Multi-fidelity Hierarchical Neural Processes

Dongxia Wu
University of California, San Diego
La Jolla, CA, USA
dowu@ucsd.edu

Matteo Chinazzi
Northeastern University
Boston, MA, USA
m.chinazzi@northeastern.edu

Alessandro Vespignani
Northeastern University
Boston, MA, USA
a.vespignani@northeastern.edu

Yi-An Ma
University of California, San Diego
La Jolla, CA, USA
yianma@ucsd.edu

Rose Yu
University of California, San Diego
La Jolla, CA, USA
roseyu@ucsd.edu

ABSTRACT

Science and engineering fields use computer simulation extensively. These simulations are often run at multiple levels of sophistication to balance accuracy and efficiency. Multi-fidelity surrogate modeling reduces the computational cost by fusing different simulation outputs. Cheap data generated from low-fidelity simulators can be combined with limited high-quality data generated by an expensive high-fidelity simulator. Existing methods based on Gaussian processes rely on strong assumptions of the kernel functions and can hardly scale to high-dimensional settings. We propose Multi-fidelity Hierarchical Neural Processes (`MF-HNP`), a unified neural latent variable model for multi-fidelity surrogate modeling. `MF-HNP` inherits the flexibility and scalability of Neural Processes. The latent variables transform the correlations among different fidelity levels from observations to latent space. The predictions across fidelities are conditionally independent given the latent states. It helps alleviate the error propagation issue in existing methods. `MF-HNP` is flexible enough to handle non-nested high dimensional data at different fidelity levels with varying input and output dimensions. We evaluate `MF-HNP` on epidemiology and climate modeling tasks, achieving competitive performance in terms of accuracy and uncertainty estimation. In contrast to deep Gaussian Processes [6] with only low-dimensional (< 10) tasks, our method shows great promise for speeding up high-dimensional complex simulations (over $7,000$ for epidemiology modeling and $45,000$ for climate modeling).

KEYWORDS

multi-fidelity surrogate modeling, neural processes, deep learning

ACM Reference Format:
Dongxia Wu, Matteo Chinazzi, Alessandro Vespignani, Yi-An Ma, and Rose Yu. 2022. Multi-fidelity Hierarchical Neural Processes. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3534678.3539364

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08.
https://doi.org/10.1145/3534678.3539364

1 INTRODUCTION

In scientific and engineering applications, a computational model, often realized by simulation, characterizes the input-output relationship of a physical system. The input describes the properties and environmental conditions, and the output describes the quantities of interest. For example, in epidemiology, computational models have long been used to forecast the evolution of epidemic outbreaks and to simulate the effects of public policy interventions on the epidemic trajectory [5, 13, 24]. In the case of COVID-19 [3, 8], model inputs range across virus and disease characteristics (e.g. transmissibility and severity), non-pharmaceutical interventions (e.g. travel bans, school closures, business closures), and individual behavioral responses (e.g. changes in mobility and contact rates); while the output describes the evolution of the pandemic (e.g. the time series of the prevalence and incidence of the virus in the population).

Computational models can be simulated at multiple levels of sophistication. High-fidelity models produce accurate output at a higher cost, whereas low-fidelity models generate less accurate output at a cheaper cost. To balance the trade-off between computational efficiency and prediction accuracy, multi-fidelity modeling [30] aims to learn a surrogate model that combines simulation outputs at multiple fidelity levels to accelerate learning. Therefore, we can obtain predictions and uncertainty analysis at high fidelity while leveraging cheap low-fidelity simulations for speedup.

Since the pioneering work of Kennedy and Hagan [17] on modeling oil reservoir simulator, Gaussian processes (GPs) [36] have become the predominant tools in multi-fidelity modeling. GPs effectively serve as surrogate models to emulate the output distribution of complex physical systems with uncertainty [21, 32, 43]. However, GPs often struggle with high-dimensional data and require prior knowledge for kernel design. Multi-fidelity GPs also require a nested data structure [31] and the same input dimension at each fidelity level [6], which significantly hinders their applicability in the real world. Therefore, efforts to combine deep learning and GPs have undergone significant growth in the machine learning community [7, 35, 37, 44]. One of the most scalable frameworks of such combinations is Neural processes (NP) [10, 11, 19], which is a neural latent variable model.

Unfortunately, existing NP models are mainly designed for single-fidelity data and cannot handle multi-fidelity outputs. While we can train multiple NPs separately, one for each fidelity, this approach fails to exploit the relations among multi-fidelity models governed

by the same physical process. Furthermore, models with more fidelity levels require more training data, which leads to higher computational costs. An alternative is to learn the relationship between low- and high-fidelity model outputs and model the correlation function with NP [42]. But this approach always requires paired data at the low- and high-fidelity level. Another limitation is high dimensionality. The correlation function maps from the joint input-output space of the low-fidelity model to the high-fidelity output, which is prone to over-fitting.

In this work, we propose Multi-Fidelity Hierarchical Neural Process (MF-HNP), the first *unified* framework for scalable multi-fidelity modeling in neural processes family. Specifically, MF-HNP inherits the properties of Bayesian neural latent variable model while learning the joint distribution of multi-fidelity output. We design a unified evidence lower bound (ELBO) for the joined distribution as a training loss. The code and data are available on https://github.com/Rose-STL-Lab/Hierarchical-Neural-Processes.

In summary, our contributions include:

- A novel multi-fidelity surrogate model, Multi-fidelity Hierarchical Neural Processes (MF-HNP). Its unified framework makes it flexible to fuse data with varying input and output dimensions at different fidelity levels.
- A novel Neural Process architecture with conditional independence at each fidelity level. It fully utilizes the multi-fidelity data, reduces the input dimension, and alleviates error propagation in forecasting.
- Real-world large-scale multi-fidelity application on epidemiology and climate modeling to show competitive accuracy and uncertainty estimation performance.

2 RELATED WORK

Multi-fidelity Modeling. Multi-fidelity surrogate modeling is widely used in science and engineering fields, from climate science [15, 39] to aerospace systems [2]. The pioneering work of [17] uses GP to relate models at multiple fidelity levels with an autoregressive model. [21] proposed recursive GP with a nested structure in the input domain for fast inference. [32, 33] deals with high-dimensional GP settings by taking the Fourier transformation of the kernel function. [31] proposed multi-fidelity Gaussian processes (NARGP) but it assumes a nested structure in the input domain to enable a sequential training process at each fidelity level. An extreme case that we include in our experiment is when the data sets at low- and high-fidelity levels are disjoint. None of the high-fidelity data could be used for training, which is a failure case for NARGP. Additionally, the prediction error of the low-fidelity model will propagate to high-fidelity output and explode as the number of fidelity levels increases. [43] proposed a Multi-Fidelity High-Order GP model to speed up the physical simulation. They extended the classical Linear Model of Coregionalization (LMC) to nonlinear case and placed a matrix GP prior on the weight functions. Their method is designed for high-dimensional outputs rather than both high-dimensional inputs and outputs. Deep Gaussian processes (DGP) [6] designs a single objective to optimize the kernel parameters at each fidelity level jointly. However, the DGP architecture introduces a constraint that requires the inputs at each fidelity level to be defined by the same domain with the same dimension. Moreover, DGP is still based on GPs, which are not scalable for applications with high-dimensional data. In contrast, NP is flexible and much more scalable.

Deep learning has been applied to multi-fidelity modeling. For example, [12] uses deep neural networks to combine parameter-dependent output quantities. [27] propose a composite neural network for multi-fidelity data from inverse PDE problems. [26] propose Bayesian neural nets for multi-fidelity modeling. [9] use transfer learning to fine-tune the high-fidelity surrogate model with the deep neural network trained with low-fidelity data. [6, 14] propose deep Gaussian process to capture nonlinear correlations between fidelities, but their method cannot handle the case where different fidelities have data with different dimensions. Tangentially, multi-fidelity methods have also recently been investigated in Bayesian optimization, active learning, and bandit problems [16, 22, 23, 34].

Neural Processes. Neural Processes (NPs) [10, 18, 25, 38] provide scalable and expressive alternatives to GPs for modeling stochastic processes. However, none of the existing NP models can efficiently incorporate multi-fidelity data. Earlier work by [35] combines multi-fidelity GP with deep learning by placing a GP prior on the features learned by deep neural networks. However, their model remains closer to GPs. Quite recently, [42] proposed multi-fidelity neural process with physics constraints (MFPC-Net). They use NP to learn the correlation between multi-fidelity data by mapping both the input and output of the low-fidelity model to the high-fidelity model output. But their model requires paired data and cannot utilize the remaining unpaired data at the low-fidelity level.

3 BACKGROUND

3.1 Muti-Fidelity Modeling

Formally, given input domain $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and output domain $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$, a model is a (stochastic) function $f : \mathcal{X} \to \mathcal{Y}$. Evaluations of f incur computational costs $c > 0$. The computational costs c are much higher at higher fidelity level. Therefore, we assume that a limited amount of expensive high-fidelity data is available for training. In multi-fidelity modeling, we have a set of functions $\{f_1, \cdots, f_K\}$ that approximate f with increasing accuracy and computational cost. We aim to learn a surrogate model $\hat{f}_K$ that combines information from low-fidelity models with a small amount of data from high-fidelity models.

Given parameters x_k at fidelity level k, we query the simulator to obtain data set from different scenarios $\mathcal{D}_k \equiv \{x_{k,i}, [y_{k,i}]_{s=1}^S\}_i$, where $[y_{k,i}]_{s=1}^S$ are S samples generated by $f_k(x_{k,i})$ for scenario i. In epidemic modeling, for example, each scenario corresponds to a different effective reproduction number of the virus, contact rates between individuals, or the effects of policy interventions. For each scenario, we simulate multiple epidemic trajectories as samples from the stochastic function. We aim to learn a deep surrogate model that approximates the data distribution $p(y_K^t | x_K^t, \mathcal{D}_1^c, \mathcal{D}_2^c, ..., \mathcal{D}_K^c)$ at the highest fidelity level K over the target set y_K^t, given context sets at different fidelity levels $\mathcal{D}_k^c \subset \mathcal{D}_k$ and the corresponding x_K^t.

For simplicity, we use two levels of fidelity, but our framework can be generalized easily. Let us denote the low-fidelity data as $\mathcal{D}_l \equiv \{x_{l,i}, [y_{l,i}]_{s=1}^S\}_i$ and high-fidelity data as $\mathcal{D}_h \equiv \{x_{h,i}, [y_{h,i}]_{s=1}^S\}_i$. If

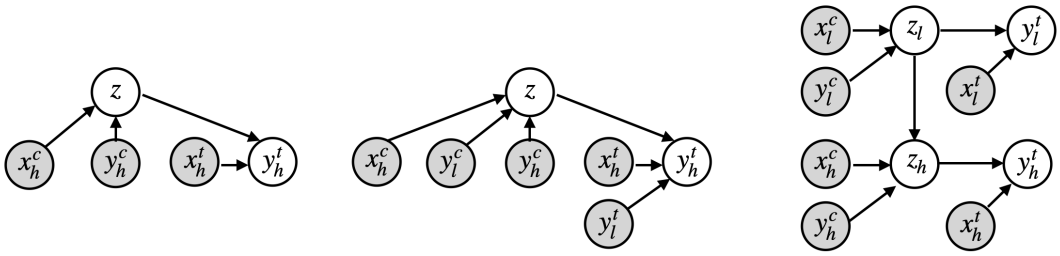

Figure 1: Graphical models for Single-Fidelity Neural Process (left), Multi-Fidelity Neural Process (middle), Multi-Fidelity Hierarchical Neural Process (right). Shaded circles denote observed variables and hollow circle represent latent variables. The directed edges represent conditional dependence.

$\mathcal{D}_h \subset \mathcal{D}_l$, the data domain has the nested structure. If $\mathcal{D}_h = \mathcal{D}_l$, we say the low- and high-fidelity data sets are paired. Low-fidelity data can be split into context sets $\mathcal{D}_l^c \equiv \{x_{l,n}^c, [y_{l,n}^c]_{s=1}^S\}_{n=1}^{N_l}$ and target sets $\mathcal{D}_l^t \equiv \{x_{l,m}^t, [y_{l,m}^t]_{s=1}^S\}_{m=1}^{M_l}$. Similarly, high-fidelity data can be split into context sets $\mathcal{D}_h^c \equiv \{x_{h,n}^c, [y_{h,n}^c]_{s=1}^S\}_{n=1}^{N_h}$ and target sets $\mathcal{D}_h^t \equiv \{x_{h,m}^t, [y_{h,m}^t]_{s=1}^S\}_{m=1}^{M_h}$.

3.2 Neural Processes

Neural processes (NPs) [11] are the family of conditional latent variable models for implicit stochastic processes ($\mathcal{SP}s$) [41]. NPs are in between GPs and neural networks (NNs). Like GPs, NPs can represent distributions over functions and estimate the uncertainty of the predictions. But they are more scalable in high dimensions and can easily adapt to new observations. According to Kolmogorov Extension Theorem [29], NPs meet exchangeability and consistency conditions to define $\mathcal{SP}s$. Formally, NP includes local latent variables $z \in \mathbb{R}^{d_z}$ and global latent variables θ and is trained by the context set $\mathcal{D}^c \equiv \{x_n^c, [y_n^c]_{s=1}^S\}_{n=1}^N$ and target sets $\mathcal{D}^t \equiv \{x_m^t, [y_m^t]_{s=1}^S\}_{m=1}^M$. Learning the posterior of z and θ is equivalent to maximizing the following posterior likelihood:

$$\prod_{s=1}^{S} p(y_{s,1:M}^t | x_{1:M}^t, \mathcal{D}^c, \theta) =$$
$$\prod_{s=1}^{S} \int p(z_s | \mathcal{D}^c, \theta) \prod_{m=1}^{M} p(y_{s,m}^t | z_s, x_m^t, \theta) dz_s$$

We omit the sample index s in what follows.

Approximate Inference. Since marginalizing over the local latent variables z is intractable, the NP family [11, 19] introduces approximate inference on latent variables and derives the corresponding evidence lower bound (ELBO) for the training process.

$$\log p(y_{1:M}^t | x_{1:M}^t, \mathcal{D}^c, \theta) \geq$$
$$\mathbb{E}_{q_\phi(z|\mathcal{D}^c \cup \mathcal{D}^t)} \Big[\sum_{m=1}^{M} \log p(y_m^t | z, x_m^t, \theta) + \log \frac{q_\phi(z|\mathcal{D}^c)}{q_\phi(z|\mathcal{D}^c \cup \mathcal{D}^t)} \Big]$$

Note that this variational approach approximates the intractable true posterior $p(z|\mathcal{D}^c, \theta)$ with the approximate posterior $q_\phi(z|\mathcal{D}^c)$. This approach is also an amortized inference method as the global parameters ϕ are shared by all context data points. It is efficient during the test time (no per-data-point optimization) [40].

NPs use NNs to represent $q_\phi(z|\mathcal{D}^c)$, and $p(y_m^t|z, x_m^t, \theta)$. $q_\phi()$ is referred as the encoder network (Enc, determined by the parameters ϕ). $p(.|\theta)$ is referred as the decoder network (Dec, determined by parameters θ). These two networks assume that the latent variable z and the outputs y follow the factorized Gaussian distribution determined by mean and variance.

$$q_\phi(z|\mathcal{D}^c) = \mathcal{N}(z|\mu_z, \text{diag}(\sigma_z^2))$$
$$\mu_z = \text{Enc}_{\mu_z,\phi}(\mathcal{D}^c), \quad \sigma_z^2 = \text{Enc}_{\sigma_z^2,\phi}(\mathcal{D}^c)$$
$$p(y_m^t|z, x_m^t, \theta) = \mathcal{N}(y_m^t|\mu_y, \text{diag}(\sigma_y^2))$$
$$\mu_y = \text{Dec}_{\mu_y,\theta}(z, x_m^t), \quad \sigma_y^2 = \text{Dec}_{\sigma_y^2,\theta}(z, x_m^t)$$

Context Aggregation. Context aggregation aggregates all context points $\mathcal{D}^c$ to infer latent variables z. To meet the exchangeability condition, the context information acquired by NPs should be invariant to the order of the data points. Garnelo et al. [10, 11], Kim et al. [18] use mean aggregation (MA). They map the data pair(x_n^c, y_n^c) to a latent representation $r_n = \text{Enc}_{r,\phi}(x_n^c, y_n^c) \in \mathbb{R}^{d_r}$, then apply the mean operation to the entire set $\{r_n\}_{n=1}^N$ to obtain the aggregated latent representation $\bar{r}$. $\bar{r}$ can be mapped to μ_z and σ_z^2 to represent the posterior $q_\phi(z|\mathcal{D}^c)$ with an additional neural network encoder. MA uses two encoder networks. $\text{Enc}_{r,\phi}(x_n^c, y_n^c) \in \mathbb{R}^{d_r}$ maps the data pair(x_n^c, y_n^c) to r_n for context aggregation. $\text{Enc}_{z,\phi}(\bar{r}) \in \mathbb{R}^{d_z}$ maps $\bar{r}$ to μ_z and σ_z^2 for latent parameter inference.

Volpp et al. [40] proposed Bayesian aggregation (BA), which merges these two steps. They define a probabilistic observation model $p(r|z)$ for r depended on z, and update $p(z)$ posterior using the Bayes rule $p(z|r_n) = p(r_n|z)p(z)|p(r_n)$ given latent observation $r_n = \text{Enc}_{r,\phi}(x_n^c, y_n^c)$. The corresponding factorized Gaussian for the inference step:

$$p(r_n|z) = \mathcal{N}(r_n|z, \text{diag}(\sigma_{r_n}^2))$$
$$r_n = \text{Enc}_{r,\phi}(x_n^c, y_n^c)$$
$$\sigma_{r_n}^2 = \text{Enc}_{\sigma_{r_n}^2,\phi}(x_n^c, y_n^c)$$

They use a factorized Gaussian prior $p_0(z) \equiv \mathcal{N}(z|\mu_{z,0}, \text{diag}(\sigma_{z,0}^2))$ to derive the parameters of posterior $q_\phi(z|\mathcal{D}^c)$:

$$\sigma_z^2 = \left[(\sigma_{z,0}^2)^{\ominus} + \sum_{n=1}^{N}(\sigma_{r_n}^2)^{\ominus})\right]^{\ominus},$$
$$\mu_z = \mu_{z,0} + \sigma_z^2 \odot \sum_{n=1}^{N}(r_n - \mu_{z,0}) \oslash (\sigma_{r_n}^2).$$

Compared with MA, which treats every context sample equally, BA uses observation variance $\sigma_{r_n}^2$ to weigh the importance of each latent representation r_n. BA also represents a permutation-invariant operation on $\mathcal{D}^c$.

4 METHODOLOGY

In this section, we introduce our proposed Multi-fidelity Hierarchical Neural Processes (MF-HNP) model in three subsections. The first section discusses the unique architecture of hierarchical neural processes for the multi-fidelity problem. Then, we develop the corresponding approximate inference method with a unified ELBO. Finally, we introduce 3 ELBO variants for scalable training.

4.1 Multi-fidelity Hierarchical Neural Processes

Our high-level goal is to train a deep surrogate model to mimic the behavior of a complex stochastic simulator at the highest level of fidelity. MF-HNP inherits the properties of Bayesian neural latent variable model while learning the joint distribution of multi-fidelity output. It adopts a single objective function for multi-fidelity training. It reduces the input dimension and alleviates error propagation by introducing the hierarchical structure in the dependency graph.

Figure 1 compares the graphical model of MF-HNP with Multi-fidelity Neural Process (MF-NP) [42] and Single-Fidelity Neural Process (SF-NP). SF-NP assumes that the high-fidelity data is independent of the low-fidelity data and reduces the model to vanilla NP setting. Details of SF-NP and MF-NP are shown in Appendix A. MF-HNP assignes latent variables z_l and z_h at each fidelity level. The prior of z_h is conditioned on z_l, parameterized by a neural network. We use Monte Carlo (MC) sampling method to approximate the posterior of z_l and z_h to calculate the ELBO.

One key feature of MF-HNP is that the model outputs at each fidelity level are conditionally independent given the corresponding latent state. This design transforms the correlations between fidelity levels from the input and output space to the latent space. Specifically, compared with MF-NP where $\hat{y}_h$ depends on (x_h, y_l) input pairs given z, $\hat{y}_h$ only depends on input x_h given z_h in MF-HNP. It helps MF-HNP to significantly reduce the high-fidelity input dimension. In addition, local latent variables at each level of fidelity enable MF-HNP to perform both inference and generative modeling separately at each fidelity level. It means MF-HNPcan fully utilize the low-fidelity data and is applicable to arbitrary multi-fidelity data sets. As MF-HNPcan reduce the input dimension and fully utilize the training data, its prediction performance is significantly improved with limited training data.

Note that in two fidelity setup, MF-HNP is related to Doubly Stochastic Variational Neural Process (DSVNP) model proposed by Wang and Van Hoof [41] which introduces local latent variables together with the global latent variables. Different from DSVNP, MF-HNP gives latent variables with separable representations. z_l, z_h represent the low- and high-fidelity functional, respectively.

4.2 Unified ELBO

We design a unified ELBO as the objective for MF-HNP. Unlike vanilla NPs, we need to infer the latent variables z_l and z_h at each fidelity level instead of the global z. For the two-fidelity level setup, we use two encoders $q_{\phi_l}(z_l|\mathcal{D}_l^c)$, $q_{\phi_h}(z_h|z_l, \mathcal{D}_h^c)$, and two decoders $p(y_l^t|z_l, x_l^t, \theta_l)$, $p(y_h^t|z_h, x_h^t, \theta_h)$. These four networks approximate the distributions of the latent variables z_l, z_h and outputs y_l and y_h. Assuming a factorized Gaussian distribution, we can parameterize the distributions by their mean and variance.

$$q_{\phi_l}(z_l|\mathcal{D}_l^c) = \mathcal{N}(z_l|\mu_{z_l}, \mathrm{diag}(\sigma_{z_l}^2))$$
$$q_{\phi_h}(z_h|z_l, \mathcal{D}_h^c) = \mathcal{N}(z_h|\mu_{z_h}, \mathrm{diag}(\sigma_{z_h}^2))$$
$$p(y_{l,m}^t|z_l, x_{l,m}^t, \theta_l) = \mathcal{N}(y_{l,m}^t|\mu_{l,m}, \mathrm{diag}(\sigma_{y_l}^2))$$
$$p(y_{h,m}^t|z_h, x_{h,m}^t, \theta_h) = \mathcal{N}(y_{h,m}^t|\mu_{h,m}, \mathrm{diag}(\sigma_{y_h}^2))$$

where

$$\mu_{z_l} = \mathrm{Enc}_{\mu_{z_l},\phi_l}(\mathcal{D}_l^c), \quad \sigma_{z_l}^2 = \mathrm{Enc}_{\sigma_{z_l}^2,\phi_l}(\mathcal{D}_l^c)$$
$$\mu_{z_h} = \mathrm{Enc}_{\mu_{z_h},\phi_h}(z_l, \mathcal{D}_h^c), \quad \sigma_{z_h}^2 = \mathrm{Enc}_{\sigma_{z_h}^2,\phi_h}(z_l, \mathcal{D}_h^c)$$
$$\mu_{y_l} = \mathrm{Dec}_{\mu_{y_l},\theta_l}(z_l, x_{l,m}^t), \quad \sigma_{y_l}^2 = \mathrm{Dec}_{\sigma_{y_l}^2,\theta_l}(z_l, x_{l,m}^t)$$
$$\mu_{y_h} = \mathrm{Dec}_{\mu_{y_h},\theta_h}(z_h, x_{h,m}^t), \quad \sigma_{y_h}^2 = \mathrm{Dec}_{\sigma_{y_h}^2,\theta_h}(z_h, x_{h,m}^t)$$

We derive the unified ELBO containing these four terms:

$$\begin{aligned}
&\log p(y_l^t, y_l^t|x_l^t, x_h^t, \mathcal{D}_l^c, \mathcal{D}_h^c, \theta) \\
&\geq \mathbb{E}_{q_\phi(z_l,z_h|\mathcal{D}_l^c\cup\mathcal{D}_l^t,\mathcal{D}_h^c\cup\mathcal{D}_h^t)}\big[\log p(y_l^t, y_h^t|z_l, z_h, x_l^t, x_h^t, \theta) \\
&+ \log \frac{q_\phi(z_l, z_h|\mathcal{D}_l^c, \mathcal{D}_h^c)}{q_\phi(z_l, z_h|\mathcal{D}_l^c \cup \mathcal{D}_l^t, \mathcal{D}_h^c \cup \mathcal{D}_h^t)}\big] \\
&= \mathbb{E}_{q_{\phi_h}(z_h|z_l,\mathcal{D}_h^c\cup\mathcal{D}_h^t)q_{\phi_l}(z^l|\mathcal{D}_l^c\cup\mathcal{D}_l^t)}\big[\log p(y_h^t|z_h, x_h^t, \theta_h) \\
&+ \log p(y_l^t|z_l, x_l^t, \theta_l) + \log \frac{q_{\phi_h}(z_h|z_l, \mathcal{D}_h^c)}{q_{\phi_h}(z_h|z^l, \mathcal{D}_h^c \cup \mathcal{D}_h^t)} \\
&+ \frac{q_{\phi_l}(z_l|\mathcal{D}_l^c)}{q_{\phi_l}(z_l|\mathcal{D}_l^c \cup \mathcal{D}_l^t)}\big]
\end{aligned} \tag{1}$$

The derivation is based on the conditional independence of MF-HNParchitecture shown in Figure 1.

4.3 Scalable Training

To calculate the ELBO in Equation 1 for the proposed MF-HNP model, we use Monte Carlo (MC) sampling to optimize the following objective function:

$$\begin{aligned}
\mathcal{L}_{MC} = \frac{1}{K}\sum_{k=1}^{K}\big[&\frac{1}{S}\sum_{s=1}^{S}\log p(y_h^t|x_h^t, z_h^{(s)}, z_l^{(k)}) \\
&- \mathrm{KL}[q(z_h|z_l^{(k)}, \mathcal{D}_h^c, \mathcal{D}_h^t))\|p(z_h|z_l^{(k)}, \mathcal{D}_h^c)]\big] \\
&+ \frac{1}{K}\sum_{k=1}^{K}\log p(y_l^t|x_l^t, z_l^{(k)}) - \mathrm{KL}\big[q(z_l|\mathcal{D}_l^c, \mathcal{D}_l^t)\|p(z_l|\mathcal{D}_l^c)\big]
\end{aligned}$$

Table 1: Comparison of different NP models at high-fidelity level.

Neural Processes Family	Prior Distribution	Posterior Distribution	Generative model
SF-NP [11]	$q(z_h\|\mathcal{D}_h^c)$	$p(z\|\mathcal{D}_h^c,\mathcal{D}_h^t)$	$p(y_h^t\|x_h^t,z)$
MF-NP [42]	$q(z_h\|\mathcal{D}_h^c)$	$p(z\|\mathcal{D}_h^c,\mathcal{D}_h^t)$	$p(y_h^t\|x_h^t,y_l^t,z)$
MF-HNP(as)	$q(z_h\|z_l^{(s)},\mathcal{D}_h^c)$	$p(z_h\|z_l^{(s)},\mathcal{D}_h^c,\mathcal{D}_h^t)$	$p(y_h^t\|x_h^t,z_h)$
MF-HNP(mean)	$q(z_h\|\mu_{z_l},\mathcal{D}_h^c)$	$p(z_h\|\mu_{z_l},\mathcal{D}_h^c,\mathcal{D}_h^t)$	$p(y_h^t\|x_h^t,z_h)$
MF-HNP(mean,std)	$q(z_h\|\mu_{z_l},\sigma_{z_l},\mathcal{D}_h^c)$	$p(z_h\|\mu_{z_l},\sigma_{z_l},\mathcal{D}_h^c,\mathcal{D}_h^t)$	$p(y_h^t\|x_h^t,z_h)$

where the latent variables $z_l^{(k)}$ and $z_h^{(s)}$ are sampled by $q_{\phi_l}(z_l|\mathcal{D}_l^c)$ and $q_{\phi_h}(z_h|z_l^{(k)},\mathcal{D}_h^c)$ respectively. This standard MC sampling method requires nested sampling. For data sets with multiple fidelity levels, it is computationally challenging.

An alternative way is to use ancestral sampling [41] (denoted by MF-HNP(AS)) for scalable training and write the estimation as:

$$\begin{aligned}\mathcal{L}_{AS} = \frac{1}{S}\sum_{s=1}^{S}\Big[&\log p(y_h^t|x_h^t,z_h^{(s)},z_l^{(s)})\\ &-\mathrm{KL}[q(z_h|z_l^{(s)},\mathcal{D}_h^c,\mathcal{D}_h^t))\|p(z_h|z_l^{(s)},\mathcal{D}_h^c)]\Big]\\ &+\frac{1}{K}\sum_{k=1}^{K}\log p(y_l^t|x_l^t,z_l^{(k)})-\mathrm{KL}\big[q(z_l|\mathcal{D}_l^c,\mathcal{D}_l^t)\|p(z_l|\mathcal{D}_l^c)\big]\end{aligned} \tag{2}$$

We also design two different techniques to infer z_h using either low-level mean of latent variables μ_{z_l} (denoted by MF-HNP(MEAN)) or both low-level mean and standard deviation $(\mu_{z_l},\sigma_{z_l}^2)$(denoted by MF-HNP(MEAN,STD)). The corresponding ELBOs are:

$$\begin{aligned}\mathcal{L}_{\mu} = \frac{1}{S}\sum_{s=1}^{S}&\log p(y_h^t|x_h^t,z_h^{(s)},\mu_{z_l})\\ &-\mathrm{KL}[q(z_h|\mu_{z_l},\mathcal{D}_h^c,\mathcal{D}_h^t))\|p(z_h|\mu_{z_l},\mathcal{D}_h^c]\\ &+\frac{1}{K}\sum_{k=1}^{K}\log p(y_l^t|x_l^t,z_l^{(k)})-\mathrm{KL}\big[q(z_l|\mathcal{D}_l^c,\mathcal{D}_l^t)\|p(z_l|\mathcal{D}_l^c)\big]\end{aligned} \tag{3}$$

$$\begin{aligned}\mathcal{L}_{\mu,\sigma} = \frac{1}{S}\sum_{s=1}^{S}&\log p(y_h^t|x_h^t,z_h^{(s)},\mu_{z_l},\sigma_{z_l})\\ &-\mathrm{KL}[q(z_h|\mu_{z_l},\sigma_{z_l},\mathcal{D}_h^c,\mathcal{D}_h^t))\|p(z_h|\mu_{z_l},\sigma_{z_l},\mathcal{D}_h^c]\\ &+\frac{1}{K}\sum_{k=1}^{K}\log p(y_l^t|x_l^t,z_l^{(k)})-\mathrm{KL}\big[q(z_l|\mathcal{D}_l^c,\mathcal{D}_l^t)\|p(z_l|\mathcal{D}_l^c)\big]\end{aligned} \tag{4}$$

We include Equation 2, Equation 3, and Equation 4 as the training loss functions for ablation study. The comparison of different NP models including SF-NP, MF-NP, MF-HNP variants for high-fidelity level inference and output generation is shown in Table 1.

5 EXPERIMENTS

We benchmark the performance of different methods on two multi-fidelity modeling tasks: stochastic epidemiology modeling and climate forecasting. Epidemiology modeling is age-stratified and climate (temperature) modeling is on a regular grid.

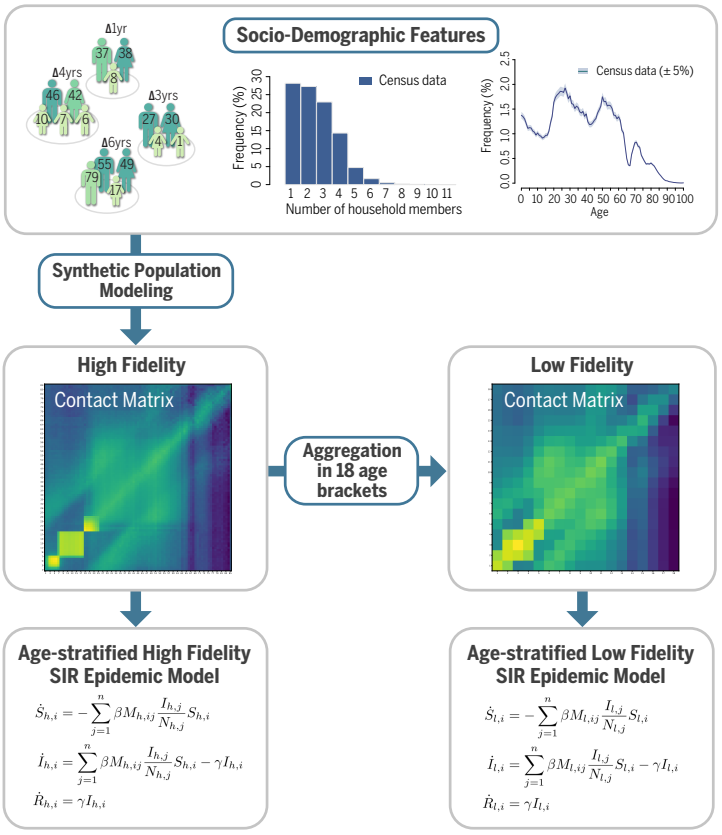

Figure 2: AS-SIR Modeling Framework: First, high-fidelity population-level contact matrices are generated using macro (census) and micro (survey) data [28]. Second, low-fidelity contact matrices are obtained by grouping individuals in fewer age brackets. Distinct age-stratified SIR models are used to simulate the epidemic at the two fidelity levels.

5.1 Experiment Setup.

For all experiments, we compare our proposed MF-HNP model with both the GP and NP baselines.

- GP baselines include the nonlinear autoregressive multi-fidelity GP regression model (NARGP) [31] and single-fidelity Gaussian Processes (SF-GP) which assumes that the data are independent at each fidelity level.
- NP baselines include single-fidelity Neural Processes (SF-NP) and multi-fidelity Neural Processes (MF-NP) [42].

- For our proposed MF-HNP model, we provide 3 variants to approximate inference for ablation study, including inference by low-level mean of latent variables (MF-HNP(MEAN)), low-level mean and standard deviation of latent variables (MF-HNP(MEAN,STD)), and ancestral sampling method (MF-HNP(AS)). Details have been discussed in Section 4.3.

For NP models, we also consider two different context aggregation methods discussed in Section 3.2, including mean context aggregation and Bayesian context aggregation. Both are applied to generate latent variables z at each fidelity level. For the NARGP and MF-NP baseline, they only work for the data with nested data structure based on their model architecture and assumption [31]. For MF-NP, it requires both low-fidelity simulation output y^l and high-fidelity input x^h as model input. Therefore, we assume that y^l is known for the validation and test set for MF-NP, which means MF-NP requires more data compared with MF-HNPand other baselines.

We report the mean absolute error (MAE) for accuracy estimation. For uncertainty estimation, we use mean negative log-likelihood (NLL). For age-stratified Susceptible-Infectious-Recovered (AS-SIR) experiment, we perform a log transformation on the number of infections in the output space to deal with the long-tailed distribution. NLL for AS-SIR experiment is calculated in the log space, while MAE is calculated in the original space. For climate modeling experiment, both NLL and MAE are measured in the original space. We calculate the NLL based on the Gaussian distribution determined by model outputs of mean and standard deviation, and MAE between the mean predictions and the truth.

5.2 Age-Stratified SIR Compartmental Model

We use an age-stratified Susceptible-Infectious-Recovered (AS-SIR) epidemic model:

$$\dot{S}_i = -\lambda_i S_i, \quad \dot{I}_i = \lambda_i S_i - \gamma I_i, \quad \dot{R}_i = \gamma I_i$$

where S_i, I_i, and R_i denote the number of susceptible, infected, and recovered individuals of age i, respectively. The age-specific force of infection is defined by λ_i and it is equal to:

$$\lambda_i = \beta \sum_j M_{i,j} \frac{I_j}{N_j},$$

where β denotes the transmissibility rate of the infection, N_j is the total number of individuals of age j, and $M_{i,j}$ is the overall age-stratified contact matrices describing the average number of contacts with individuals of age j for an individual of age i.

This model assumes heterogeneous mixing between age groups, where the population-level contact matrices M are generated using highly detailed macro (census) and micro (survey) data on key socio-demographic features [28] to realistically capture the social mixing differences that exist between different countries/regions of the world and that will affect the spread of the virus.

Dataset. We include overall 109 scenarios at different locations in China, U.S., Europe. The data in China is at the province level. The data in the U.S. is at state level. The data in Europe is at the country level. For each scenario, we generate 30 samples for 100 day's new infection prediction at low- and high-fidelity levels based on the corresponding initial conditions, R_0, age-stratified population, and the overall age-stratified contact matrices. The high-fidelity data, as shown in Figure 2, has 85 age groups. The size of the age-stratified contact matrices $M_{h,ij}$ is 85×85. For low-fidelity data, we aggregate the data and obtain 18 age groups, resulting in a contact matrix $M_{l,ij}$ of size 18×18.

We randomly split 31 scenarios for training candidate set, 26 scenarios for the validation set and 52 scenarios for test set at both fidelity levels. In the nested data set case, we first randomly select 26 scenarios from the training candidate set as the training set at low-fidelity level, then randomly select 5 scenarios from them as the training set at high-fidelity level. In the non-nested data set case, we randomly split 26 scenarios as the training set at low-fidelity level and 5 scenarios as the training set at high-fidelity level. The validation and test set are both at high-fidelity level.

Performance Analysis. Table 2 compares the prediction performance for 2 GP methods and 10 NP methods for 100 day ahead infection forecasting. The performance is reported in MAE and NLL over 100 days. MF-HNP(MEAN)-BA has the best prediction performance in terms of MAE for both the scenario with nested data structure and non-nested data structure. GP baselines SF-GP and NARGP have similar worst MAE, which means the low-fidelity data does not help NARGP learn useful information. Because in high-dimensions, the strict assumption of no observation noise at low-fidelity level does not hold for NARGP.

For NP baselines, MF-NP-(MA/BA) baselines have worse accuracy performance compared with the SF-NP-(MA/BA) baselines. This is due to the limited number of paired training data that MF-NP can utilize. The small number of training data plus the high-dimensional input and output space makes it difficult for MF-NP to learn the correct pattern for model predictions. For all NP models, we find Bayesian aggregation improves the performance. With respect to different hierarchical inference methods of MF-HNP. Table 2 shows MF-HNP(AS) and MF-HNP(MEAN) have superior performance compared to MF-HNP(MEAN,STD) in terms of both NLL and MAE.

Figure 3 visualizes the prediction results of two randomly selected scenarios in the nested dataset. It shows the truth, our MF-HNP prediction together with two other baselines representing the best GP baseline and the best NP baseline in four age groups (10,30,50,70). In this experiment, the best GP is NARGP and the best NP is SF-NP. One interesting finding is that although SF-GP has the best NLL score, the visualization shows its prediction is very conservative by generating a large confidence interval, which is not informative. On the contrary, MF-HNPprediction is able to generate a narrower confidence interval while covering the truth at the same time (shown in Figure 3).

When switching to non-nest data set, the MF-HNP model is still reliable for this much harder task. In fact, the MAE performance of MF-HNP(MEAN)-BA is even better.

5.3 Climate Model for Temperature.

We further test our method on the multi-fidelity climate dataset provided by Hosking [15]. The dataset includes low-fidelity and high-fidelity climate model temperature simulations over a region in Peru. The left part of Figure 4 shows the region of interest.

Dataset. The low-fidelity data is generated by low-fidelity Global Climate Model with spatial resolution 14×14 [20]. The high-fidelity

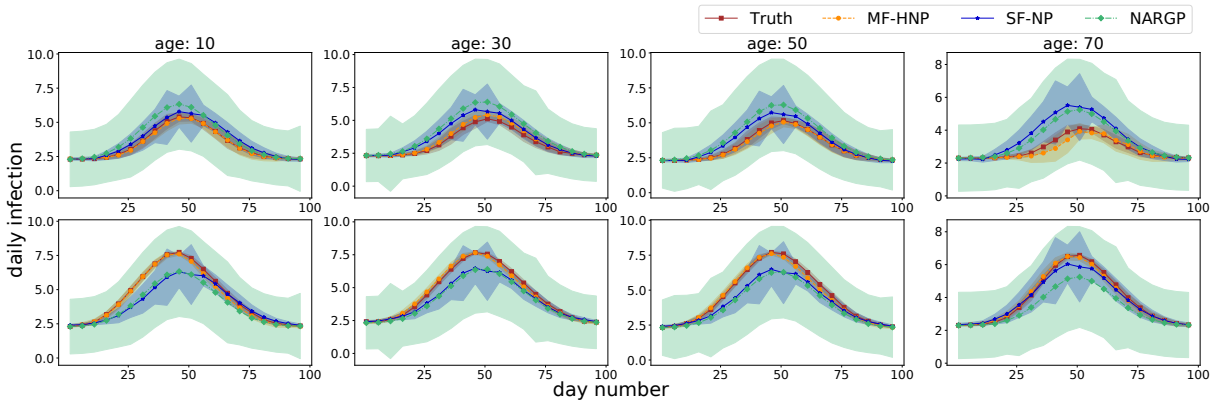

Figure 3: 100 days ahead infectious incidence compartment forecasting of randomly selected scenario at each row, analyzed in 4 age groups. Natural log scale for y axis.

Table 2: Prediction performance comparison on Age-Stratified SIR data sets.

Data	Method	MAE (nested) ↓	NLL (nested) ↓	MAE (non-nested) ↓	NLL (non-nested) ↓
nested	SF-GP	342.99 ± 0.04	**1.71** ± 0.06	342.99 ± 0.04	**1.71** ± 0.06
	NARGP	342.72 ± 0.13	1.78 ± 0.1	×	×
	SF-NP-MA	333.41 ± 100.73	6.14 ± 4.11	333.41 ± 100.73	6.14 ± 4.11
	MF-NP-MA	341.08 ± 0.18	6.5 ± 0.58	×	×
	MF-HNP(mean)-MA	257.39 ± 24.17	11.09 ± 11.93	249.5 ± 25.82	10.58 ± 11.33
	MF-HNP(mean,std)-MA	257.0 ± 23.13	9.26 ± 9.38	254.04 ± 18.0	13.04 ± 14.84
	MF-HNP(as)-MA	266.17 ± 16.13	10.59 ± 11.06	262.61 ± 10.68	11.66 ± 12.71
	SF-NP-BA	294.3 ± 75.81	36.35 ± 46.5	294.3 ± 75.81	36.35 ± 46.5
	MF-NP-BA	340.22 ± 1.51	4.34 ± 2.23	×	×
	MF-HNP(mean)-BA	**201.56** ± 61.15	1.97 ± 0.44	**199.75** ± 64.51	1.95 ± 0.5
	MF-HNP(mean,std)-BA	229.09 ± 77.44	8.24 ± 9.54	203.05 ± 65.84	6.66 ± 7.19
	MF-HNP(as)-BA	205.26 ± 49.1	2.69 ± 1.0	205.43 ± 43.79	3.24 ± 1.59

data is generated by high-fidelity Regional Climate Model [1] with spatial resolution 87×87. The example is shown in Figure 4. Both include monthly data from 1980 to 2018 over the same region (latitude range: $(-7.5, -10.7)$, longitude range: $(280.5, 283.7)$).

The task is to use 6 month data as input to generate the next 6 month predictions as output. We randomly split 119 scenarios for training candidate set, 50 scenarios for validation set, and 50 scenarios for the test set at both fidelity level. In the nested data set case, we first randomly select 87 scenarios from the training candidate set as the training set at low-fidelity level, then randomly select 32 scenarios from them as the training set at high-fidelity level. In the non-nested data set case, we randomly split 87 scenarios as the training set at low-fidelity level and 32 scenarios as the training set at high-fidelity level. The validation and test set are both at high-fidelity level.

Performance Analysis. Table 3 compares the prediction performance for 2 GP methods and 10 NP methods to predict the next 6 months temperature based on the past 6 months temperature data. The performance is reported in MAE and NLL. The results of this task are consistent with what we found in AS-SIR infection prediction task. `MF-HNP` has significantly better performance compared with either GP or NP baselines. But this time `MF-HNP`(MC)-BA is the most accurate one with or without a nested data structure. Considering both MAE and NLL, we still recommend using `MF-HNP`(MC)-BA and `MF-HNP`(MEAN)-BA.

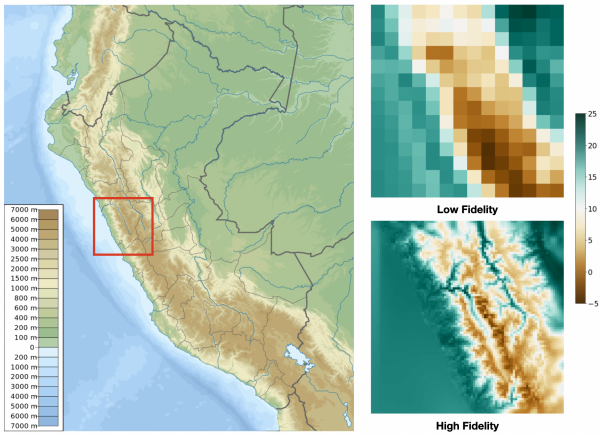

Figure 4: Left: Region of interest [4]. Upper Right: sample from low-fidelity temperature model. Lower Right: sample from high-fidelity temperature model.

Figure 5 is the visualization of predictions among the best `MF-HNP` variant, GP and NP baselines on a randomly selected scenario in the test set. To highlight the performance difference, we visualize the residual between the predictions and the truth from 1 to 6 months

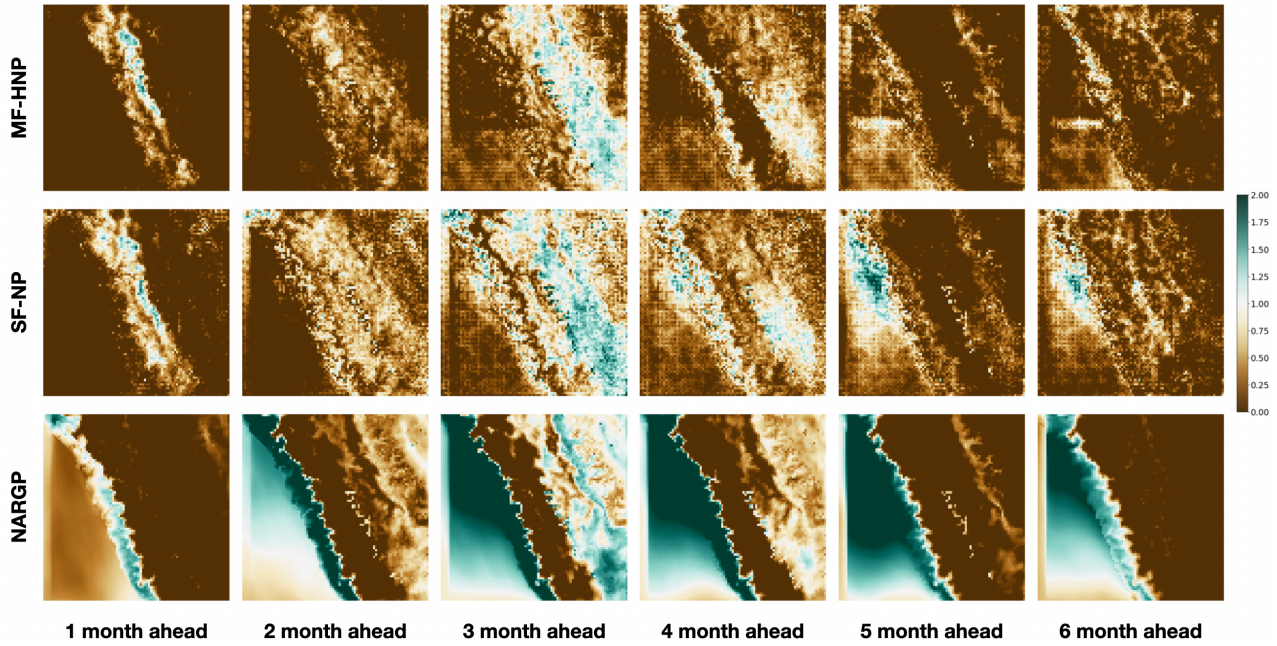

Figure 5: MF-HNP vs. SF-NP vs. NARGP for 6 month ahead temperature prediction residual.

Table 3: Prediction performance comparison on climate data sets.

Method	MAE (nested) ↓	NLL (nested) ↓	MAE (non-nested) ↓	NLL (non-nested) ↓
SF-GP	0.91 ± 0.365	2.288 ± 0.004	0.91 ± 0.365	2.288 ± 0.004
NARGP	0.91 ± 0.365	2.3 ± 0.006	×	×
SF-NP-MA	0.778 ± 0.01	**1.489** ± 0.026	0.778 ± 0.01	**1.489** ± 0.026
MF-NP-MA	0.902 ± 0.005	1.889 ± 0.012	×	×
MF-HNP(mean)-MA	0.765 ± 0.004	1.535 ± 0.059	0.788 ± 0.029	1.666 ± 0.174
MF-HNP(mean,std)-MA	0.773 ± 0.011	1.592 ± 0.057	0.768 ± 0.027	1.607 ± 0.089
MF-HNP(as)-MA	0.758 ± 0.024	1.578 ± 0.079	0.769 ± 0.02	1.594 ± 0.098
SF-NP-BA	0.751 ± 0.052	1.546 ± 0.133	0.751 ± 0.052	1.546 ± 0.133
MF-NP-BA	0.954 ± 0.019	1.909 ± 0.028	×	×
MF-HNP(mean)-BA	0.706 ± 0.049	1.549 ± 0.164	0.714 ± 0.027	1.58 ± 0.061
MF-HNP(mean,std)-BA	0.717 ± 0.045	1.606 ± 0.106	0.695 ± 0.03	1.548 ± 0.068
MF-HNP(as)-BA	**0.678** ± 0.026	1.506 ± 0.027	**0.68** ± 0.009	1.58 ± 0.012

ahead predictions. Higher value means lower accuracy. It can be found that MF-HNP outperforms all the baselines for the predictions for each month.

6 CONCLUSION & LIMITATION

We propose Multi-Fidelity Hierarchical Neural Process (MF-HNP), the first *unified* framework for scalable multi-fidelity surrogate modeling in the neural processes family. Our model is more flexible and scalable compared with existing multi-fidelity modeling approaches. Specifically, it no longer requires a nested data structure for training and supports varying input and output dimensions at different fidelity levels. Moreover, the latent variables introduce conditional independence for different fidelity levels, which alleviates the error propagation issue and improves the accuracy and uncertainty estimation performance. We demonstrate the superiority of our method on two real-world large-scale multi-fidelity applications: age-stratified epidemiology modeling and temperature outputs from different climate models.

Regarding future work, it is natural to extend our multi-fidelity Hierarchical Neural Process to active learning setup. Instead of passively training the neural processes, we can proactively query the simulator, gather training data, and incrementally improve the surrogate model performance.

ACKNOWLEDGEMENT

This work was supported in part by U.S. Department Of Energy, Office of Science under Award #DE-SC0022255, U. S. Army Research Office under Grant W911NF-20-1-0334, Facebook Data Science Research Awards, AWS Machine Learning Research Award, Google Faculty Award, NSF Grants #2037745, NSF-SCALE MoDL-2134209, and NSF-CCF-2112665 (TILOS). D.W. acknowledges support from the HDSI Ph.D. Fellowship. M.C. and A.V. acknowledge

support from grants HHS/CDC 5U01IP0001137 and HHS/CDC 6U01IP001137. The findings and conclusions in this study are those of the authors and do not necessarily represent the official position of the funding agencies, the National Institutes of Health, or the U.S. Department of Health and Human Services.

REFERENCES

[1] Edward Armstrong, Peter O Hopcroft, and Paul J Valdes. 2019. Reassessing the value of regional climate modeling using paleoclimate simulations. *Geophysical Research Letters* 46, 21 (2019), 12464–12475.

[2] Loïc Brevault, Mathieu Balesdent, and Ali Hebbal. 2020. Overview of Gaussian process based multi-fidelity techniques with variable relationship between fidelities, application to aerospace systems. *Aerospace Science and Technology* 107 (2020), 106339.

[3] Matteo Chinazzi, Jessica T Davis, Marco Ajelli, Corrado Gioannini, Maria Litvinova, Stefano Merler, Ana Pastore y Piontti, Kunpeng Mu, Luca Rossi, Kaiyuan Sun, et al. 2020. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* (2020).

[4] Wikimedia Commons. 2022. File:Peru physical map.svg — Wikimedia Commons, the free media repository. https://commons.wikimedia.org/w/index.php?title=File:Peru_physical_map.svg&oldid=618434504 [Online; accessed 9-February-2022].

[5] Estee Y Cramer, Velma K Lopez, Jarad Niemi, Glover E George, Jeffrey C Cegan, Ian D Dettwiller, William P England, Matthew W Farthing, Robert H Hunter, Brandon Lafferty, et al. 2022. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US. *PNAS* (2022).

[6] Kurt Cutajar, Mark Pullin, Andreas Damianou, Neil Lawrence, and Javier González. 2019. Deep gaussian processes for multi-fidelity modeling. *arXiv preprint arXiv:1903.07320* (2019).

[7] Andreas Damianou and Neil D Lawrence. 2013. Deep gaussian processes. In *Artificial intelligence and statistics*. PMLR, 207–215.

[8] Jessica T Davis, Matteo Chinazzi, Nicola Perra, Kunpeng Mu, Ana Pastore y Piontti, Marco Ajelli, Natalie E Dean, Corrado Gioannini, Maria Litvinova, Stefano Merler, et al. 2021. Cryptic transmission of SARS-CoV-2 and the first COVID-19 wave. *Nature* 600, 7887 (2021), 127–132.

[9] Subhayan De, Jolene Britton, Matthew Reynolds, Ryan Skinner, Kenneth Jansen, and Alireza Doostan. 2020. On transfer learning of neural networks using bi-fidelity data for uncertainty propagation. *International Journal for Uncertainty Quantification* 10, 6 (2020).

[10] Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. 2018. Conditional neural processes. In *International Conference on Machine Learning*. PMLR, 1704–1713.

[11] Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. 2018. Neural processes. *arXiv preprint arXiv:1807.01622* (2018).

[12] Mengwu Guo, Andrea Manzoni, Maurice Amendt, Paolo Conti, and Jan S Hesthaven. 2022. Multi-fidelity regression using artificial neural networks: efficient approximation of parameter-dependent output quantities. *Computer methods in applied mechanics and engineering* 389 (2022), 114378.

[13] M. Elizabeth Halloran, Neil M. Ferguson, Stephen Eubank, Ira M. Longini, Derek A. T. Cummings, Bryan Lewis, Shufu Xu, Christophe Fraser, Anil Vullikanti, Timothy C. Germann, Diane Wagener, Richard Beckman, Kai Kadau, Chris Barrett, Catherine A. Macken, Donald S. Burke, and Philip Cooley. 2008. Modeling targeted layered containment of an influenza pandemic in the United States. *Proceedings of the National Academy of Sciences* 105, 12 (2008), 4639–4644. https://doi.org/10.1073/pnas.0706849105 arXiv:https://www.pnas.org/content/105/12/4639.full.pdf

[14] Ali Hebbal, Loic Brevault, Mathieu Balesdent, El-Ghazali Talbi, and Nouredine Melab. 2021. Multi-fidelity modeling with different input domain definitions using Deep Gaussian Processes. *Structural and Multidisciplinary Optimization* 63, 5 (2021), 2267–2288.

[15] S. Hosking. 2020. Multifidelity climate modelling, GitHub. https://github.com/scotthosking/mf_modelling.

[16] Kirthevasan Kandasamy, Gautam Dasarathy, Jeff Schneider, and Barnabás Póczos. 2017. Multi-fidelity bayesian optimisation with continuous approximations. In *International Conference on Machine Learning*. PMLR, 1799–1808.

[17] Marc C Kennedy and Anthony O'Hagan. 2000. Predicting the output from a complex computer code when fast approximations are available. *Biometrika* 87, 1 (2000), 1–13.

[18] Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. 2018. Attentive Neural Processes. In *International Conference on Learning Representations*.

[19] Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. 2019. Attentive neural processes. *arXiv preprint arXiv:1901.05761* (2019).

[20] NOAA Geophysical Fluid Dynamics Laboratory. 2009. Climate modeling, Geophysical Fluid Dynamics Laboratory. https://www.gfdl.noaa.gov/climate-modeling.

[21] Loic Le Gratiet and Josselin Garnier. 2014. Recursive co-kriging model for design of computer experiments with multiple levels of fidelity. *International Journal for Uncertainty Quantification* 4, 5 (2014).

[22] Shibo Li, Robert Kirby, and Shandian Zhe. 2021. Batch Multi-Fidelity Bayesian Optimization with Deep Auto-Regressive Networks. *Advances in Neural Information Processing Systems* 34 (2021).

[23] Shibo Li, Wei Xing, Robert Kirby, and Shandian Zhe. 2020. Multi-fidelity Bayesian optimization via deep neural networks. *Advances in Neural Information Processing Systems* 33 (2020), 8521–8531.

[24] Eric T. Lofgren, M. Elizabeth Halloran, Caitlin M. Rivers, John M. Drake, Travis C. Porco, Bryan Lewis, Wan Yang, Alessandro Vespignani, Jeffrey Shaman, Joseph N. S. Eisenberg, Marisa C. Eisenberg, Madhav Marathe, Samuel V. Scarpino, Kathleen A. Alexander, Rafael Meza, Matthew J. Ferrari, James M. Hyman, Lauren A. Meyers, and Stephen Eubank. 2014. Opinion: Mathematical models: A key tool for outbreak response. *Proceedings of the National Academy of Sciences* 111, 51 (2014), 18095–18096. https://doi.org/10.1073/pnas.1421551111 arXiv:https://www.pnas.org/content/111/51/18095.full.pdf

[25] Christos Louizos, Xiahan Shi, Klamer Schutte, and Max Welling. 2019. The Functional Neural Process. *Advances in Neural Information Processing Systems* (2019).

[26] Xuhui Meng, Hessam Babaee, and George Em Karniadakis. 2021. Multi-fidelity Bayesian neural networks: Algorithms and applications. *J. Comput. Phys.* 438 (2021), 110361.

[27] Xuhui Meng and George Em Karniadakis. 2020. A composite neural network that learns from multi-fidelity data: Application to function approximation and inverse PDE problems. *J. Comput. Phys.* 401 (2020), 109020.

[28] Dina Mistry, Maria Litvinova, Ana Pastore y Piontti, Matteo Chinazzi, Laura Fumanelli, Marcelo FC Gomes, Syed A Haque, Quan-Hui Liu, Kunpeng Mu, Xinyue Xiong, et al. 2021. Inferring high-resolution human mixing patterns for disease modeling. *Nature communications* 12, 1 (2021), 1–12.

[29] Bernt Øksendal. 2003. Stochastic differential equations. In *Stochastic differential equations*. Springer, 65–84.

[30] Benjamin Peherstorfer, Karen Willcox, and Max Gunzburger. 2018. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *Siam Review* 60, 3 (2018), 550–591.

[31] Paris Perdikaris, Maziar Raissi, Andreas Damianou, Neil D Lawrence, and George Em Karniadakis. 2017. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 473, 2198 (2017), 20160751.

[32] Paris Perdikaris, Daniele Venturi, and George Em Karniadakis. 2016. Multifidelity information fusion algorithms for high-dimensional systems and massive data sets. *SIAM Journal on Scientific Computing* 38, 4 (2016), B521–B538.

[33] Paris Perdikaris, Daniele Venturi, Johannes O Royset, and George Em Karniadakis. 2015. Multi-fidelity modelling via recursive co-kriging and Gaussian–Markov random fields. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 471, 2179 (2015), 20150018.

[34] Daniel J Perry, Robert M Kirby, Akil Narayan, and Ross T Whitaker. 2019. Allocation strategies for high fidelity models in the multifidelity regime. *SIAM/ASA Journal on Uncertainty Quantification* 7, 1 (2019), 203–231.

[35] Maziar Raissi and George Karniadakis. 2016. Deep multi-fidelity Gaussian processes. *arXiv preprint arXiv:1604.07484* (2016).

[36] Carl Edward Rasmussen. 2003. Gaussian processes in machine learning. In *Summer school on machine learning*. Springer, 63–71.

[37] Hugh Salimbeni and Marc Peter Deisenroth. 2017. Doubly Stochastic Variational Inference for Deep Gaussian Processes. In *NIPS*.

[38] Gautam Singh, Jaesik Yoon, Youngsung Son, and Sungjin Ahn. 2019. Sequential Neural Processes. *Advances in Neural Information Processing Systems* 32 (2019), 10254–10264.

[39] Mario Miguel Valero, Lluís Jofre, and Ricardo Torres. 2021. Multifidelity prediction in wildfire spread simulation: Modeling, uncertainty quantification and sensitivity analysis. *Environmental Modelling & Software* 141 (2021), 105050.

[40] Michael Volpp, Fabian Flürenbrock, Lukas Grossberger, Christian Daniel, and Gerhard Neumann. 2020. Bayesian context aggregation for neural processes. In *International Conference on Learning Representations*.

[41] Qi Wang and Herke Van Hoof. 2020. Doubly stochastic variational inference for neural processes with hierarchical latent variables. In *International Conference on Machine Learning*. PMLR, 10018–10028.

[42] Yating Wang and Guang Lin. 2020. MFPC-Net: Multi-fidelity Physics-Constrained Neural Process. *arXiv preprint arXiv:2010.01378* (2020).

[43] Zheng Wang, Wei Xing, Robert Kirby, and Shandian Zhe. 2021. Multi-Fidelity High-Order Gaussian Processes for Physical Simulation. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 847–855.

[44] Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. 2016. Deep kernel learning. In *Artificial intelligence and statistics*. PMLR, 370–378.

Table 4: Hyperparameters for NP baselines and our proposed MF-HNP model, including learning rate, batch size, and patience.

	LEARNING RATE	BATCH SIZE	PATIENCE
AS-SIR	$1e^{-3}$	128	1000
CLIMATE	$5e^{-3}$	32	250

A NEURAL PROCESSS BASELINES

A.1 Single-Fidelity Neural Processes (SF-NP).

A simple way to apply NP to the multi-fidelity problem is to train NP only using the data at high-fidelity level only assuming it is not correlated with the data at the low-fidelity level. We name it as Single-Fidelity Neural Processes baseline (SF-NP). During the training process, the high-level training data can be randomly split into context set $\mathcal{D}_h^c$ and target set $\mathcal{D}_h^t$. We use the corresponding evidence lower bound (ELBO) as the training loss function:

$$\log p(y^t_{h,1:M}|x^t_{h,1:M}, \mathcal{D}^c_h, \theta) \geq$$
$$\mathbb{E}_{q_\phi(z|\mathcal{D}^c_h \cup \mathcal{D}^t_h)}\Big[\sum_{m=1}^{M} \log p(y^t_{h,m}|z, x^t_{h,m}, \theta) + log\frac{q_\phi(z|\mathcal{D}^c_h)}{q_\phi(z|\mathcal{D}^c_h \cup \mathcal{D}^t_h)}\Big]$$

where $p(\theta)$ is a decoder in a neural network and q_ϕ indicates a encoder to infer the latent variable z.

A.2 Multi-Fidelity Neural Processes (MF-NP).

Multi-Fidelity Neural Processes (MF-NP) [42] assume a comprehensive correlation between multi-fidelity models y_h and y_l can be represented as:

$$y_h(x) = \mathcal{G}(y_l(x)) + \delta(x),$$

where $\mathcal{G}$ is a nonlinear function mapping the low-fidelity data to high-fidelity data, and $\delta(x)$ is space dependent bias between fidelity levels. To train MF-NP model, we take data pairs $(x, y_l(x))$ as the input to predict the corresponding $y_h(x)$. The corresponding context sets $\mathcal{D}^c_l \equiv \{x^c_{h,n}, y^c_{l,n}, y^c_{h,n}\}_{n=1}^{N_l}$ and target sets $\mathcal{D}^t_l \equiv \{x^t_{h,m}, y^t_{l,m}, y^t_{h,n}\}_{m=1}^{M_l}$. The ELBO for the training process is:

$$\log p(y^t_{h,1:M}|x^t_{h,1:M}, y^t_{l,1:M}, \mathcal{D}^c_h, \theta) \geq$$
$$\mathbb{E}_{q_\phi(z|\mathcal{D}^c_h \cup \mathcal{D}^t_h)}\Big[\sum_{m=1}^{M} \log p(y^t_{h,m}|z, x^t_{h,m}, y^t_{l,m}, \theta) +$$
$$log\frac{q_\phi(z|\mathcal{D}^c_h)}{q_\phi(z|\mathcal{D}^c_h \cup \mathcal{D}^t_h)}\Big]$$

Since this method requires $(x, y_l(x), y_h(x))$ for input and output, it can not fully utilize the training data at low-fidelity level which $y_h(x)$ is unknown. Furthermore, MF-NP requires a nested data structure, which means the training inputs of high-fidelity level need to be a subset of the training inputs of low-fidelity level. On the contrary, if the training inputs at the different fidelity level are disjoint, no data set can be used for training.

B EXPERIMENT DETAILS

For GP baselines, we use RBF kernels. The optimal learning rate is $5e^{-2}$ for both AS-SIR and climate modeling tasks. We train 2000 epochs with patience equal to 100 to ensure convergence. For NP baselines and our proposed MF-HNP model, the hyperparameters can be found in Table 4.