

Comparing Data Assimilation and Likelihood-Based Inference on Latent State Estimation in Agent-Based Models

Blas Kolic¹, Corrado Monti², Gianmarco De Francisci Morales², and Marco Pangallo²

¹Universidad Carlos III, Ronda de Toledo, 1, 28005, Madrid, Spain

²CENTAI, Corso Inghilterra, 3, 10138, Turin, Italy

Abstract

In this paper, we present the first systematic comparison of Data Assimilation (DA) and Likelihood-Based Inference (LBI) in the context of Agent-Based Models (ABMs). These models generate observable time series driven by evolving, partially-latent microstates. Latent states need to be estimated to align simulations with real-world data—a task traditionally addressed by DA, especially in continuous and equation-based models such as those used in weather forecasting. However, the nature of ABMs poses challenges for standard DA methods. Solving such issues requires adaptation of previous DA techniques, or ad-hoc alternatives such as LBI. DA approximates the likelihood in a model-agnostic way, making it broadly applicable but potentially less precise. In contrast, LBI provides more accurate state estimation by directly leveraging the model’s likelihood, but at the cost of requiring a hand-crafted, model-specific likelihood function, which may be complex or infeasible to derive. We compare the two methods on the Bounded-Confidence Model, a well-known opinion dynamics ABM, where agents are affected only by others holding sufficiently similar opinions. We find that LBI better recovers latent agent-level opinions, even under model mis-specification, leading to improved individual-level forecasts. At the aggregate level, however, both methods perform comparably, and DA remains competitive across levels of aggregation under certain parameter settings. Our findings suggest that DA is well-suited for aggregate predictions, while LBI is preferable for agent-level inference.

1 Introduction

Agent-Based Models (ABMs) have become indispensable tools for studying complex systems across various disciplines, including economics, epidemiology, ecology, and sociology [26, 20, 18, 22]. By explicitly representing individual agents, each with distinct behaviors, interactions, and adaptive rules, ABMs capture how macro-scale patterns emerge from micro-level heterogeneity. This granularity enables researchers to explore counterfactual scenarios, test policy interventions, and uncover mechanisms driving phenomena such as market crashes, disease spreading, or cultural shifts. Unlike aggregate models, ABMs preserve the interpretability of individual decisions while accommodating nonlinearity and path dependence. Their flexibility makes them particularly valuable in the social sciences, where human behavior often defies simplistic averaging assumptions.

A paradigmatic example is opinion dynamics, where ABMs such as the Bounded-Confidence Model (BCM) simulate how agents influence one another’s views [14]. In the BCM, agents iteratively adjust their opinions only when interacting with others whose beliefs lie within a fixed confidence bound [4]. This simple rule generates rich macro-level outcomes, such as polarization, fragmentation, or consensus, that can be understood in terms of single agent-level trajectories.

Despite their simplicity, calibrating such models to real-world data poses a critical challenge: while interactions and their outcomes may be observable, the latent microstates—the evolving opinions of agents—

are typically inaccessible. Here, we interpret calibration as the broad task of aligning a simulation-based model to real-world data [2, 27, 15], whether through estimating or tuning a few *global* parameters [6, 19, 8, 1, 11, 12, 17], or through initializing and tracking agent-level latent attributes and variables [16, 24, 9, 12]. Traditionally, most of the research effort has concentrated on parameter estimation [27, 5, 3], mostly due to a lack of appropriate methodologies for estimating latent states. Yet, both parameter tuning and latent state estimation are paramount to the broader calibration goal. Reconstructing the latent states is thus essential for both validating ABMs, interpreting mechanisms, and generating accurate forecasts [17].

Latent state inference in ABMs faces three key challenges. First, the high dimensionality of microstates (e.g., opinions of thousands of agents) complicates inference, especially when observations are sparse or aggregate. Second, ABMs often blend deterministic rules (e.g., the BCM’s interaction threshold) with stochastic elements, creating hybrid dynamics that resist traditional analytical methods. Third, observation granularity varies widely: real-world data may capture single interactions (e.g., social network ties), agent-level summaries (e.g., individual survey responses), or population-level statistics (e.g., polling averages), each requiring distinct inference approaches.

To address these challenges, two methodologies have recently gained traction: Data Assimilation (DA), which sequentially integrates observations into simulations [9], and Likelihood-Based Inference (LBI), which optimizes parameters and states against a probabilistic model [16]. A systematic comparison of these methods for ABMs is missing, leaving practitioners without guidance on their trade-offs. In this paper, we bridge this gap by evaluating DA and LBI for latent state estimation and forecasting in the BCM. We focus on three questions:

- **State Recovery:** Can DA and LBI accurately reconstruct agent-level latent opinions? In a system where the state evolution equations are deterministic, such as the BCM, this task is equivalent to estimating the initial opinions of the agents.
- **Forecasting Accuracy:** How does latent state re-

covery affect forecasting errors for observable variables (i.e., interactions)? This task is challenging in systems with feedback loops, where inaccuracies in initial estimates can propagate nonlinearly.

- **Robustness:** How do these methods perform under model mis-specification, such as noise-corrupted states or erroneous confidence bounds?

Our results show that LBI outperforms DA in recovering agent-level latent states, even under misspecification, which in turn leads to better individual-level forecasts. However, both methods perform similarly at the aggregate level, suggesting that DA may suffice for macro-scale predictions while LBI is preferable for micro-level inference. This degree of reliability is striking since DA operates without model-specific likelihoods. These findings provide practical guidance for ABM calibration and highlight trade-offs between methodological complexity and accuracy.

2 Results

We evaluate the performance of both Data Assimilation (DA) and Likelihood-Based Inference (LBI) on a deterministic, synchronous variant of the Bounded-Confidence Model (BCM) of opinion dynamics [4] running on a fully-connected network. Agents iteratively converge in their opinions based on interactions with neighbors whose opinions lie within a fixed, known confidence bound ϵ , following the deterministic update rules of Deffuant et al. [4]. The steady state may result in either consensus, polarization, or fragmentation, depending on ϵ and the initial conditions.

To evaluate latent state recovery, we infer agents’ opinions using observed data from the first 25% of a simulated trajectory (250 out of 1000 time steps). Reconstruction accuracy is quantified by comparing estimated opinions to ground truth at the final observed timestep ($t = 250$). The BCM’s deterministic dynamics reduce state recovery to estimating the initial conditions $\mathbf{x}(0)$. Forecasting performance is evaluated by initializing the BCM with the reconstructed states at $t = 250$, simulating forward for $t > 250$, and considering the observable variable, i.e., the interactions. Forecast errors are measured at three aggrega-

tion levels: edge-level accuracy in predicting pairwise interactions, agent-level deviation in the number of interactions with its neighborhood, and graph-level discrepancies in macro-scale aggregate metrics, which indicate polarization or consensus.

We further assess the robustness of the methods by introducing two forms of mis-specification: (*i*) noise-corrupted states, where stochasticity is artificially injected into the latent opinion update, and (*ii*) incorrect confidence bounds, where the assumed value of ϵ during inference is incorrect.

We evaluate the ability of DA (blue in the figures) and LBI (red) to reconstruct latent opinion trajectories in the deterministic Bounded-Confidence Model (BCM), using only agent-level interaction events as observations. We assess their performance in reconstructing the agent opinions and in forecasting future agent interactions using the reconstructed state.

2.1 Latent State Inference

In Figure 1, we compare reconstructions of noise-free BCM trajectories under two regimes: *polarization* ($\epsilon = 0.2$) and *consensus* ($\epsilon = 0.3$). Across both regimes, LBI consistently recovers latent opinions more accurately than DA. This can be seen qualitatively in the closer alignment of reconstructed and true trajectories, and quantitatively in the reconstruction errors. For polarization, the error is substantially lower under LBI (mean = 0.09, IQR = 0.04–0.13) compared to DA (mean = 0.28, IQR = 0.26–0.31). Similarly, for consensus, LBI achieves a mean error of 0.09 (IQR = 0.07–0.11), while DA yields 0.21 (IQR = 0.19–0.22). LBI significantly outperforms DA in both regimes. However, DA still captures the qualitative nature of the dynamics, correctly reproducing polarization and consensus patterns (see Figure SI4 for errors on sorted states, which highlight qualitatively good reconstructions). This robustness is notable given that DA operates without model-specific likelihoods, making it broadly applicable across agent-based models.

Robustness to mis-specification. In Figure 2, we examine robustness when the latent trajectories deviate from the idealized model, either due to noise

in the agent states or parameter mis-specification. We refer to Figure SI5 to observe the original noisy trajectories to be reconstructed. With weak noise ($\sigma = 0.0004$, top row), the underlying trajectories remain visible, and both methods capture the overall dynamics, though LBI achieves substantially lower errors. For polarization, the mean reconstruction error under LBI is 0.09 (IQR = 0.04–0.13), compared to 0.28 (IQR = 0.26–0.31) for DA. For consensus, LBI again improves performance with an error of 0.09 (IQR = 0.07–0.11), versus 0.21 (IQR = 0.19–0.22) for DA.

Under strong noise ($\sigma = 0.0016$, middle row), where the true trajectories are almost entirely obscured, errors increase for both methods, but LBI remains clearly superior. In the polarization regime, LBI yields 0.13 (IQR = 0.09–0.15) versus 0.33 (IQR = 0.33–0.34) for DA, while in the consensus regime the errors are 0.14 (IQR = 0.11–0.15) and 0.23 (IQR = 0.22–0.25), respectively. These results are expected noting how the ground truth trajectories are affected by such high level of noise (see Figure SI5).

Finally, when the confidence bound ϵ is mis-specified (bottom row, polarization and consensus swapped between the true and inferred models), LBI still achieves lower reconstruction errors and adapts the inferred dynamics to the observed interactions by effectively compressing the opinion space (see the bottom-right panel). In the mis-specified scenario, the mean reconstruction error was 0.26 (IQR = 0.25–0.28) for DA and 0.11 (IQR = 0.06–0.12) for LBI when using $\epsilon = 0.3$ (consensus) to predict $\epsilon = 0.2$ (polarization). In the reverse case, errors were 0.20 (IQR 0.19–0.21) for DA and 0.09 (IQR 0.07–0.12) for LBI. These values are very similar to those obtained under correct specification. Interestingly, DA reconstructions in this setting are slightly improved relative to the well-specified case, though they remain less accurate than those of LBI. Overall, these results show that while both methods degrade under strong noise, LBI is consistently more robust to both observational noise and structural mis-specification.

Inspecting the individual traces from Figures 1 and 2, LBI faithfully recovers the true initial conditions, producing trajectories that remain close to the ground truth. In contrast, DA yields approximate

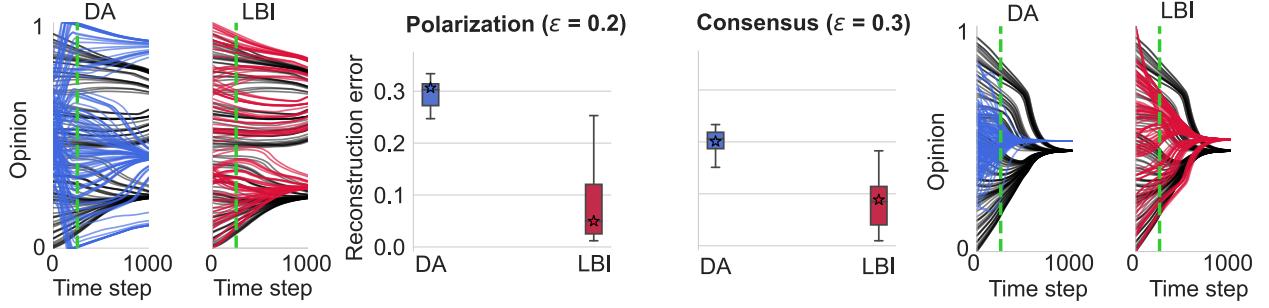


Figure 1: Estimation results for LBI and DA for two levels of ϵ . The box plots in the center represent the reconstruction error on the y-axis at the end of training ($T = 250$) for DA and LBI (x-axis). Beside each box, we depict one of the corresponding traces, with the true (in black) and estimated (blue for DA, red for LBI) positions in one single estimation experiment (whose error is represented with a star in the bar plot, for reference). In each trace plot, the x-axis represents time, and the y-axis represents the opinion of each agent.

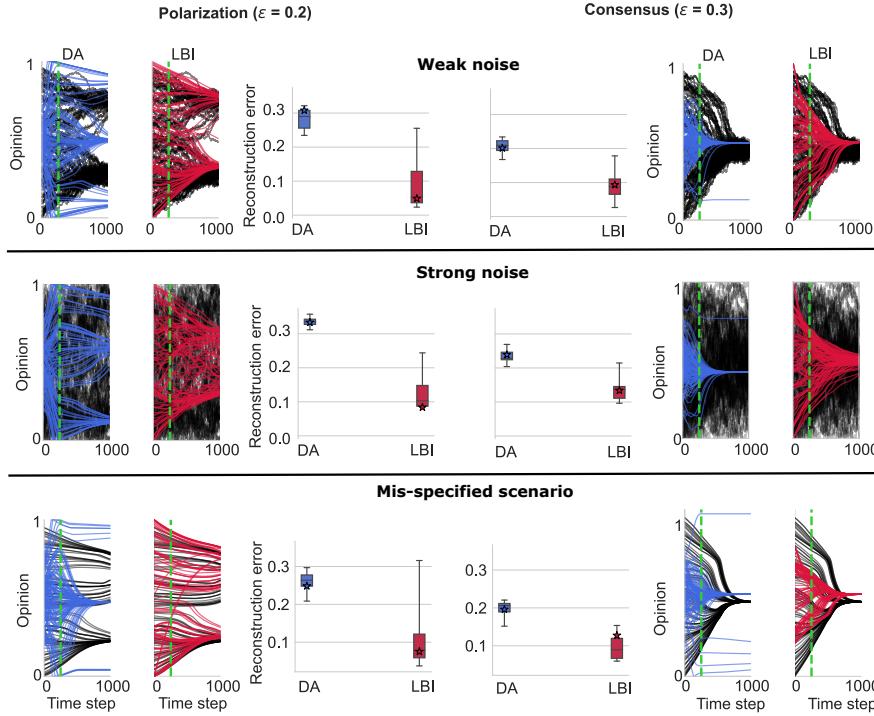


Figure 2: Estimation results for LBI and DA under noisy and mis-specified scenarios. Each row corresponds to a different setting: weak noise ($\sigma = 0.0004$, top), strong noise ($\sigma = 0.0016$, middle), and mis-specified confidence bound ϵ (bottom, polarization and consensus regimes swapped). Otherwise, settings are identical to those of Figure 1.

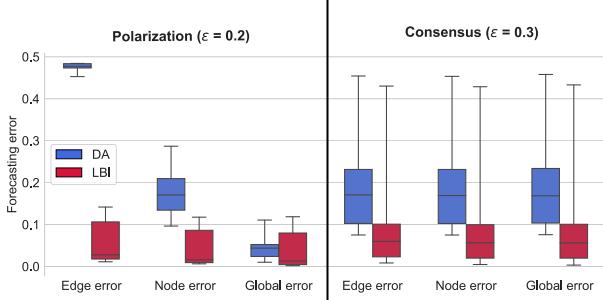


Figure 3: Forecasting accuracy of DA and LBI in the observation space. Boxplots show normalized mean absolute error (MAE) at three levels of granularity: edge (interaction-level), node (number of interactions per agent), and global (total number of interactions over time). Results are reported for polarization ($\epsilon = 0.2$, left) and consensus ($\epsilon = 0.3$, right). In polarization, LBI achieves substantially lower forecasting errors across all metrics, while in consensus, the gap narrows, reflecting that a shared latent space makes reconstruction easier.

reconstructions that nevertheless capture the main qualitative features of the dynamics. This highlights the trade-off in which LBI excels at precise state recovery, while DA provides coarser but still reliable reconstructions of the system’s macroscopic behavior.

2.2 Forecasting agent interactions

Beyond reconstructing latent trajectories, we evaluate how well each method forecasts system behavior in the observation space. Specifically, we assess the predictive accuracy on out-of-sample intervals ($t > 250$) at three levels of granularity: interaction-level (edge error), agent-level (node error), and system-wide (global error) (see Section 4.5 for details). We summarize our results in Figure 3 (see also Figure SI7 for full forecast trajectories of both methods and Figure SI8 for results measured with Brier score).

As expected, LBI achieves lower forecasting errors than DA across all metrics, particularly in the polarization scenario. At the interaction level, DA reaches a mean error of 0.48 (IQR = 0.47–0.48) compared to

0.06 (IQR = 0.02–0.11) for LBI. At the agent level, the errors are 0.17 (0.13–0.21) versus 0.05 (0.01–0.09), respectively. In the consensus regime ($\epsilon = 0.3$), both methods improve, and the gap narrows: DA attains a mean error of 0.20 (IQR = 0.10–0.23), while LBI achieves 0.12 (IQR = 0.02–0.10). These differences reflect a fundamental contrast between regimes: in consensus, all agents converge toward the same latent space, which simplifies reconstruction and forecasting. As a result, DA —despite lacking model-specific likelihoods—has a much easier time tracking the dynamics and becomes competitive with LBI, particularly at aggregated (global) levels.

Importantly, DA remains robust despite its more general formulation. Without relying on the explicit likelihood, it still reproduces key patterns of the observed dynamics, and at coarse levels of aggregation, its forecasts often approach those of LBI. This flexibility makes DA especially attractive in scenarios where the true generative model is unknown or only partially specified. By contrast, LBI offers sharper precision when the likelihood is available, particularly under more complex dynamics like the polarization scenario, where the reconstruction problem is more demanding.

3 Discussion

This study presents the first systematic comparison of Data Assimilation (DA) and Likelihood-Based Inference (LBI) for latent state estimation in agent-based models. Using the Bounded-Confidence Model (BCM) of opinion dynamics as a testbed, we demonstrate that LBI achieves superior accuracy in recovering agent-level latent states and generating individual-level forecasts, even under model mis-specification. By contrast, DA remains competitive for aggregate-level predictions, accurately capturing the macroscopic dynamics of polarization or consensus despite its model-agnostic formulation. This divergence reveals a fundamental methodological trade-off: while LBI delivers higher precision through model-specific likelihoods, DA offers broader applicability through ensemble-based approximations that require no explicit likelihood function. Practitioners should thus weigh the need for agent-level precision against methodological

complexity and model compatibility.

Interestingly, noise-corrupted latent states only degrade performance at high intensity levels. This result suggests that both methods tolerate moderate stochasticity inherent in real-world systems. More notably, LBI exhibits unexpected robustness to parametric misspecification of the confidence bound. When the true confidence bound ($\epsilon = 0.3$) exceeds the assumed value ($\epsilon = 0.2$) during inference, LBI infers opinions closer together than reality. This ‘shrinkage’ effectively compensates for the narrower assumed interaction threshold, thus ensuring agents remained within inferred ϵ -neighborhoods during updates. Consequently, while absolute state estimates diverge, the dynamics remain probabilistically consistent with observations: LBI can implicitly correct parametric errors through state adaptation. However, this resilience relies on symmetries inherent in the BCM (e.g., invariance around $x = 0.5$). Whether similar adaptability extends to ABMs lacking such symmetries—such as those with heterogeneous agents or asymmetric interaction rules—remains an open question worth exploring.

While DA performs remarkably well for aggregate forecasting, its applicability to complex ABMs still faces important limitations. The Ensemble Kalman Filter assumes near-linear dynamics and Gaussian uncertainties, which often misalign with the high-dimensional, discrete, or heterogeneous state spaces characteristic of social ABMs. Adapting DA to these settings will require tailored methodological innovations.

As a proof of concept, our analysis is limited to one ABM (BCM). While illustrative, broader generalizations require validation across diverse ABMs (e.g., in economics, epidemiology, or ecology). Nevertheless, this work fills a critical gap in the ABM calibration literature by providing the first structured comparison of DA and LBI for latent state inference. We hope it catalyzes more systematic evaluations, ultimately informing standardized calibration pipelines for computational social science.

4 Methods

4.1 Bounded-Confidence Model

The opinion dynamics model used in this work is an Agent-Based Model within the family of the Bounded-Confidence Model (BCM). In particular, we use the deterministic, synchronous version of the model by Deffuant et al. [4]. The Deffuant model, as all BCMS, captures the phenomenon whereby individuals tend to be influenced only by those whose opinions are close to their own, reflecting a “bounded confidence” assumption. This model operates according to the following key principles:

- Agents and opinions: Each agent holds an opinion, represented as a real number within a specified range. Without loss of generality, the range is typically $[0, 1]$, where the two extremes represent the polar opposites along some axis (e.g., left-right political spectrum, pro-anti abortion or gun control, or skeptical-believer in human-made climate change). The initial opinions are distributed uniformly at random across this range.
- Confidence Bound Parameter (ϵ): This is the confidence threshold. An agent only considers the opinions of others within an ϵ -distance of their own opinion. This distance defines the “social neighborhood” of each agent.
- Opinion updating: As most other ABMs, the model works in discrete rounds. At each iteration, each agent updates its opinion proportionally to the difference between their opinions within its ϵ -bound neighborhood. This mechanism models the idea that agents are influenced by others within their confidence bound.
- Convergence dynamics: Over time, the population tends to evolve toward clusters of consensus. Multiple clusters may form if the initial opinions are too dispersed or if ϵ is relatively small, thus leading to polarized groups with limited influence across clusters, or even complete opinion fragmentation.

The Deffuant model has been instrumental in understanding how social fragmentation, echo chambers, and consensus can emerge from simple local interaction rules. It is widely used in modeling political

polarization, the spread of ideas, and behavior on social networks.

More formally, the update equations of the Deffuant models are as follows. Let $\mathbf{x}(t) \in [0, 1]^N$ represent the opinion vector of the system of N agents at time t . Let $x_i(t)$ represent its i -th component, i.e., the opinion of agent i at the same time. Let the bounded confidence threshold be $0 < \epsilon \leq 1$. Agent i considers the opinions of agent j only if $|x_i(t) - x_j(t)| \leq \epsilon$. We denote this condition with an indicator function $y_{ij}(t) = \mathbb{1}(|x_i(t) - x_j(t)| \leq \epsilon)$, which we call the *observation*. At each time step, the opinion of agent i is updated based on the opinions of all agents within its bounded confidence interval. Let $\Gamma_i(t)$ denote the set of agents whose opinions fall within agent i 's confidence interval at time t (*confidence set*)

$$\Gamma_i(t) = \{j : |x_i(t) - x_j(t)| \leq \epsilon\}.$$

Then, the opinion update rule for each agent i is

$$x_i(t+1) = x_i(t) + \mu \sum_{j \in \Gamma_i(t)} (x_j(t) - x_i(t)), \quad (1)$$

where $\mu \in [0, 0.5]$ is a *convergence rate* parameter that regulates the speed of the convergence of the system and its dynamics, but does not affect its steady state as this version of the model is synchronous [10, 13] (differently from the original asynchronous model where a single pair of agents interacts at each time step [4]).

Note that the confidence set depends on the absolute value of the difference between agents' opinions, and thus is invariant to translation (modulo border effects) and reflection. Similarly, the opinion updates are proportional to the signed difference in opinions. Thus, the opinion trajectories of a system with initial opinion vector \mathbf{x} and the ones with initial opinion vector $\mathbf{x}' = 1 - \mathbf{x}$ are symmetric around the midpoint of the opinion space $x = 0.5$. As such, they are indistinguishable when we only have access to the observations \mathbf{y} , and we are unable to directly observe the opinions of the agents \mathbf{x} , as explained next.

4.2 Latent and Observable Variables

Data-driven agent-based models (ABMs) are attractive due to their flexibility in simulating complex

systems of interacting agents while achieving strong performance in real-world applications [25]. However, we rarely observe agent-level states directly, as they often represent hard-to-measure constructs. Instead, we might observe agent-level actions or summary statistics of the system and track them over time.

For example, consider measuring user interactions over time on a social platform such as Reddit. This system represents an evolving interaction network—who replies to whom and when—while the underlying opinions or attitudes driving these interactions are latent [16]. A model such as the BCM captures the evolution of these latent opinions that are not directly accessible from the data. This naturally leads us to distinguish between *observable* variables, denoted by \mathbf{y} —such as the evolving interaction network—and *latent* variables, denoted by \mathbf{x} —such as the agent's opinions described by the BCM.

Formally, we can think of our ABM as a dynamical system \mathbf{f} that describes the evolution of the (latent) agent states

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{x}(\tau \leq t)), \quad (2)$$

where $\mathbf{x}(t)$ represents the agent states at times t and $\mathbf{x}(\tau \leq t)$ the full history up to that point.

Observations are then derived through an observation operator, \mathbf{h} , applied to the latent states at a given time

$$\mathbf{y}(t) = \mathbf{h}(\mathbf{x}(t)). \quad (3)$$

In general ABMs, $\mathbf{x}(t)$ might also depend on $\mathbf{y}(\tau < t)$. Here, we assume that that \mathbf{y} is a measure of \mathbf{x} , so it does not affect its dynamics.

In our previous BCM example, the opinions of the agents represent the latent state \mathbf{x} , the network interactions are the observable variables \mathbf{y} , the update Equation (1) describes the system \mathbf{f} , and the confidence bound ϵ determines the observation operator \mathbf{h} . Our main goal is to infer the latent trajectory $\mathbf{x} = (\mathbf{x}(0), \dots, \mathbf{x}(t))$ from the observed data $\mathbf{y} = (\mathbf{y}(0), \dots, \mathbf{y}(t))$. We can frame this task as computing the posterior distribution $p(\mathbf{x} | \mathbf{y})$ according to Bayes rule

$$p(\mathbf{x} | \mathbf{y}) \propto p(\mathbf{y} | \mathbf{x})p(\mathbf{x}), \quad (4)$$

where $p(\mathbf{y} | \mathbf{x})$ is the *likelihood* of the observed data given the model and $p(\mathbf{x})$ is the *prior* over the ABM trajectories. If the ABM is deterministic, the prior reduces to the distribution over initial states $\mathbf{x}(0)$. The posterior provides a unifying theoretical foundation from which different inference strategies have been developed [9, 16, 12]. This work focuses on two main approaches: Data Assimilation (DA) and Likelihood-Based Inference (LBI).

DA focuses on approximating the full posterior distribution by making assumptions about the likelihood and the model structure. This approach relaxes the necessity of crafting a model-specific likelihood. However, the quality of DA depends on how well the approximated likelihood and structure match the true data-generating process. Section 4.3 discusses the specifics of DA.

In contrast, LBI seeks the *most likely latent state trajectory* $\hat{\mathbf{x}}$ by maximizing the log-likelihood

$$\hat{\mathbf{x}} = \arg \max_{\mathbf{x}} \log p(\mathbf{y} | \mathbf{x}). \quad (5)$$

The main challenge here is deriving a model-specific likelihood [11]. When successful, the maximum likelihood trajectory will represent a realistic explanation of the data stemming from the latent states. Note that the maximum likelihood estimate can be understood in terms of Bayesian statistics as the maximum a posteriori (MAP) estimate, i.e., the mode of the posterior distribution, when the priors are non-informative. Section 4.4 discusses the details of LBI.

4.3 Data Assimilation: The Ensemble Kalman Filter

The Ensemble Kalman Filter (EnKF) is a widely used data assimilation (DA) technique for estimating the latent states of dynamical systems given noisy observations [7]. In the context of the BCM, these methods provide a practical approach to infer individual agent opinions $\mathbf{x}(t)$ based on observed interactions $\mathbf{y}(t)$. In contrast to likelihood-based methods that require explicit computation of the likelihood function, the EnKF employs an ensemble of model realizations to approximate likelihood and state distributions and update estimates recursively.

The EnKF is a sequential filtering approach that estimates the latent state of a system by propagating an ensemble of initial conditions sampled from some prior distribution through the system dynamics and updating their estimates based on new observations. It consists of two main steps: the *forecasting step* and the *analysis step*.

First, we sample an ensemble of N initial states $\hat{\mathbf{x}}^k(t) \sim p_{\mathbf{x}}$, where $p_{\mathbf{x}}$ is some prior distribution, and $\hat{\mathbf{x}}^k(t)$ represents the estimated state for ensemble member i at time t .

In the forecasting step, each ensemble member propagates independently according to the BCM model:

$$\check{\mathbf{x}}^k(t) := \mathbf{f}(\hat{\mathbf{x}}^k(t-1)) + \eta_i, \quad i = 1, \dots, N_e, \quad (6)$$

where $\mathbf{f}(\mathbf{x})$ represents the BCM's opinion update (Equation (1)), $\eta_i \sim \mathcal{N}(0, \Sigma(t))$ is the model uncertainty represented by unbiased Gaussian noise with covariance matrix $\Sigma(t)$, and the diacritic in $\check{\mathbf{x}}$ represents a temporary estimation which is carried to the analysis step.

We perform the analysis step when the observation $\mathbf{y}(t)$ becomes available by updating the ensemble members according to the correction matrix, $\mathbf{K}(t)$, called the *Kalman gain*:

$$\hat{\mathbf{x}}^k(t) = \check{\mathbf{x}}^k(t) + \mathbf{K}(t) [\mathbf{y}(t) - \mathbf{h}(\check{\mathbf{x}}^k(t))] + \nu_i, \quad (7)$$

where \mathbf{h} , the observation operator from Equation (3), maps the latent opinions to observed interactions, and $\nu_i \sim \mathcal{N}(0, \mathbf{R}(t))$ is a Gaussian perturbation term with the covariance matrix $\mathbf{R}(t)$ representing the observational noise. Removing the explicit dependence on time t for readability, the *Kalman gain* is given by:

$$\mathbf{K} = \tilde{\mathbf{X}} \left(\mathbf{h}(\tilde{\mathbf{X}}) \right)^T \left[\mathbf{h}(\tilde{\mathbf{X}}) \left(\mathbf{h}(\tilde{\mathbf{X}}) \right)^T + \mathbf{R} \right]^{-1}, \quad (8)$$

where $\tilde{\mathbf{x}}(t) = [\check{\mathbf{x}}^k(t) - \langle \check{\mathbf{x}}^k(t) \rangle_k]$ is the zero-mean ensemble matrix. The estimated latent state at time t is the average over the ensemble:

$$\hat{\mathbf{x}}(t) = \langle \hat{\mathbf{x}}^k(t) \rangle_k, \quad (9)$$

and the approximated posterior of the trajectory is given by the higher moments over the ensemble of particles.

Given the new observation $\mathbf{y}(t)$, the Kalman gain determines how much we should adjust the forecasted state $\hat{\mathbf{x}}(t)$ to obtain the corrected estimate $\tilde{\mathbf{x}}(t)$. In Bayesian terms, this corresponds to an optimal trade-off between the prior uncertainty (from the forecast step) and the observational uncertainty. The update step effectively applies Bayes' rule in a Gaussian setting, where the new state estimate is obtained by weighing the prior state estimate and the likelihood of the observation. This results in a posterior distribution that balances prior knowledge with new data.

This approach allows EnKF to efficiently incorporate new observations while accounting for model uncertainty and ensemble spread. However, it relies on the assumption of Gaussian uncertainties and approximately linearizable dynamics to work under optimal conditions. In the BCM setting, the observation operator \mathbf{h} matching agent opinions to interactions is highly non-linear, as interactions occur according to a step function of the distance of agents' opinions. Despite this limitation, we use the EnKF to estimate opinions and evaluate its performance compared to LBI.

4.4 Likelihood-Based Inference: Gradient Descent with Automatic Differentiation

The problem of parameter estimation in opinion dynamics can be approached as a *likelihood-based inference* task, where the goal is to infer the initial opinions of agents from observed interactions over time. The Bounded-Confidence Model by Deffuant describes the evolution of opinions based on pairwise interactions, governed by a confidence bound ϵ and a convergence rate μ . Rather than simulation—that can be seen as the *forward* pass—we estimate $x(0)$ given observed interaction data. Since opinions are latent and only interactions are observed, the solution involves optimizing $x(0)$ such that the resulting opinion trajectories best explain the observed interaction patterns \mathbf{y} , as expressed by Equation (5).

We formulate the task as a *maximum likelihood estimation (MLE)* problem, where the probability of observed interactions is modeled as a function of pair-

wise opinion differences. We derive a differentiable loss function by using the log-likelihood of the interaction data, which can be seen as a *binary cross-entropy loss* comparing predicted interaction probabilities to observed interactions. In fact, the binary cross-entropy corresponds exactly to the log-likelihood of the system under the estimate $\hat{\mathbf{x}}$:

$$\begin{aligned} \log p(\mathbf{y} | \mathbf{x}) = & \sum_{i,j} \left(\mathbf{y}_{i,j} (|\hat{\mathbf{x}}_i(t) - \hat{\mathbf{x}}_j(t)| - \epsilon) \right. \\ & \left. + (1 - \mathbf{y}_{i,j}) (1 - |\hat{\mathbf{x}}_i(t) - \hat{\mathbf{x}}_j(t)| - \epsilon) \right) \end{aligned} \quad (10)$$

This setting makes the problem similar to a supervised learning task, where the estimated logits $(|\hat{\mathbf{x}}_i(t) - \hat{\mathbf{x}}_j(t)| - \epsilon)$ can be seen as a score used to predict whether the agents interact.

These logits, once we fix the parameters of the model, are purely a function of $\hat{\mathbf{x}}$, which in turn is a deterministic function of the past interactions and of $\hat{\mathbf{x}}(0)$. The latent states $\mathbf{x}(0)$, in fact, define the entire opinion trajectory via the BCM update equations. At this point, the whole training consists in optimizing $\hat{\mathbf{x}}(0)$ via *gradient-based methods* such as *RMSprop* or *Adam*.

Auto-differentiation enables efficient computation of gradients through the sequence of opinion updates, ensuring stable and effective convergence. However, in order for this solution to work efficiently, it requires to rewrite the update of opinions as a tensorial operation. In other words, we represent the update process as a sequence of vectorized operations rather than iterating over individual agents. Given the initial opinion vector $\mathbf{x}(0)$, the opinion evolution can be expressed as a sequence of transformations governed by a time-dependent adjacency structure. At each time step t , this interaction structure is represented as a sparse $N \times N$ adjacency matrix $A(t)$, s.t. $A_{i,j}(t) = 1$ iff $j \in \Gamma_i(t)$. This allows us to rewrite the opinion update rule from Equation (1) as the matrix operation

$$\mathbf{x}(t+1) = \mu A(t)^\top \mathbf{x}(t) + (1 - \mu A(t)^\top \mathbf{1}) \circ \mathbf{x}(t),$$

where $\mathbf{1}$ is a vector of ones, and \circ represents the Hadamard (element-wise) product. This formulation

highlights the two key components of the update: (i) each agent retains a fraction of its previous opinion, weighted by the total influence received, and (ii) each agent absorbs a weighted sum of the opinions of its influencing neighbors. By iterating this transformation over T steps, the full opinion trajectory can be expressed as a deterministic function of the initial condition. This vectorized formulation enables efficient computation using tensor operations, thus making it well-suited to auto-differentiation and optimization within machine-learning frameworks such as PyTorch. This way, the loss function computed from an interaction at time t will be back-propagated over each opinion vector, defined as a function of the previous one, and finally updating the free variable $\mathbf{x}(0)$. Finally, to stabilize results and avoid convergence to suboptimal local minima, multiple random restarts can be employed during optimization, as well as standard gradient-descent regularization techniques such as weight decay.

4.5 Evaluation Protocol

We run the model described in Section 4.1 for various settings, to generate a variety of ground-truth data traces that we use to evaluate the performance of DA and LBI. The model has two parameters: the confidence threshold ϵ and the convergence rate μ . The latter simply determines the speed of convergence of the system to the steady state, so we choose $\mu = 10^{-4}$ in order to have sufficiently slow convergence to be able to observe the transient and learn the agent states both with DA and LBI. The confidence threshold ϵ is instead fundamental to the ABM dynamics. We choose two values of ϵ that lead to different behaviors. Specifically, we run simulations with $\epsilon = 0.2$, which is small enough to generate multiple opinion clusters (*polarization scenario*), and we also consider a larger $\epsilon = 0.3$, which is large enough to make agents converge on a single consensus opinion (*consensus scenario*). In addition to these two parameters, we explore the effect of adding a noise term to the opinion update, Equation (1). In addition to the deterministic scenario described in Section 4.1, we consider five noise levels, ranging from $2^0 \times 10^{-4}$ to $2^4 \times 10^{-4}$. These magnitudes encompass all noise magnitudes from a

small alteration of the deterministic trajectories to completely noisy dynamics (these values should be interpreted in comparison with the convergence rate $\mu = 10^{-4}$). Finally, to explore the effect of stochasticity in the model, we consider 10 seeds of the random number generator, which govern both the initial conditions of the opinions $\mathbf{x}(0)$ and the realizations of the noise. Summarizing, we run the ABM for 2 values of ϵ , 6 values of noise (including zero noise), and 10 different random seeds, resulting in a total of 120 distinct simulations.

We run all simulations for $N = 100$ agents and a complete interaction network, in the sense that all agents potentially interact with all other agents, leading to a total of $E = N(N - 1)/2 = 4950$ edges. We run the ABM for a total of $T = 1000$ time steps and record:

- the ground truth opinions $\mathbf{x}(t)$;
- the edge indicator variable $y_{ij}(t)$ for whether agents (i, j) interact at time t ;
- how many interactions agent i has with all other agents at time t , namely node-level interactions $y_i(t) = \sum_j y_{ij}(t)$
- how many interactions all agents have with all other agents at time t , namely global-level interactions $y(t) = \frac{1}{2} \sum_i y_i(t)$.

To infer the latent opinions, we use ground truth data up to time step 250 (or 25% of the simulation length) as input for both inference algorithms, DA and LBI. Our results do not depend strongly on the specific choice of time step 250. We select this time step to strike a balance between giving enough data to the inference algorithms and not having already reached convergence to the steady state. Any other step that is neither too early or too late in the simulation would lead to similar results. We denote estimates by $\hat{\mathbf{x}}(t)$. With our latest estimates at $t = 250$ for each algorithm, we then run the ABM up to the final time step $T = 1000$, without any adjustment to the latent opinions, and use these data traces to evaluate the out-of-sample forecasting capabilities of both the DA and LBI algorithms. In the next sections, we discuss in more detail the evaluation for the inference and forecasting tasks.

To explore the effect of mis-specification, we run both the DA and LBI algorithms with the correct value of ϵ and with the mis-specified value of ϵ (i.e., $\epsilon = 0.2$ in the inference when $\epsilon = 0.3$ in the ground truth, and vice versa). Noise represents a further source of mis-specification: we never account for it in the inference algorithms; therefore, the larger the noise, the stronger the mis-specification.

A final dimension for evaluation is what exact ground truth data are fed to the DA and LBI inference algorithms. By construction, LBI can only handle the finest-grained, edge-level observations $y_{ij}(t)$. Instead, DA can use edge-level information, but also more aggregated node level ($y_i(t)$) and global level ($y(t)$) information. In conclusion, we infer the latent variables for 240 data traces via LBI (120 traces with well-specified ϵ and 120 traces with mis-specified ϵ) and 720 data traces via DA (240 traces with edge-level, node-level, and global-level information each).

4.5.1 Inference

We focus on the latent opinion inference at the initial step ($t = 0$) and at the final step of the inference part ($t = 250$). As an evaluation metric, we use the Mean Absolute Error (MAE). So, the inference error at time t averaged over all agents is

$$\mathcal{E}(t) = \frac{1}{N} \sum_i |x_i(t) - \hat{x}_i(t)|. \quad (11)$$

According to the metric above, the inference algorithms must infer the correct opinion of all agents i . However, as discussed in Section 4.1, there is a symmetry around $x = 0.5$ which prevents identification of the exact value of \mathbf{x} . This is also known as label switching symmetry in probabilistic models [23]. To account for this symmetry, we consider a symmetric version of $\mathcal{E}(t)$, giving

$$\mathcal{E}_{\text{symm}}(t) = \frac{1}{N} \sum_i \min(|x_i(t) - \hat{x}_i(t)|, |x_i(t) - (1 - \hat{x}_i(t))|). \quad (12)$$

In the absence of edge-level information, it may be possible to get the correct distribution of opinions but not to assign the correct opinion to the correct

agent. To compare the distributions, we sort both the ground truth and the estimate:

$$\mathcal{E}_{\text{sort}}(t) = \frac{1}{N} \sum_i |\text{sort}(\mathbf{x}(t))_i - \text{sort}(\hat{\mathbf{x}}(t))_i|. \quad (13)$$

4.5.2 Forecasting

To recap, using the DA- and LBI-inferred opinions at time step 250, we run the ABM for 750 further time steps. Then, we compare the obtained trace to the ground truth time series to evaluate the out-of-sample forecasting capabilities of DA and LBI. This measure is likely correlated to how well the methods reconstruct the latent opinions at time step 250. However, due to the nonlinearity of the complex system, different errors in the reconstructed opinions may lead to more severe divergence from the ground truth over the simulation.

As indicators of forecasting quality, we may be interested in how well the inference algorithms reconstruct the edge-level interactions $y_{ij}(t)$, the node-level interactions $y_i(t)$, or the global interactions $y(t)$. When considering edge-level interactions, we quantify the forecasting error by the MAE. Therefore, letting $\hat{p}_{ij}(t)$ denote the forecast probability of an interaction between (i, j) at t when starting the ABM from the inferred opinions at time step 250, the MAE score at time t is

$$\mathcal{F}_{\text{edge}}(t) = \frac{1}{E} \sum_{ij} |y_{ij}(t) - \hat{p}_{ij}(t)|. \quad (14)$$

We choose the MAE as the main evaluation metric, which is an axiomatically good scoring metric for quantification tasks [21].

To evaluate how well we match node-level interactions, we aggregate the probabilities of interactions at the node level both in the ground truth and in the simulations following from latent opinions inference, leading to a probabilistic count error metric

$$\mathcal{F}_{\text{node}}(t) = \frac{1}{N} \sum_i \left| y_i(t) - \sum_j \hat{p}_{ij}(t) \right|. \quad (15)$$

Finally, the match to total interactions is computed

by summing all probabilities of interactions, giving

$$\mathcal{F}_{\text{global}}(t) = \left| y(t) - \sum_{ij} \hat{p}_{ij}(t) \right|. \quad (16)$$

References

- [1] M. Benedetti, G. Catapano, F. De Sclavis, M. Favorito, A. Glielmo, D. Magnanimi, and A. Muci. Black-it: A ready-to-use and easy-to-extend calibration kit for agent-based models. *Journal of Open Source Software*, 7(79):4622, 2022. [2](#)
- [2] T. Brenner and C. Werker. A taxonomy of inference in simulation models. *Computational Economics*, 30:227–244, 2007. [2](#)
- [3] K. Cranmer, J. Brehmer, and G. Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020. [2](#)
- [4] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch. Mixing Beliefs among Interacting Agents. *Advances in Complex Systems*, 03(01n04):87–98, 2000. doi: 10.1142/S0219525900000078. [1](#), [2](#), [6](#), [7](#)
- [5] G. Fagiolo, M. Guerini, F. Lamperti, A. Moneta, and A. Roventini. Validation of agent-based models in economics and finance. *Computer simulation validation: fundamental concepts, methodological frameworks, and philosophical perspectives*, pages 763–787, 2019. [2](#)
- [6] J. Grazzini and M. Richiardi. Estimation of ergodic agent-based models by simulated minimum distance. *Journal of Economic Dynamics and Control*, 51:148–165, 2015. [2](#)
- [7] P. L. Houtekamer and F. Zhang. Review of the ensemble kalman filter for atmospheric data assimilation. *Monthly Weather Review*, 144(12):4489–4532, 2016. [8](#)
- [8] D. Kim, T.-S. Yun, I.-C. Moon, and J. W. Bae. Automatic calibration of dynamic and heterogeneous parameters in agent-based models. *Autonomous Agents and Multi-Agent Systems*, 35(2):46, 2021. [2](#)
- [9] B. Kolic, J. Sabuco, and J. D. Farmer. Estimating initial conditions for dynamical systems with incomplete information. *Nonlinear Dynamics*, 108(4):3783–3805, 2022. [2](#), [8](#)
- [10] N. Lanchier. The critical value of the deffuant model equals one half. *ALEA*, 9(2):383–402, 2012. [7](#)
- [11] J. Lenti, C. Monti, and G. De Francisci Morales. Likelihood-Based Methods Improve Parameter Estimation in Opinion Dynamics Models. In *International Conference on Web Search and Data Mining*, WSDM, pages 350–359. ACM, Mar. 2024. doi: 10.1145/3616855.3635785. [2](#), [8](#)
- [12] J. Lenti, F. Silvestri, and G. De Francisci Morales. Variational Inference of Parameters in Opinion Dynamics Models, Mar. 2024. [2](#), [8](#)
- [13] H.-L. Li. A straightforward proof of the critical value in the Hegselmann-Krause model: up to one-half, Aug. 2024. URL <http://arxiv.org/abs/2408.03718>. arXiv:2408.03718 [math]. [7](#)
- [14] J. Lorenz. Continuous opinion dynamics under bounded confidence: A survey. *International Journal of Modern Physics C*, 18(12):1819–1838, 2007. [1](#)
- [15] T. Lux and R. C. Zwinkels. Empirical validation of agent-based models. In *Handbook of computational economics*, volume 4, pages 437–488. Elsevier, 2018. [2](#)
- [16] C. Monti, G. De Francisci Morales, and F. Bonchi. Learning opinion dynamics from social traces. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 764–773, 2020. [2](#), [7](#), [8](#)
- [17] M. Pangallo and R. M. del Rio-Chanona. Data-driven economic agent-based models. *arXiv preprint arXiv:2412.16591*, 2024. [2](#)

- [18] M. Pangallo, A. Aleta, R. M. del Rio-Chanona, A. Pichler, D. Martín-Corral, M. Chinazzi, F. Lafond, M. Ajelli, E. Moro, Y. Moreno, A. Vespignani, and J. D. Farmer. The unequal effects of the health–economy trade-off during the covid-19 pandemic. *Nature Human Behaviour*, pages 1–12, 2023. [1](#)
- [19] D. Platt. A comparison of economic agent-based model calibration methods. *Journal of Economic Dynamics and Control*, 113:103859, 2020. [2](#)
- [20] S. F. Railsback and V. Grimm. *Agent-based and individual-based modeling: a practical introduction*. Princeton University Press, 2019. [1](#)
- [21] F. Sebastiani. Evaluation measures for quantification: an axiomatic approach. *Information Retrieval Journal*, 23(3):255–288, June 2020. ISSN 1386-4564, 1573-7659. doi: 10.1007/s10791-019-09363-y. URL <https://link.springer.com/10.1007/s10791-019-09363-y>. [11](#)
- [22] M. Starnini, F. Baumann, T. Galla, D. Garcia, G. Iñiguez, M. Karsai, J. Lorenz, and K. Sznajd-Weron. Opinion dynamics: Statistical physics and beyond. *arXiv preprint arXiv:2507.11521*, 2025. [1](#)
- [23] M. Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000. [11](#)
- [24] D. Tang and N. Malleson. Data assimilation with agent-based models using markov chain sampling. *arXiv preprint arXiv:2205.01616*, 2022. [2](#)
- [25] S. Wiese, J. Kaszowska-Mojsa, J. Dyer, J. Moran, M. Pangallo, F. Lafond, J. Muellbauer, A. Călinescu, and J. D. Farmer. Forecasting macroeconomic dynamics using a calibrated data-driven agent-based model. *arXiv preprint arXiv:2409.18760*, 2024. [7](#)
- [26] U. Wilensky and W. Rand. *An introduction to agent-based modeling: modeling natural, social, and engineered complex systems with NetLogo*. Mit Press, 2015. [1](#)
- [27] P. Windrum, G. Fagiolo, and A. Moneta. Empirical validation of agent-based models: Alternatives and prospects. *Journal of Artificial Societies and Social Simulation*, 10(2):8, 2007. [2](#)

A Appendix

In this appendix, we present additional figures that complement the results discussed in the main text.

A.1 Latent State Reconstruction

Aggregate reconstruction. Figure 1 shows results for the task of latent state reconstruction, assuming one is interested in reconstructing the latent state of each agent individually. In other words, the true state of an agent i is compared to the reconstructed state of the same agent. However, in some contexts one might be interested in reconstructing a faithful *distribution* of latent states, disregarding their individual identity. To measure the quality of LBI and DA in this context, we adopt in this section a *sorted metric*, where each agent’s reconstructed latent state is compared against the agent with the same rank in the original set of latent states. That is, we measure the reconstruction error between the *sorted* vector of true and reconstructed latent states. Figure SI4 shows these results. In this context, the two methods perform similarly, with outcomes essentially tied in the polarization scenario and LBI showing a slight advantage in the consensus scenario.

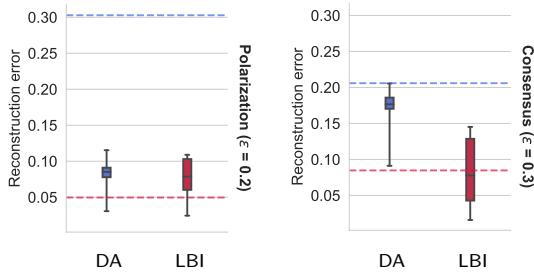


Figure SI4: Reconstruction error using the sorted metric, which compares opinion distributions without enforcing identity of agents, for the polarization (left) and consensus (right) scenarios without any mis-specification. Dashed horizontal lines indicate the median error measured by the unsorted error metric (shown in Figure 1).

System trajectories in the presence of noise.

In Figure 2, we have shown results under a particular type of mis-specification, i.e., noise in the original latent trajectories. To further illustrate this scenario and aid comprehension, we report in Figure SI5 such ground truth trajectories under our three settings of noise.

DA observation operators. In the main text, for LBI we focused on the “edge” observation operator. Figure SI6 shows the performance obtained in this task by DA by each of its three possible observation operators (edge-, node-, and global-level), with LBI also reported as reference.

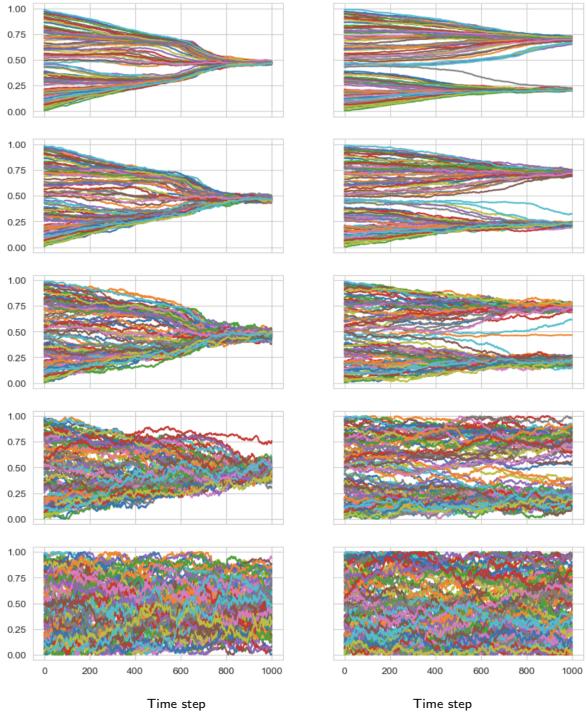


Figure SI5: Ground-truth trajectories of the Bounded-Confidence Model under different noise intensities (top corresponds to *no noise*, bottom to the *strong noise* setting). Left column represents the consensus scenario and right column the polarization scenario. In each plot, we represent opinion evolution.

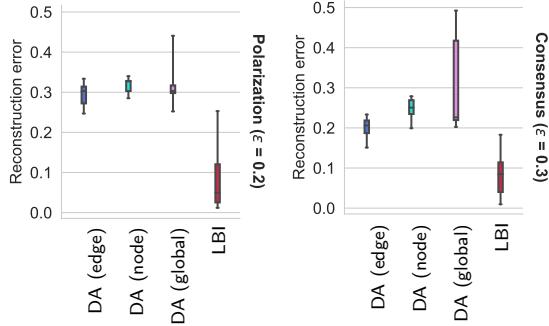


Figure SI6: Comparison of reconstruction performance when DA uses different observation operators (edge-, node-, and global-level), for the polarization (left) and consensus (right) scenarios without any mis-specification. LBI is also reported as reference.

A.2 Forecasting Accuracy

We next examine how well DA and LBI recover future dynamics once trained on partial trajectories. Here, we provide extended analyses of this forecasting task. In Figure SI7, we show explicitly the time series of the quantity to forecast (i.e., the total number of interactions) in the ground truth and in the reconstruction by LBI and DA. Finally, we also report additional metrics, obtained by aggregating errors at the node-, edge- and total level. Figure SI8 shows the same result of Figure 3 but measured with the Brier score.

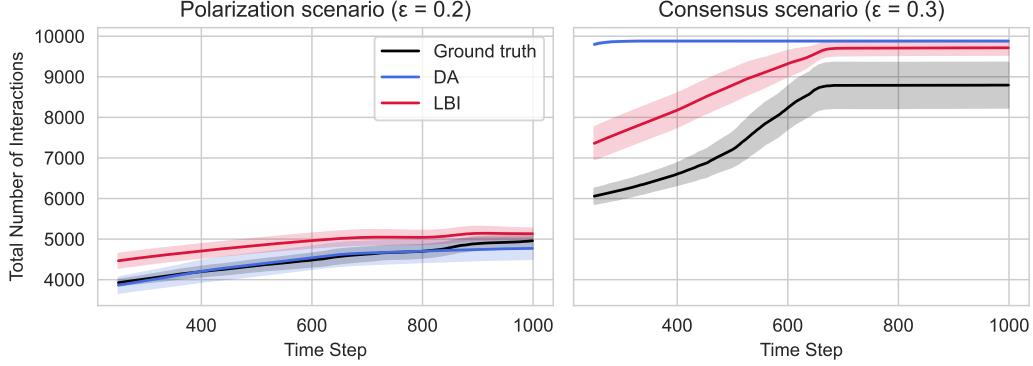


Figure SI7: Forecasted system-wide trajectories of total number of interactions under (left) polarization ($\epsilon = 0.2$) and (right) consensus ($\epsilon = 0.3$) scenarios. We compare ground truth (black), DA (blue), and LBI (red). Shaded regions indicate variability across runs. In the polarization scenario, both methods qualitatively track system behavior. Instead, in the consensus scenario only LBI is able to make substantially accurate forecasts.

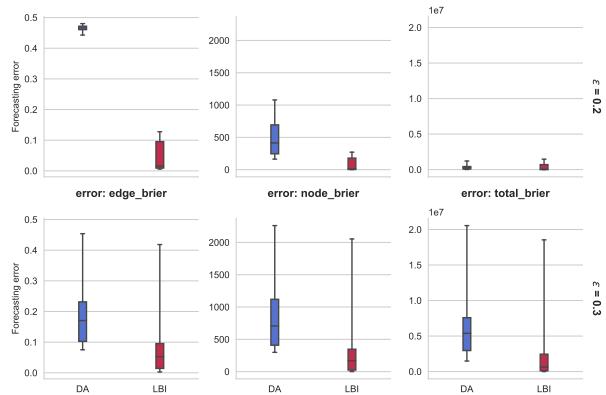


Figure SI8: Forecasting accuracy measured by Brier score at three levels of granularity: edge-level, node-level, and system-wide. Higher values indicate more accurate forecasts.