted preceding stage, ensuring that the optimizer always starts on solid ground. Stage 5 (joint tuning) then unfreezes all 38 parameters to accommodate any misalignments — since the best fit for an upstream stage may not be globally optimal once downstream parameters are included.

This staged approach does not equally generalize to all SD models: it relies on the ability to identify meaningful, loosely coupled parameter groups, which is straightforward in feedforward cascade structures but more difficult in models with rich feedback loops. For the present hypertension model, however, the cascade architecture provides a natural decomposition.

To verify that staging improves convergence without sacrificing fit quality, we compared the 5-stage design against a joint-only baseline that optimizes all 38 parameters simultaneously from the start, using the same 500 LHS starting points and the same total evaluation budget (2,000 evaluations). Figure 2 shows convergence curves (median payoff of the top-50 solutions vs. cumulative evaluations) for all 20 cities. Two patterns are visible. First, the staged design (red) reaches lower payoff values earlier: by evaluation 400 (the end of Stage 4), the staged curve is already well below the joint-only curve (blue) in all 20 cities, confirming that decomposing the 38-parameter problem into smaller, causally ordered subproblems accelerates convergence. Second, by the end of the 2,000-evaluation budget, the two curves converge to comparable payoff levels – the joint-only design eventually catches up but does not achieve a better final fit. This indicates that staging improves speed without compromising quality: the same fit is reached faster.

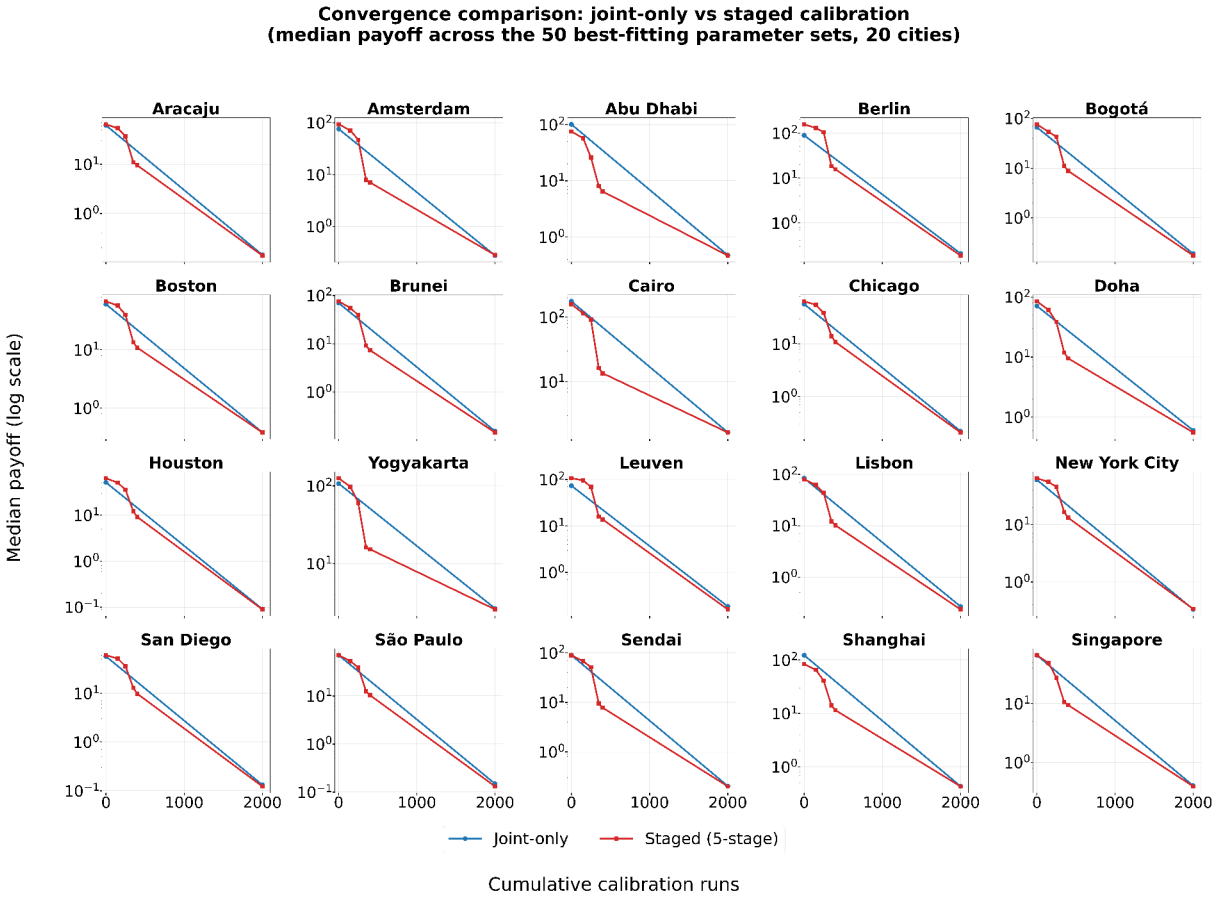

Figure 2: Convergence comparison — joint-only (blue) vs staged (red) calibration across 20 cities. Each curve shows the median payoff (log scale) across the 50 best-fitting parameter sets as a function of cumulative function evaluations.

3 Results

We report two main findings. First, the calibration pipeline produces good fits across all 20 cities: the simulated HTN cascade trajectories closely reproduce the historical data. Second, the multi-start design reveals that many model parameters are structurally non-identifiable, i.e., the same compensating parameter pairs recur across all or most of the 20 cities, indicating that data alone cannot uniquely determine some individual parameter values even when the aggregate fit is good. This second finding has direct implication for how calibration results should be interpreted and communicated to public health stakeholders.

3.1 Calibration quality

All 20 cities achieved calibration results where the simulated cascade trajectories are broadly consistent with the historical data. We assess fit quality using the Theil inequality coefficient (U), a normalized measure of simulation-to-data discrepancy that ranges from 0 (perfect fit) to 1 (maximum error), and its decomposition into bias (UM), variation (US), and covariation (UC) components (Sterman, 2000). Table 2 shows these metrics for Houston as a representative example; all 16 indicators achieve Theil U < 0.015, with a median of 0.008. The decomposition reveals that different indicator groups have different error profiles: the total HTN population stocks are bias-dominated (UM = 0.65–0.77), the cascade flow stocks (undiagnosed, diagnosed, treated) are variation-dominated (US = 0.51–0.63), and the CVD indicators are covariation-dominated (UC = 0.87–0.90). In all cases, the absolute error remains small (all MAPE < 2%). The same goodness-of-fit assessment was performed for all 20 cities, with similar patterns.

Table 2. Theil inequality statistics for Houston (best-fit parameter set).

Indicator	Theil U	MAPE	Bias UM	Variation US	Covariation UC
Total HTN [M]	0.002	0.3%	0.77	0.01	0.22
Total HTN [F]	0.002	0.3%	0.65	0.08	0.26
Undiagnosed [M]	0.006	0.9%	0.00	0.51	0.49
Diagnosed [M]	0.014	1.9%	0.03	0.55	0.42
Treated [F]	0.003	0.4%	0.16	0.63	0.21
Controlled [M]	0.011	1.8%	0.09	0.28	0.62
CVD incidence [M]	0.003	0.5%	0.08	0.02	0.90
CVD deaths [M]	0.007	1.1%	0.00	0.13	0.87
…					

Note: Selected indicators shown. Full table: 16 indicators, all Theil U < 0.015. MAPE = mean absolute percentage error.

3.2 Parameter non-identifiability

Despite achieving comparable aggregate fit across the 500 solutions, the underlying parameter values are diverse. We use Houston as a representative example to investigate this diversity, then confirm the findings across all 20 cities.

Rather than relying on a single best-fit parameter set, we retain the top-50 solutions (ranked by final payoff) as an ensemble of "behavioral" parameter sets in the sense of the Generalized Likelihood Uncertainty Estimation (GLUE) framework (Beven & Binley, 1992; Nott et al., 2012), which treats all parameter sets achieving an acceptable fit as equally plausible. We begin by examining the distribution of these top-50 parameter sets. Figure 3 shows the normalized parameter values (0 = lower bound, 1 = upper bound of the configured feasible range) for each of the 38 parameters, sorted by inter-quartile

range (IQR). At the top of the figure, parameters such as cvBETA and the CVD event death rates cluster tightly. The historical data effectively constrain their values. Moving down, however, the spread increases. At the bottom, several parameters span nearly the entire feasible range: the initial ratio for post-CVD uncontrolled population and the CVD-HTN ratio both reach an IQR of 1.00, meaning the top-50 solutions disagree on these parameters almost completely. In total, 10 of the 38 parameters show an IQR above 0.5. This indicates that, for a substantial share of parameters, the calibration data alone cannot distinguish between very different values – they are not individually identifiable from the available time-series data.

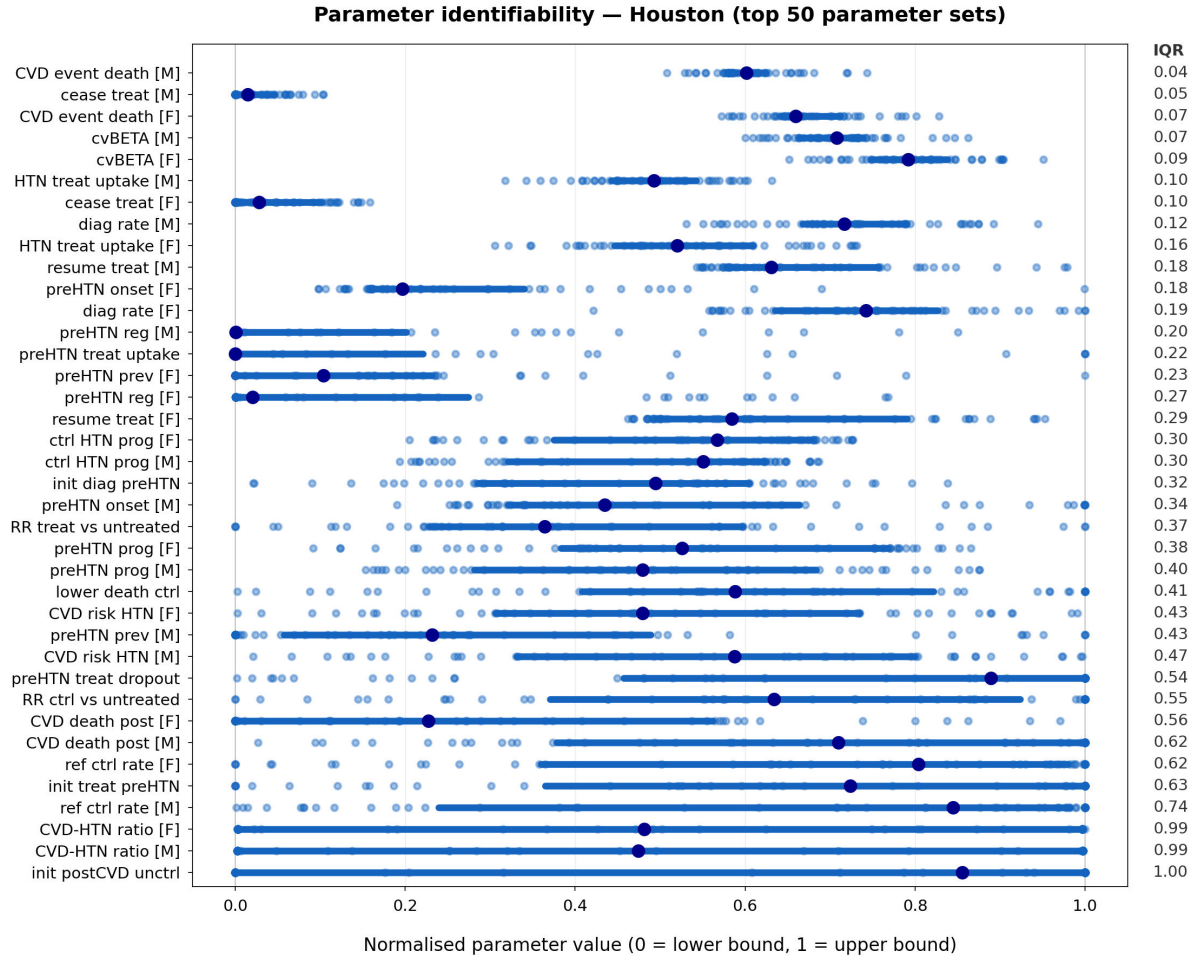

Figure 3: Parameter identifiability for Houston — top-50 solutions, sorted by IQR (tightest at top, widest at bottom).

However, this is not the full story. Many of these poorly identified parameters are not free-floating independently: they compensate each other in pairs. Figure 4 shows scatter plots of eight correlated parameter pairs among the top-50 Houston solutions, with male (blue circles) and female (red triangles) overlaid in the same panel. The pairs are ordered from strongest to weakest correlation ($|\rho|$ ranging from 0.97 to 0.43); the threshold is set at $|\rho| \geq 0.43$ to include eight structurally distinct pair types for visualization. For example, the cease treatment rate and resume treatment rate show a tight positive correlation (ρ = +0.95 for Female, +0.92 for Male): when one increases, the other increases proportionally, keeping the net dropout flow nearly constant. Similarly, the diagnosis rate and HTN treatment uptake compensate each other (ρ = +0.84 for Female, +0.67 for Male) – a higher diagnosis rate paired with lower treatment uptake can produce the same number of treated patients as the reverse. In these cases, the data constrain

the ratio or difference between paired parameters, even though each parameter individually spans a wide range. The parameters are identifiable collectively, just not individually.

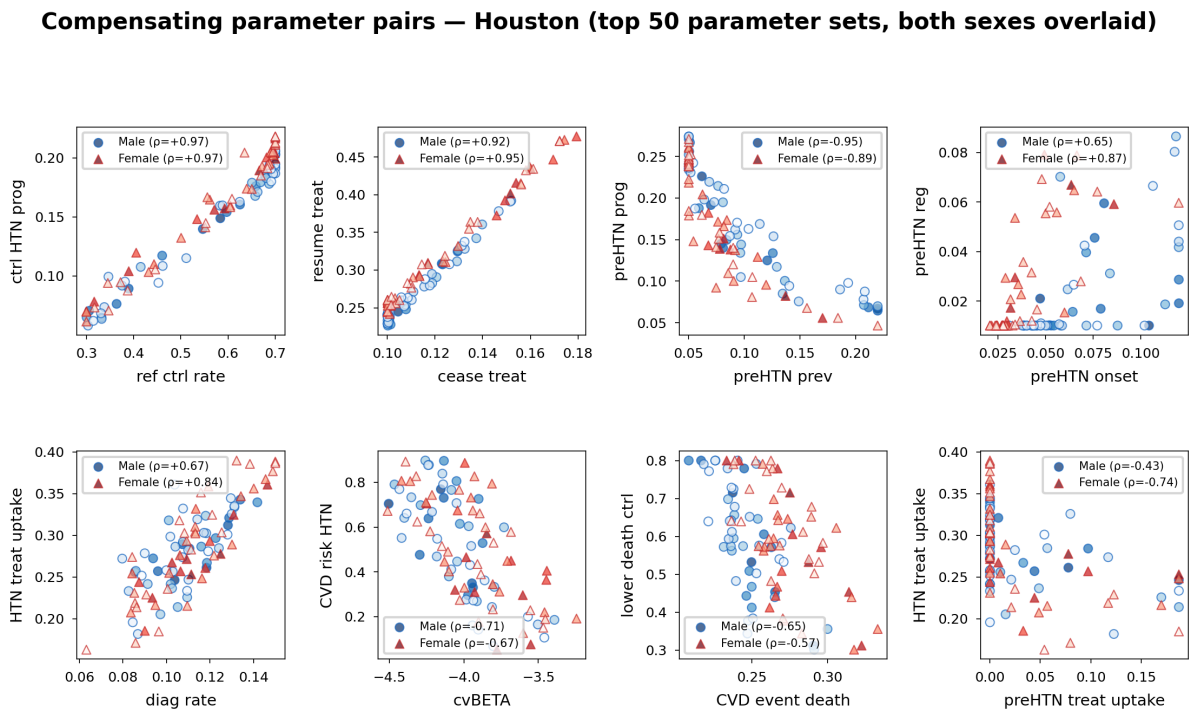

Figure 4: Compensating parameter pairs in Houston — top-50 solutions, both sexes overlaid (blue circles = Male, red triangles = Female). Color intensity indicates payoff (darker = better fit). Eight structurally distinct pair types shown, ordered from strongest to weakest correlation.

To map the full pairwise correlation structure, Figure 5 shows a dendrogram of all 38 parameters, where the branching height indicates the correlation distance ($1 - |\rho|$). Pairs that merge at the lowest levels, such as cease treatment and resume treatment, control rate and controlled HTN progression, and prevalence and progression of pre-hypertension, are the most tightly coupled. Dashed reference lines at $|\rho| = 0.9$, 0.7, and 0.5 help the reader gauge the strength of each cluster. This dendrogram provides a complete picture of the correlation structure that the individual scatter plots in Figure 4 illustrate for selected pairs.

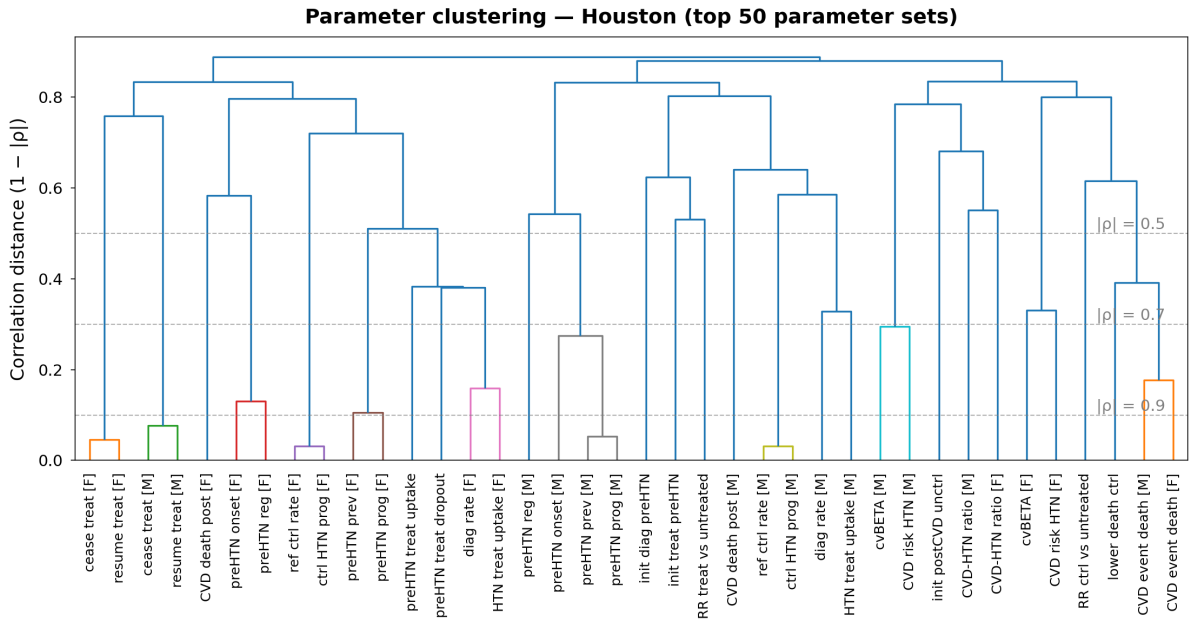

Figure 5: Parameter clustering by correlation distance (Houston, top-50 solutions). Dashed lines indicate $|\rho| = 0.9$, 0.7, and 0.5 thresholds.

A natural question is whether these compensating pairs are specific to Houston or reflect a more general property. Figure 6 addresses this by counting, for each parameter pair with $|\rho| > 0.5$, in how many of the 20 cities this compensation occurs among the top-50 solutions. The top pairs, cvBETA ↔ CVD risk HTN, preHTN prevalence ↔ preHTN progression, diagnosis rate ↔ HTN treatment uptake, and cease treatment ↔ resume treatment appear in all or nearly all 20 cities. This consistency across epidemiologically diverse settings indicates that the compensation is structural: it arises from the cascade architecture itself (e.g., two sequential flow rates whose product determines the net flow into a stock) rather than from any particular city's data.

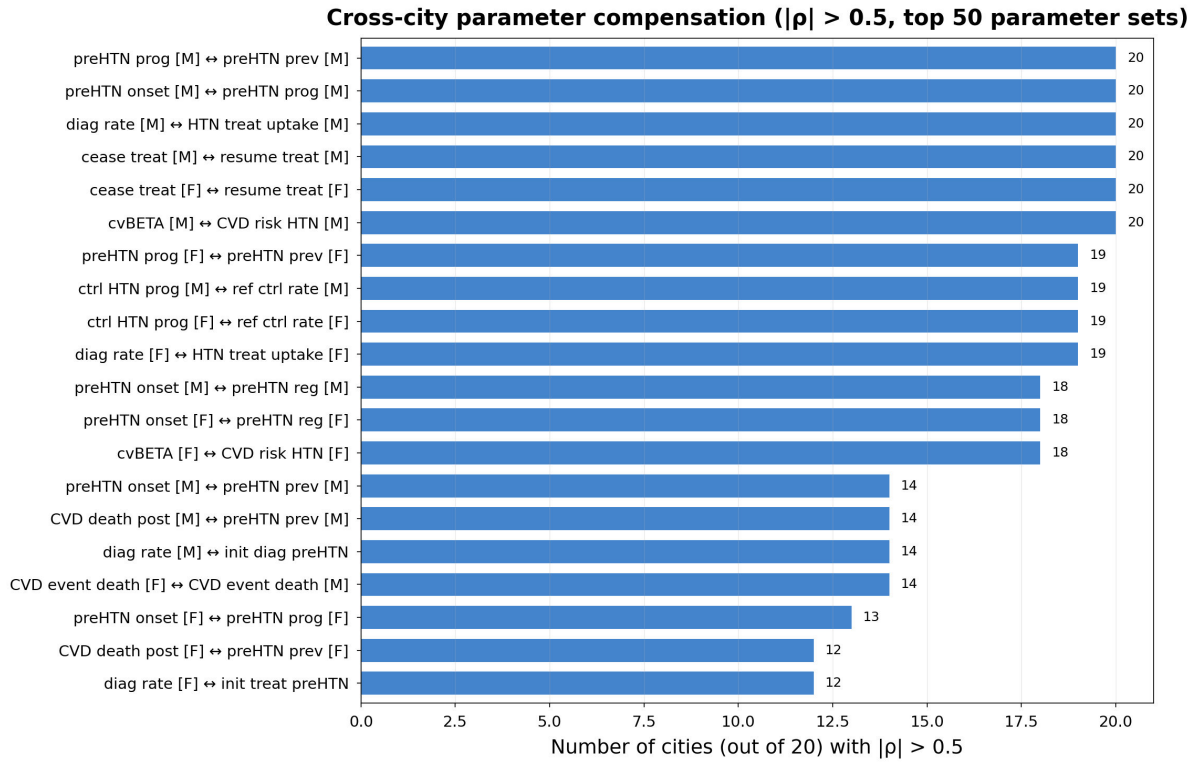

Figure 6: Cross-city parameter compensation — number of cities (out of 20) in which each parameter pair shows |ρ| > 0.5 among the top-50 solutions.

This cross-city consistency has a practical implication for multi-city calibration strategies. In settings where parameter non-identifiability is data-specific (i.e., different parameters are poorly identified in different cities), one can leverage cross-city coupling which shares information across cities to narrow each city's uncertainty, as demonstrated by Rahmandad et al. (2021) for COVID-19 behavioral parameters across 92 nations. However, the present results show that the same parameter pairs compensate in the same way across all 20 cities. Cross-city coupling would therefore provide limited additional constraint on these structurally non-identifiable pairs, because no city's data resolves the ambiguity that the model structure creates.

This highlights a boundary of what calibration data alone can achieve: the compensating pairs are a structural property of the model, not a data-specific artifact. We discuss the implications of this finding, including for cross-country calibration strategies, in Section 4. Since these calibrated models are actively being used to support city-level intervention planning within the CARDIO4Cities initiative, resolving these ambiguities through expert review (step 3) and stakeholder engagement (step 4) becomes a natural and necessary next stage of the workflow.

4 Discussion

4.1 Calibration at scale using open data

The results demonstrate that a single SD model architecture, combined with publicly available global data and a multi-start calibration pipeline, can produce city-specific parameterizations across 20 diverse settings. The calibration pipeline requires no local survey data, manual tuning, or city-specific model modifications; only the city's population size and country assignment are needed to generate a complete set of inputs. This is relevant for programs seeking to prioritize hypertension interventions across many settings simultaneously (WHO, 2023).

4.2 The shape of the solution space

The multi-start design reveals that the objective-function landscape is characterized by extended valleys rather than isolated minima, which means the "best" parameter set out of 500 is not meaningfully superior to the 50th-best (Raue et al., 2009; Rahmandad et al., 2015, Ch. 2). The contribution of the present analysis is the empirical documentation of this phenomenon across 20 cities: the same compensating pairs and identifiability patterns recur across epidemiologically diverse settings, suggesting that the non-identifiability is structurally rooted in the cascade architecture rather than an artifact of any particular city's data.

However, the cross-city consistency may also partly reflect the nature of the data sources. The calibration targets (NCD-RisC and GBD) are not raw survey observations but outputs of statistical models that impose their own structure. NCD-RisC, for instance, uses a hierarchical Bayesian model that borrows information across countries (Zhou et al., 2021), so the "data" for different countries are correlated by construction. The apparent universality of compensating pairs therefore has two contributing sources – our SD model's cascade architecture and the shared structure of the upstream data-generating models – which we cannot cleanly separate.

This has implications for cross-country calibration strategies. Methods that leverage cross-country coupling to reduce per-country uncertainty (e.g., Rahmandad et al., 2021) assume that each country's data provides independent information. When targets are derived from shared hierarchical models, this independence is weakened: additional cities provide diminishing returns because the data already encodes cross-country information borrowing. More broadly, when calibrating to model-derived data across multiple sites, cross-site consistency in parameter estimates should be interpreted with caution – it may reflect the upstream models' priors as much as the underlying epidemiological reality.

4.3 Public health implications

From a public health perspective, the calibration workflow enables simulation-based decision support across multiple cities: each city receives a calibrated simulator that can project how changes in screening, treatment, or control programs would affect hypertension and CVD outcomes at the population level. This enables an initial, evidence-informed conversation with local health authorities even in settings where detailed local data are not yet available. The identifiability diagnostics further strengthen this conversation: by transparently communicating which model parameters are well-constrained and which are not, stakeholders can understand where the model's projections are robust and where additional local data or expert judgment would improve confidence.

4.4 Practical implications

First, staged calibration improves convergence but does not resolve non-identifiability. Our five-stage design, which activates parameters in causal order, helps the optimizer avoid poor local minima and improves the fraction of starts that converge to well-fitting solutions. However, it does not narrow the ridges in parameter space, because these ridges are a property of the model structure, not of the optimization algorithm.

Second, projection uncertainty depends on which parameters drive the scenario. If a policy scenario primarily affects a well-identified parameter (e.g., the control rate, which is directly constrained by observed data), then projections from different well-fitting parameter sets will be similar. If the scenario affects a poorly identified parameter (e.g., the pre-HTN onset rate), then the 500-solution ensemble should be propagated through the scenario to capture the resulting projection spread (the full projection landscape). This indicates a future research direction which studies the relationship between scenario uncertainty and parameter identifiability.

Third, parameter identifiability diagnostics are inexpensive to compute from multi-start results. The IQR ranking (Figure 3) and Spearman correlation matrix require only the stored parameter values and payoffs from the multi-start ensemble with no additional simulations needed. We suggest including these diagnostics as a reporting element alongside calibration fit metrics in SD modeling studies that employ multi-start optimization.

4.5 Limitations

We identify three limitations of the present work. First, the current analysis covers an initial batch of 20 cities, constrained by available time and resources. The NCD-RisC dataset provides hypertension cascade estimates for approximately 200 countries and regions, and the workflow is designed to accommodate further expansion. However, expanding to more cities does not guarantee improved parameter identification: the data sources are themselves model-generated with cross-country correlations encoded via hierarchical Bayesian methods, so additional cities provide diminishing returns on parameter constraint (see Section 4.2).

Second, the 500-start LHS design provides better coverage of the parameter space than single- or few-start calibration, but it is not equivalent to a formal posterior distribution. The top-K ensemble is an approximation whose relationship to the true posterior depends on assumptions about the likelihood function (the GLUE framework; Beven & Binley, 1992; Nott et al., 2012). Full Bayesian estimation (e.g., Rahmandad et al., 2021; Andrade & Duggan, 2021) would provide more rigorous uncertainty quantification but at substantially higher computational cost.

Third, the city-level scaling assumes that country-level hypertension rates apply within each city, which may not hold for cities with demographics that differ systematically from the national average. Subnational data, where available, could improve calibration targets. This limitation is specific to the current stage, which focuses on standardizing the multi-city calibration workflow so that a public data-informed simulator can facilitate initial engagement with local stakeholders, from whom subnational data and region-specific expert knowledge can be obtained to further contextualize the simulator. In the meantime, we acknowledge that this limitation and its impact on projection scenarios must be transparently communicated to users and stakeholders.

4.6 Future work

We identify two directions for future research. First, the present work relies on the goodness of fit of simulation data to historical data – the sum of log-transformed squared errors subject to parameter bounds (Sterman, 2000; Oliva, 2003). This treats all parameter combinations that achieve equal fit as equivalent. However, modelers routinely possess domain knowledge that distinguishes plausible from implausible parameter sets: biological rates should not oscillate widely year-to-year, and male and female rates for the same physiological process should not diverge by orders of magnitude. In full Bayesian estimation, this knowledge is encoded as prior distributions (Andrade & Duggan, 2021; Rahmandad et al., 2021). However, Bayesian methods add substantial modeling overhead as they require the specification of (often hierarchical) priors and incur high computational cost to sample the posterior distribution. For the majority of SD practitioners using standard optimization techniques, a useful middle path could be to add penalty terms to the calibration objective function that encode specific domain knowledge as soft constraints. We have explored this direction with a sex-coupling penalty term, encoding the prior knowledge that male and female rates for the same process should not diverge excessively. In preliminary experiments on Houston, a sex-coupling penalty weight of 1.0 reduced the sex divergence by 16% (from 0.28 to 0.24, measured as the absolute normalized difference |theta_M - theta_F| / range averaged across all sex-paired parameters and top-50 solutions) while increasing the HTN+CVD payoff by only 9% (from 0.64 to 0.70). This suggests that tighter sex coupling can be achieved without substantial loss of fit.

However, penalty terms have technical and methodological limitations. Technically, the weight of a penalty term relative to the log-SSE is set without a clear, direct link to domain knowledge, and the working solution may lack a principled interpretation. Methodologically, penalty terms do not carry a formal meaning of uncertainty as in the full Bayesian framework. More comprehensive empirical study is needed to characterize and assess the usefulness of penalty terms in comparison with full Bayesian methods.

Second, we have identified that some parameters in the hypertension model are structurally non-identifiable and recommend that such diagnostics be routinely reported in model calibration studies. Beyond documenting which parameters compensate and to what extent, it is also informative to explain why, that is, to provide a structural explanation of parameter non-identifiability. Some cases are straightforward, such as two sequential flow rates whose product determines the net flow into a stock, while others may be more implicit. An investigation into these underlying structural causes would help modelers pre-assess their model's identifiability properties before collecting historical data and be more aware of the model's limitations when communicating results.

5 Conclusions

We calibrated a sex-stratified hypertension cascade-of-care and CVD model across 20 cities using 500 Latin Hypercube starting points per city, staged Powell optimization, and exclusively publicly available data. The calibration achieved good fit across all cities (all Theil U < 0.015, all MAPE < 2%).

The multi-start design revealed that the parameter space exhibits structural non-identifiability: compensating parameter pairs, such as cease/resume treatment rates and diagnosis/treatment uptake rates, recur across all 20 cities, indicating that the non-identifiability is rooted in the model's cascade architecture rather than in any particular city's data. This finding has two practical implications. First, reporting only the best-fit parameter set from a single optimization run overstates the confidence in individual parameter values; multi-start diagnostics (IQR rankings, correlation dendrograms) should accompany calibration results. Second, cross-city parameter coupling provides limited additional constraint on structurally non-identifiable pairs, because the same pairs compensate everywhere.

We propose a four-step workflow, including literature-informed bounds, data-based multi-start calibration, expert review, and stakeholder engagement, in which the identifiability diagnostics from step 2 directly inform which parameter pairs require expert judgment in step 3. The calibration pipeline, data preparation scripts, and all diagnostic tools are available for adaptation to additional cities.

Acknowledgements

This research was funded by Novartis Foundation as part of the CARDIO4Cities initiative. We gratefully acknowledge their support.

Calculations were performed at sciCORE (http://scicore.unibas.ch/), the scientific computing center at the University of Basel.

References

Andrade, J. & Duggan, J. (2021). A Bayesian approach to calibrate system dynamics models using Hamiltonian Monte Carlo. System Dynamics Review, 37(4), 283–309.

Beven, K. & Binley, A. (1992). The future of distributed models: Model calibration and uncertainty prediction. Hydrological Processes, 6(3), 279–298.

D'Agostino, R.B., Vasan, R.S., Pencina, M.J., et al. (2008). General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation, 117(6), 743-753.

Institute for Health Metrics and Evaluation (IHME). (2021). Global Burden of Disease Study 2021 (GBD 2021) Results. Seattle: IHME. https://vizhub.healthdata.org/gbd-results/

isee systems. (2020). Stella Architect: Sensitivity analysis and optimization. Technical documentation. https://www.iseesystems.com/

Kim, S.J., Lee, J., Nam, C.M., et al. (2011). Progression rate from new-onset pre-hypertension to hypertension in Korean adults. Circulation Journal, 75(1), 135-140.

Loo, P. S., Silveira, M., Baxter, Y. C., Avezum, Á., Drager, L. F., Bortolotto, L. A., Boch, J., & Cobos Muñoz, D. (2026b). Hypertension management in São Paulo: Insights from a computational simulation approach. The Lancet Regional Health – Americas. Advance online publication. https://doi.org/10.1016/j.lana.2026.101449

Loo, P.S., Socha, A., Silveira, M., Baxter, Y.C., Avezum, Á., Drager, L.F., … & Cobos Munoz, D. (2026a). Mapping the Dynamic Complexity of Hypertension Management in São Paulo, Brazil. International Journal of Public Health, 70, 1608895.

McKay, M.D., Beckman, R.J., & Conover, W.J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics, 21(2), 239–245.

Nott, D.J., Fan, Y., Marshall, L., & Sisson, S.A. (2012). Generalized likelihood uncertainty estimation (GLUE) and approximate Bayesian computation: What's the connection? Water Resources Research, 48(12), W12602.

Oliva, R. (2003). Model calibration as a testing strategy for system dynamics models. European Journal of Operational Research, 151(3), 552–568.

Park, Y., Kim, H., Jang, S., & Koh, C. (2011). Predictors of adherence to medication in older Korean patients with hypertension. European Journal of Cardiovascular Nursing, 12(1), 17-24.

Pereira, M., Lunet, N., Paulo, C., et al. (2012). Incidence of hypertension in a prospective cohort study of adults from Porto, Portugal. BMC Cardiovascular Disorders, 12, 114.

Rahmandad, H. & Sterman, J.D. (2012). Reporting guidelines for simulation-based research in social sciences. System Dynamics Review, 28(4), 396–411.

Rahmandad, H., Lim, T.Y., & Sterman, J.D. (2021). Behavioral dynamics of COVID-19: estimating underreporting, multiple waves, and adherence fatigue across 92 nations. System Dynamics Review, 37(1), 5–31.

Rahmandad, H., Oliva, R., & Osgood, N.D. (Eds.) (2015). Analytical Methods for Dynamic Modelers. MIT Press.

Raue, A., Kreutz, C., Maiwald, T., et al. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics, 25(15), 1923–1929.

Reiker, T., Des Rosiers, S., Boch, J., et al. (2023). Population health impact and economic evaluation of the CARDIO4Cities approach to improve urban hypertension management. PLOS Global Public Health, 3(4), e0001480.

Saric, J., Aerts, A., Barboza, J., et al. (2026). CARDIO4Cities: from global evaluation framework to local monitoring in Dakar and Sao Paulo. BMC Public Health. https://doi.org/10.1186/s12889-026-26710-z

Sterman, J.D. (2000). Business Dynamics: Systems Thinking and Modeling for a Complex World. McGraw-Hill.

United Nations. (2024). World Population Prospects 2024. https://population.un.org/wpp/

Vasan, R.S., Larson, M.G., Leip, E.P., et al. (2001). Assessment of frequency of progression to hypertension in non-hypertensive participants in the Framingham Heart Study. The Lancet, 358(9294), 1682-1686.

World Health Organization. (2023). Global report on hypertension: the race against a silent killer. WHO, Geneva.

Zhou, B., Carrillo-Larco, R.M., Danaei, G., et al. (NCD-RisC). (2021). Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: a pooled analysis of 1201 population-representative studies with 104 million participants. The Lancet, 398(10304), 957–980.