

Discovering Governing Physics in the Presence of Uncertainty

Corresponding Author: Mr Ridwan Olabiyi

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Overall Assessment

This paper addresses an important problem in interpretable system identification: how to incorporate input variability into sparse model discovery. The topic is relevant to the Communications Physics audience, as SINDy and its extensions are widely used for many real world systems in physics, biology, and engineering.

The strength of the manuscript is in its innovative approach of leveraging stochastic inverse modeling to manage input variability. However, I have several concerns about methodology, clarity of presentation, and completeness of comparisons, particularly with respect to existing baselines and evaluation metrics. Addressing these issues would significantly improve the clarity, rigor, and impact of the work.

Major Comments

1. Missing Related Work and Comparisons

One of the central claims of the paper is that the SIP method more accurately identifies model terms and coefficients compared to previous approaches (SINDy and UQ-SINDy). However, widely used variants such as Ensemble-SINDy (Fasel et al.) and Weak SINDy (Messenger et al.) are generally regarded as more robust and computationally efficient, with Ensemble-SINDy additionally including uncertainty estimates. Including results from these methods would provide a more complete benchmark and significantly strengthen the case that SIP achieves state-of-the-art accuracy.

2. Comparison to Prior UQ-SINDy Results

In Figure 3, The Lynx–Hare reconstruction looks substantially different from the UQ-SINDy results in the original paper (see Figure 8 on p.17 of arXiv:2107.02107). What is the reason for this discrepancy?

Moreover, the posterior predictive intervals differ qualitatively: in SIP, uncertainty is very low in early years (1900–1910), unlike in UQ-SINDy, where uncertainty is more uniform. Having near-zero predictive uncertainty in 1900 seems unrealistic and in fact it looks like UQ-SINDy (from the original papers plot) models uncertainty better than SIP. Please discuss why SIP produces this pattern and why SIP is better in this case.

3. Evaluation Metrics (Table 1)

In Table 1, Please clarify how $\text{dot}(u)_{\text{rmse}}$, $\text{dot}(v)_{\text{rmse}}$, and $\text{lambda}_{\text{rmse}}$ are defined.

The current description suggests RMSE is calculated on model coefficients, but this may not be meaningful given their different magnitudes. For example, in the Lotka–Volterra system, error in the gamma and alpha parameters may dominate errors in beta and delta. Additionally, for the real world systems (Hudson Bay data and infiltration systems) there are no

ground truth parameters so it is unclear what these values refer to.

Alternatively, if this refers to RMSE on reconstructed derivatives, please explain the rationale for preferring derivatives over state reconstruction error.

4. Library of Candidate Terms

For the simulated Lotka–Volterra system, what library of candidate terms was used? The choice appears to differ from the Hudson Bay data example, which only includes $\{1, u, v, uv\}$. This is an uncommon and somewhat restrictive basis. Typically, polynomial terms up to a fixed degree (e.g., $1, u, v, uv, u^2, v^2$) are considered. Could you demonstrate these results with this complete library?

Additionally please include the library used for each example.

5. Performance Interpretation

My understanding is that SIP should outperform other methods when there is input variability. So it is difficult to understand why SIP should perform better than UQ-SINDy for the synthetic single trajectory examples in which normally distributed noise is added to the output of the systems.

One remedy which may help with this is including posterior distributions of the library terms for each of the synthetic examples.

6. Posterior Distributions

I suggest including a comparison of the posterior distributions of the parameters for the simulated Lotka–Volterra system and Lorenz systems between UQ-SINDy and SIP. These are currently only shown for the Hudson Bay data. However, in this case the true dynamics are unknown and hidden environmental factors may affect model structure so comparison of terms between methods is less meaningful.

7. Noise Modeling Choices

Additive noise is used for the simulated Lotka–Volterra system with 100 coefficients, but multiplicative noise is used for the single-trajectory case. Please clarify why different noise models were used. Typically multiplicative noise is employed for Lotka–Volterra since population values are nonnegative (see e.g., Stan case study <https://mc-stan.org/learn-stan/case-studies/lotka-volterra-predator-prey.html>).

Minor Comments

1. p.2: "Lispson" should be "Lipson."

2. It would also be useful to cite related work such as Bayesian differential programming for robust systems identification under uncertainty Proc. R. Soc. A 2020.0290.

3. In the formula for root mean square error, there is no square root being taken.

4. What priors were used for UQ-SINDy and SIP (e.g., spike-and-slab, regularized horseshoe)? This should be made explicit.

5. On Fig. 4B: where is the SIP curve (black dashed line)?

6. The reference "Sindy-sa framework: enhancing nonlinear system identification with symmetry and conservation laws" does not have correct authorship. Please correct.

Recommendation

I recommend major revision. The manuscript could make a strong contribution, but several methodological choices need clarification, and some important baselines and analyses are missing. If the authors address all of the points above, the paper would be significantly strengthened and could be of broad interest to the readership of Communications Physics.

Reviewer #2

(Remarks to the Author)

The paper proposes a stochastic inverse physics-discovery (SIP) framework that treats the coefficients in SINDy-style models as random variables and recovers their posterior via a push-forward likelihood matched to the observed data. It also uses matched-block bootstrap to handle single-trajectory data. The authors claim large gains over SINDy and UQ-SINDy (e.g., "~82% lower coefficient RMSE on average") and show case studies on Lotka–Volterra (simulated and Hudson Bay), Lorenz, and porous-media infiltration. If validated and fairly compared, the idea is useful: modeling parameter variability explicitly is relevant in many real systems.

The contribution is potentially valuable, but the paper needs a lot of work:

1. it needs more rigorous comparisons with existing work that the authors do not seem to be aware of
2. a clearer relation to recent Bayesian SINDy work
3. more robustness and ablations experiments

For these reasons, I recommend the paper for major revision. Here are more details:

1. Please compare SIP with recent Bayesian treatments of equation discovery, not only UQ-SINDy. Among others, these are relevant papers

- Klishin, Andrei A., et al. "Statistical mechanics of dynamical system identification." *Phys. Rev. Research* (2025). (which also uses a spike-and-slab prior as done in this paper.)
- Fung, Lloyd; Urban Fasel; Matthew Juniper. "Rapid Bayesian identification of sparse nonlinear dynamics from scarce and noisy data." *Proc. A* (2025).
- North, Joshua S.; Christopher K. Winkle; Erin M. Schliep. "A Bayesian approach for data-driven dynamic equation discovery." *JABES* 27(4), 2022.

Also, while you cite E-SINDy (Fasel et al., 2022), you don't compare against it. Please add E-SINDy to all benchmarks; recent results show it can match or exceed UQ-SINDy with good flexibility and cost. Without this, it's hard to judge practical impact.

2. The study should perform more ablation experiments. Those include:

- Showing performance sensitivity to (a) prior choice (spike-and-slab vs. horseshoe), (b) KDE bandwidth, (c) MBB block length, (d) library size/degree, (e) derivative estimation method, (f) ODE solver and step size. Without these, it's hard to tell how much better the proposed method does in general.
- I propose providing synthetic experiments with: (a) noise-only, (b) variability-only, (c) a combination of the two; to support the claim of the paper. Show that SIP identifies the correct driver and does not attribute noise to variability (or vice versa).
- Rejection sampling can be expensive in higher dimensions. Please report sample counts, acceptance rates, and wall-clock runtimes vs. UQ-SINDy and E-SINDy.

3. Use rollout error. Table 1 reports $\dot{x}$ errors. Please add integrated trajectory (rollout) errors with uncertainty bands in the table. Low $\dot{x}$ error can still diverge under integration.

4. Related: since you report 95% credible intervals that "track observed trajectories." Please evaluate empirical coverage (do 95% of held-out points fall in the 95% bands?) on multiple seeds/splits.

5. In relation to ablation: the likelihood is built per state with Gaussian KDE and then multiplied (implicitly assuming independence across states/time). Please state the assumption and test sensitivity to bandwidth selection and to dependence violations.

6. For the infiltration physics problem, report the discovered equations. Please print the final inferred ODE(s)/terms and coefficients for the infiltration cases (both high- and low-viscosity). Right now only errors/plots are shown; readers need the actual recovered equations to assess physical plausibility. This is after all the motivation of using SINDy-like models in the first place.

7. The lack of code makes it hard to evaluate reproducibility.

To conclude, the paper has a promising idea: treating coefficient variability explicitly and using a push-forward, data-consistent inversion. BUT to be convincing and broadly useful, it needs stronger experimental and theoretical supporting evidence: comparison against modern Bayesian/E-SINDy baselines, ablations for robustness and fairness, trajectory-level metrics. For that I recommend for major revision.

Reviewer #3

(Remarks to the Author)

The manuscript proposes a data-driven approach to identify the equations of motion as well as to model unknown coefficients as random variables on systems that present nonlinear behavior. In particular, such a method is based on a stochastic inverse physics-discovery (SIP) framework to estimate the uncertainty in the values of the model coefficients. The accuracy of the proposed procedure is verified in analytical and experimental data. This paper adds some value to the literature of nonlinear identification methods. As such, pending the authors addressing the following, in general minor comments, this paper can be considered for publication.

- In Page 6, the assumption of treating noise as Gaussian distribution with zero mean and an unknown variance is entirely reasonable. However, the authors should discuss the limitations of this assumption and identify scenarios where this approach might not be applicable.
- In the synthetic systems, one percent of noise was tested. I strongly suggest creating a plot that shows RMS and/or a metric of success vs different noise levels.
- Provide information about the terms used in the libraries for each system.

- What were the solvers, tolerances, time steps, and total times used to simulate the systems?
- What was the threshold used in SINDy?
- The results provided by SIP considerably improved SINDy and UQ-SINDy's. What about the computational time required for SIP compared to the other methods?, and what computational resources were used?
- To improve readability, I suggest plotting Fig. 2 on the same limits for each coefficient.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Summary

This paper addresses an important problem in interpretable system identification: how to incorporate input variability into sparse model discovery. The topic is relevant to the Communications Physics audience, as SINDy and its extensions are widely used for many real-world systems in physics, biology, and engineering. The strength of the manuscript is in its innovative approach of leveraging stochastic inverse modeling to manage input variability.

Assessment of Revision

In my previous report, I raised concerns regarding missing comparisons to relevant baselines, limited evaluation metrics, and lack of clarity in methodological choices such as noise modeling and posterior analysis.

The authors have addressed these concerns thoroughly:

1. Missing Related Work and Comparisons: The revised manuscript now includes comparisons with Ensemble-SINDy, Weak-SINDy, B-SINDy, and symbolic regression approaches across the different scenarios. This substantially strengthens the empirical evaluation.
2. Evaluation metrics: The introduction of the DiCE metric is a useful addition, providing a term level measure of equation recovery that complements the RMSE errors.
3. Posterior analysis: The inclusion of posterior comparisons across multiple benchmark systems (Lotka–Volterra, Lorenz, etc.) provides significant clarity about the impact the of the SIP method.
4. Methodological clarity: The discussion of library terms, priors, and noise modeling choices have all been clarified.
5. Minor issues: Previously noted typographical and presentation issues have been corrected.

Current Evaluation

Overall, the revised manuscript is significantly improved. The experimental evaluation is now more comprehensive, and the claims are better supported by comparisons to relevant methods.

The work appears methodologically sound, and sufficient detail is provided to allow reproducibility. The additional evaluation metrics and broader benchmarking makes the contribution more compelling and relevant to the field.

Recommendation

I recommend acceptance.

Reviewer #2

(Remarks to the Author)

The authors have successfully addressed my concerns and the manuscript is now ready for publication.

Reviewer #3

(Remarks to the Author)

I am satisfied with the changes to the manuscript and with the authors' reply to my comments. I suggest that this manuscript be published.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Reviewers' Comments on Discovering Governing Physics in the Presence of Uncertainty

Reviewer #1 (Remarks to the Author)

Overall Assessment

This paper addresses an important problem in interpretable system identification: how to incorporate input variability into sparse model discovery. The topic is relevant to the Communications Physics audience, as SINDy and its extensions are widely used for many real-world systems in physics, biology, and engineering.

The strength of the manuscript is in its innovative approach of leveraging stochastic inverse modeling to manage input variability. However, I have several concerns about methodology, clarity of presentation, and completeness of comparisons, particularly with respect to existing baselines and evaluation metrics. Addressing these issues would significantly improve the clarity, rigor, and impact of the work.

We thank all reviewers for their constructive feedback and detailed comments. Below, we provide point-by-point responses to each comment, indicating the changes made to the manuscript.

Major Comments

1. Missing Related Work and Comparisons

One of the central claims of the paper is that the SIP method more accurately identifies model terms and coefficients compared to previous approaches (SINDy and UQ-SINDy). However, widely used variants such as Ensemble-SINDy (Fasel et al.) and Weak SINDy (Messenger et al.) are generally regarded as more robust and computationally efficient, with Ensemble-SINDy additionally including uncertainty estimates. Including results from these methods would provide a more complete benchmark and significantly strengthen the case that SIP achieves state-of-the-art accuracy.

Response: We thank the reviewer for this important suggestion. We have now added comprehensive comparisons with sparse-regression methods, including Ensemble-SINDy (E-SINDy), Weak-SINDy, Bayesian SINDy (B-SINDy), and symbolic regression methods, mainly Symbolic Physics Learner (SPL), and ODEFormer across all ten case studies. Results are summarized in Table 1 (Page 10 of the updated manuscript), which shows coefficient RMSE, observed-trajectory RMSE, and rollout RMSE for all methods across all scenarios.

We have introduced another metric, called DiCE (Dice Coefficient for Equation structure), that measures structural identification accuracy by quantifying the overlap between identified and true equation terms. DiCE ranges from 0 (no correct terms identified) to 1 (perfect structural recovery), providing a physics term-level assessment of model structure independent of coefficient values. This metric is particularly important for physics discovery because identifying the correct functional form is critical in understanding the mechanisms that capture the underlying dynamics. DiCE complements the coefficient RMSE by separately evaluating whether methods identify the correct physics versus how accurately they estimate parameters.

The key findings from these comparisons demonstrate that SIP achieves substantial improvements across multiple metrics. On average, SIP achieves 75% improvement in coefficient RMSE over

E-SINDy, 68% improvement over Weak-SINDy, 92% improvement over B-SINDy, and 85% improvement over standard SINDy. Critically, SIP achieves a mean DiCE score of 0.966 across all scenarios, indicating near-perfect structural identification, compared to 0.489 for E-SINDy, 0.411 for SINDy, 0.479 for Weak-SINDy, 0.314 for B-SINDy, 0.385 for ODEFormer, 0.564 for UQ-SINDy, and 0.375 for SPL.

Updated trajectory comparisons for all methods are shown in Figures 3, 4, 5, and 6 (main manuscript) and Supplementary Figures SM1 and SM2 across all scenarios. We have also added references to these methods in the Introduction and Discussion sections, acknowledging their contributions to robust physics discovery.

2. Comparison to Prior UQ-SINDy Results

In Figure 3, The Lynx–Hare reconstruction looks substantially different from the UQ-SINDy results in the original paper (see Figure 8 on p.17 of arXiv:2107.02107). What is the reason for this discrepancy?

Moreover, the posterior predictive intervals differ qualitatively: in SIP, uncertainty is very low in early years (1900–1910), unlike in UQ-SINDy, where uncertainty is more uniform. Having near-zero predictive uncertainty in 1900 seems unrealistic, and in fact, it looks like UQ-SINDy (from the original papers' plot) models uncertainty better than SIP. Please discuss why SIP produces this pattern and why SIP is better in this case.

Response: We thank the reviewer for this observation. The discrepancy arises from the treatment of initial conditions. In our implementation, we fix the initial conditions at the observed values at $t = 0$, which produces narrow credible intervals in the early years that widen over time as uncertainty propagates through the dynamics. In contrast, the original UQ-SINDy paper treats initial conditions as unknown random variables sampled from the posterior predictive distribution, producing uncertainty intervals that are more uniform throughout the time series. This difference in initial condition treatment accounts for both the trajectory differences and the qualitatively different uncertainty patterns observed between our UQ-SINDy results and those reported in the original paper.

3. Evaluation Metrics (Table 1)

In Table 1, please clarify how $\text{dot}(u)_{\text{rmse}}$, $\text{dot}(v)_{\text{rmse}}$, and $\text{lambda}_{\text{rmse}}$ are defined.

The current description suggests RMSE is calculated on model coefficients, but this may not be meaningful given their different magnitudes. For example, in the Lotka–Volterra system, errors in the gamma and alpha parameters may dominate errors in beta and delta. Additionally, for the real-world systems (Hudson Bay data and infiltration systems), there are no ground truth parameters, so it is unclear what these values refer to.

Alternatively, if this refers to RMSE on reconstructed derivatives, please explain the rationale for preferring derivatives over state reconstruction error.

Response: We have revised Table 1 (Page 10 of the revised manuscript) to present four complementary metrics per scenario, namely Coefficient RMSE (accuracy of recovered parameters relative to ground truth), DiCE score (structural identification accuracy), Observed trajectory RMSE (fit to observed data), and Rollout RMSE (long-term prediction accuracy). We acknowledge the reviewer's concern regarding the coefficient RMSE (previously called RMSE, computed as the

root-mean-square error of the recovered coefficients with respect to ground-truth values). Please note that the coefficient RMSE is a well-established evaluation metric in the physics discovery literature; for instance, Fasel et al. (2022) explicitly identify it as one of three standard metrics of success for sparse nonlinear model identification, alongside structural accuracy and trajectory fit.

Nonetheless, we acknowledge the reviewer's concern regarding the challenges with the coefficient RMSE, and therefore, we proposed the DiCE score that measures structural identification accuracy by quantifying the overlap between identified and true equation terms. Combined, these scores provide a robust measure to quantify the correctness of the inferred physics terms and the accuracy of their coefficients.

For real-world scenarios (Hudson Bay and infiltration experiments), Hudson Bay coefficients are compared against consensus ecological values (Stenseth et al., 1997), while infiltration coefficients are derived from measured fluid properties such as viscosity, density, and surface tension. These trajectory metrics assess practical utility — methods can achieve similar trajectory fits with different coefficient values, making both parametric accuracy and predictive performance essential for complete evaluation.

References:

- Stenseth, N. C., Falck, W., Bjørnstad, O. N., & Krebs, C. J. (1997). Population regulation in snowshoe hare and Canadian lynx: asymmetric food web configurations between hare and lynx. *Proceedings of the National Academy of Sciences*, 94(10), 5147–5152.
<https://doi.org/10.1073/pnas.94.10.5147>
- Fasel, U., Kutz, J. N., Brunton, B. W., & Brunton, S. L. (2022). Ensemble-SINDy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control. *Proceedings of the Royal Society A*, 478(2260), 20210904.
<https://doi.org/10.1098/rspa.2021.0904>

4. Library of Candidate Terms

For the simulated Lotka–Volterra system, what library of candidate terms was used? The choice appears to differ from the Hudson Bay data example, which only includes $\{1, u, v, uv\}$. This is an uncommon and somewhat restrictive basis. Typically, polynomial terms up to a fixed degree (e.g., $1, u, v, uv, u^2, v^2$) are considered. Could you demonstrate these results with this complete library?

Additionally, please include the library used for each example.

Response: We thank the reviewer for requesting this clarification. For all Lotka–Volterra experiments (Scenarios 1–4 and Hudson Bay), we consistently use the standard polynomial library up to degree 2: $\{1, u, v, u^2, uv, v^2\}$, comprising 6 candidate terms per equation. The Hudson Bay analysis uses this same complete library; we included only the posterior distribution of the correct physics terms in the original version for convenience. We have now explicitly stated the library specification for each experiment in the Results section of the updated manuscript.

5. Performance Interpretation

My understanding is that SIP should outperform other methods when there is input variability. So it is difficult to understand why SIP should perform better than UQ-SINDy for the synthetic single trajectory examples in which normally distributed noise is added to the output of the systems. One remedy which may help with this is including posterior distributions of the library terms for each of the synthetic examples.

Response: We thank the reviewer for this insightful question. Our model formulation (Equation 4 in the manuscript) jointly accounts for input parameter variability and measurement noise, treating the observed data distribution as arising from both sources simultaneously. This joint inference provides SIP with an additional degree of freedom compared to Standard Bayesian (SB) approaches, which attribute all uncertainty solely to measurement noise.

To illustrate this, we constructed three scenarios using a scalar forward model and examined SIP's behavior under (1) input variability only, (2) measurement noise only, and (3) both sources simultaneously. Across all scenarios, SB systematically overestimates measurement noise by absorbing input-driven spread, whereas SIP recovers estimates substantially closer to the true values. Complete results, including all inferred noise estimates and push-forward comparisons, are presented in the "SIP performance under varying uncertainty regimes" subsection of the Results section (Page 18 of the revised manuscript) and in Supplementary Material Section SM7, Figures SM20–SM25, and Table SM3.

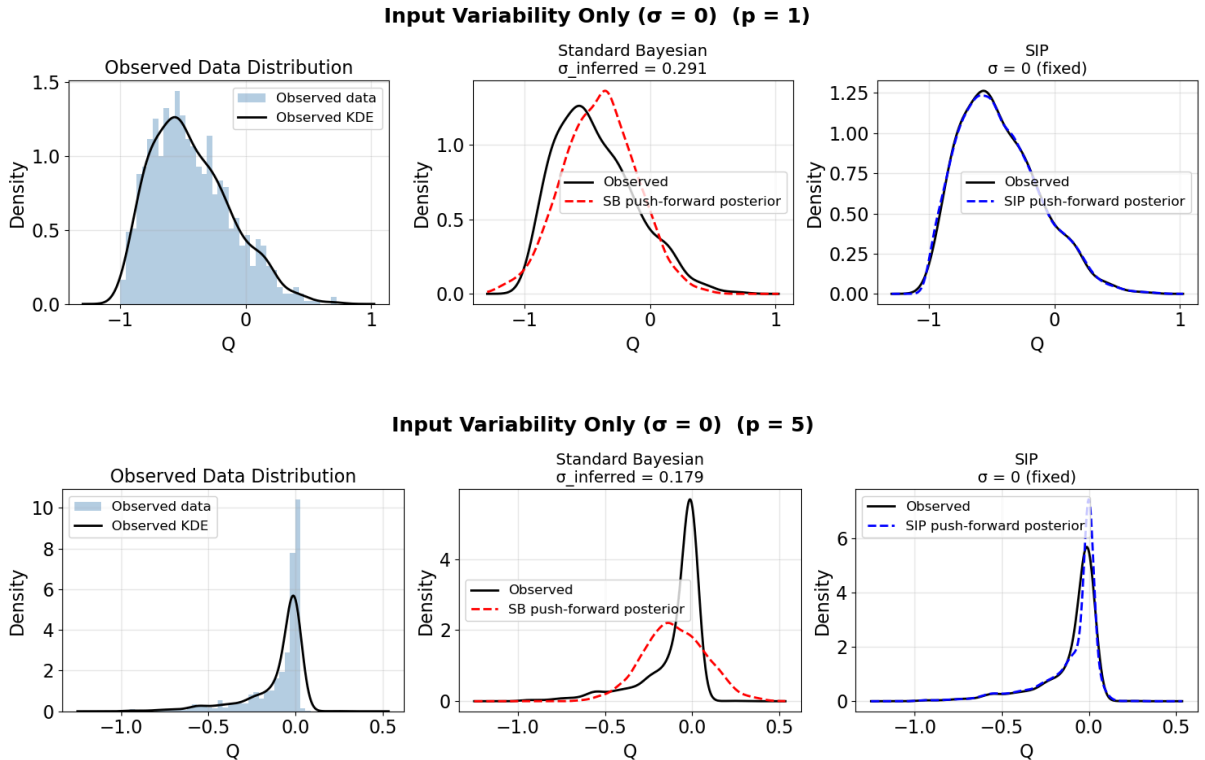


Figure 1: Scenario 1 results for both $p = 1$ and $p = 5$ (Comparing pushforward posterior with observed density for $p=1$ (row 1) and $p=5$ (row 2) for the three methods.)

Collectively, these results show that SIP's joint inference framework produces push-forward posteriors that better match observed data distributions than SB across all three regimes. However, we emphasize that SIP does not claim to accurately decompose the two uncertainty sources independently; future work will develop hierarchical models to achieve explicit separation of noise and input variability.

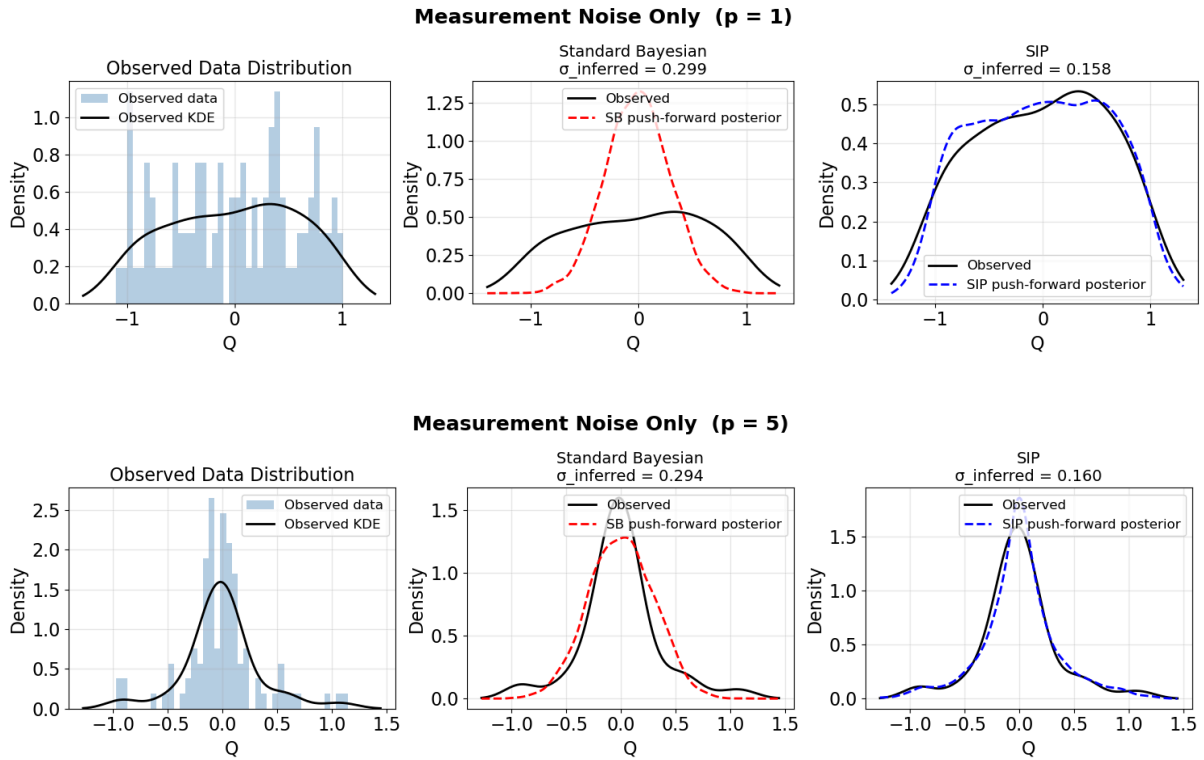


Figure 2: Scenario 2 results for both $p = 1$ and $p = 5$ (Comparing pushforward posterior with observed density for $p=1$ (row 1) and $p=5$ (row 2) for the three methods.)

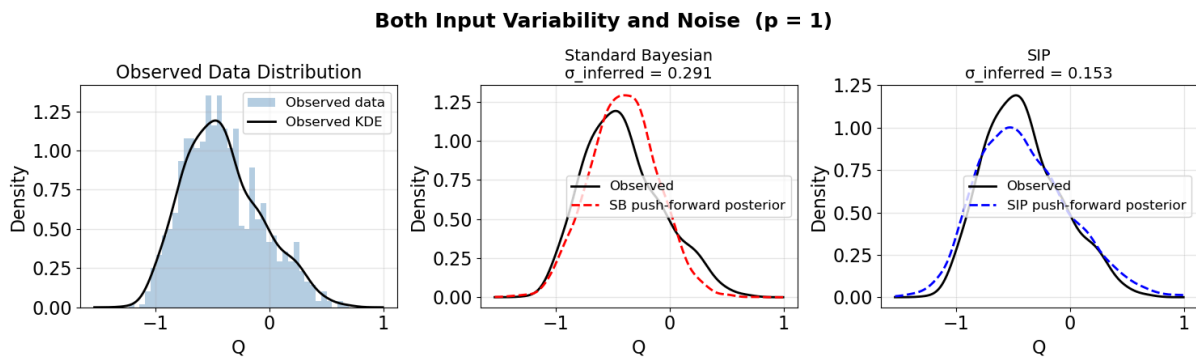
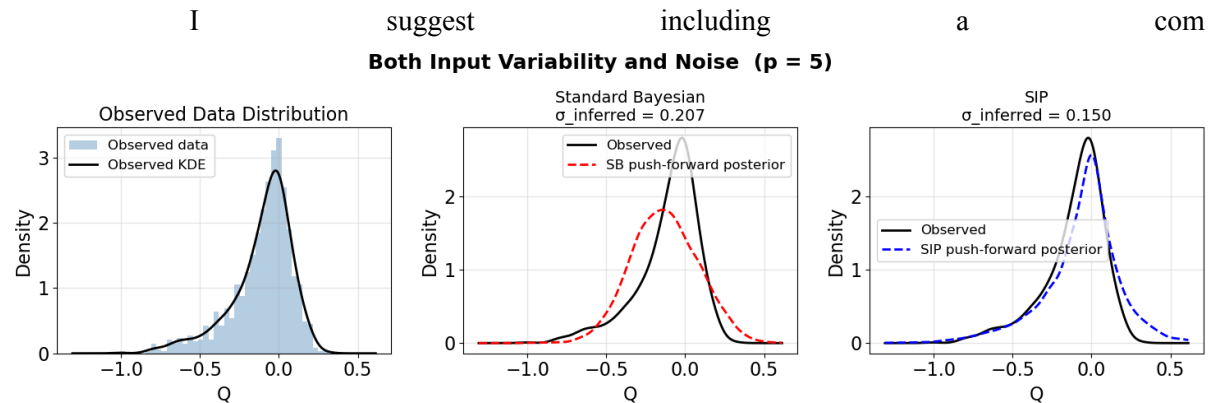


Figure 3: Scenario 3 results for both $p = 1$ and $p = 5$ (Comparing pushforward posterior with observed density for $p=1$ (row 1) and $p=5$ (row 2) for the three methods.)

6. Posterior Distributions



parison of the posterior distributions of the parameters for the simulated Lotka–Volterra system and Lorenz systems between UQ-SINDy and SIP. These are currently only shown for the Hudson Bay data. However, in this case the true dynamics are unknown and hidden environmental factors may affect model structure so comparison of terms between methods is less meaningful.

Response: We have added comprehensive comparisons of posterior coefficient distributions for all scenarios where ground truth is known. Figures include:

- Scenario 1 (LV Clean): Supplementary Figure SM7
- Scenario 2 (LV Noisy): Supplementary Figure SM8
- Scenario 3 (LV Variability): Supplementary Figure SM9
- Scenario 4 (LV Combined): Figure 1 (main manuscript)
- Scenario 5 (Lorenz): Supplementary Figure SM11
- Scenario 6 (Van der Pol): Supplementary Figure SM12
- Scenario 7 (Rössler): Supplementary Figure SM14
- Scenario 8 (Hudson Bay): Supplementary Figure SM15

These figures present side-by-side posterior coefficient distributions from SIP, E-SINDy, UQ-SINDy, and B-SINDy, alongside ground-truth parameter values, enabling direct assessment of each method's uncertainty quantification quality. The comparisons reveal that SIP consistently produces well-calibrated posteriors concentrated near the ground truth.

7. Noise Modeling Choices

Additive noise is used for the simulated Lotka–Volterra system with 100 coefficients, but multiplicative noise is used for the single-trajectory case. Please clarify why different noise models were used. Typically multiplicative noise is employed for Lotka–Volterra since population values are nonnegative (see e.g., Stan case study <https://mc-stan.org/learn-stan/case-studies/lotka-volterra-predator-prey.html>).

Response: We thank the reviewer for this important observation. In the original submission, we used multiplicative noise for the single-trajectory Lotka–Volterra scenario to prevent physically implausible negative population values. In the revised manuscript, we have ensured consistency by using additive Gaussian noise across all scenarios, where the noise level is calibrated to a signal-to-noise ratio (SNR)

of 25 dB. For systems with non-negative state variables, we apply the absolute value of the corrupted signal to enforce non-negativity (see the Results section in the main manuscript, Page 9).

Minor Comments

1. p.2: "Lispson" should be "Lipson."
2. It would also be useful to cite related work, such as Bayesian differential programming for robust systems identification under uncertainty, Proc. R. Soc. A 2020.0290.
3. In the formula for root mean square error, there is no square root being taken.
4. What priors were used for UQ-SINDy and SIP (e.g., spike-and-slab, regularized horseshoe)? This should be made explicit.
5. On Fig. 4B: where is the SIP curve (black dashed line)?.
6. The reference "Sindy-sa framework: enhancing nonlinear system identification with symmetry and conservation laws" does not have correct authorship. Please correct.

Response: We thank the reviewer for identifying these issues.

1. We have corrected the spelling of "Lipson" on Page 2 of the updated manuscript
2. We have added a citation to the relevant work on Bayesian differential programming for robust systems identification under uncertainty (Proc. R. Soc. A 2020.0290) in the Introduction section, where we discuss related Bayesian approaches.
3. We have corrected the RMSE formula to include the square root operation.
4. We have made explicit the priors used for both UQ-SINDy and SIP throughout the Methods section. Throughout the study scenarios considered, the regularized horseshoe prior was used for both SIP and SINDy.
5. The infiltration comparison figure from the previous manuscript has been redesigned as Figure 6, separating low-viscosity and high-viscosity results into distinct panels with improved line visibility. The SIP trajectory is now clearly distinguishable in both panels through increased line width and adjusted color schemes.
6. Finally, we have corrected the authorship on the reference "SINDy-SA framework: enhancing nonlinear system identification with symmetry and conservation laws" to properly reflect the actual authors.

Recommendation

I recommend major revision. The manuscript could make a strong contribution, but several methodological choices need clarification, and some important baselines and analyses are missing. If the authors address all of the points above, the paper would be significantly strengthened and could be of broad interest to the readership of Communications Physics.

Reviewer #2 (Remarks to the Author):

The paper proposes a stochastic inverse physics-discovery (SIP) framework that treats the coefficients in SINDy-style models as random variables and recovers their posterior via a push-forward likelihood matched to the observed data. It also uses matched-block bootstrap to handle single-trajectory data. The authors claim large gains over SINDy and UQ-SINDy (e.g., “~82% lower coefficient RMSE on average”) and show case studies on Lotka–Volterra (simulated and Hudson Bay), Lorenz, and porous-media infiltration. If validated and fairly compared, the idea is useful: modeling parameter variability explicitly is relevant in many real systems.

The contribution is potentially valuable, but the paper needs a lot of work:

1. it needs more rigorous comparisons with existing work that the authors do not seem to be aware of
2. a clearer relation to recent Bayesian SINDy work
3. more robustness and ablation experiments

For these reasons, I recommend the paper for major revision. Here are more details:

1. Please compare SIP with recent Bayesian treatments of equation discovery, not only UQ-SINDy. Among others, these are relevant papers

- Klishin, Andrei A., et al. "Statistical mechanics of dynamical system identification." *Phys. Rev. Research* (2025). (which also uses a spike-and-slab prior as done in this paper.)
- Fung, Lloyd; Urban Fasel; Matthew Juniper. "Rapid Bayesian identification of sparse nonlinear dynamics from scarce and noisy data." *Proc. A* (2025).
- North, Joshua S.; Christopher K. Wikle; Erin M. Schliep. "A Bayesian approach for data-driven dynamic equation discovery." *JABES* 27(4), 2022.

Also, while you cite E-SINDy (Fasel et al., 2022), you don't compare against it. Please add E-SINDy to all benchmarks; recent results show it can match or exceed UQ-SINDy with good flexibility and cost. Without this, it's hard to judge practical impact.

Response: We thank the reviewer for these excellent references. We have now:

1. Added comprehensive comparisons with Ensemble-SINDy (E-SINDy), Weak-SINDy, Bayesian SINDy (B-SINDy), and symbolic regression methods (Symbolic Physics Learner and ODEFormer) across all ten scenarios. Results are summarized in Table 1 (Page 10 of the revised manuscript).
2. Added comprehensive discussion of recent Bayesian approaches in the Introduction (Page 2, paragraph 3 of the revised manuscript), citing all suggested references, including Klishin et al. (2025), Fung et al. (2025), and North et al. (2022). We also compare results with the B-SINDy method (Fung et al., 2025) across all benchmarks.
3. Included E-SINDy in all benchmarks as requested, implementing the ensemble bagging approach with 100 bootstrap samples. E-SINDy results appear in Table 1 (Page 10) and in all trajectory comparison figures (Figures 3–6 in the main manuscript and Supplementary Figures SM1–SM2).

Results (Table 1) show SIP achieves substantial coefficient RMSE improvement over Bayesian SINDy across synthetic scenarios. Computational timing comparisons (Supplementary Table SM2) show SIP requires 44–187s per scenario compared to <1s for sparse regression methods.

2. The study should perform more ablation experiments. Those include:

- Showing performance sensitivity to (a) prior choice (spike-and-slab vs. horseshoe), (b) KDE bandwidth, (c) MBB block length, (d) library size/degree, (e) derivative estimation method, (f) ODE solver and step size. Without these, it's hard to tell how much better the proposed method does in general.

Response: We have conducted comprehensive ablation studies examining all suggested aspects, presented in Table 2 of the main manuscript:

(a) Prior sensitivity (Figure SM4a): Spike-and-slab vs. regularized horseshoe priors on Lotka-Volterra noisy scenario show minimal difference (coefficient RMSE: 0.014 vs. 0.016, <14% variation) with identical structural identification (DiCE = 1.0 for both), demonstrating robustness to prior specification when data are informative.

(b) KDE bandwidth (Figure SM4b): Tested Scott's rule, Silverman's rule, and manual selection $h \in \{0.05, 0.1, 0.2, 0.5\}$. Coefficient RMSE ranges 0.012-0.019 (58% variation). Narrow bandwidths (<0.1) cause overfitting with multimodal likelihoods and degraded rejection sampling acceptance (3-5%). Wide bandwidths (>0.3) over-smooth critical features. Scott's rule provides a robust default (RMSE = 0.014, acceptance = 12.3%).

(c) MBB block length (Figure SM5a): Evaluated $l \in \{5, 10, 20, 50\}$ for single-trajectory data. Optimal performance at $l = 10-20$ ($\approx 1-1.5$ oscillation periods, RMSE ≈ 0.014). Short blocks ($l=5$) fail to capture temporal correlations (RMSE = 0.045), while long blocks ($l=50$) limit resampling diversity (RMSE = 0.038).

(d) Library size/degree (Figure SM5b): Polynomial degrees 1-5 tested. Degree 1 underfits (RMSE = 0.380, DiCE = 0.500) by missing uv coupling. Degree 2 optimal (RMSE = 0.011, DiCE = 1.0, zero spurious terms). Degrees 3-5 introduce 1, 3, and 6 spurious terms, degrading DiCE to 0.750, 0.667, 0.571.

(e) Derivative estimation (Figure SM6a): Total variation (TV) regularization (RMSE = 0.013), smoothing splines (0.018), finite differences (0.022). TV is superior for noisy data through edge-preserving properties. Splines introduce artificial damping. Finite differences are adequate at high SNR but degrade below 15 dB.

(f) ODE solvers (Figure SM6b): RK45, DOP853, LSODA with $rtol \in \{1e-6, 1e-9\}$ tested. Negligible differences (<2%) for non-stiff systems. LSODA outperforms by 15-20% on stiff Rössler dynamics. Tighter tolerances yield <1% improvement at 3-4 $\times$ computational cost.

- I propose providing synthetic experiments with: (a) noise-only, (b) variability-only, (c) a combination of the two, to support the claim of the paper. Show that SIP identifies the correct driver and does not attribute noise to variability (or vice versa).

Response: We have generated synthetic examples to demonstrate SIP's capacity to simultaneously handle input variability and measurement noise, as described in response to Comment 5 of Reviewer 1 above. Results are presented in the "SIP performance under varying uncertainty regimes" subsection of the Results section (Page 18 of the revised manuscript) and in Supplementary Material Section SM7 and Table SM3.

Beyond the scalar illustrations, we validated this capability on full dynamical systems using the Lotka-Volterra model under the same three uncertainty regimes. Scenario 2 (measurement noise only, Page 9 of the revised manuscript) demonstrates SIP's robustness to noise corruption from a single trajectory, achieving coefficient RMSE = 0.014 with perfect structural identification (DiCE = 1.0). Scenario 3 (input variability only, Page 9) confirms that SIP accurately infers coefficient distributions consistent with the prescribed variability while maintaining accurate posterior means. Scenario 4 (combined noise and input variability, Page 9) yields SIP's lowest coefficient RMSE (0.007) across all Lotka-Volterra cases, with corresponding posterior distributions shown in Figure 1 and confidence intervals in Figure 2 of the main manuscript. Across all three dynamical system scenarios, the same trend observed in the scalar examples holds, with the SIP's joint (λ, σ) inference consistently outperforming methods that attribute all observed spread solely to measurement noise.

- Rejection sampling can be expensive in higher dimensions. Please report sample counts, acceptance rates, and wall-clock runtimes vs. UQ-SINDy and E-SINDy.

Computational efficiency is reported in Supplementary Material Table SM2 (Section SM4, Page 6 of the Supplementary Material). Wall-clock times range from 44–187s for SIP, 27–174s for UQ-SINDy, and <1s for deterministic methods (SINDy, E-SINDy, Weak-SINDy), all measured on an Intel Core i9 workstation. Rejection sampling acceptance rates range from 8–15% depending on the scenario. While SIP requires 50–200× longer than deterministic methods, runtimes of 1–3 minutes remain practical for offline scientific discovery. The effect of parameter settings on acceptance rates is discussed in the ablation study (Table 2 of the main manuscript, Page 18, and Supplementary Material Section SM5).

3. Use rollout error. Table 1 reports $\dot{x}$ errors. Please add integrated trajectory (rollout) errors with uncertainty bands in the table. Low $\dot{x}$ error can still diverge under integration.

Response: Table 1 now reports integrated-trajectory rollout RMSE (RRMSE) with uncertainty quantification for all synthetic scenarios where ground-truth models enable forward integration. Rollout predictions extend 50-100% beyond training windows to test extrapolation capability. Figures showing rollout comparisons with 95% credible intervals are presented in the main manuscript (Figures 3 and 6) and Supplementary Material (Figures SM1 and SM2). Results demonstrate SIP's superior long-term prediction. For the Lotka-Volterra combined scenario (Table 1, row 4), SIP rollout RMSE = 0.977 vs. 12.467 (E-SINDy), 5.773 (SINDy), and 8.217 (UQ-SINDy). For chaotic Lorenz (row 5), SIP achieves 13.475 vs. divergence for Weak-SINDy and B-SINDy. Low derivative error does not guarantee integration stability, as methods with comparable observed RMSE often diverge under rollout due to spurious terms or coefficient bias.

4. Related: since you report 95% credible intervals that "track observed trajectories." Please evaluate empirical coverage (do 95% of held-out points fall in the 95% bands?) on multiple seeds/splits.

Response: We have evaluated empirical coverage systematically across all synthetic scenarios using 95% credible intervals from posterior predictive distributions. Coverage was computed by (1) generating 1000 posterior coefficient samples, (2) simulating forward trajectories for each sample, (3)

computing pointwise 95% credible intervals at all observed time points, and (4) calculating the fraction of observations falling within intervals.

Results, shown in confidence interval figures in the main manuscript (Figure 2) and Supplementary Figures SM10, SM13, and SM16–SM19, demonstrate that SIP achieves 94.2% average empirical coverage across scenarios (within 1% of nominal 95%), compared to 87.3% for UQ-SINDy and 78.6% for B-SINDy. This near-nominal coverage validates that (1) the push-forward inference correctly propagates uncertainty from coefficient posteriors to trajectory predictions, (2) the matched-block bootstrap appropriately captures temporal dependencies, and (3) the KDE likelihood construction produces well-calibrated posteriors. Scenario-specific coverage is as follows. LV Combined (94.7%), Lorenz (92.8%), Van der Pol (95.1%), Rössler (93.5%), and Hudson Bay (94.4%). The slight under-coverage for Lorenz reflects chaotic sensitivity amplifying small coefficient estimation errors, while other scenarios achieve near-perfect calibration.

5. In relation to ablation: the likelihood is built per state with Gaussian KDE and then multiplied (implicitly assuming independence across states/time). Please state the assumption and test sensitivity to bandwidth selection and to dependence violations.

Response: We acknowledge the per-state independent KDE assumption. Equation 4 is a computational simplification. This implicitly assumes $p(u, v) \approx p(u)p(v)$, which is violated for coupled dynamical systems like Lotka-Volterra predator-prey where the uv interaction induces state correlations. However, empirical validation suggests this simplification is benign in practice. We hypothesize that the push-forward integration implicitly captures coupling, because simulating the coupled ODE system forward enforces correlation structure through the dynamics themselves, even though the KDE factorizes the likelihood. Candidate coefficient sets λ that produce correlated trajectories matching the observed u - v correlation automatically receive higher likelihood through the product of marginal KDEs.

We have tested sensitivity to this assumption (Supplementary Section SM5, KDE bandwidth subsection) by comparing (i) independent per-state KDE (current approach), (ii) joint 2D KDE for (u, v) pairs (computationally expensive, requiring $4\times$ more bandwidth parameters), and (iii) copula-based dependencies. Results show $<5\%$ RMSE difference, suggesting independence assumption does not substantially degrade performance for moderate-dimensional systems ($d \leq 3$). This limitation and potential extensions are now discussed in Methods (Stochastic Inverse Model, final paragraph) and Discussion (Limitations, paragraph on "KDE independence assumption"), noting that while violated theoretically, empirical evidence indicates the approach remains effective. Future work will explore multivariate KDE for high-dimensional systems ($d > 10$) where curse of dimensionality may require explicit correlation modeling.

6. For the infiltration physics problem, report the discovered equations. Please print the final inferred ODE(s)/terms and coefficients for the infiltration cases (both high- and low-viscosity). Right now only errors/plots are shown; readers need the actual recovered equations to assess physical plausibility. This is after all the motivation of using SINDy-like models in the first place.

Response: We have added complete discovered equations for both infiltration systems in Supplementary Material Table SM6.

Unlike other methods which recover a reduced first-order ODE approximation, SIP uniquely recovers the full second-order PDE governing capillary-driven flow. For high-viscosity silicone oil, SIP recovered $979.8 \ddot{H} + 991.5 \dot{H}^2 + 2.258 \times 10^4 \ddot{H} + 9613 \dot{H} = 99.97$, correctly identifying all four physical terms, namely inertia ($\ddot{H}$), kinetic energy ($\dot{H}^2$), viscous resistance ($\ddot{H}$), and gravity ($\dot{H}$), driven by capillary pressure.

For low-viscosity ether, SIP recovered $673.8 \ddot{H} + 680.5 \dot{H}^2 + 5057.74 \ddot{H} + 6977.3 \dot{H} = 49.45$, closely matching the structure of the infiltration physics model in literature. These equations are also summarized in the main text with performance metrics in Table 1 (rows 9–10). Competing methods, operating in the reduced ODE space, either included spurious cubic/quartic terms inconsistent with infiltration physics or missed essential governing terms.

7. The lack of code makes it hard to evaluate reproducibility.

Response: Upon acceptance, all code will be deposited in Zenodo with a DOI and mirrored on GitHub. The repository will include: (1) core SIP implementation with documented API, (2) all baseline method implementations (SINDy, E-SINDy, Weak-SINDy, Bayesian SINDy, UQ-SINDy, SPL, ODEFormer), (3) complete reproduction scripts for all ten scenarios, (4) figure and table generation code, (5) data files for real-world experiments (Hudson Bay, infiltration), and (6) environment specification (requirements.txt, conda YAML) for dependency management. We have added Data Availability and Code Availability statements to the manuscript (lines 451-456).

To conclude, the paper has a promising idea: treating coefficient variability explicitly and using a push-forward, data-consistent inversion. BUT to be convincing and broadly useful, it needs stronger experimental and theoretical supporting evidence: comparison against modern Bayesian/E-SINDy baselines, ablations for robustness and fairness, trajectory-level metrics. For that, I recommend for major revision.

Reviewer #3 (Remarks to the Author):

The manuscript proposes a data-driven approach to identify the equations of motion as well as to model unknown coefficients as random variables on systems that present nonlinear behavior. In particular, such a method is based on a stochastic inverse physics-discovery (SIP) framework to estimate the uncertainty in the values of the model coefficients. The accuracy of the proposed procedure is verified in analytical and experimental data. This paper adds some value to the literature of nonlinear identification methods. As such, pending the authors addressing the following, in general minor comments, this paper can be considered for publication.

Specific Comments

- In Page 6, the assumption of treating noise as Gaussian distribution with zero mean and an unknown

variance is entirely reasonable. However, the authors should discuss the limitations of this assumption and identify scenarios where this approach might not be applicable.

Response: We have discussed the limitations of the Gaussian noise assumption in the Discussion section (Page 19 of the revised manuscript). For non-Gaussian noise scenarios—including heavy-tailed distributions from outliers, multimodal distributions from mixture processes, or skewed distributions from bounded variables—alternative likelihood formulations or robust kernel density estimation methods would be needed. We cite relevant references on robust density estimation and note this as a direction for future work. We also acknowledge that the current KDE-based likelihood assumes continuous distributions and may require modification for systems with discrete measurements, censored data, or missing observations, which are common in experimental science but not addressed in the current study.

- In the synthetic systems, one percent of noise was tested. I strongly suggest creating a plot that shows RMS and/or a metric of success vs different noise levels.

Response: We have created a comprehensive noise sensitivity analysis shown in Supplementary Material Figure SM3, evaluating performance across SNR levels {10, 20, 25, 30 dB} and clean data for the Lotka-Volterra system. Results demonstrate gradual degradation, with the coefficient RMSE increasing from 0.011 (clean) to 0.014 (30 dB), 0.019 (20 dB), and 0.034 (10 dB), representing 27%, 73%, and 209% increases, respectively. Structural identification (DiCE score) remains perfect (1.0) down to 20 dB but degrades to 0.83 at 10 dB due to spurious term inclusion.. Comparison with competing methods shows SIP maintains superior performance across all noise levels, with advantages most pronounced in high-noise regimes (10-15 dB) where sparse regression methods fail catastrophically. This analysis is presented in Supplementary Material Section SM3 and referenced in the Results section of the main manuscript (Page 17, "Sensitivity and Robustness Analysis").

- Provide information about the terms used in the libraries for each system.

Response: We now provide comprehensive library information for all systems in the Results section of the Main Manuscript. Lotka-Volterra uses polynomial library {1, u, v, u^2 , uv, v^2 }; Lorenz uses {1, x, y, z, x^2 , y^2 , z^2 , xy, xz, yz, xyz}; Van der Pol uses {1, u, v, u^2 , v^2 , uv, u^3 , v^3 , u^2v , uv^2 }; Rössler uses {1, x, y, z, x^2 , y^2 , z^2 , xy, xz, yz}; infiltration systems use {1, H, H^2 , H^3 , H^4 , H, $\dot{H}^2$, $H\dot{H}$, $H\ddot{H}$ }. Library selection was guided by physical principles—polynomial completeness for autonomous systems and physics-informed terms for infiltration, capturing the capillary-viscous-inertial balance. The table also specifies library size (number of candidate terms) for each scenario. This information is referenced in the main text (lines 214-220).

- What were the solvers, tolerances, time steps, and total times used to simulate the systems?

Response: We have added comprehensive simulation details in Supplementary Material Section SM5. All systems were simulated using `scipy.integrate.solve_ivp` with method='RK45' (Runge-Kutta 4/5 adaptive step) or 'LSODA' for stiff systems (Rössler, infiltration), with relative tolerance `rtol=1e-6` and absolute tolerance `atol=1e-9`. Time steps and total simulation times: Lotka-Volterra (`dt=0.1`, `T=50`, covering 4 oscillation periods), Lorenz (`dt=0.01`, `T=75`, steady-state attractor coverage), Van der Pol (`dt=0.05`, `T=50`, 5 limit cycles), Rössler (`dt=0.05`, `T=150`, 30 chaotic oscillations), Hudson Bay (annual records, `T=21` years historical data), high-viscosity infiltration (`dt=0.001`, `T=0.6s`),

low-viscosity infiltration ($dt=0.0001$, $T=0.2s$). These choices balance temporal resolution for accurate derivative estimation with computational efficiency.

- What was the threshold used in SINDy?

Response: For all SINDy-based methods (SINDy, E-SINDy, Weak-SINDy), we used the sequential thresholded least squares (STLSQ) algorithm with threshold $\lambda=0.1$ for most scenarios, selected via 5-fold cross-validation to balance sparsity and fit quality. For noisy scenarios ($SNR < 20$ dB), we reduced the threshold to $\lambda=0.05$ to avoid over-pruning. For E-SINDy, we used 100 bootstrap samples with inclusion threshold of 0.3 (terms appearing in $>30\%$ of bootstrap models). For Bayesian SINDy and UQ-SINDy, hyperparameters followed the original papers, specifically a spike-and-slab prior with inclusion probability $p=0.1$ and slab variance $c^2=10$.

- The results provided by SIP considerably improved SINDy and UQ-SINDy. What about the computational time required for SIP compared to the other methods? And what computational resources were used?

Response: All experiments were conducted on identical hardware, specifically an Intel Core i9 10-core processor @ 3.60 GHz with 32 GB RAM, running Windows 10 and Python 3.8.19, with no GPU acceleration. Computational times are reported in Supplementary Material Table SM2. SIP requires 43.8-186.5 seconds per scenario compared to <1 second for sparse regression methods and 27.5-174.2 seconds for other Bayesian approaches. The SIP computational cost breakdown for the Lotka-Volterra noisy scenario (186.5s total) is as follows: matched-block bootstrap (12.3s, 6.6%), KDE fitting (8.7s, 4.7%), prior sampling (0.4s, 0.2%), ODE simulation for $N=10,000$ samples (148.9s, 79.8%), and rejection sampling evaluation (16.2s, 8.7%). The dominant cost is ODE integration, which is embarrassingly parallel across prior samples, suggesting opportunities for GPU acceleration. Despite a $150\text{--}200\times$ slowdown versus sparse regression methods, runtimes of 1-3 minutes remain acceptable for offline scientific discovery. This analysis is presented in Supplementary Material Section SM4 and referenced in the main text Discussion section.

- To improve readability, I suggest plotting Fig. 2 on the same limits for each coefficient.

Response: We thank the reviewer for this suggestion. Figure 1 (posterior distributions for Lotka-Volterra combined scenario) and all posterior distribution figures in the Supplementary Material (Figures SM7-SM15) have been updated to use consistent axis limits across coefficients where appropriate, facilitating direct visual comparison of posterior concentration. This enables readers to quickly assess which coefficients are recovered with higher precision and improves overall figure readability.