Multi-Fidelity Residual Neural Processes for Scalable Surrogate Modeling

Ruijia Niu [1] **Dongxia Wu** [1] **Kai Kim** [1] **Yi-An Ma** [2] **Duncan Watson-Parris** [2 3] **Rose Yu** [1]

Abstract

Multi-fidelity surrogate modeling aims to learn an accurate surrogate at the highest fidelity level by combining data from multiple sources. Traditional methods relying on Gaussian processes can hardly scale to high-dimensional data. Deep learning approaches utilize neural network based encoders and decoders to improve scalability. These approaches share encoded representations across fidelities without including corresponding decoder parameters. This hinders inference performance, especially in out-of-distribution scenarios when the highest fidelity data has limited domain coverage. To address these limitations, we propose Multi-fidelity Residual Neural Processes (`MFRNP`), a novel multi-fidelity surrogate modeling framework. `MFRNP` explicitly models the residual between the aggregated output from lower fidelities and ground truth at the highest fidelity. The aggregation introduces decoders into the information sharing step and optimizes lower fidelity decoders to accurately capture both in-fidelity and cross-fidelity information. We show that `MFRNP` significantly outperforms state-of-the-art in learning partial differential equations and a real-world climate modeling task. Our code is published at: `github.com/Rose-STL-Lab/MFRNP`.

1. Introduction

From engineering to climate science, a computational model, often realized by simulation, is frequently used to characterize the input-output relationship of a physical system. These computational models can be simulated at multiple levels of sophistication. The high-fidelity simulators are more accurate but resource demanding, whereas lower-fidelity models are less accurate but more computationally efficient. For example, in climate science, people incorporate real-world observations with computational models to calibrate simulations (Hersbach et al., 2020; Karami & Kashef, 2020). While the calibrated data has refined details and accuracy, the domain coverage is very limited and the simulation process is computation heavy. Multi-fidelity surrogate modeling (Peherstorfer et al., 2018) aims to balance the computation efficiency-accuracy trade-off by utilizing data across fidelities to learn an accurate surrogate at the highest fidelity.

Gaussian Processes (GPs) (Seeger, 2004) are a popular tool for surrogate modeling. Recent works have attempted to extend GP to multi-fidelity setting (Perdikaris et al., 2016; Wang et al., 2021). However, they inherit the limited scalability from GP due to the inversion of the kernel matrix (Williams & Rasmussen, 1995; Rasmussen, 2003). To solve this issue, many have proposed deep learning-based surrogate models (Damianou & Lawrence, 2013; Raissi & Karniadakis, 2016; Salimbeni & Deisenroth, 2017; Wilson et al., 2016). Neural Processes (NPs) (Wang & Lin, 2020; Hebbal et al., 2021) stand out as one of the most appealing approaches regarding inference performance and scalability. NPs are capable of encoding fidelity-specific data into low-dimensional latent representations and use them to improve inference performance at the highest fidelity, alleviating the scalability issue from high-dimensional data.

For example, Deep Multi-fidelity Active Learning (DMFAL) (Li et al., 2022b;a) proposed to pass information from lower fidelities to higher fidelities with encoded hidden representations. This method requires a hierarchical structure in the latent space passing information from low to high fidelity level, which can lead to the error propagation issue. Disentangled Multi-fidelity Deep Bayesian Active Learning (D-MFDAL) (Wu et al., 2023) alleviates this issue by redesigning the NP using local and global hidden representations. Nevertheless, these methods rely on latent representations from only the encoders for cross-fidelity information sharing. However, the decoder parameters varies across fidelities and are not shared. At the highest fidelity, shared representations are decoded with different parameters, making the decoded output inherently inaccurate. This significantly limits the inference performance, especially

[1]Department of Computer Science and Engineering, University of California San Diego, La Jolla, California, USA [2]Halıcıoğlu Data Science Institute, University of California San Diego, La Jolla, California, USA [3]Scripps Institution of Oceanography, University of California San Diego, La Jolla, California, USA. Correspondence to: Rose Yu <roseyu@ucsd.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

when the model needs to extrapolate with out-of-distribution (OOD) inputs when the training data has limited domain coverage at the highest fidelity.

In this work, we introduce a novel multi-fidelity surrogate modeling framework, Multi-fidelity Residual Neural Process (MFRNP), to address the aforementioned issues. MFRNP aggregates the predictions from surrogate models across lower-fidelity levels and employs an NP surrogate to capture the residual between the aggregated prediction and the ground truth at the highest fidelity level. By directly utilizing the outputs from lower-fidelity surrogates to share information, MFRNP includes decoders in cross-fidelity information sharing, improving accuracy while maintaining scalability. Moreover, we developed a tailored Evidence Lower Bound, named Residual-ELBO, to serve as our loss function. This novel loss function ensures the highest fidelity latent variable z_K depends on all the other latent variables and decoder parameters across fidelities. Thus, MFRNP ensures accurate information aggregation from lower fidelities to promote residual modeling at the highest fidelity. To summarize, our contributions include:

- A novel multi-fidelity surrogate model, Multi-fidelity Residual Neural Process (MFRNP). Its architecture shares input-specific information from lower fidelities, tackles the varying decoder problem with no error propagation yet preserving scalability.
- A novel Residual-ELBO to simultaneously promote learning across fidelities and optimize lower fidelity decoders for residual modeling at the highest fidelity.
- Superior performance in global scale real-world climate modeling and numerous benchmark tasks on partial differential equations. MFRNP outperform the state-of-the-art baseline by $\sim$**90%** in average.

2. Background

Multi-Fidelity Modeling. Multi-fidelity modeling aims to capture the complex mapping from low-dimensional input variables $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ to high-dimensional output $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ of the function $f : \mathcal{X} \to \mathcal{Y}$. For systems with K fidelities where $K > 1$, the cost c_k of evaluating $f_k \in \{f_1, \cdots, f_K\}$ increases with the fidelity level $(c_1 < \cdots < c_K)$ as f_k conveying more detailed information in approximating f. Our goal is to learn a deep surrogate model $\hat{f}_K$ to approximate f_K by combining data from K fidelities, each with N samples (input parameters) $\{x_{k,n}, y_{k,n}\}_{k=1,n=1}^{K,N}$.

Neural Processes. Combining Gaussian Processes (GPs) and neural networks (NNs), Neural Processes (NPs) (Garnelo et al., 2018b) constitute a family of latent variable models for implicit stochastic processes (Wang & Van Hoof, 2020). NPs represent distributions over functions and estimate prediction uncertainties like GPs while featuring scalability in high dimensions (Jha et al., 2022).

Formally, NP consists of latent variables $z \in \mathbb{R}^{d_z}$ and model parameters θ, trained on context set $\mathcal{D}^c \equiv \{x_n^c, y_n^c\}_{n=1}^N$ and target sets $\mathcal{D}^t \equiv \{x_m^t, y_m^t\}_{m=1}^M$. $\mathcal{D}^c$ and $\mathcal{D}^t$ are randomly split from the training set $\mathcal{D}$. Learning the posterior of z and θ equals to maximize the posterior likelihood below:

$$p(y_{1:M}^t | x_{1:M}^t, \mathcal{D}^c, \theta) = \int p(z|\mathcal{D}^c, \theta) \prod_{m=1}^{M} p(y_m^t | z, x_m^t, \theta) dz \tag{1}$$

Due to the intractability of marginalizing over latent variables z, the NP family utilize approximate inference, deriving the approximated evidence lower bound (ELBO):

$$\log p(y_{1:M}^t | x_{1:M}^t, \mathcal{D}^c, \theta) \gtrapprox \mathbb{E}_{q_\phi(z|\mathcal{D}^c \cup \mathcal{D}^t)} \Big[\sum_{m=1}^{M} \log p(y_m^t | z, x_m^t, \theta) + \log \frac{q_\phi(z|\mathcal{D}^c)}{q_\phi(z|\mathcal{D}^c \cup \mathcal{D}^t)} \Big] \tag{2}$$

This variational approach approximates the intractable true posterior $p(z|\mathcal{D}^c, \theta)$ with the approximate posterior $q_\phi(z|\mathcal{D}^c)$. Here, ϕ parameterize the encoder and θ parameterize the decoder of the model. Implementation-wise, each pair $\{x_n^c, y_n^c\} \in \mathcal{D}^c$ is first encoded as a latent representation r_n^c, forming a latent representation set $\{r_n^c\}_{n=1}^N$, then aggregated to parameterise the latent variable z. For simplicity, we denote $\{x_n\}_{n=1}^N$ as X and $\{y_n\}_{n=1}^N$ as Y.

3. Methodology

The high-level goal of MFRNP is to explicitly model the residual between the aggregated output from the lower fidelities and the ground truth in the highest fidelity. It optimizes the aggregation to simultaneously promote in-fidelity learning and accurate cross-fidelity information sharing.

For each fidelity $k \in \{1...K-1\}$, MFRNP learns a NP surrogate $\hat{f}_k$ to approximate f_k, producing an output $\hat{y}_k$. For the highest fidelity K, MFRNP learns a NP to approximate the residual function $R(x) = f_K(x) - [\frac{1}{K-1} \sum_{k=1}^{K-1} \hat{f}_k(x)]$. The lower fidelity predictions are aggregated by linearly interpolating to the resolution at the highest fidelity and taking the average. We choose to average over other aggregation methods because this stabilizes the aggregation and ensures the equal contribution of lower fidelity surrogates. We explore other aggregation techniques in section 5.4. Unlike previous work where all the information from a fidelity is shared via the latent variable during encoding, MFRNP shares information via the decoded outputs. This explicitly involves the decoder at each fidelity to improve information-sharing. The input-specific information-sharing

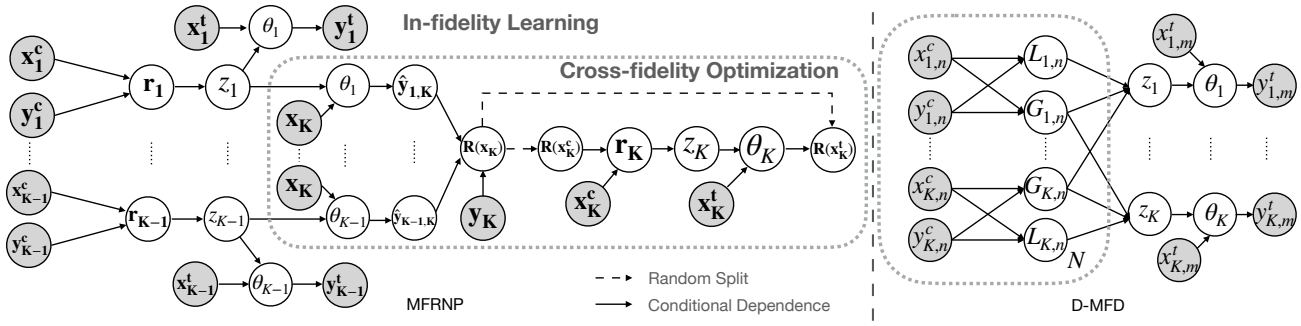

Figure 1. Graphical model comparison of MFRNP against the state-of-the-art D-MFD baseline model. Shaded circles denote observed variables. Right: D-MFD disentangles the latent representations $r_{k,n}$ into local and global representations $L_{k,n}$ and $G_{k,n}$ to infer z_k. Each $G_{k,n}$ is from a different fidelity level, while no information about the decoder parameters θ_k is shared, making $G_{k,n}$ inaccurate regarding θ_i where $i \neq k$. Left: MFRNP dynamically constructs $\mathcal{D}_K^{train} = \{X_K, R(X_K)\}$ in cross-fidelity optimization step and learn fidelity-specific information with in-fidelity learning step. The residual function R makes z_K dependant on $z_{1:K-1}$ and $\theta_{1:K-1}$ without inflating any dimensions. Thus, MFRNP optimizes lower fidelity decoders for better information sharing. We use bold letters in our model graph to denote a set of variables.

scheme further tackles the error propagation issue and ensures the accuracy and full exploitation of lower-fidelity information.

Approximating the Residual Function. The core idea of MFRNP is to make the highest fidelity latent variable z_K dependent on all other latent variables as well as the decoder parameters at lower fidelities. We introduce the dependency by approximating $R(x)$ at the highest fidelity, which models the residual between ground truth and the proposed aggregation from lower fidelities given input x. Intuitively, $\hat{f}_{k\in\{1,\dots,K-1\}}$ is optimized with two objectives. As shown in Figure 1, the in-fidelity learning step captures fidelity-specific information while the cross-fidelity optimization step encourages lower fidelity surrogates to propose accurate and informative aggregations w.r.t the highest fidelity given input x_K.

During inference, given target input x^t, MFRNP approximates $f_K(x^t)$ by aggregating predictions from lower fidelities $(1...K-1)$ and adding it with the residual term from fidelity K : $\hat{f}_K(x^t) = R(x^t) + [\frac{1}{K-1}\sum_{k=1}^{K-1}\hat{f}_k(x^t)]$. MFRNP generates $\hat{y}_K^t$ by fully exploiting the rich information in $\{\hat{y}_{1\cdots K-1}\}$. In the OOD setup where $\mathcal{D}_K^{train}$ and $\mathcal{D}_K^{test}$ covers different input domains, MFRNP can effectively explore the regions beyond the $\mathcal{D}_K^{train}$, thus enhancing its input-domain extrapolation capabilities in modeling f_K.

Residual-ELBO. We design a Residual-ELBO (R-ELBO) for MFRNP. For each fidelity $k \in \{1 \cdots K-1\}$, we infer the latent variable z_k with the NN encoder $q_{\phi_k}(z_k|\mathcal{D}^c)$ and decoder $p_{\theta_k}(y_k^t|z_k, x_k^t)$. At fidelity K, $\mathcal{D}_K$ is dynamically constructed with $\{X_K, R(X_K)\}$. Here X_K denotes a set of input variables x_K. We infer the latent variable z_K with the NN encoder $q_{\phi_K}(z_K|z_{1..K-1}, \theta_{1..K-1}, \mathcal{D}_K^c)$ and decoder $p_{\theta_k}(R(X_K)|z_K, X_K, \theta_K)$, where $R(X_K)$ depends on the proposed aggregations of lower fidelity predictions via ancestral sampling (Wang & Van Hoof, 2020). We derive the R-ELBO for $K > 1$ fidelities in two terms:

$$
\begin{aligned}
&\log p(R(x_K^t)|x_K^t, \mathcal{D}_{1:K}^c \cup \mathcal{D}_{1:K-1}^t, \theta_K) \\
\geq &\mathbb{E}_{q_{\phi_K}(z_K|\mathcal{D}_{1:K}^c \cup \mathcal{D}_{1:K}^t)} \\
&[\log p(R(x_K^t)|z_K, x_K^t, \theta_K) \\
&+ \log \frac{q_{\phi_K}(z_K|\mathcal{D}_{1:K}^c \cup \mathcal{D}_{1:K-1}^t)}{q_{\phi_K}(z_K|\mathcal{D}_{1:K}^c \cup \mathcal{D}_{1:K}^t)}] \\
= &\mathbb{E}_{q_{\phi_K}(z_K|z_{1:K-1}, \theta_{1:K-1}, \mathcal{D}_K^c \cup \mathcal{D}_K^t))} \\
&[\log p(R(x_K^t)|z_K, x_K^t, \theta_K) \\
&+ \log \frac{q_{\phi_K}(z_K|z_{1:K-1}, \theta_{1:K-1}, \mathcal{D}_K^c)}{q_{\phi_K}(z_K|z_{1:K-1}, \theta_{1:K-1}, \mathcal{D}_K^c \cup \mathcal{D}_K^t))}]
\end{aligned} \tag{3}
$$

$$
\begin{aligned}
&\log p(y_{1:K-1}^t|x_{1:K-1}^t, \mathcal{D}_{1:K-1}^c, \theta_{1:K-1}) \\
\geq &\mathbb{E}_{q_\phi(z_{1:K-1}|\mathcal{D}_{1:K-1}^c \cup \mathcal{D}_{1:K-1}^t)} \\
&[\log p(y_{1:K-1}^t|z_{1:K-1}, x_{1:K-1}^t, \theta_{1:K-1}) \\
&+ .. + \log \frac{q_\phi(z_{1:K-1}|\mathcal{D}_{1:K-1}^c)}{q_\phi(z_{1:K-1}|\mathcal{D}_{1:K-1}^c \cup \mathcal{D}_{1:K-1}^t)}] \\
= &\mathbb{E}_{q_{\phi_1}(z_1|\mathcal{D}_1^c \cup \mathcal{D}_1^t)..q_{\phi_{K-1}}(z_{K-1}|\mathcal{D}_{K-1}^c \cup \mathcal{D}_{K-1}^t)} \\
&[\log p(y_1^t|z_1, x_1^t, \theta_1) \\
&+ .. + \log p(y_{K-1}^t|z_{K-1}, x_{K-1}^t, \theta_{K-1}) \\
&+ \log \frac{q_{\phi_1}(z_1|\mathcal{D}_1^c)}{q_{\phi_1}(z_1|\mathcal{D}_1^c \cup \mathcal{D}_1^t)} \\
&+ .. + \frac{q_{\phi_{K-1}}(z_{K-1}|\mathcal{D}_{K-1}^c)}{q_{\phi_{K-1}}(z_{K-1}|\mathcal{D}_{K-1}^c \cup \mathcal{D}_{K-1}^t)}]
\end{aligned} \tag{4}
$$

Equation 4 is a unified ELBO accounting for learning from fidelity specific datasets at lower fidelities. Equation 3 is the ELBO for the residual function at the highest fidelity. This term introduces dependency to every fidelity and optimizes the output aggregations to better approximate $R(x)$.

For training, we calculate the R-ELBO with Monte Carlo (MC) sampling and ancestral sampling (AS) to optimize the objective function below:

$$\begin{aligned}\mathcal{L}_{MC}^{\hat{f}} &= \sum_{k=1}^{K-1}\Big[\tfrac{1}{S}\sum_{s=1}^{S}\log p(y_k^t|x_k^t, z_k^{(s)}) \\ &\quad -\mathrm{KL}[q(z_k|\mathcal{D}_k^c, \mathcal{D}_k^t)\|p(z_k|\mathcal{D}_k^c)]\Big] \qquad (5)\end{aligned}$$

$$\begin{aligned}\mathcal{L}_{MC}^{R} &= \tfrac{1}{S}\sum_{s=1}^{S}\log p(R(x_K^t)|x_K^t, z_K^{(s)}) \\ &\quad -\mathrm{KL}[q(z_K|\mathcal{D}_K^c, \mathcal{D}_K^t)\|p(z_K|\mathcal{D}_K^c)] \qquad (6)\end{aligned}$$

$$\mathcal{L}_{MC} = \mathcal{L}_{MC}^{R} + \mathcal{L}_{MC}^{\hat{f}} \qquad (7)$$

where the time for $q_\phi(z|\mathcal{D}^c)$ to sample $z_k^{(s)}$ scales linearly with the number of fidelity levels.

Training & Inference. As shown in Algorithm 1, MFRNP calculates $\mathcal{L}_{MC}^{\hat{f}}$ by predicting the set $\hat{Y}_k^t$ given input target set X_k^t for every fidelity $k \in \{1..K-1\}$. For $\mathcal{L}_{MC}^{R}$, the AS steps are highlighted in orange in Algorithm 1.

MFRNP introduce decoder to the information aggregation step by predicting with the highest fidelity level input set X_K at lower fidelity levels. We note the prediction as set $\{\hat{o}_{k,K}\}$ where $k \in \{1..K-1\}$, representing surrogate predictions given X_K at fidelity k. We linearly interpolate each $\{\hat{o}_{k,K}\}$ to math the dimension of Y_K. Then we dynamically construct the training dataset at fidelity K as $\mathcal{D}'_K = \{X_K, R(X_K)\}$. Finally, we randomly split $\mathcal{D}'_K$ into $\mathcal{D}'^c_K, \mathcal{D}'^t_K$ to obtain $\mathcal{L}_{MC}^{R}$ and perform back propagation with $\mathcal{L}_{MC}$. Here, $\mathcal{L}_{MC}^{R}$ encourages MFRNP to optimize lower fidelity surrogates for residual modeling at the highest fidelity, while $\mathcal{L}_{MC}^{\hat{f}}$ regulates lower fidelity levels to learn from fidelity-specific datasets.

The inference process is demonstrated in Algorithm 2 where the AS steps are highlighted in orange. Given input set $\{x_{i,K}\}$, MFRNP propose aggregated predictions $\{a_{i,K}\}$ from lower fidelities and the predicted residual $\hat{R}_i$ at fidelity K. The final prediction is given as $\{\hat{y}_{i,K}\} = \{a_{i,K}\} + \{\hat{R}_i\}$.

Algorithm 1 MFRNP Training Process

Input: Dataset $\mathcal{D}_{1\cdots K}$, number of fidelities K.
for $k = 1$ **to** $K-1$ **do**
  Randomly split $\mathcal{D}_k$ into $\{\mathcal{D}_k^c, \mathcal{D}_k^t\}$
  Sample $\{z_{i,k}\}$ where $z_{i,k} \sim q_k(.|\mathcal{D}_k^c)$
  Predict $\{\hat{y}_{i,k}^t\}$ where $\hat{y}_{i,k}^t \sim p_k(.|z_{i,k}, x_{i,k}^t)$
  Sample $\{z'_{i,k}\}$ where $z'_{i,k} \sim q_k(.|\mathcal{D}_k^c)$
  Predict $\{\hat{o}_{i,k}\}$ where $\hat{o}_{i,k} \sim p_k(.|z'_{i,k}, x_{i,K})$
  Linearly interpolate $\{\hat{o}_{i,k}\}$ to the resolution at K.
end for
Get the residual set $\{R_i\}$ at K where $R_i = y_{i,K} - \frac{\sum(\hat{o}_{i,1},..,\hat{o}_{i,K-1})}{K-1}$
Random-split $\mathcal{D}'_K = \{x_{i,K}, R_i\}$ into $\{\mathcal{D}'^c_K, \mathcal{D}'^t_K\}$
Sample $\{z_{i,K}\}$ where $z_{i,K} \sim q_K(.|\mathcal{D}_K^c)$
Predict $\{\hat{R}_i^t\}$ where $\hat{R}_i^t \sim p_K(.|z_{i,K}^t, x_{i,K}^t)$
Back propagate with $\mathcal{L}_{MC} = \mathcal{L}_{MC}^{R} + \mathcal{L}_{MC}^{\hat{f}}$ in Eqn 7

Algorithm 2 MFRNP Inference Process

Input: Latent variables $z_{1..K}$, input $\{x_{i,K}\}$
for $k = 1$ **to** $K-1$ **do**
  Sample $\{z'_{i,k}\}$ from z_k.
  Predict $\{\hat{o}_{i,k}\}$ where $\hat{o}_{i,k} \sim p_k(.|z'_{i,k}, x_{i,K})$
  Linearly interpolate $\{\hat{o}_{i,k}\}$ to the resolution at K.
end for
Obtain aggregation $\{a_{i,K}\} = \frac{\sum(\hat{o}_{i,1},..,\hat{o}_{i,K-1})}{K-1}$
Sample $\{z'_{i,K}\}$ from z_K.
Predict $\{\hat{R}_i\}$ where $\hat{R}_i \sim p_K(.|z_{i,K}, x_{i,K})$
Return $\{\hat{y}_{i,K}\} = \{a_{i,K}\} + \{\hat{R}_i\}$

4. Related Work

Multi-fidelity Modeling. Multi-fidelity surrogate modeling is prevalent across scientific and engineering domains, including applications in aerospace systems (Brevault et al., 2020) and climate science (Hosking, 2020; Valero et al., 2021). The foundational work of Kennedy & O'Hagan (2000) employs GPs to connect models of varying fidelity levels, introducing an autoregressive model. Le Gratiet & Garnier (2014) introduces a recursive GP with a nested structure in the input domain to facilitate rapid inference. Perdikaris et al. (2015; 2016) addresses high-dimensional GP scenarios by leveraging the Fourier transformation of the kernel function. Perdikaris et al. (2017) puts forth the concept of multi-fidelity Gaussian processes (NARGP), assuming a nested structure in the input domain for sequential training at each fidelity level.

Wang et al. (2021) presents a Multi-Fidelity High-Order GP model for accelerating physical simulations. They extend the Linear Model of Coregionalization (LMC) to the nonlinear case, incorporating a matrix GP prior on the weight functions. Deep Gaussian processes (DGPs) (Cutajar et al., 2019) formulate a unified objective to optimize kernel parameters jointly across fidelity levels. DGPs face scalability challenges with high-dimensional data. Infinite-Fidelity

Coregionalization (IFC) from (Li et al., 2022c) models the output space as a continuous function of fidelity and input based on neural ODEs, allowing the model to extrapolate to higher fidelities. However, IFC faces scalability issue due to the computation-demanding ODE solver.

Multi-fidelity modeling has been integrated with deep learning. Guo et al. (2022) employs deep neural networks to merge parameter-dependent output quantities. Wang et al. (2023) utilizes diffusion model to fuse fidelity information into the diffusion-denoising process to solve PDE problems with a conditional scoring model. Meng & Karniadakis (2020) proposes a composite neural network for multi-fidelity data in inverse PDE problems, while Meng et al. (2021) introduces Bayesian neural nets for multi-fidelity modeling. De et al. (2020) employs transfer learning to fine-tune high-fidelity surrogate models using deep neural networks trained on low-fidelity data. Despite advancements, existing deep GP models (Cutajar et al., 2019; Hebbal et al., 2021) struggle with cases where different fidelities involve data with varying dimensions. Additionally, multi-fidelity methods have been investigated in Bayesian optimization, active learning, and bandit problems (Li et al., 2020; Song et al., 2019; Li et al., 2022a; Perry et al., 2019; Kandasamy et al., 2017).

Neural Processes. Neural Processes (NPs) (Garnelo et al., 2018a; Kim et al., 2018; Louizos et al., 2019; Singh et al., 2019) emerge as scalable and expressive alternatives to GPs for modeling stochastic processes. Previous work by Raissi & Karniadakis (2016) combines multi-fidelity GPs with deep learning, placing a GP prior on features learned by deep neural networks. Wang & Lin (2020) proposes a multi-fidelity neural process with physics constraints (MFPC-Net), leveraging NPs to capture correlations between multi-fidelity data. Nonetheless, this model requires paired data and cannot utilize unpaired data at the low-fidelity level.

The recent work of Wu et al. (2023) proposes to disentangle the latent variable at each fidelity into local and global representations and share the global part across fidelities. However, the fidelity-specific decoder parameters are not included in information sharing. Thus, the highest fidelity decoder expresses shared representations differently, hindering the inference performance, especially in OOD scenarios.

Climate Modeling. Climate modeling is a central component of modern climate science and the primary tool for predicting future climate states (Flato et al., 2014). Various modeling centers around the world have developed distinct climate models. The semi-independent development process has led to many plausible, but disagreeing, climate models representing the same earth system (Knutti et al., 2010; Flato et al., 2014). Averaging these models often leads to improved results compared to using individual models (Lambert & Boer, 2001; Gleckler et al., 2008; Knutti et al., 2010). Proper aggregation of different climate models for a consensus estimate is therefore an important topic (Tebaldi & Knutti, 2007). Averaging all models with equal or varying weights has been the most common approach (Giorgi & Mearns, 2002; Abramowitz et al., 2019), known as (weighted) ensemble averaging (EA). However, EA techniques often do not retain much spatial information and can cause severe blurring, corrupting regional signals.

To address this problem, various alternative approaches have been developed based on Bayesian hierarchical models, regression and machine learning that all use observational data to improve model aggregation. DeepESD (Baño-Medina et al., 2022) utilize CNNs to learn the mapping from the EA of 8 regional climate models to the observation-calibrated data (Dee et al., 2011) at finer resolution. Similarly, NNGPR (Harris et al., 2023) utilizes deep kernel GPs to model the residual between the EA of 16 low-resolution global climate models and the up-to-date observation-calibrated ERA5-reanalysis dataset (Hersbach et al., 2020). A lot of progress have been made in downscaling the climate simulators with observational data (Kotamarthi et al., 2021), yet lacking the method to accurately infer the long-term climate scenarios directly from climate drivers.

5. Experiments

5.1. Datasets

We include 6 Partial Differential Equation benchmarks, a more complicated fluid simulation task and a real-world climate modeling task for earth surface temperature prediction.

Partial Differential Equations (PDEs). We include Heat and Poisson's equations (Olsen-Kettle, 2011) from computational physics. We test `MFRNP` for predicting the spatial solution fields of these equations. We use numerical solvers to generate the ground-truth data with dense and coarse meshes for different fidelity levels. For both Heat and Poisson's equations with 2 fidelity setting, we use 16×16 and 32×32 meshes as low and high fidelities. For three fidelity scenarios, we run the solvers with 64×64 meshes at the highest fidelity level. For five fidelity scenarios, we run the solvers with 96×96 and 128×128 meshes as the two additional fidelity levels. The heat equation has an input dimension of 3, corresponding to the thermal diffusivity coefficient and boundary conditions at the two edges. The Poisson's equation has an input dimension of 5, corresponding to the 4 boundary conditions and the magnitude of flow at the centered point source.

Fluid Simulation. This task simulates the dynamics of a circular smoke cloud propelled by an inflow force within a 50×50 grid. The spatial-temporal solution field is derived

Task (Full)	DMF	NARGP	MFHNP	D-MFD	SF-NP	MFRNP
Heat 2	0.138 ± 4.0e-8	0.31±2.12e-6	0.026±4.01e-5	0.015±1.42e-5	0.308±6.38e-05	**0.005±3.27e-4**
Heat 3	0.137±1.23e-7	0.309±3.46e-6	0.111±4.82e-6	0.108±4.85e-8	0.307±8.62e-05	**0.0039±2.94e-4**
Heat 5	0.135±2.55e-4	0.306±1.14e-4	0.115±1.22e-2	0.106±1e-4	0.306±6.90e-05	**0.0045±2.94e-4**
Poisson 2	0.107 ± 6.58e-5	0.585±9.84e-5	0.093±2.55e-4	0.07 ±2.99e-4	0.575±1.39e-04	**0.0076±7.49e-4**
Poisson 3	0.121±1.47e-5	0.58 ±1.02e-4	0.335±2.37e-5	0.101±1.81e-4	0.572±1.95e-04	**0.0073±5.25e-4**
Poisson 5	0.101±2.24e-3	0.571±1.29e-4	0.299±8.84e-3	0.279±3.35e-3	0.571±1.82e-04	**0.0046±1.2e-4**
Fluid	0.275±4.59e-7	0.353±9.28e-4	0.234±4.82e-6	0.207±1.31e-5	0.383±5.53e-05	**0.129±8.19e-4**

Table 1. Performance (nRMSE) comparison of 6 different models applied to the Heat and Poisson simulators with two, three, five fidelities and fluid simulation with Navier-Stokes equation with two fidelities. The full setting means same domain coverage of $\mathcal{D}_K^{train}$ and $\mathcal{D}_K^{test}$.

Task (OOD)	DMF	NARGP	MFHNP	D-MFD	SF-NP	MFRNP
Heat 2	0.168±1.36e-4	0.313±1.30e-4	0.033±1.37e-2	0.213±1.65e-3	0.312±1.14e-04	**0.005±1.33e-4**
Heat 3	0.163±6.04e-4	0.309±5.00e-4	0.143±5.88e-3	0.141±4.94e-3	0.310±7.03e-05	**0.004±3.82e-4**
Heat 5	0.187±8.60e-4	0.308±1.05e-4	0.15±2.72e-3	0.145±2.87e-3	0.308±5.25e-05	**0.012±1.05e-2**
Poisson 2	0.183±6.85e-4	0.749±1.47e-5	0.103±2.03e-2	0.214±4.28e-2	0.749±1.79e-03	**0.017±2.72e-3**
Poisson 3	0.186±8.06e-4	0.744±8.24e-5	0.189±1.14e-2	0.2±1e-2	0.745±7.62e-04	**0.018±1.27e-3**
Poisson 5	0.16±4.47e-4	0.743±3.07e-4	0.399±1.07e-2	0.375±8.56e-3	0.744±1.85e-03	**0.013±2.76e-4**

Table 2. Performance (nRMSE) comparison of 6 different models applied to the Heat and Poisson simulators with two, three, five fidelities. The OOD setting here indicates $\mathcal{D}_K^{test}$ is OOD w.r.t the training domain at fidelity K.

through the application of the incompressible Navier-Stokes equations and the Boussinesq approximation (Holl et al., 2020; Chorin, 1968). Initiated with the introduction of an incompressible static smoke cloud of radius 5 at the lower center of the grid, a persistent inflow force is subsequently applied at the center of the cloud. The input parameters encompass two variables governing the magnitude of the inflow force along the x and y directions. The output is the initial component of the velocity field following a temporal evolution of 30 steps with the inflow force. The simulation generates low-fidelity ground-truth with 32×32 mesh, and high-fidelity with 64×64 mesh.

Climate Modeling: Earth Surface Temperature. In this task, we take one step further from previous works in Section 4 with multi-fidelity surrogate modeling to directly learn the mapping from climate drivers (Watson-Parris et al., 2022) to the observation-calibrated ERA5-reanalysis data (Hersbach et al., 2020), together with 13 low-resolution computational climate model predictions from the ScenarioMIP project(O'Neill et al., 2016) at global scale. We provide details about the computational models used in Appendix A.1. We group these climate data into 9 fidelities based on their original resolutions. The fidelity dimensions are: 144×192, 160×320, 192×288, 180×288, 120×180, 143×144, 80×96, 192×384 and 721×1440 at the highest fidelity level, representing air temperature at 2 meters above the earth surface. The climate drivers as inputs here consists of 12 variables representing the total emission of greenhouse gases (CO2, CH4) and aerosol gases (BC, SO2) from year 1850 to 2015. Each of the aerosol gas is split into 5 signals via principal component analysis (Wold et al., 1987; Watson-Parris et al., 2022). For years after 2015, we test 4 hypothetical global gas emission scenarios: ssp126, ssp245, ssp370, ssp585. Larger numbers correspond to more total gas emissions, leading to more severe climate change. Ssp126 represents the condition in which future gas emissions are well controlled and gradually decrease. This scenario makes the gas emission data as input to have similar domain coverage as of training data. Whereas in the other three scenarios, future gas emissions would keep increasing, introducing out-of-distribution inputs at the highest fidelity.

5.2. Experiment Setup

We implement MFRNP with PyTorch (Paszke et al., 2019) and compare the average nRMSE in three random runs with following baselines. Details are provided in Appendix A.3.

- **SF-NP.** The naive single-fidelity neural process trained on the highest fidelity data, as the lower performance bound for multi-fidelity modeling.
- **DMF**(Li et al., 2021) performs multi-fidelity learning based on a multi neural network (NN) structure. Each NN corresponds to one fidelity and the NN on the next fidelity adapts the latent output from the NN at current

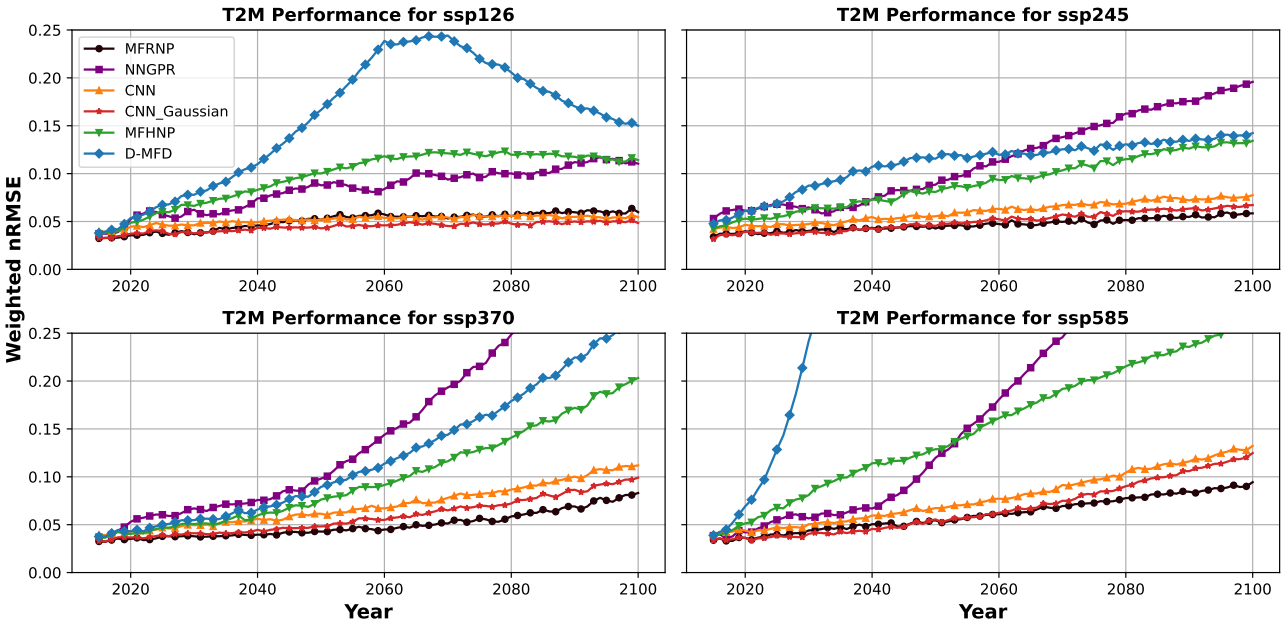

Figure 2. Perfect model test performance for climate modeling task. Measured in latitude-weighted nRMSE. MFRNP outperforms other models in 3 out of the 4 scenarios and maintains consistent performance across all the scenarios from year $2015 \sim 2100$.

fidelity to propagate information.

- **NARGP**(Perdikaris et al., 2017) utilizes multi-fidelity Gaussian Processes by sequentially train Gaussian Processes from low to high fidelity levels under the assumption of a nested input domain.
- **MFHNP**(Wu et al., 2022) learns distribution over functions at each fidelity level with a latent variable. They designed a hierarchical structure of the latent variables from low to high fidelity levels to pass information.
- **D-MFD**(Wu et al., 2023) tackles the error propagation problem in MFHNP and the inter-fidelity NN overfitting problem in DMF by introducing local and global latent representations at each fidelity.

PDEs Setup. For PDE tasks, we consider two data composition scenarios.

- **In-distribution (Full) Scenario.** In this setting, the training set $\mathcal{D}^{train}$ and testing set $\mathcal{D}^{test}$ at each fidelity level have the same coverage on dataset domain $\mathcal{D}$.
- **Out-of-distribution (OOD) Scenario.** This scenario simulates real-world conditions where the highest fidelity data has limited domain coverage. We evaluate model trained using $\mathcal{D}^{train} \subset \mathcal{D}$ at the highest fidelity level, and test the model performance on $\mathcal{D}^{test}$ consists of data covering the rest of $\mathcal{D}$ s.t. $\mathcal{D}^{train} \cap \mathcal{D}^{test} = \emptyset$ and $\mathcal{D}^{train} \cup \mathcal{D}^{test} = \mathcal{D}$.

We construct the OOD scenario with Heat and Poisson's equations. We did not include Fluid data because the low fidelity mesh is not detailed enough to capture the differences between $\mathcal{D}^{train}$ and $\mathcal{D}^{test}$. Under the OOD setting, information covering the input domain of $\mathcal{D}^{test}$ is missing in all the fidelities, making it impossible for models to extrapolate with lower fidelity information. For Heat equation, we test the OOD scenario of the thermal diffusivity coefficient by limiting the corresponding input parameter to cover 80% of original domain. We uniformly sample X in the constrained domain and obtain corresponding Y with numerical solver as $\mathcal{D}^{train}$. $\mathcal{D}^{test}$ is constructed with the rest 20% of original domain. Similarly, for Poisson's equation, we construct $\mathcal{D}^{train}$ and $\mathcal{D}^{test}$ with the $80/20$ coverage split on 3 boundary conditions. Details are demonstrated in Appendix A.2.

Climate Modeling Setup. In this task, we also compare MFRNP with recent works in the climate science community of downscaling global climate projections. We include deep kernel GPs from NNGPR (Harris et al., 2023) and Gaussian CNN from DeepSED (Baño-Medina et al., 2022). Since no previous work simultaneously downscales while modeling the mapping between climate drivers and scenarios, to ensure fair comparison under the same task, we set up a naive extension for these methods. We first fit 13 linear surrogate

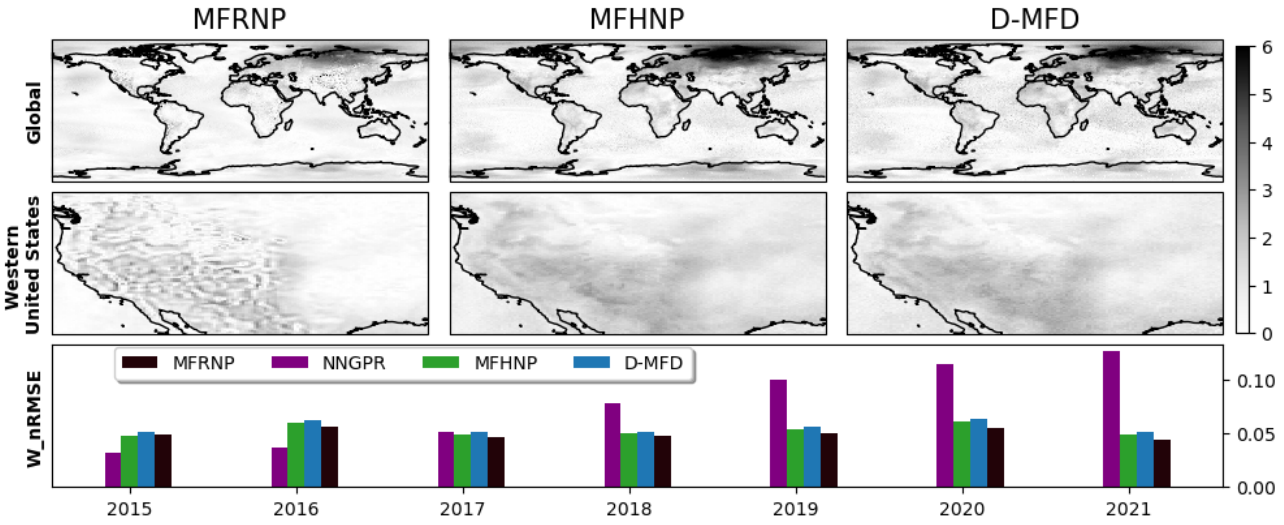

Figure 3. Absolute residual plot (°C) between ground truth and model predictions on ERA5-reanalysis data at year 2020, with global resolution of 721×1440. The first row presents global view, second row focuses on smaller geographic regions. The third row shows the latitude weighted nRMSE to measure prediction accuracy from year 2015 to 2021. MFRNP outperforms other models staring from 2017.

models with the climate drivers as input X, and the computational climate model predictions as target Y. Then we train the models in NNGPR and DeepSED to downscale the averaged predictions from linear surrogates given X.

We use 13 "perfect model" tests (Knutti et al., 2017) as the performance metric. In each test, one climate model is held out as the ground truth and we only use its historical data ($1950 \sim 2015$) for training at the highest fidelity. The remaining 12 climate models are grouped at lower fidelity levels based on their resolutions. We compare the average latitude weighted nRMSE across 13 tests. Thus, we are able to empirically measure the performance with physics based future projections. We then include the ERA5-reanalysis data as the highest fidelity level to measure the performance on a refined resolution. We use year $1950 \sim 2014$ as the highest fidelity training data and $2015 \sim 2021$ for testing.

5.3. Results

PDEs Performance. We test the performance of MFRNP and baselines across 7 tasks under the full setting and 6 tasks under the OOD setting. Table 1 shows the full setting performance and Table 2 shows the OOD setup performance. Among all the tasks and settings, our model consistently outperforms baselines by an average of $\sim 90\%$ in nRMSE.

The results indicate that MFRNP is more efficient in utilizing information across fidelities to generate good predictions. Unlike other baselines for which the performance generally gets worse as the number of fidelities increase, the performance of MFRNP in the 5 fidelity tasks is even better than the corresponding 2 fidelity tasks in 3 out of 4 five-fidelity experiments. This indicates MFRNP gathers lower fidelity information efficiently enough to outrun the increasing task complexity in modeling Heat and Poisson's equations.

Climate Modeling Performance. As shown in Figure 2, for the perfect model test, MFRNP performs consistently across all emission scenarios and outperforms other baselines in 3 out of the 4 scenarios. The perfect model test empirically shows the capability of MFRNP to efficiently aggregate information from lower fidelities and its versatility for OOD scenarios. Although CNN and Gaussian CNN perform relatively well, they have very limited scalability and cannot scale to the global ERA5-reanalysis dataset. We also had memory issues running NARGP and DMF with 9 fidelities for this task. Among the baseline models that scale to the ERA5-reanalysis dataset (MFHNP, D-MFD and NNGPR), the performance of MFRNP is significantly better.

For the ERA5-reanalysis data, MFRNP also outperforms other baseline models, as shown in Figure 3. Our model keeps a consistent performance from year $2015 \sim 2021$, while the performance of other baseline models worsens for further future predictions. MFRNP makes a cleaner residual plot compared to the other models. Although we can only evaluate performance up to year 2021 due to observation limitations, we have empirically shown the long term prediction consistency of MFRNP from the perfect model test and we expect it to further outperform other baselines for years after 2021 with the ERA5-reanalysis data.

MFRNP has shown superior performance across various gas emission scenarios from year 2015 to 2100 via the perfect

model test. Together with the accurate reanalysis predictions, we have demonstrated that MFRNP is not only capable of fully exploiting lower fidelity information for reliable predictions, but also able to provide great scalability with consistent performance for real-world tasks.

5.4. Ablation Study

Latent Aggregation with Residual. We evaluate the impact of including decoders in the information sharing step. The same experiment settings are maintained for consistency in the comparison. We create a variant of MFRNP, namely MFRNP-H, for which the surrogate models for fidelity levels below the highest are implemented with a hierarchical structure that shares information via the latent space (Wu et al., 2022). We use the surrogate output at fidelity $K-1$ as the aggregated information and model the residual between this aggregation and the highest fidelity ground truth.

Task (Full)	MFRNP-H	MFRNP
Heat 3	0.013±4.46e-4	**0.0039±2.94e-4**
Heat 5	0.013±2.76e-4	**0.0045±2.94e-4**
Poisson 3	0.138±6.1e-3	**0.0073±5.25e-4**
Poisson 5	0.184±1.77e-2	**0.0046±1.2e-4**

Table 3. Performance (nRMSE) comparison of MFRNP-H and MFRNP applied to the Heat and Poisson simulators with three, five fidelities under Full setup.

Task (OOD)	MFRNP-H	MFRNP
Heat 3	0.047±2.72e-4	**0.004±3.82e-4**
Heat 5	0.055±1.87e-4	**0.012±1.05e-2**
Poisson 3	0.128±1.3e-2	**0.018±1.27e-3**
Poisson 5	0.151±7.9e-3	**0.013±2.76e-4**

Table 4. Performance (nRMSE) comparison of MFRNP-H and MFRNP applied to the Heat and Poisson simulators with three, five fidelities under OOD setup.

Our results in Table 3 and Table 4 show that MFRNP significantly outperforms MFRNP-H. This indicates including decoders in the information aggregation step promotes cross-fidelity information sharing and yields better learning capabilities for both Full and OOD setup.

Weighted Averaging. We explore the impact of non-uniform averaging in this study on Heat and Poisson dataset with 5 fidelities under Full setting. We tested MFRNP with uniform averaging for simplicity and to avoid introducing additional hyperparameters. Intuitively, higher fidelity levels should be assigned higher weights since they provide more detailed predictions. We set the weighted average scheme to correlate with fidelity level as following:

$$\begin{aligned}\text{Aggregation} = &\, 0.1 \times \hat{Y}_{k=1} + 0.2 \times \hat{Y}_{k=2} \\ &+ 0.3 \times \hat{Y}_{k=3} + 0.4 \times \hat{Y}_{k=4}\end{aligned} \tag{8}$$

Task	Uniform Averaging	Weighted Averaging
Heat5	0.0045±2.94e-4	**0.0042±1.19e-3**
Poisson5	**0.0046±1.2e-4**	0.0103±2.02e-3

Table 5. Comparison of Uniform Averaging and Weighted Averaging on Heat5 and Poisson5 tasks, measured with nRMSE.

As shown in Table 5, for Heat5, using increasing weights grants 6.67% performance improvement. However, for Poisson5, using increasing weights doubles the nRMSE. Fine tuning the fidelity weights could potentially bring performance improvements, but it would require dataset-specific weight adjustments.

6. Discussion & Conclusion

In this paper, we present Multi-fidelity Residual Neural Processes (MFRNP), a novel Neural Process-based multi-fidelity surrogate model. MFRNP utilizes a residual modeling framework, which allows MFRNP to leverage the rich input space coverage from lower fidelities while preserving accuracy from the highest fidelity data. Our tailored Residual-ELBO loss promotes learning across fidelities and simultaneously optimizes the lower fidelity decoders for accurate information sharing. Experimental results on partial differential equations and climate modeling demonstrate that MFRNP outperforming state-of-the-art methods by more than 90%, highlighting scalability, efficiency in information fusion, and versatility of MFRNP for real-world surrogate modeling.

The limitation of MFRNP mainly lies in encoder and decoder complexity. Due to ancestral samplings step, the inference time is doubled. Although this additional time penalty is small compared with the actual simulators, it is a potential concern for computationally expensive encoders and decoders, such as neural ODEs. Moreover, uniform averaging may cause loss of detailed information from lower fidelities. For future work, we plan to address this limitation and incorporate physics-informed structure to better model expensive physics simulations.

Acknowledgments

This work was supported in part by the U.S. Army Research Office under Army-ECASE award W911NF-07-R-0003-03, the U.S. Department Of Energy, Office of Science, IARPA HAYSTAC Program, NSF Grants SCALE MoDL-2134209, CCF-2112665 (TILOS), #2205093, #2146343, #2134274, CDC-RFA-FT-23-0069, as well as DARPA AIE FoundSci.

Impact Statement

The work in this paper aims to advance the performance and scalability of multi-fidelity surrogate modeling. The model proposed has potential impact in climate science, computational chemistry, structural engineering and many other communities, all of which, we foresee potential positive impacts in accelerating physics simulations and scientific discoveries.

References

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R., and Schmidt, G. A. Esd reviews: Model dependence in multi-model climate ensembles: Weighting, sub-selection and out-of-sample testing. *Earth System Dynamics*, 10(1): 91–105, 2019.

Baño-Medina, J., Manzanas, R., Cimadevilla, E., Fernández, J., González-Abad, J., Cofiño, A. S., and Gutiérrez, J. M. Downscaling multi-model climate projection ensembles with deep learning (deepesd): contribution to cordex eur-44. *Geoscientific Model Development*, 15(17):6747–6758, 2022.

Bi, D., Dix, M., Marsland, S., O'farrell, S., Sullivan, A., Bodman, R., Law, R., Harman, I., Srbinovsky, J., Rashid, H. A., et al. Configuration and spin-up of access-cm2, the new generation australian community climate and earth system simulator coupled model. *Journal of Southern Hemisphere Earth Systems Science*, 70(1):225–251, 2020.

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., Bekki, S., Bonnet, R., Bony, S., Bopp, L., et al. Presentation and evaluation of the ipsl-cm6a-lr climate model. *Journal of Advances in Modeling Earth Systems*, 12(7):e2019MS002010, 2020.

Brevault, L., Balesdent, M., and Hebbal, A. Overview of gaussian process based multi-fidelity techniques with variable relationship between fidelities, application to aerospace systems. *Aerospace Science and Technology*, 107:106339, 2020.

Byun, Y.-H., Lim, Y.-J., Shim, S., Sung, H. M., Sun, M., Kim, J., Kim, B.-H., Lee, J.-H., and Moon, H. Nims-kma kace1. 0-g model output prepared for cmip6 scenariomip. 2019.

Cherchi, A., Fogli, P. G., Lovato, T., Peano, D., Iovino, D., Gualdi, S., Masina, S., Scoccimarro, E., Materia, S., Bellucci, A., et al. Global mean climate and main patterns of variability in the cmcc-cm2 coupled model. *Journal of Advances in Modeling Earth Systems*, 11(1):185–209, 2019.

Chorin, A. J. Numerical solution of the navier-stokes equations. *Mathematics of computation*, 22(104):745–762, 1968.

Cutajar, K., Pullin, M., Damianou, A., Lawrence, N., and González, J. Deep gaussian processes for multi-fidelity modeling. *arXiv preprint arXiv:1903.07320*, 2019.

Damianou, A. and Lawrence, N. D. Deep gaussian processes. In *Artificial intelligence and statistics*, pp. 207–215. PMLR, 2013.

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D., DuVivier, A., Edwards, J., Emmons, L., Fasullo, J., Garcia, R., Gettelman, A., et al. The community earth system model version 2 (cesm2). *Journal of Advances in Modeling Earth Systems*, 12(2):e2019MS001916, 2020.

De, S., Britton, J., Reynolds, M., Skinner, R., Jansen, K., and Doostan, A. On transfer learning of neural networks using bi-fidelity data for uncertainty propagation. *International Journal for Uncertainty Quantification*, 10(6), 2020.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, d. P., et al. The era-interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the royal meteorological society*, 137(656):553–597, 2011.

Dunne, J. P., Horowitz, L., Adcroft, A., Ginoux, P., Held, I., John, J., Krasting, J. P., Malyshev, S., Naik, V., Paulot, F., et al. The gfdl earth system model version 4.1 (gfdl-esm 4.1): Overall coupled model description and simulation characteristics. *Journal of Advances in Modeling Earth Systems*, 12(11):e2019MS002015, 2020.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., et al. Evaluation of climate models. In *Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, pp. 741–866. Cambridge University Press, 2014.

Garnelo, M., Rosenbaum, D., Maddison, C., Ramalho, T., Saxton, D., Shanahan, M., Teh, Y. W., Rezende, D., and Eslami, S. A. Conditional neural processes. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2018a.

Garnelo, M., Schwarz, J., Rosenbaum, D., Viola, F., Rezende, D. J., Eslami, S., and Teh, Y. W. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018b.

Giorgi, F. and Mearns, L. O. Calculation of average, uncertainty range, and reliability of regional climate changes from aogcm simulations via the "reliability ensemble averaging"(rea) method. *Journal of climate*, 15(10):1141–1158, 2002.

Gleckler, P. J., Taylor, K. E., and Doutriaux, C. Performance metrics for climate models. *Journal of Geophysical Research: Atmospheres*, 113(D6), 2008.

Guo, M., Manzoni, A., Amendt, M., Conti, P., and Hesthaven, J. S. Multi-fidelity regression using artificial neural networks: efficient approximation of parameter-dependent output quantities. *Computer methods in applied mechanics and engineering*, 389:114378, 2022.

Gutjahr, O., Putrasahan, D., Lohmann, K., Jungclaus, J. H., von Storch, J.-S., Brüggemann, N., Haak, H., and Stössel, A. Max planck institute earth system model (mpi-esm1. 2) for the high-resolution model intercomparison project (highresmip). *Geoscientific Model Development*, 12(7): 3241–3281, 2019.

Harris, T., Li, B., and Sriver, R. Multimodel ensemble analysis with neural network gaussian processes. *The Annals of Applied Statistics*, 17(4):3403–3425, 2023.

Hebbal, A., Brevault, L., Balesdent, M., Talbi, E.-G., and Melab, N. Multi-fidelity modeling with different input domain definitions using deep gaussian processes. *Structural and Multidisciplinary Optimization*, 63(5):2267–2288, 2021.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730): 1999–2049, 2020.

Holl, P., Koltun, V., and Thuerey, N. Learning to control pdes with differentiable physics. *arXiv preprint arXiv:2001.07457*, 2020.

Hosking, S. Multifidelity climate modelling, github. https://github.com/scotthosking/mf_modelling, 2020.

Jha, S., Gong, D., Wang, X., Turner, R. E., and Yao, L. The neural process family: Survey, applications and perspectives. *arXiv preprint arXiv:2209.00517*, 2022.

Kandasamy, K., Dasarathy, G., Schneider, J., and Póczos, B. Multi-fidelity bayesian optimisation with continuous approximations. In *International Conference on Machine Learning*, pp. 1799–1808. PMLR, 2017.

Karami, Z. and Kashef, R. Smart transportation planning: Data, models, and algorithms. *Transportation Engineering*, 2:100013, 2020.

Kennedy, M. C. and O'Hagan, A. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000.

Kim, H., Mnih, A., Schwarz, J., Garnelo, M., Eslami, A., Rosenbaum, D., Vinyals, O., and Teh, Y. W. Attentive neural processes. In *International Conference on Learning Representations*, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A. Challenges in combining projections from multiple climate models. *Journal of Climate*, 23(10):2739–2758, 2010.

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., and Eyring, V. A climate model projection weighting scheme accounting for performance and interdependence. *Geophysical Research Letters*, 44(4): 1909–1918, 2017.

Kotamarthi, R., Hayhoe, K., Wuebbles, D., Mearns, L. O., Jacobs, J., and Jurado, J. *Downscaling techniques for high-resolution climate projections: From global change to local impacts*. Cambridge University Press, 2021.

Lambert, S. J. and Boer, G. J. Cmip1 evaluation and intercomparison of coupled climate models. *Climate Dynamics*, 17:83–106, 2001.

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

Li, S., Xing, W., Kirby, R., and Zhe, S. Multi-fidelity bayesian optimization via deep neural networks. *Advances in Neural Information Processing Systems*, 33: 8521–8531, 2020.

Li, S., Wang, Z., Kirby, R. M., and Zhe, S. Deep multi-fidelity active learning of high-dimensional outputs, 2021.

Li, S., Phillips, J., Yu, X., Kirby, R., and Zhe, S. Batch multi-fidelity active learning with budget constraints. In *Advances in Neural Information Processing Systems*, 2022a.

Li, S., Wang, Z., Kirby, R., and Zhe, S. Deep multi-fidelity active learning of high-dimensional outputs. In *International Conference on Artificial Intelligence and Statistics*, pp. 1694–1711. PMLR, 2022b.

Li, S., Wang, Z., Kirby, R., and Zhe, S. Infinite-fidelity coregionalization for physical simulation. *Advances in Neural Information Processing Systems*, 35:25965–25978, 2022c.

Louizos, C., Shi, X., Schutte, K., and Welling, M. The functional neural process. *Advances in Neural Information Processing Systems*, 2019.

Meng, X. and Karniadakis, G. E. A composite neural network that learns from multi-fidelity data: Application to function approximation and inverse pde problems. *Journal of Computational Physics*, 401:109020, 2020.

Meng, X., Babaee, H., and Karniadakis, G. E. Multi-fidelity bayesian neural networks: Algorithms and applications. *Journal of Computational Physics*, 438:110361, 2021.

Olsen-Kettle, L. Numerical solution of partial differential equations. *Lecture notes at University of Queensland, Australia*, 2011.

O'Neill, B. C., Tebaldi, C., Van Vuuren, D. P., Eyring, V., Friedlingstein, P., Hurtt, G., Knutti, R., Kriegler, E., Lamarque, J.-F., Lowe, J., et al. The scenario model intercomparison project (scenariomip) for cmip6. *Geoscientific Model Development*, 9(9):3461–3482, 2016.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Peherstorfer, B., Willcox, K., and Gunzburger, M. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *Siam Review*, 60(3):550–591, 2018.

Perdikaris, P., Venturi, D., Royset, J. O., and Karniadakis, G. E. 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):20150018, 2015.

Perdikaris, P., Venturi, D., and Karniadakis, G. E. Multifidelity information fusion algorithms for high-dimensional systems and massive data sets. *SIAM Journal on Scientific Computing*, 38(4):B521–B538, 2016.

Perdikaris, P., Raissi, M., Damianou, A., Lawrence, N. D., and Karniadakis, G. E. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2198):20160751, 2017.

Perry, D. J., Kirby, R. M., Narayan, A., and Whitaker, R. T. Allocation strategies for high fidelity models in the multifidelity regime. *SIAM/ASA Journal on Uncertainty Quantification*, 7(1):203–231, 2019.

Raissi, M. and Karniadakis, G. Deep multi-fidelity gaussian processes. *arXiv preprint arXiv:1604.07484*, 2016.

Rasmussen, C. E. Gaussian processes in machine learning. In *Summer school on machine learning*, pp. 63–71. Springer, 2003.

Salimbeni, H. and Deisenroth, M. P. Doubly stochastic variational inference for deep gaussian processes. In *NIPS*, 2017.

Seeger, M. Gaussian processes for machine learning. *International journal of neural systems*, 14(02):69–106, 2004.

Seland, Ø., Bentsen, M., Olivié, D., Toniazzo, T., Gjermundsen, A., Graff, L. S., Debernard, J. B., Gupta, A. K., He, Y.-C., Kirkevåg, A., et al. Overview of the norwegian earth system model (noresm2) and key climate response of cmip6 deck, historical, and scenario simulations. *Geoscientific Model Development*, 13(12):6165–6200, 2020.

Sellar, A. A., Jones, C. G., Mulcahy, J. P., Tang, Y., Yool, A., Wiltshire, A., O'connor, F. M., Stringer, M., Hill, R., Palmieri, J., et al. Ukesm1: Description and evaluation of the uk earth system model. *Journal of Advances in Modeling Earth Systems*, 11(12):4513–4558, 2019.

Singh, G., Yoon, J., Son, Y., and Ahn, S. Sequential neural processes. *Advances in Neural Information Processing Systems*, 32:10254–10264, 2019.

Song, J., Chen, Y., and Yue, Y. A general framework for multi-fidelity bayesian optimization with gaussian processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3158–3167. PMLR, 2019.

Stouffer, R. Ipcc ddc: Ua mcm-ua-1-0 model output prepared for cmip6 scenariomip. 2019.

Tebaldi, C. and Knutti, R. The use of the multi-model ensemble in probabilistic climate projections. *Philosophical transactions of the royal society A: mathematical, physical and engineering sciences*, 365(1857):2053–2075, 2007.

Valero, M. M., Jofre, L., and Torres, R. Multifidelity prediction in wildfire spread simulation: Modeling, uncertainty quantification and sensitivity analysis. *Environmental Modelling & Software*, 141:105050, 2021.

Volodin, E. Possible climate change in russia in the 21st century based on the inm-cm5-0 climate model. *Russian Meteorology and Hydrology*, 47(5):327–333, 2022.

Wang, Q. and Van Hoof, H. Doubly stochastic variational inference for neural processes with hierarchical latent variables. In *International Conference on Machine Learning*, pp. 10018–10028. PMLR, 2020.

Wang, Y. and Lin, G. Mfpc-net: Multi-fidelity physics-constrained neural process. *arXiv preprint arXiv:2010.01378*, 2020.

Wang, Z., Xing, W., Kirby, R., and Zhe, S. Multi-fidelity high-order gaussian processes for physical simulation. In *International Conference on Artificial Intelligence and Statistics*, pp. 847–855. PMLR, 2021.

Wang, Z., Li, S., Fang, S., and Zhe, S. Diffusion-generative multi-fidelity learning for physical simulation. *arXiv preprint arXiv:2311.05606*, 2023.

Watson-Parris, D., Rao, Y., Olivié, D., Seland, Ø., Nowack, P., Camps-Valls, G., Stier, P., Bouabid, S., Dewey, M., Fons, E., et al. Climatebench v1. 0: A benchmark for data-driven climate projections. *Journal of Advances in Modeling Earth Systems*, 14(10):e2021MS002954, 2022.

Williams, C. and Rasmussen, C. Gaussian processes for regression. *Advances in neural information processing systems*, 8, 1995.

Wilson, A. G., Hu, Z., Salakhutdinov, R., and Xing, E. P. Deep kernel learning. In *Artificial intelligence and statistics*, pp. 370–378. PMLR, 2016.

Wold, S., Esbensen, K., and Geladi, P. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

Wu, D., Chinazzi, M., Vespignani, A., Ma, Y.-A., and Yu, R. Multi-fidelity hierarchical neural processes. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2029–2038, 2022.

Wu, D., Niu, R., Chinazzi, M., Ma, Y., and Yu, R. Disentangled multi-fidelity deep bayesian active learning. *arXiv preprint arXiv:2305.04392*, 2023.

Wu, T., Yu, R., Lu, Y., Jie, W., Fang, Y., Zhang, J., Zhang, L., Xin, X., Li, L., Wang, Z., et al. Bcc-csm2-hr: a high-resolution version of the beijing climate center climate system model. *Geoscientific Model Development*, 14(5): 2977–3006, 2021.

Yukimoto, S., Kawai, H., Koshiro, T., Oshima, N., Yoshida, K., Urakawa, S., Tsujino, H., Deushi, M., Tanaka, T., Hosaka, M., et al. The meteorological research institute earth system model version 2.0, mri-esm2. 0: Description and basic evaluation of the physical component. *Journal of the Meteorological Society of Japan. Ser. II*, 97(5): 931–965, 2019.

A. Appendix

A.1. Climate Modeling Task Dataset

We list the 13 computational climate models used here, together with the literatures. They are all included in the ScenarioMIP project (O'Neill et al., 2016). ACCESS-CM2(Bi et al., 2020); BCC-CSM2-MR(Wu et al., 2021); CMCC-CM2-SR5(Cherchi et al., 2019); GFDL-ESM4(Dunne et al., 2020); INM-CM5-0(Volodin, 2022); IPSL-CM6A-LR(Boucher et al., 2020); KACE-1-0-G(Byun et al., 2019); MCM-UA-1-0(Stouffer, 2019); MRI-ESM2-0(Yukimoto et al., 2019); CESM2(Danabasoglu et al., 2020); MPI-ESM(Gutjahr et al., 2019); NorESM2(Seland et al., 2020); UKESM1(Sellar et al., 2019).

A.2. OOD Dataset Construction

Generally, we follow the experiment setup as the full scenario. The following describes the difference in building the highest fidelity dataset $\mathcal{D}_K$. For Heat equation, we uniformly sample X_{train} in the training scope: ((0,0.8),(-1,0),(0.01,0.1)) and use numerical solver to generate the corresponding Y_{train}. We then uniformly sample X_{test} in the testing scope ((0.8, 1),(-1,0),(0.01,0.1)) and use numerical solver to generate Y_{test}.

For the Poisson's equation, we uniformly sample X_{train} in the training scope: ((0.1, 0.74),(0.1, 0.74),(0.1, 0.74),(0.1, 0.9),(0.1, 0.9)) and use numerical solver to generate the corresponding Y_{train}. We then uniformly sample X_{test} in the testing scope ((0.74, 0.9),(0.74, 0.9),(0.74, 0.9),(0.1, 0.9),(0.1, 0.9)) and use numerical solver to generate Y_{test}.

A.3. Experimental Setup

Metrics. We use latitude weighted nRMSE as shown in Equation 9 to measure model performance in the climate modeling task. For PDE tasks, we use nRMSE as shown in Equation 10.

$$W_nRMSE = \frac{\sqrt{\sum_{i=1}^{N}\left(\frac{\cos(\text{lat}_i)}{\frac{\sum_{j=1}^{N}\cos(\text{lat}_j)}{N}}\right) * \left(\frac{(y_i-\hat{y}_i)^2}{N}\right)}}{std(\{y_{1:N}\})} \tag{9}$$

$$nRMSE = \frac{\sqrt{\sum_{i=1}^{N}\left(\frac{(y_i-\hat{y}_i)^2}{N}\right)}}{std(\{y_{1:N}\})} \tag{10}$$

PDE task Training Configurations. For training, we use Adam optimizer (Kingma & Ba, 2014) with base learning rate of $1e-3$. We use 10% of training data set as validation set. For Heat and Poisson's equation, we run our model with latent dimension and encoder/decoder dimension of 32, and run our model with maximum epoch of 50000 and patience 10000. We use learning rate decay of 0.85 and stepsize 10000. We set the highest fidelity weight to 2 and lower fidelities to 1 in loss calculation to focus more on optimizing toward the highest fidelity. For Poisson5, using lower fidelity weight of 0.25 further improves performance. For context-target split, we randomly select $20\% \sim 25\%$ of training data as our context set, the rest as our target set for each fidelity. For fluid simulation, we follow the same setup but use latent dimension and encoder/decoder dimension of 128. We set patience to 5000 with learning rate decay of 0.01. We normalize the data before training and measure nRMSE on the de-normalized space. For the baseline models, we follow the same setup as above. All models are trained on NVIDIA A100 GPU with 80GB memory.

Climate Modeling Training Configurations. We follow the similar setup as above, but set the highest fidelity weight to 5, latent and hidden dimensions to 512 to incorporate more fidelities and data at higher dimensions. We do not normalize the data before training. For the climate methods introduced from climate science community (NNGPR, CNN, CNN_Gaussian), we use their original model setup. The linear surrogates for low-resolution climate simulators are built with 4 layers with ReLU activation. Dimensions are 12, $\lfloor\frac{(lat*lon)}{128}\rfloor$, $\lfloor\frac{(lat*lon)}{32}\rfloor$, $\lfloor\frac{(lat*lon)}{4}\rfloor$. Here, (lat, lon) refers to the climate simulator data dimensions. These models are run until convergence.