ROBUST AMORTIZED BAYESIAN INFERENCE WITH SELF-CONSISTENCY LOSSES ON UNLABELED DATA

Aayush Mishra
Department of Statistics
TU Dortmund University, Germany
equal contribution

Daniel Habermann
Department of Statistics
TU Dortmund University, Germany
equal contribution

Marvin Schmitt
Independent Scientist

Stefan T. Radev
Department of Cognitive Science
Rensselaer Polytechnic Institute, USA
Email: radevs@rpi.edu

Paul-Christian Bürkner
Department of Statistics
TU Dortmund University, Germany
Email: paul.buerkner@gmail.com

ABSTRACT

Neural amortized Bayesian inference (ABI) can solve probabilistic inverse problems orders of magnitude faster than classical methods. However, neural ABI is not yet sufficiently robust for widespread and safe applicability. In particular, when performing inference on observations outside of the scope of the simulated data seen during training, for example, because of model misspecification, the posterior approximations are likely to become highly biased. Due to the bad pre-asymptotic behavior of current neural posterior estimators in the out-of-simulation regime, the resulting estimation biases cannot be fixed in acceptable time by just simulating more training data. In this paper, we propose a semi-supervised approach that enables training not only on (labeled) simulated data generated from the model, but also on unlabeled data originating from any source, including real-world data. To achieve the latter, we exploit Bayesian self-consistency properties that can be transformed into strictly proper losses without requiring knowledge of true parameter values, that is, without requiring data labels. The results of our initial experiments show remarkable improvements in the robustness of ABI on out-of-simulation data. Notably, inference remains accurate even when the observed data lies far outside both the labeled and unlabeled training distributions. If our findings generalize to other scenarios and model classes, our method could offer a significant step forward towards robust neural ABI.

1 INTRODUCTION

Theory-driven computational models (i.e., process, simulation, or mechanistic models) are highly influential across numerous branches of science (Lavin et al., 2021). The utility of computational models largely stems from their ability to fit real data x and extract information about hidden parameters θ. Bayesian methods have been instrumental for this task, providing a principled framework for uncertainty quantification and inference (Gelman et al., 2013). However, gold-standard Bayesian methods, such as Gibbs or Hamiltonian Monte Carlo samplers (Brooks et al., 2011), remain notoriously slow. Moreover, these methods are rarely feasible for fitting complex models (Dax et al., 2021) or even simpler models in big data settings with many thousands of data points in a single dataset (Blei et al., 2017), or when thousands of independent datasets require repeated model re-fits (von Krause et al., 2022).

In recent years, deep learning methods have helped address some of these efficiency challenges (Cranmer et al., 2020). In particular, *amortized Bayesian inference* (ABI; Gershman & Goodman, 2014; Le et al., 2017; Gonçalves et al., 2020; Radev et al., 2023; Gloeckler et al., 2023; Elsemüller et al., 2024; Zammit-Mangion et al., 2024) has received considerable attention for its potential to automate Bayesian workflows by training generative neural networks on model simulations, subsequently enabling near-instant downstream inference on real data. However, due to the reliance on pre-trained neural networks, ABI methods can become unreliable when applied to data that is unseen or sparsely encountered during training. In particular, posterior samples from amortized methods may deviate significantly from samples obtained with gold-standard MCMC samplers when there is a mismatch between the simulator used in training and the real data (Ward et al., 2022; Schmitt et al., 2023; Siahkoohi et al., 2023; Gloeckler et al., 2023; Frazier et al., 2024). This lack of robustness limits the widespread and safe applicability of ABI methods.

In this work, we propose an approach that incorporates both supervised and unsupervised components to strengthen the robustness of ABI on out-of-simulation data. The supervised part learns from a "labeled" set of parameters and corresponding synthetic (simulated) observations, $\{\theta, x\}$, while the unsupervised part leverages an "unlabeled" data set of real observations $\{x^*\}$ without requiring the explicit parameters. Combining these two components leads to a *semi-supervised approach*, which contributes to an emerging stream of research concerned with increasing the robustness of ABI. Our approach can be understood as a form of meta-learning (Hospedales et al., 2021; Finn et al., 2017), where the objective is to optimize for a set of tasks utilizing minimal data. In the context of ABI, inference on data from different distributions (simulated and real-world data) constitutes such a set of tasks. By incorporating multiple data sources, we can quickly adapt to new and unseen data, improving the ability to generalize even when training data is scarce.

In contrast to other methods built to enhance the robustness of ABI, our approach does not require ground-truth parameters θ^* (Wehenkel et al., 2024), post hoc corrections (Ward et al., 2022; Siahkoohi et al., 2023), or specific adversarial defenses (Gloeckler et al., 2023), nor does it entail a loss of amortization (Ward et al., 2022; Huang et al., 2023) or generalized Bayesian inference (Gao et al., 2023; Pacchiardi et al., 2024). To achieve this, we expand on previous work on *self-consistency losses* (Schmitt et al., 2024; Ivanova et al., 2024), demonstrating notable robustness gains even for as few as four real-world observations.

2 METHODS

2.1 BAYESIAN SELF-CONSISTENCY

Self-consistency leverages a simple symmetry in Bayes' rule to enforce more accurate posterior estimation even in regions with sparse data (Schmitt et al., 2024; Ivanova et al., 2024). Crucially, it incorporates likelihood (when available) or a surrogate likelihood during training, thereby providing the networks with *additional information* beyond the standard simulation-based loss typically employed in ABI (see below).

Following Schmitt et al. (2024), we will focus on the marginal likelihood based on neural posterior or likelihood approximation. Under exact inference, the marginal likelihood is independent of the parameters θ. That is, the Bayesian self-consistency ratio of likelihood-prior product and posterior is constant across any set of parameter values $\theta^{(1)}, \ldots, \theta^{(L)}$,

$$p(x) = \frac{p(x \mid \theta^{(1)}) \, p(\theta^{(1)})}{p(\theta^{(1)} \mid x)} = \cdots = \frac{p(x \mid \theta^{(L)}) \, p(\theta^{(L)})}{p(\theta^{(L)} \mid x)}. \tag{1}$$

However, replacing $p(\theta \mid x)$ with a neural estimator $q(\theta \mid x)$ (likewise for the likelihood) leads to undesired variance in the marginal likelihood estimates across different parameter values on the right-hand-side (Schmitt et al., 2024). Since this variance is a proxy for *approximation error*, we can directly minimize it via backpropagation along with any other ABI loss to provide further training signal and reduce errors guided by density information. Our proposed semi-supervised formulation builds on these advantageous properties.
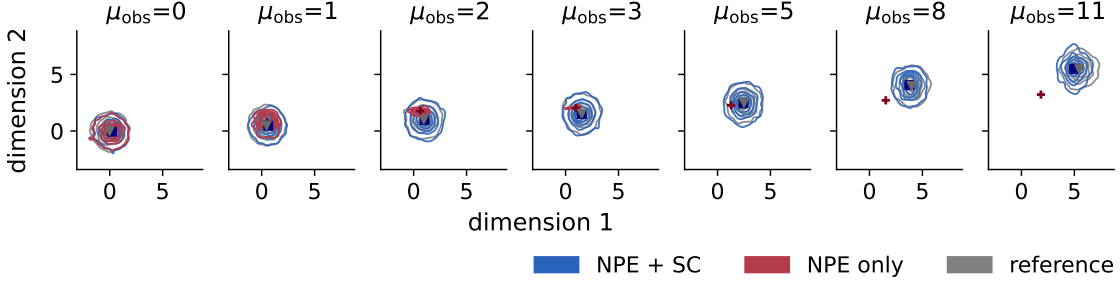
Figure 1: Contour plot of the normal means problem using standard NPE (red) or our semi-supervised approach (NPE + SC, blue), with the analytic posterior in gray. Symbols indicate posterior mean estimates (red cross: NPE only; blue square: NPE + SC; gray triangle: reference). Each subplot shows posterior inference on observed data that are increasingly distant from the labeled training data ($\mu_{\text{prior}} = 0$). Only the first two dimensions of the 10-dimensional posterior are shown. While standard NPE collapses to zero variance for $\mu_{\text{obs}} \geq 2$, adding the self-consistency loss preserves accurate posterior estimates even far beyond both training spaces ($\mu_{\text{obs}} > 3$). Training was performed using the default configuration (see Section 4.1).

2.2 SEMI-SUPERVISED AMORTIZED BAYESIAN INFERENCE

The formulation in Eq. (1) is straightforward, but practically never used in traditional sampling-based methods (e.g., MCMC) because they do not provide a closed-form for the approximate posterior density $q(\theta \mid x)$. In contrast, we can readily evaluate $q(\theta \mid x)$ in ABI when using a neural density estimator that allows efficient density computation (e.g., normalizing flows, Kobyzev et al. (2020)). Thus, we can formulate a family of *semi-supervised losses* of the form:

$$(q^*, h^*) = \underset{q,h}{\operatorname{argmin}} \, \mathbb{E}_{(\theta,x)\sim p(\theta,x)} \left[S(q(\theta \mid h(x)), \theta) \right] + \lambda \cdot \mathbb{E}_{x^* \sim p^*(x)} \left[C\left(\frac{p(x^* \mid \theta)\, p(\theta)}{q(\theta \mid h(x^*))} \right) \right], \quad (2)$$

where S is a strictly proper score (Gneiting & Raftery, 2007) and C is a self-consistency score (Schmitt et al., 2024). The neural networks to be optimized are a generative model q and (potentially) a summary network h extracting lower dimensional sufficient statistics from the data. We will call the first loss component, $\mathbb{E}_{(\theta,x)\sim p(\theta,x)} \left[S(q(\theta \mid h(x)), \theta) \right]$, the (standard) simulation-based loss, as it forms the basis for standard ABI approaches using simulation-based learning. E.g., this is the maximum likelihood loss for normalizing flows (Kobyzev et al., 2020; Papamakarios et al., 2021) or a vector-field loss for flow matching (Liu et al., 2023; Lipman et al., 2023). We will refer to the second loss component as the (Bayesian) self-consistency loss.

In practice, we approximate the expectations in Eq. (2) with finite amounts of simulated and real training data. That is, for N instances $(\theta_n, x_n) \sim p(\theta, x)$ and M instances $x_m^* \sim p^*(x)$, we employ

$$(q^*, h^*) = \underset{q,h}{\operatorname{argmin}} \, \frac{1}{N} \sum_{n=1}^{N} \left[S(q(\theta_n \mid h(x_n)), \theta_n) \right] + \lambda \cdot \frac{1}{M} \sum_{m=1}^{M} \left[C\left(\frac{p(x_m^* \mid \theta)\, p(\theta)}{q(\theta \mid h(x_m^*))} \right) \right]. \quad (3)$$

Asymptotically for $N \to \infty$, that is, for infinite training data generated from the simulator $p(\theta, x)$, a universal density estimator (Draxler et al., 2024) minimizing a strictly proper simulation-based loss (Gneiting & Raftery, 2007) is sufficient to ensure perfect posterior approximation for any data. By this, we mean that the posterior approximation becomes identical to the posterior we would obtain if we could analytically solve Bayes' Theorem $p(\theta \mid x) = p(x \mid \theta)p(\theta)/p(x)$. This *analytic posterior* is sometimes also referred to as true or correct posterior. In practice, the posterior is rarely analytic, but we can still verify the accuracy

of an approximation by comparing it with the results of a gold-standard approach (if available), such as a sufficiently long, converged MCMC run (Magnusson et al., 2024).

While neural posterior approximation is perfect asymptotically, its pre-asymptotic performance, that is, when training $q(\theta \mid h(x))$ only on a finite amount of simulated data, can become arbitrarily bad: For any kind of atypical data x^* that is outside the data space implied by $p(\theta, x)$, for instance, when the model is misspecified, the posterior approximation $q(\theta \mid h(x^*))$ may be arbitrarily far away from the analytical posterior $p(\theta \mid x^*)$ (Schmitt et al., 2023). As a result, a simulation-based loss is insufficient to achieve robust ABI in practice. This is where the self-consistency loss comes in: As we will show, adding the latter during training can greatly improve generalization to atypical data at inference time, rendering ABI much more robust.

One particular choice for C is the variance over parameters on the log scale of the Bayesian self-consistency ratio (Schmitt et al., 2024):

$$C\left(\frac{p(x^* \mid \theta)\,p(\theta)}{q(\theta \mid h(x^*))}\right) = \mathrm{Var}_{\theta \sim p_C(\theta)}\left[\log p(x^* \mid \theta) + \log p(\theta) - \log q(\theta \mid h(x^*))\right], \tag{4}$$

where $p_C(\theta)$ can be any proposal distribution over the parameter space, for example, the prior $p(\theta)$ or even the current approximate posterior $q_t(\theta \mid h(x^*))$ as given in a training iteration or snapshot t. Notably, the choice of $p_C(\theta)$ can influence training dynamics considerably, with the empirical consequences being difficult to anticipate (Schmitt et al., 2024). In pratice, we approximate the variance $\mathrm{Var}_{\theta \sim p_C(\theta)}$ by the empirical variance $\mathrm{Var}_{l=1}^{L}$ computed over L samples $\theta^{(l)} \sim p_C(\theta)$.

2.3 Self-Consistency Losses Are Strictly Proper

Below, we discuss the strict properness of Bayesian self-consistency losses, which underline their widespread usefulness. To simplify the notation, we denote posterior approximators simply as $q(\theta \mid x)$ without considering architectural details such as the use of summary networks $h(x)$. All theoretical results and their proofs remain the same if x is replaced by $h(x)$ as long as the summary network is expressive enough to learn sufficient statistics from x.

Proposition 1. *Let C be a score that is globally minimized if and only if its functional argument is constant across the support of the posterior $p(\theta \mid x)$ almost everywhere. Then, C applied to the Bayesian self-consistency ratio with known likelihood*

$$C\left(\frac{p(x \mid \theta)\,p(\theta)}{q(\theta \mid x)}\right) \tag{5}$$

is a strictly proper loss. That is, it is globally minimized if and only if $q(\theta \mid x) = p(\theta \mid x)$ almost everywhere.

In particular, the variance loss (4) fulfills the assumptions of Proposition 1.

Proposition 2. *The loss* (4) *based on the variance of the log Bayesian self-consistency ratio is strictly proper if the support of $p_C(\theta)$ encompasses the support of $p(\theta \mid x)$.*

The proofs of Propositions 1 and 2 are provided in Appendix A. The strict properness extends to semi-supervised losses of the form (2), which combine standard simulation-based losses with self-consistency losses.

Proposition 3. *Under the assumptions of Proposition 1, the semi-supervised loss* (2) *is strictly proper for any choice of $p^*(x)$.*

The proof of Proposition 3 follows immediately from the fact the sum of strictly proper losses is strictly proper. Importantly, since Proposition 3 holds independently of $p^*(x)$, it holds both in the case of a well-specified model, where $p^*(x) = p(x)$, and also in case of any model misspecification or domain shift where

$p^*(x) \neq p(x)$. That is, there is *no trade-off* in the semi-supervised loss (2) between the standard simulation-based loss and the self-consistency loss, since they are both globally minimized for the same target.

Lastly, for completeness, we can also define strictly proper self-consistency losses for likelihood instead of posterior approximations.

Proposition 4. *Consider the case where the posterior $p(\theta \mid x)$ is known and the likelihood is estimated by $q(x \mid \theta)$. Then, under the assumptions of Proposition 1, Bayesian self-consistency ratio losses of the form*

$$C\left(\frac{q(x \mid \theta)\, p(\theta)}{p(\theta \mid x)}\right) \tag{6}$$

are strictly proper. That is, they are globally minimized if and only if $q(x \mid \theta) = p(x \mid \theta)$ almost everywhere.

The proof of Proposition 4 proceeds in the same manner as for Proposition 1, just exchanging likelihood and posterior. Clearly, strict properness does not necessarily hold if *both* posterior and likelihood are unknown or approximate. This is because any pair of approximators $q(\theta \mid x)$ and $q(x \mid \theta)$ that satisfy $q(\theta \mid x) \propto q(x \mid \theta)\, p(\theta)$ minimize the self-consistency loss regardless of their relation to the accurate posterior $p(\theta \mid x)$ and likelihood $p(x \mid \theta)$. For example, the choices $q(\theta \mid x) = p(\theta)$ and $q(x \mid \theta) \propto 1$ minimize the self-consistency loss, but may be arbitrarily far away from their actual target distributions.

In other words, if both likelihood and posterior are unknown, the self-consistency loss needs to be coupled with another loss component, such as the maximum likelihood loss, to enable joint learning of both approximators $q(\theta \mid x)$ and $q(x \mid \theta)$ (Schmitt et al., 2024). Nevertheless, the self-consistency loss still yields notable improvements: as demonstrated in our experiments, the semi-supervised loss (2) considerably enhances the robustness of ABI *even when both the posterior and likelihood are unknown*.

3 RELATED WORK

The robustness of ABI and simulation-based inference methods more generally has been the focus of multiple recent studies (e.g., Frazier et al., 2020; Frazier & Drovandi, 2021; Frazier et al., 2024; Dellaporta et al., 2022; Ward et al., 2022; Gloeckler et al., 2023; Huang et al., 2023; Gao et al., 2023; Siahkoohi et al., 2023; Kelly et al., 2024; Wehenkel et al., 2024; Pacchiardi et al., 2024; Schmitt et al., 2023). These efforts can be broadly classified into two categories: (a) analyzing or detecting simulation gaps and (b) mitigating the impact of simulation gaps on posterior estimates.

Since our work falls into the latter category, we briefly discuss methods aimed at increasing the robustness of fully amortized approaches. For instance, Gloeckler et al. (2023) explore efficient regularization techniques that trade off some posterior accuracy to enhance the robustness of posterior estimators against adversarial attacks. Ward et al. (2022) and Siahkoohi et al. (2023) apply *post hoc* corrections based on real data, utilizing MCMC and the reverse Kullback-Leibler divergence, respectively. Differently, Gao et al. (2023) propose a departure from standard Bayesian inference by minimizing the expected distance between simulations and observed data, akin to generalized Bayesian inference with scoring rules (Pacchiardi et al., 2024). Perhaps the closest work in spirit to ours is Wehenkel et al. (2024), which introduces the use of additional training information in the form of a (labeled) calibration set (x^*, θ^*) that contains observables from the real data distribution as well as the corresponding ground-truth parameters.

In contrast to the methods above, our approach (a) avoids trade-offs between accuracy and robustness, (b) requires no modifications to the neural estimator after training, therefore fully maintaining inference speed, (c) affords proper Bayesian inference, and (d) does not assume known ground truth parameters for a calibration set. Thus, it can be viewed as one of the first instantiations of *semi-supervised* ABI.

4 CASE STUDIES

4.1 MULTIVARIATE NORMAL MODEL

We first illustrate the usefulness of our proposed self-consistency loss on the following multivariate normal model, also known as the normal means problem. This model has also been considered in previous studies investigating the robustness of neural ABI (Schmitt et al., 2023). It allows us to explore the behavior of NPE trained with the semi-supervised loss (2) in a simple, controlled setting where the true posterior is available in closed-form. The prior and likelihood are given by

$$\theta \sim \text{Normal}(\mu_{\text{prior}}, \sigma_{\text{prior}}^2 I_D), \qquad x^{(k)} \sim \text{Normal}(\theta, \sigma_{\text{lik}}^2 I_D) \tag{7}$$

The parameters $\theta \in \mathbb{R}^D$ are sampled from a D-dimensional multivariate normal distribution with mean vector μ_{prior} and diagonal covariance matrix $\sigma_{\text{prior}}^2 I_D$. Here, we fix $\mu_{\text{prior}} = 0$ and $\sigma_{\text{prior}}^2 = 1$. On this basis, K independent, synthetic data points $x^{(k)} \in \mathbb{R}^D$ are sampled from a D-dimensional multivariate normal distribution with mean vector θ and diagonal covariance matrix $\sigma_{\text{lik}}^2 I_D$. We fix $\sigma_{\text{lik}}^2 = K$ such that the total information in x remains constant, independent of K, which simplifies comparisons across observations of varying number of data points. More details on the training setup and employed neural architectures can be found in Appendix A.2.

In our numerical experiments, we study the influence of several aspects of the normal model on the performance of NPE. To prevent combinatorial explosion, we vary the factors below separately, with all other factors fixed to their default configuration (highlighted in **bold**): (1) parameter dimensionality ($D = 2, \mathbf{10}, 100$), (2) number of unlabeled observations for the self-consistency loss $\{x_m^*\}_{m=1}^M$ ($M = 1, 4, \mathbf{32}$), (3) mean μ^* of the unlabeled observations x_m^* ($\mu^* = 0, 1, 2, \mathbf{3}, 5$), (4) inclusion of a summary network ($K = 10$) or **not** ($\mathbf{K = 1}$), (5) likelihood function (**known**, estimated).

Results   In Figure 1, we depict the results obtained from (a) standard NPE (trained on the simulation-based loss only), (b) our semi-supervised NPE (with the self-consistency loss on known likelihood), and (c) the gold-standard (analytic) reference. We see that standard NPE already completely fails for $x_{\text{obs}} \sim N(\mu_{\text{obs}} = 2, 0.01 I_D)$, and subsequently also for any larger values $\mu_{\text{obs}} > 2$. In contrast, adding the self-consistency loss to obtain our semi-supervised approach achieves almost perfect posterior estimation. This holds true even in cases where x_{obs} is multiple standard deviations away from *all* the training data, that is, from both the labeled dataset $\{(\theta_n, x_n)\}_{n=1}^N$ and the unlabeled dataset $\{x_m^*\}_{m=1}^M$. These results indicate that the self-consistency criterion can provide strong robustness gains even far outside the typical space of training data.

In Figure 3 (Appendix A.3), we report the posterior mean and standard deviation bias as well as the maximum mean discrepancy (MMD) between the approximate and true posterior for the above five factors. When varying the parameter dimensionality, including the self-consistency loss yields nearly perfect posterior approximation up to 10 dimensions—even with extreme deviations from the initial training data. It also significantly improves estimates even for 100 dimensions while we do see larger discrepancies from the true posterior. The dataset size factor shows robust gains, with clear improvements over the standard simulation-based loss even when using as few as four unlabeled observations (versus 1024 labeled ones). Varying the mean μ_{obs} of the new observations shows that, as long as the data used for evaluating the self-consistency loss is not identical to the training data (i.e., as long as $\mu_{\text{obs}} \neq 0$), including the self-consistency loss component enables accurate posterior approximation far outside the typical space of the training data.

In Figure 4 in Appendix A.3, we see that the benefits of self-consistency persist when the posterior is conditioned on more than one data point per observation ($K = 10$), that is, in the presence of a summary network. Further, we still see clear benefits of adding the self-consistency loss even when the likelihood is estimated by a neural likelihood approximator $q(x \mid \theta)$, trained jointly with the posterior approximator $q(\theta \mid x)$ on the same training data. However, with an estimated likelihood, posterior bias, especially bias in the posterior standard deviation, and MMD distance to the true posterior are larger than in the known likelihood case.
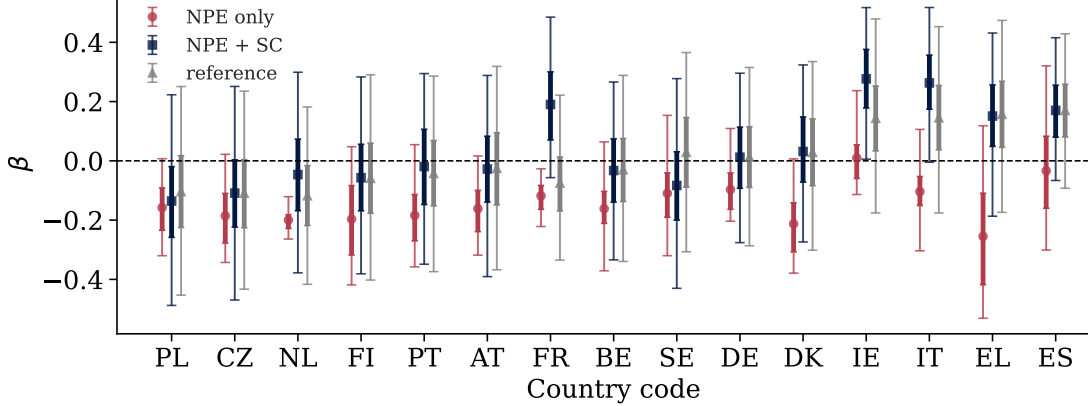
Figure 2: Comparison of posterior estimates for 15 countries (ISO 3166 alpha-2 codes) among standard NPE (red circles), NPE + self-consistency loss (blue squares), and Stan (reference; gray triangles). Central 50% (thick lines) and 95% (thin lines) posterior intervals of the autoregressive component β are shown, sorted by lower 5% quantile as per Stan. The self-consistency loss was evaluated on data from $M = \mathbf{8}$ countries during training, greatly enhancing ABI's robustness in both no-misspecification scenarios and real-data evaluations.

4.2 Forecasting Air Passenger Traffic: An Autoregressive Model with Predictors

Table 1: Posterior metrics for NPE and NPE augmented with self-consistency loss (NPE + SC) relative to Stan. For each parameter, the absolute bias in posterior means and standard deviations are reported along with the Wasserstein distance between the posteriors, using Stan as reference. The self-consistency loss was evaluated on data from $M = \mathbf{8}$ countries during training. Metrics are averaged over all 15 countries.

Parameter	$|\mu - \mu_{\text{Stan}}|$		$|\sigma - \sigma_{\text{Stan}}|$		Wasserstein distance	
	NPE	NPE+SC	NPE	NPE+SC	NPE	NPE+SC
α	0.078	**0.013**	0.033	**0.020**	0.085	**0.034**
β	0.152	**0.030**	0.054	**0.004**	0.161	**0.054**
γ	0.087	**0.005**	0.057	**0.035**	0.153	**0.068**
δ	0.052	**0.041**	0.037	**0.030**	0.119	**0.064**
$\log(\sigma)$	0.214	**0.148**	0.049	**0.010**	0.304	**0.170**

We apply our self-consistency loss to analyze trends in European air passenger traffic data provided by Eurostat (2022a;b;c). Using this case study, we highlight that the strong robustness gains in the normal means problem also translate to real-world scenarios and model classes that are challenging to estimate in a simulation-based inference setting. We observe that approximators trained with the standard simulation-based loss alone yield incorrect posterior estimates for several countries. In contrast, approximators trained also with our self-consistency loss provide highly similar results to Stan as a gold-standard reference.

We retrieved time series of annual air passenger counts between 15 European countries (departures) and the USA (destination) from 2004 to 2019 and fit the following autoregressive process of order 1:

$$y_{j,t+1} \sim \text{Normal}(\alpha_j + y_{j,t}\beta_j + u_{j,t}\gamma_j + w_{j,t}\delta_j, \sigma_j), \tag{8}$$

where the target quantity $y_{j,t+1}$ is the difference in air passenger traffic for country j between time $t + 1$ and t. To predict $y_{j,t+1}$ we use two additional predictors: $u_{j,t}$ is the annual household debt of country j at

time t, measured in % of gross domestic product (GDP) and $w_{j,t}$ is the real GDP per capita. The parameters α_j are country-level intercepts, β_j are the autoregressive coefficients, γ_j are the regression coefficients of household debt and δ_j are the regression coefficients of GDP per capita, and σ_j is the standard deviation of the noise term. This model was previously used within ABI in Habermann et al. (2024). As commonly done for autoregressive models, we regress on time period differences to mitigate non-stationarity. This is critical for simulation-based inference because when $\beta_j > 1$, exponential growth quickly produces unrealistic air traffic volumes. Moreover, amortizing over covariate spaces—such as varying GDP per capita between countries—can lead to model misspecification if such fluctuations are underrepresented in training. Training used a simulation budget of $N = 1024$, with the self-consistency loss evaluated on real data from $M \in \{4, 8, 15\}$ countries. Further details on training are in Appendix A.4.

Results   In Figure 2, we show exemplary results from standard NPE, our semi-supervised NPE ($M = 8$), and Stan as reference. We see that standard NPE is highly inaccurate for many countries, whereas our semi-supervised approach is in strong agreement with the reference for all but one country. As shown in Table 1, adding the self-consistency loss ($M = 8$) strongly improves posterior estimates for all five parameters across all metrics, on average across countries. The complete results can be found in Appendix A.5.

5   DISCUSSION

In this paper, we demonstrated that Bayesian self-consistency losses can drastically improve the robustness of neural amortized Bayesian inference (ABI) on out-of-simulation data. Performing accurate inference on such data – outside of the space seen during simulation-based training (e.g., because of model misspecification) – has been one of the major challenges of ABI since its inception: conventional ABI approaches are known to dramatically fail in such cases (Schmitt et al., 2023; Gloeckler et al., 2023; Huang et al., 2023), as we also illustrated in our experiments. In contrast, when adding the self-consistency loss and training it on *unlabeled* out-of-simulation data, we obtained near-unbiased posterior estimation far beyond the training data spaces. That is, ABI became highly accurate even for data sets far away from any data seen during training by either of the two loss components – the standard simulation-based loss and the self-consistency loss. The self-consistency loss was initially proposed to improve training efficiency for models with slow simulators (Schmitt et al., 2024; Ivanova et al., 2024), but it was not considered in combination with (potentially real) out-of-simulation data. Importantly, as self-consistency loss does not require data labels (i.e., true parameter values), we can use any amount of *real data* during training to improve the robustness of ABI.

We believe that incorporating self-consistency loss trained on out-of-simulation data offers a major contribution to neural ABI. To the best of our knowledge, no existing ABI approach achieves similar degrees of robustness as our method presented here. The strong robustness gains persisted even in relatively high-dimensional models (tested up to 100 parameters). Additionally, using a *learned* (i.e., approximate) rather than a known likelihood density also increased the robustness significantly. Our variance-based self-consistency loss relies on fast density evaluations during training, keeping times competitive. This makes free-form methods such as flow matching (Lipman et al., 2023) less practical due to their need for numerical integration. As a result, joint learning of posteriors and likelihoods, along with efficient self-consistency losses for free-form flows, remains an open research challenge.

ACKNOWLEDGMENTS

REFERENCES

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Maximilian Dax, Stephen R Green, Jonathan Gair, Jakob H Macke, Alessandra Buonanno, and Bernhard Schölkopf. Real-time gravitational wave science with neural posterior estimation. *Physical review letters*, 127(24):241103, 2021.

Charita Dellaporta, Jeremias Knoblauch, Theodoros Damoulas, and François-Xavier Briol. Robust bayesian inference for simulator-based models via the mmd posterior bootstrap. In *International Conference on Artificial Intelligence and Statistics*, pp. 943–970. PMLR, 2022.

Felix Draxler, Stefan Wahl, Christoph Schnörr, and Ullrich Köthe. On the universality of coupling-based normalizing flows. *arXiv preprint arXiv:2402.06578*, 2024.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in neural information processing systems*, 32, 2019.

Lasse Elsemüller, Hans Olischläger, Marvin Schmitt, Paul-Christian Bürkner, Ullrich Köthe, and Stefan T Radev. Sensitivity-aware amortized Bayesian inference. *Transactions on Machine Learning Research (TMLR)*, 2024.

Eurostat. International extra-eu air passenger transport by reporting country and partner world regions and countries, doi:10.2908/avia_paexcc, 2022a.

Eurostat. Household debt, consolidated including Non-profit institutions serving households - % of GDP, doi:10.2908/TIPSD22, 2022b.

Eurostat. Real gdp per capita, doi:10.2908/SDG_08_10, 2022c.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

David T. Frazier and Christopher Drovandi. Robust Approximate Bayesian Inference With Synthetic Likelihood. *Journal of Computational and Graphical Statistics*, 30(4):958–976, October 2021. doi: 10.1080/10618600.2021.1875839.

David T. Frazier, Christian P. Robert, and Judith Rousseau. Model misspecification in approximate Bayesian computation: consequences and diagnostics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):421–444, April 2020. doi: 10.1111/rssb.12356.

David T. Frazier, Ryan Kelly, Christopher Drovandi, and David J. Warne. The Statistical Accuracy of Neural Posterior and Likelihood Estimation. *arXiv preprint*, 2024. doi: 10.48550/arXiv.2411.12068.

Richard Gao, Michael Deistler, and Jakob H Macke. Generalized bayesian inference for scientific simulators via amortized cost estimation. *Advances in Neural Information Processing Systems*, 36:80191–80219, 2023.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. CRC Press, 2013.

Samuel Gershman and Noah Goodman. Amortized inference in probabilistic reasoning. *Proceedings of the annual meeting of the cognitive science society*, 36, 2014.

Manuel Gloeckler, Michael Deistler, and Jakob H Macke. Adversarial robustness of amortized bayesian inference. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 11493–11524, 2023.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Pedro J Gonçalves, Jan-Matthis Lueckmann, Michael Deistler, Marcel Nonnenmacher, Kaan Öcal, Giacomo Bassetto, Chaitanya Chintaluri, William F Podlaski, Sara A Haddad, Tim P Vogels, et al. Training deep neural density estimators to identify mechanistic models of neural dynamics. *Elife*, 9:e56261, 2020.

A Gretton, K. Borgwardt, Malte Rasch, Bernhard Schölkopf, and AJ Smola. A Kernel Two-Sample Test. *The Journal of Machine Learning Research*, 2012.

Daniel Habermann, Marvin Schmitt, Lars Kühmichel, Andreas Bulling, Stefan T. Radev, and Paul-Christian Bürkner. Amortized Bayesian Multilevel Models. *arXiv preprint*, 2024. doi: 10.48550/arXiv.2408.13230.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021.

Daolang Huang, Ayush Bharti, Amauri Souza, Luigi Acerbi, and Samuel Kaski. Learning robust statistics for simulation-based inference under model misspecification. *Advances in Neural Information Processing Systems*, 36:7289–7310, 2023.

Desi R Ivanova, Marvin Schmitt, and Stefan T Radev. Data-efficient variational mutual information estimation via Bayesian self-consistency. In *NeurIPS 2024 Workshop on Bayesian Decision-making and Uncertainty*, 2024.

Ryan Kelly, David J Nott, David T Frazier, David Warne, and Chris Drovandi. Misspecification-robust sequential neural likelihood for simulation-based inference. *Transactions on Machine Learning Research*, 2024(June):Article–number, 2024.

Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, 2020.

Alexander Lavin, David Krakauer, Hector Zenil, Justin Gottschlich, Tim Mattson, Johann Brehmer, Anima Anandkumar, Sanjay Choudry, Kamil Rocki, Atılım Güneş Baydin, et al. Simulation intelligence: Towards a new generation of scientific methods. *arXiv preprint arXiv:2112.03235*, 2021.

Tuan Anh Le, Atilim Gunes Baydin, and Frank Wood. Inference compilation and universal probabilistic programming. In *Artificial Intelligence and Statistics*, pp. 1338–1348. PMLR, 2017.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Måns Magnusson, Jakob Torgander, Paul-Christian Bürkner, Lu Zhang, Bob Carpenter, and Aki Vehtari. posteriordb: Testing, Benchmarking and Developing Bayesian Inference Algorithms. *arXiv preprint*, 2024. doi: 10.48550/arXiv.2407.04967.

Lorenzo Pacchiardi, Sherman Khoo, and Ritabrata Dutta. Generalized bayesian likelihood-free inference. *Electronic Journal of Statistics*, 18(2):3628–3686, 2024.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Stefan T Radev, Marvin Schmitt, Valentin Pratz, Umberto Picchini, Ullrich Köthe, and Paul-Christian Bürkner. Jana: Jointly amortized neural approximation of complex bayesian models. In *Uncertainty in Artificial Intelligence*, pp. 1695–1706. PMLR, 2023.

Marvin Schmitt, Paul-Christian Bürkner, Ullrich Köthe, and Stefan T Radev. Detecting model misspecification in amortized bayesian inference with neural networks. In *DAGM German Conference on Pattern Recognition*, pp. 541–557. Springer, 2023.

Marvin Schmitt, Desi Ivanova, Daniel Habermann, Paul-Christian Bürkner, Ullrich Köthe, and Stefan T Radev. Leveraging self-consistency for data-efficient amortized Bayesian inference. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 43723–43741. PMLR, 2024.

Ali Siahkoohi, Gabrio Rizzuti, Rafael Orozco, and Felix J Herrmann. Reliable amortized variational inference with physics-based latent distribution correction. *Geophysics*, 88(3):R297–R322, 2023.

Mischa von Krause, Stefan T Radev, and Andreas Voss. Mental speed is high until age 60 as revealed by analysis of over a million participants. *Nature human behaviour*, 6(5):700–708, 2022.

Daniel Ward, Patrick Cannon, Mark Beaumont, Matteo Fasiolo, and Sebastian Schmon. Robust neural posterior estimation and statistical model criticism. *Advances in Neural Information Processing Systems*, 35:33845–33859, 2022.

Antoine Wehenkel, Juan L Gamella, Ozan Sener, Jens Behrmann, Guillermo Sapiro, Marco Cuturi, and Jörn-Henrik Jacobsen. Addressing misspecification in simulation-based inference through data-driven calibration. *arXiv preprint arXiv:2405.08719*, 2024.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in Neural Information Processing Systems*, 2017.

Andrew Zammit-Mangion, Matthew Sainsbury-Dale, and Raphaël Huser. Neural methods for amortized inference. *Annual Review of Statistics and Its Application*, 12, 2024.

A   APPENDIX

A.1   PROOFS

Proof of Proposition 1. By assumption, C is globally minimized if and only if

$$\frac{p(x \mid \theta) \, p(\theta)}{q(\theta \mid x)} = A \tag{9}$$

for some constant A (independent of θ) almost everywhere over the posterior's support. Accordingly, any approximate posterior solution $q(\theta \mid x)$ that attains this global minimum has to be of the form

$$q(\theta \mid x) = p(x \mid \theta) \, p(\theta) \, / \, A. \tag{10}$$

By construction, $q(\theta \mid x)$ is a proper probability density function, so it integrates to 1. It follows that

$$1 = \int q(\theta \mid x) \, d\theta = \int p(x \mid \theta) \, p(\theta) \, d\theta \, / \, A = p(x) \, / \, A. \tag{11}$$

Rearranging the equation yields $A = p(x)$ and thus

$$q(\theta \mid x) = p(x \mid \theta) \, p(\theta)/p(x) = p(\theta \mid x) \tag{12}$$

almost everywhere. □

Proof of Proposition 2. The variance over a distribution $p_C(\theta)$ reaches its global minimum (i.e., zero), if and only if its argument is constant across the support of $p_C(\theta)$. Because the log is a strictly monotonic transform,

$$\log p(x^* \mid \theta) + \log p(\theta) - \log q(\theta \mid x^*) = \log A \tag{13}$$

for some constant A implies

$$\frac{p(x \mid \theta) \, p(\theta)}{q(\theta \mid x)} = A, \tag{14}$$

which is sufficient to satisfy the assumptions of Proposition 1. □

A.2   DETAILED SETUP OF THE MULTIVARIATE NORMAL CASE STUDY

From the multivariate normal model described in Section 4.1, we simulate a *labeled* training dataset with a budget of $N = 1024$, that is, N independent instances of θ_n (the "labels") with corresponding observations $x_n = \{x_n^{(k)}\}_{k=1}^K$, each consisting of K data points. This labeled training dataset $\{(\theta_n, x_n)\}_{n=1}^N$ is used for optimizing the standard simulation-based loss component. The self-consistency loss component is optimized on an additional *unlabeled* dataset $\{x_m^*\}_{m=1}^M$ of $M = 32$ independent sequences $x_m^* = \{x_m^{*(k)}\}_{k=1}^K$, which, for the purpose of this case study, are simulated from

$$x_m^{*(k)} \sim \text{Normal}(\mu^*, I_D). \tag{15}$$

Since the self-consistency loss does not need labels (i.e., the true parameters having generated x_m^*), we could have also chosen any other source for x^*, for example, real-world data. Within each training iteration t, the variance term within the self-consistency loss was computed from $L = 32$ samples $\theta^{(l)} \sim q_t(\theta \mid x_m^*)$ from the current posterior approximation.

To evaluate the accuracy and robustness of the NPEs, we perform posterior inference on completely new observations $x_\text{obs} = \{x_\text{obs}^{(k)}\}_{k=1}^K$, each consisting of K independent data points sampled from

$$x_\text{obs}^{(k)} \sim \text{Normal}(\mu_\text{obs}, \sigma_\text{obs}^2 = 0.01 I_D). \tag{16}$$

The mean values $\mu_{\mathrm{obs}} \in \{0, 1, \ldots, 11\}$ are progressively farther away from the training data. While conceptually simple and synthetic, this setting is already extremely challenging for simulation-based inference algorithms because of the large simulation gap (Schmitt et al., 2023): standard NPEs are only trained on (labeled) training data that are several standard deviations away from the observed data the model sees at inference time.

The faithfulness of the approximated posteriors $q(\theta \mid x_{\mathrm{obs}})$ are assessed by computing the bias in posterior mean and standard deviation as well as the maximum mean discrepancy (MMD) with a Gaussian kernel (Gretton et al., 2012) between the approximate and true (analytic) posterior.

The analytic posterior for the normal means problem is a conjugate normal distribution

$$p(\theta \mid x_{\mathrm{obs}}) = \mathrm{Normal}(\mu_{\mathrm{post}}, \sigma^2_{\mathrm{post}} I_D), \tag{17}$$

where μ_{post} is a D-dimensional posterior mean vector with elements

$$(\mu_{\mathrm{post}})_d = \sigma^2_{\mathrm{post}} \left(\frac{\mu_{\mathrm{prior}}}{\sigma^2_{\mathrm{prior}}} + \frac{K(\bar{x}_{\mathrm{obs}})_d}{\sigma^2_{\mathrm{lik}}} \right), \tag{18}$$

σ^2_{post} is the posterior variance (constant across dimensions) given by

$$\sigma^2_{\mathrm{post}} = \left(\frac{1}{\sigma^2_{\mathrm{prior}}} + \frac{K}{\sigma^2_{\mathrm{lik}}} \right)^{-1}, \tag{19}$$

and $(\bar{x}_{\mathrm{obs}})_d$ is the mean over the Dth dimension of the K new data points $\{x^{(k)}_{\mathrm{obs}}\}^K_{k=1}$.

For the NPEs $q(\theta \mid x)$, we use a neural spline flow (Durkan et al., 2019) with 5 coupling layers of 128 units each utilizing ReLU activation functions, L2 weight regularization with factor $\gamma = 10^{-3}$, 5% dropout and a multivariate unit Gaussian latent space. The network is trained using the Adam optimizer for 100 epochs with a batch size of 32 and a learning rate of 5×10^{-4}. These settings were the same for both the standard simulation-based loss and our proposed semi-supervised loss. For the conditions involving an estimated likelihood $q(x \mid \theta)$, we use the same configuration for the likelihood network as for the posterior network. For the summary network $h(x)$ (if included), we use a deep set architecture (Zaheer et al., 2017) with 30 summary dimensions and mean pooling, 2 equivariant layers each consisting of 2 dense layers with 64 units and a ReLU activation function. The inner and outer pooling functions also use 2 dense layers with the same configuration. The likelihood network as well as the summary network are jointly trained with the inference network using the Adam optimizer for 100 epochs with a batch size of 32 and a learning rate of 5×10^{-4}.

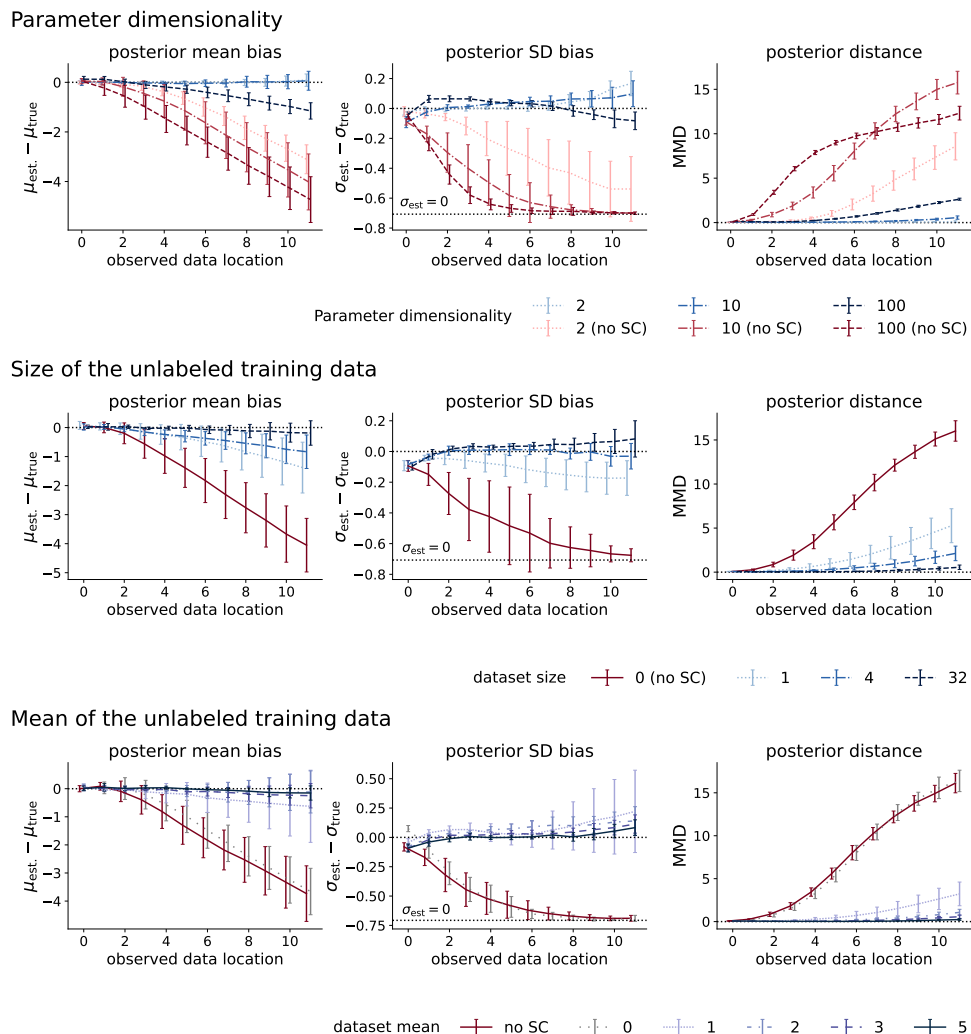A.3   COMPREHENSIVE RESULTS FOR THE MULTIVARIATE NORMAL CASE STUDY

Figure 3: Bias of posterior mean, bias of posterior standard deviation and posterior distance quantified by maximum mean discrepancy to the analytic posterior for variations of the default configuration outlined in Section 4.1. NPE approximators with the added self-consistency loss component are shown in blue, NPE approximators using just the standard simulation-based loss are shown in red. Irrespective of the varied factor and for all metrics, adding the self-consistency loss component is always a drastic improvement over the standard simulation-based loss alone. The plots show that adding the self-consistency loss component provides strong robustness gains even in high-dimensional spaces (top row) or when the self-consistency loss is evaluated on little data (center row). Variation of the mean of the unlabeled training data show that adding the self-consistency loss drastically improves posterior estimation as long as data used for evaluating the self-consistency loss is at least slightly out-of-distribution compared to the original training data ($\mu^* \geq 1$).
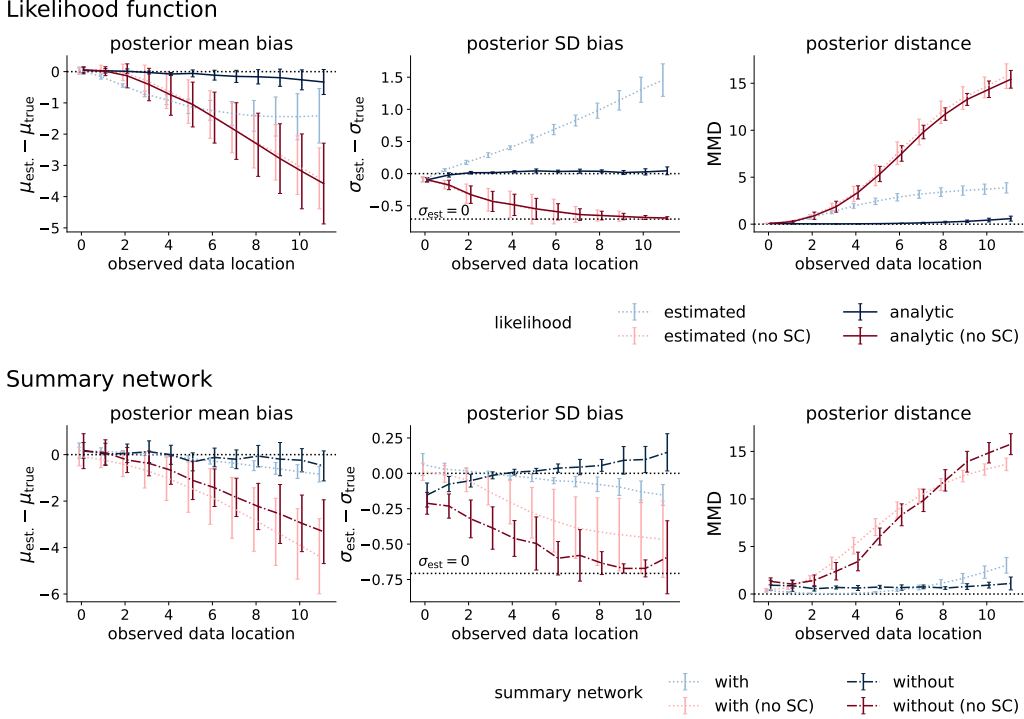
Figure 4: Bias of posterior mean, bias of posterior standard deviation and posterior distance quantified by maximum mean discrepancy to the analytic posterior when the likelihood is estimated (top row) and in presence of a summary network ($K = 10$ data points; bottom row). In the setting where the likelihood function is estimated, we observe a lower bias of the posterior mean and lower maximum mean discrepancy to the true posterior when the self-consistency loss component is added compared to the standard simulation-based loss alone. However, we do see some bias of the posterior standard deviation, although with reversed signed compared to the standard loss. The self-consistency loss provides strong robustness gains in the presence of a summary network (and known likelihood) in terms of all metrics.

A.4   DETAILED SETUP OF THE AIR TRAFFIC CASE STUDY

For the air traffic model defined in Section 4.2, we set independent priors on the parameters as follows:

$$
\begin{aligned}
\alpha_j &\sim \text{Normal}(0, 0.5) & \beta_j &\sim \text{Normal}(0, 0.2) \\
\gamma_j &\sim \text{Normal}(0, 0.5) & \delta_j &\sim \text{Normal}(0, 0.5) \\
\log(\sigma_j) &\sim \text{Normal}(-1, 0.5).
\end{aligned}
\tag{20}
$$

For the NPEs $q(\theta \mid x)$, we use a neural spline flow (Durkan et al., 2019) with 6 coupling layers of 128 units each utilizing exponential linear unit activation functions, L2 weight regularization with factor $\gamma = 10^{-3}$, 5% dropout and a multivariate unit Gaussian latent space. These settings were the same for both the standard simulation-based loss and our proposed semi-supervised loss. The simulation budget was set to $N = 1024$. For the summary network, we use a long short-term memory layer with 64 output dimensions followed by

15

two dense layers with output dimensions of 256 and 64. The inference and summary networks are jointly trained using the Adam optimizer for 100 epochs with a batch size of 32 and a learning rate of 5×10^{-4}.

A.5 COMPREHENSIVE RESULTS FOR THE AIR TRAFFIC CASE STUDY

Table 2: Posterior metrics for NPE and NPE augmented with self-consistency loss (NPE + SC) relative to Stan. For each parameter, the absolute bias in posterior means and standard deviations are reported along with the Wasserstein distance between the posteriors, using Stan as reference. The self-consistency loss was evaluated on data from $M = \mathbf{4}$ countries during training. Metrics are averaged over all 15 countries.

Parameter	$\|\mu - \mu_{\text{Stan}}\|$		$\|\sigma - \sigma_{\text{Stan}}\|$		Wasserstein distance	
	NPE	NPE+SC	NPE	NPE+SC	NPE	NPE+SC
α	0.078	**0.003**	0.033	**0.019**	0.085	**0.032**
β	0.152	**0.011**	0.054	**0.011**	0.161	**0.070**
γ	0.087	**0.047**	0.057	**0.022**	0.153	**0.112**
δ	0.052	**0.046**	0.037	**0.018**	0.119	**0.102**
$\log(\sigma)$	0.214	**0.207**	**0.049**	0.057	0.304	**0.281**

Table 3: Posterior metrics for NPE and NPE augmented with self-consistency loss (NPE + SC) relative to Stan. For each parameter, the absolute bias in posterior means and standard deviations are reported along with the Wasserstein distance between the posteriors, using Stan as reference. The self-consistency loss was evaluated on data from $M = \mathbf{15}$ countries during training. Metrics are averaged over all 15 countries.

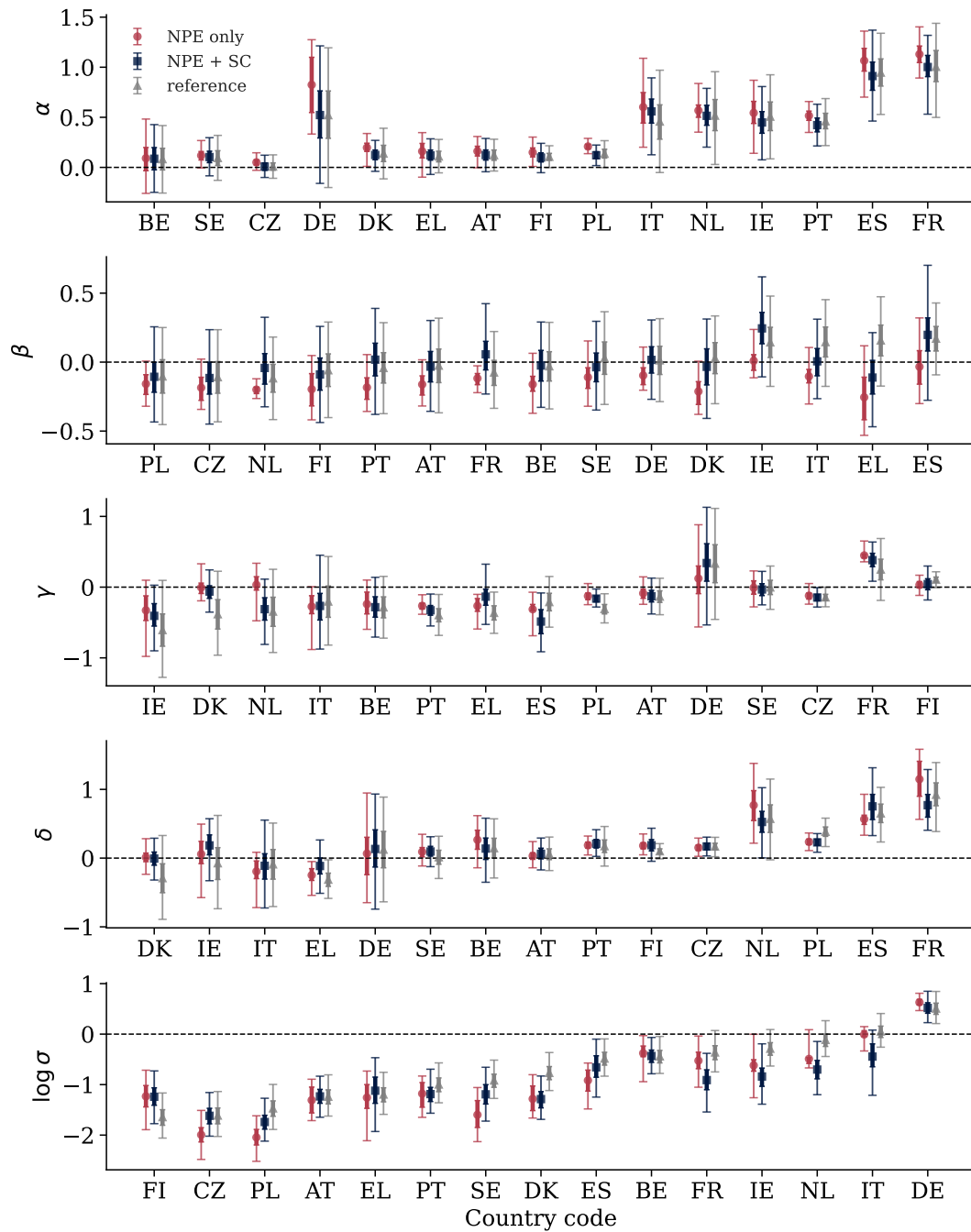| Parameter	$\|\mu - \mu_{\text{Stan}}\|$		$\|\sigma - \sigma_{\text{Stan}}\|$		Wasserstein distance	
	NPE	NPE+SC	NPE	NPE+SC	NPE	NPE+SC
α	0.078	**0.002**	0.033	**0.001**	0.085	**0.006**
β	0.152	**0.001**	0.054	**0.001**	0.161	**0.009**
γ	0.087	**0.002**	0.057	**0.006**	0.153	**0.014**
δ	0.052	**0.003**	0.037	**0.005**	0.119	**0.013**
$\log(\sigma)$	0.214	**0.002**	0.049	**0.004**	0.304	**0.011**

Figure 5: Comparison of posterior estimates between standard amortized NPE (red circles), NPE augmented by our self-consistency loss (NPE + SC; blue squares) and Stan (reference; gray triangles). The plots illustrate central 50% (thick lines) and 95% (thin lines) credible intervals of all five parameters for different countries, sorted by the lower 5% quantile according to Stan. Abbreviations follow the ISO 3166 alpha-2 codes. The self-consistency loss was evaluated on data from $M = 4$ countries during training.
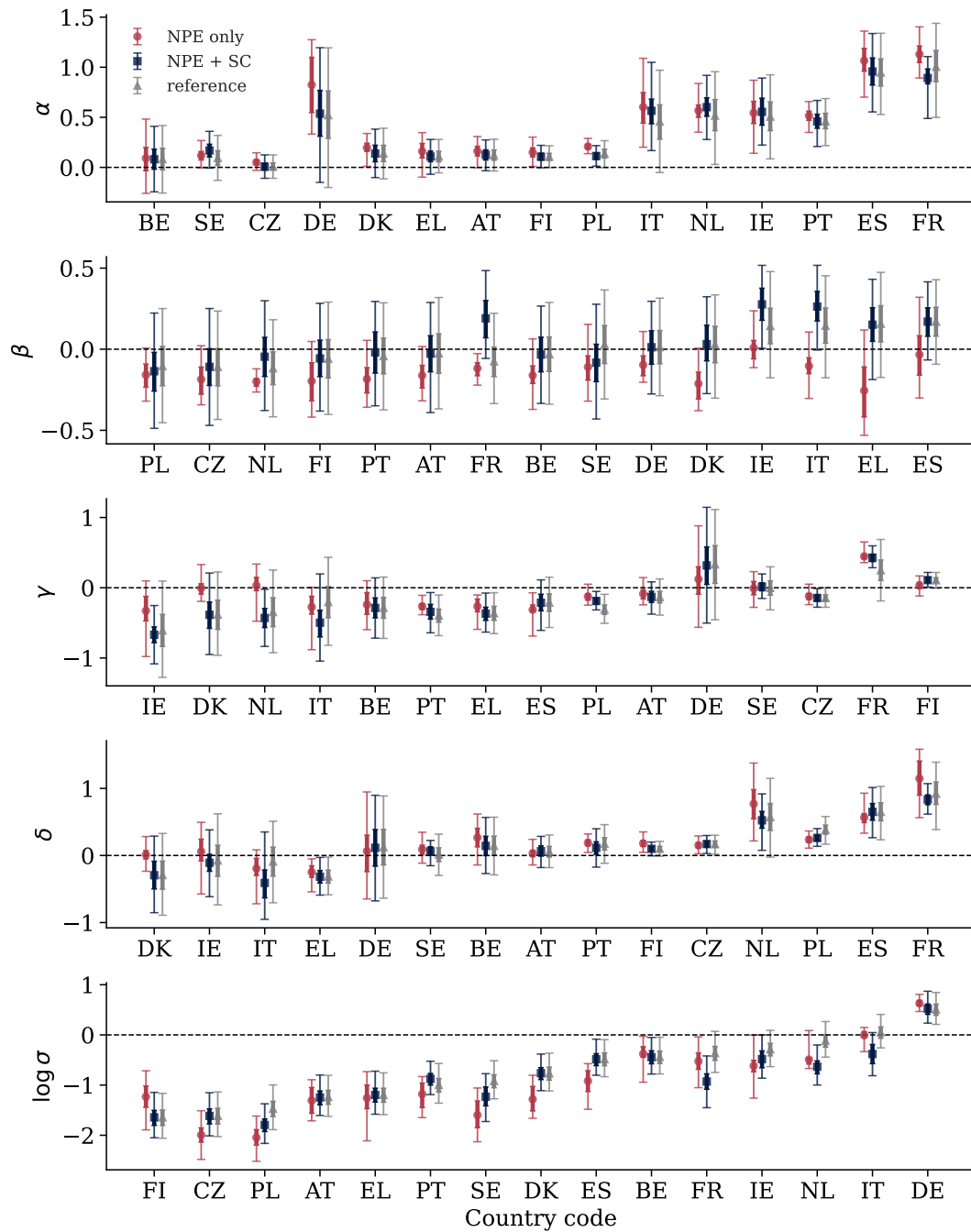
Figure 6: Comparison of posterior estimates between standard amortized NPE (red circles), NPE augmented by our self-consistency loss (NPE + SC; blue squares), and Stan (reference; gray triangles). The plots illustrate central 50% (thick lines) and 95% (thin lines) credible intervals of all five parameters for different countries, sorted by the lower 5% quantile according to Stan. Abbreviations follow the ISO 3166 alpha-2 codes. The self-consistency loss was evaluated on data from $M = \mathbf{8}$ countries during training.
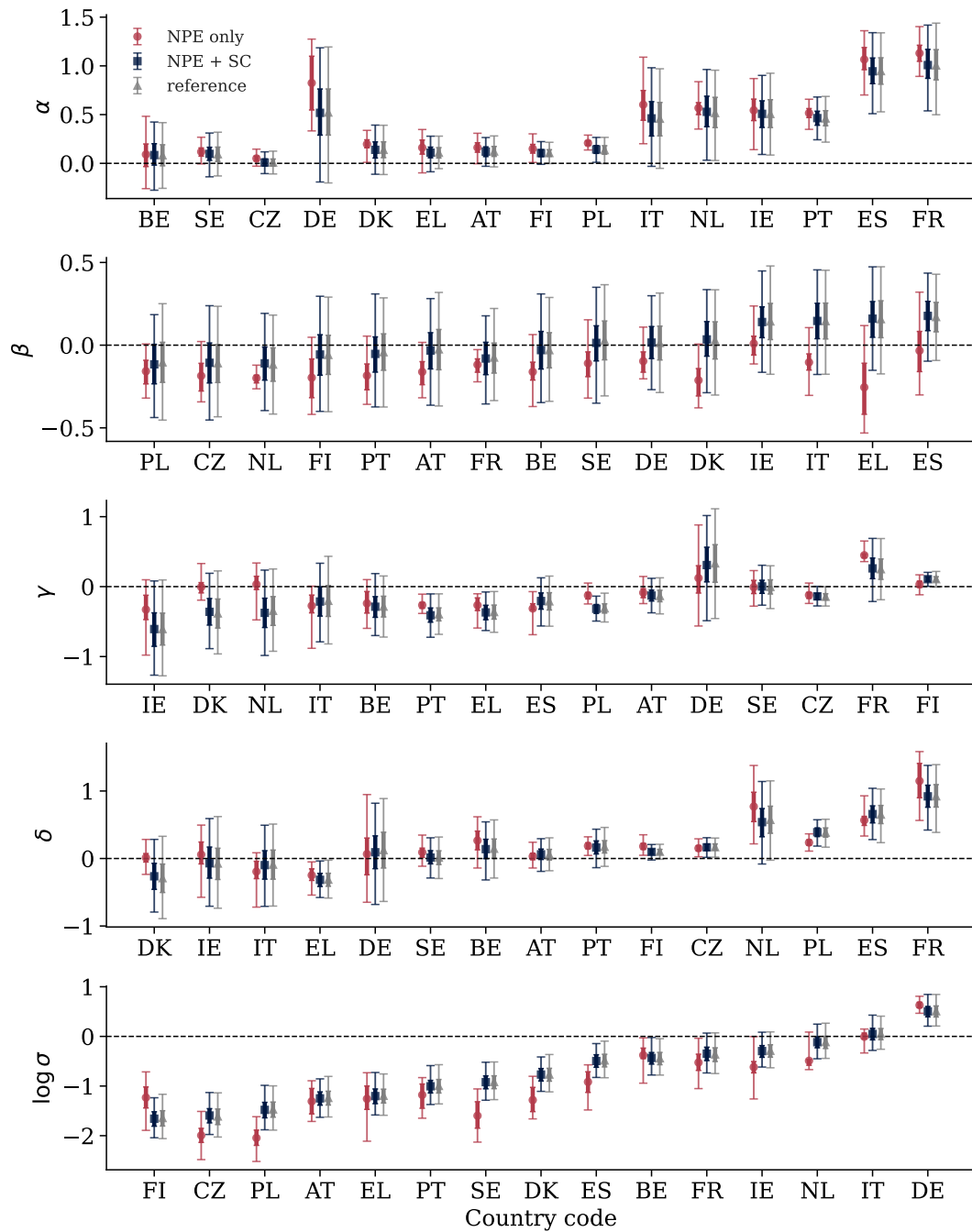
Figure 7: Comparison of posterior estimates between standard amortized NPE (red circles), NPE augmented by our self-consistency loss (NPE + SC; blue squares), and Stan (reference; gray triangles). The plots illustrate central 50% (thick lines) and 95% (thin lines) credible intervals of all five parameters for different countries, sorted by the lower 5% quantile according to Stan. Abbreviations follow the ISO 3166 alpha-2 codes. The self-consistency loss was evaluated on data from $M = \mathbf{15}$ countries during training.

19