

Global Convolutional Neural Processes

Xuesong Wang
University of New South Wales
xuesong.wang1@unsw.edu.au

Lina Yao
University of New South Wales
lina.yao@unsw.edu.au

Xianzhi Wang
University of Technology Sydney
xianzhi.wang@uts.edu.au

Hye-young Paik
University of New South Wales
h.paik@unsw.edu.au

Sen Wang
University of Queensland
sen.wang@uq.edu.au

Abstract—The ability to deal with uncertainty in machine learning models has become equally, if not more, crucial to their predictive ability itself. For instance, during the pandemic, governmental policies and personal decisions are constantly made around uncertainties. Targeting this, Neural Process Families (NPFs) have recently shone a light on prediction with uncertainties by bridging Gaussian processes and neural networks. Latent neural process, a member of NPF, is believed to be capable of modelling the uncertainty on certain points (local uncertainty) as well as the general function priors (global uncertainties). Nonetheless, some critical questions remain unresolved, such as a formal definition of global uncertainties, the causality behind global uncertainties, and the manipulation of global uncertainties for generative models. Regarding this, we build a member GloBal Convolutional Neural Process (GBCoNP) that achieves the SOTA log-likelihood in latent NPFs. It designs a global uncertainty representation $p(z)$, which is an aggregation on a discretized input space. The causal effect between the degree of global uncertainty and the intra-task diversity is discussed. The learnt prior is analyzed on a variety of scenarios, including 1D, 2D, and a newly proposed spatial-temporal COVID dataset. Our manipulation of the global uncertainty not only achieves generating the desired samples to tackle few-shot learning, but also enables the probability evaluation on the functional priors.

Index Terms—uncertainty modelling, neural processes, deep learning, few-shot learning, covid trend forecast

I. INTRODUCTION

In recent years, machine learning, especially deep learning, has shown massive success on a range of prediction tasks, such as time-series forecasting [1], geographical and spatial-temporal inference in medical science, engineering, and finance domains [2]. Nonetheless, the uncertainty of machine learning models is less considered than model predictions themselves. When the existing knowledge of a task is not abundant to present a deterministic prediction, uncertainty provides a reasonable guess interval that includes major possibilities of the predictions. In fact, uncertainty can be as equally important as the prediction capability of models, considering that it increases as the expansion of underlying factors. For example, when estimating the life expectancy of a mechanical part, it is more feasible to estimate an approximate time-range than to predict an exact time stamp at which to discard the part as there are more uncontrollable variants in the long term. Besides, modelling the uncertainty helps increase models’ tolerance to task diversities in datasets. For instance,

when exploring substances from satellite images of a planet, modelling the uncertainty helps incorporate heterogeneous substance candidates that can not be confidently distinguished from the ground truth. Uncertainty is crucially important for predictions related to the pandemic—it determines how much we could trust the classification results of COVID-19 cases [3], the anticipated virus spreading trends [4], and assessment of lockdown policies [5], all of which have a vital impact on the daily lives of many people.

Neural process families (NPF) recently shone a light on predictions with uncertainties by bridging Gaussian processes (GPs) and neural networks. Inherited from Gaussian processes, they make predictions under a joint of correlated normal distributions and present each prediction along with a confidence interval, i.e., uncertainty (See Fig 1). As neural networks show competence in deep feature representation, NPFs advance GPs in modelling complicated functions efficiently. In addition, NPFs are suitable for solving meta-learning tasks, where each task is sampled from a distribution of functions instead of a single function. We illustrate predictions with uncertainties using the examples shown in Fig 1. Given a **context set** with a cluster of observable data $(\mathbf{x}_C, \mathbf{y}_C) := (\mathbf{x}_i, \mathbf{y}_i)_{i \in C}$, an encoder network infers a function f that can generate the context set. Then, a decoder network uses the function to make predictions with uncertainties on a **target set** $(\mathbf{x}_T, \mathbf{y}_T) := (\mathbf{x}_i, \mathbf{y}_i)_{i \in T}$, where only the locations \mathbf{x}_T are unveiled. The resulting outputs follow a normal distribution $p(\mathbf{y}_T | \mathbf{x}_T, f) = \mathcal{N}(\mu_y, \sigma_y^2)$. We call the standard deviation on local target points σ_{y_T} “local uncertainty” in contrast to the global uncertainty to be introduced in latent neural processes later. Eventually, NPF optimizes the parameters in neural nets by maximizing the likelihood of actual target values \mathbf{y}_T .

Latent neural processes [6] hypothesize that the encoded function should not come from a deterministic vector but rather a distribution $f \sim p(f)$. As shown in Fig 1, two descent function samples f_1 and f_2 with periodic differences can be generated using the same context set. They both represent the local uncertainty on the target points but obviously have different priors, meaning there exists another uncertainty that determines this general prior, which we call “global uncertainty”. Despite some previous efforts [7] [6] [8] empirically illustrate such global uncertainties there remains

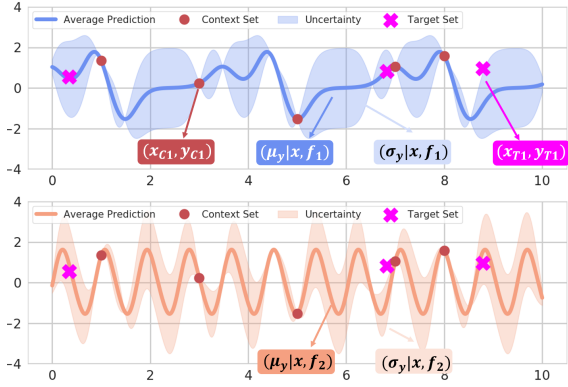


Fig. 1. Predictions with uncertainties under two function samples f_1 and f_2

three unresolved issues:

- Formalization of global uncertainties. We wonder if the global uncertainty can be quantified and compared in ways other than empirical visualization of diverse samples.
- Causality behind global uncertainties. This concerns the model-wise and data-wise factors that affect the global uncertainty.
- Manipulation of the global uncertainty for data generation. This concerns tailoring the priors and sampling a desired function once the causal effects of the global uncertainty are examined.

We propose a GloBal Convolutional Neural Processes (GB-CoNP) that make predictions with uncertainties to address the above challenges. GB-CoNP defines global uncertainty as the learnt posterior of a functional distribution conditioned on a small context set $q(z|C)$. This formalization enables comparisons among datasets and different latent models, including the causal-effect between the intra-task diversity and global uncertainty. It further enables us to discover and edit insightful semantic features with regards to global uncertainty during sample generations. Finally, we evaluate the log-likelihood of GB-CoNP with peers on extensive 1D and 2D datasets, and propose a COVID dataset using spatial-temporal uncertainty prediction. This case study is expected to enhance our understanding of the patterns in virus spread and benefit the research community. Our major contributions are three folds:

- A new discretized space for global uncertainty projection that is suitable for out-of-range prediction while maintaining the shared global prior.
- A causal-effect analysis of the global uncertainty, which is seldom discussed in previous research. Our analysis reveals dataset characteristics, such as intra-task diversity, can depict the stochasticity.
- Manipulation of the global uncertainty that empowers sampling with priors. We novelly generalize the applications to a high-dimensional spatial-temporal scenario.

II. METHODOLOGY

A. Global Uncertainty in Latent Neural Processes

We follow the notations of [7] and denote a context set by $(\mathbf{x}_C, \mathbf{y}_C) := (\mathbf{x}_i, \mathbf{y}_i)_{i \in C}$, where both the inputs and predictions are given in a meta-regression task. The context set defines a sample from a functional distribution. The objective is to predict on the target inputs \mathbf{x}_T using this function sample and maximize the likelihood of \mathbf{y}_T if the target set $(\mathbf{x}_T, \mathbf{y}_T) := (\mathbf{x}_i, \mathbf{y}_i)_{i \in T}$ comes from the same function sample. Similar to conditional variational autoencoders (CVAE), a latent NP models the prediction with a conditional distribution (1):

$$\begin{aligned} p(\mathbf{y}_T | \mathbf{x}_T, \mathbf{x}_C, \mathbf{y}_C) &:= \mathcal{N}(\mu_y, \sigma_y^2) \\ &= \int p(\mathbf{y}_T | \mathbf{x}_T, \mathbf{r}_C, \mathbf{z}) p(\mathbf{z}) d\mathbf{z} \end{aligned} \quad (1)$$

where $\mathbf{r}_C := r(\mathbf{x}_C, \mathbf{y}_C)$ is a neural network that captures the deterministic part of the functional sample. The prior latent distribution $p(\mathbf{z}) := \mathcal{N}(\mu_z, \sigma_z^2)$ captures the global uncertainty of the functional sample. As shown in Fig 2(a), \mathbf{r}_C and \mathbf{z} forms the encoder. The decoder network takes the encoded functional condition $(\mathbf{r}_C, \mathbf{z})$ and the target inputs \mathbf{x}_T to make predictions of \mathbf{y}_T . Depending on the inductive biases imposed over the relationships within the context set and between context and target sets, different neural network structures can be adopted for encoders, such as equally weighted (neural processes [6]) and attentively weighted aggregation (attentive neural processes [7]).

Since this prior knowledge of the functional distribution, i.e., global uncertainty $p(\mathbf{z})$ is intractable, previous studies turn to amortised variational posterior $p(\mathbf{z}) = \mathcal{N}(\mu_z, \sigma_z^2) \approx q(\mathbf{z} | \mathbf{r}_C) = q(\mathbf{z} | \mathbf{x}_C, \mathbf{y}_C)$ for inference. They typically pass \mathbf{r}_C

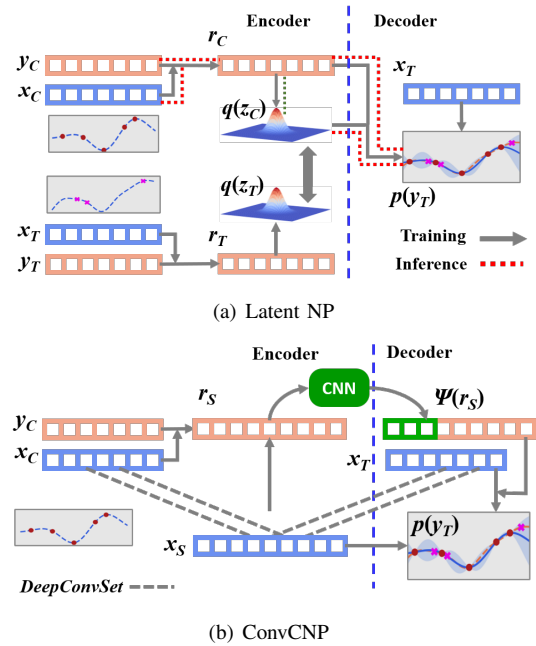


Fig. 2. Workflows for latent NP families and ConvCoNP

through an MLP to get the distributional parameters and then optimize (1) with a differentiable network:

$$p(\mathbf{y}_T|\mathbf{x}_T, \mathbf{x}_C, \mathbf{y}_C) := \int p(\mathbf{y}_T|\mathbf{x}_T, \mathbf{r}_C, \mathbf{z})q(\mathbf{z}|\mathbf{x}_C, \mathbf{y}_C) dz \quad (2)$$

A predictive evidence-lower-bound (ELBO) is given based on the variational inference of $p(\mathbf{z})$:

$$\log p(\mathbf{y}_T|\mathbf{x}_T, \mathbf{x}_C, \mathbf{y}_C) \geq \mathbb{E}_{q(\mathbf{z}|\mathbf{x}_T, \mathbf{y}_T)} \log p(\mathbf{y}_T|\mathbf{x}_T, \mathbf{r}_C, \mathbf{z}) - \mathcal{KL}(q(\mathbf{z}|\mathbf{x}_T, \mathbf{y}_T)||q(\mathbf{z}|\mathbf{x}_C, \mathbf{y}_C)) \quad (3)$$

Now, the training process is to maximize the log-likelihood $\log p(\mathbf{y}_T|\mathbf{x}_T, \mathbf{x}_C, \mathbf{y}_C)$ which equals maximizing its lower bound comprised of a conditional log-likelihood based on latent \mathbf{z} , i.e., $\log p(\mathbf{y}_T|\mathbf{x}_T, \mathbf{r}_C, \mathbf{z})$, minus a non-negative KL divergence. While CVAEs regularize the posterior with a standard normal distribution $\mathcal{KL}(q(\mathbf{z})||\mathcal{N}(0, I))$, latent NPs differ in minimizing the divergence from the posterior (obtained from the target set) to the context set during training when the target sets are accessible.

During inference, the prior $p(\mathbf{z}) = \mathcal{N}(\mu_z, \sigma_z^2)$ is replaced by $q(\mathbf{z}|\mathbf{x}_C, \mathbf{y}_C)$. We formalize the global uncertainty as:

$$p(\mathbf{z}) := \mathcal{N}(\mu_z, \sigma_z^2) = q(\mathbf{z}|\mathcal{C}), \text{ s.t. } |\mathcal{C}| < \epsilon \quad (4)$$

where ϵ comprises a very small proportion of the index set. We did not set the condition $\mathcal{C} = \emptyset$ since in real-world datasets and the empty set barely carries any function prior, but $|\mathcal{C}|$ cannot be too large, otherwise it would leave no space for uncertainty. The mean μ_z in (4) determines the sensitivity of final predictions affected by the global uncertainty, whereas the variance σ_z^2 implies the diversity of function priors on a certain task.

B. Out-of-range Predictions with Convolution

Out-of-range predictions is an essential characteristic when scaling the NPF to real-world tasks. It requires the model to generalize predictions when the testing task is out of the training range. The recently proposed Convolutional Conditional Neural Processes(ConvCNP) [9] tackle this issue with convolutions and outperform peer NPF members. ConvCNPs assume “translational invariance”, i.e., the prediction pattern near a local context data is transferable to the rest of the input space, which can be well addressed by convolution. They omit the latent representation and only decode \mathbf{r} and \mathbf{x}_T . However, instead of directly encoding \mathbf{r} from the raw context inputs \mathbf{x}_C , they introduce a discretization of the indefinite input space \mathbf{x}_S that incorporates context and target inputs $\mathbf{x}_C, \mathbf{x}_T$ and map the deterministic representation on \mathbf{x}_S :

$$p(\mathbf{y}_T|\mathbf{x}_T, \mathbf{x}_C, \mathbf{y}_C) := \mathcal{N}(\mu_y, \sigma_y^2) = p(\mathbf{y}_T|\mathbf{x}_T, \psi(\mathbf{r}_S)) \quad (5)$$

where $\mathbf{r}_S := r(\mathbf{x}_C, \mathbf{x}_S, \mathbf{y}_C)$ is the deterministic representation. $r(\mathbf{x}_C, \mathbf{x}_S, \mathbf{y}_C)$ is a *DeepConvSet* that resembles a simplified attention. Given a query of a discretized set \mathbf{x}_S , *DeepConvSet* projects the value \mathbf{y}_C to the space \mathcal{S} based on the similarity between the query \mathbf{x}_S and the key \mathbf{x}_C . Then, \mathbf{r}_S is passed through a convolutional neural network $\psi(\cdot)$, which achieves the inductive bias of translational invariance (shown in Fig

2(b)). This guarantees that the function description only focuses on a local range of inputs so that the model is still effective even when the testing inputs are out of the training range. The *DeepConvSet* : $\mathcal{C} \mapsto \mathcal{S}$ and the CNN module constitute the encoder. The decoder is another *DeepConvSet* whose keys and values are \mathbf{x}_S and \mathbf{r}_S with queries \mathbf{x}_T . Without the latent representation, the training and inference share the same workflow.

C. GloBal Convolutional Neural Process(GBCoNP)

While ConvCNPs outperform NPF members on many scenarios, we believe the latent distribution $p(\mathbf{z})$ has a great impact on maintaining the stochasticity for NPs, particularly on the global uncertainty. If the function sample representation \mathbf{r}_S is deterministic, the resulting prediction will be precise yet “dull” with single distribution parameters μ_y and σ_y^2 ; In contrast, latent NPs can sample different \mathbf{z} values which correspond to diverse priors over the functions. Each prior is able to generate a cluster of (μ_y, σ_y^2) . ConvCNP can potentially be tailored to a latent NP, given that the mapping function to the space \mathcal{S} can be latent and shared between the context and target set. Therefore, we introduce a member named GloBal Convolutional Neural Process (GBCoNP) that also adopts amortised variational inference on the global uncertainty (shown in Fig 3).

The predictions of GBCoNP follow a conditional distribution with a latent path:

$$\begin{aligned} p(\mathbf{y}_T|\mathbf{x}_T, \mathbf{x}_C, \mathbf{y}_C) &:= \mathcal{N}(\mu_y, \sigma_y^2) \\ &= \int p(\mathbf{y}_T|\mathbf{x}_T, \psi(\mathbf{r}_S, \mathbf{z}))p(\mathbf{z}) dz \end{aligned} \quad (6)$$

where $\psi(\cdot)$ is the convolution module applied on both the deterministic and the latent representation. $\mathbf{r}_S := r(\mathbf{x}_C, \mathbf{x}_S, \mathbf{y}_C)$ is a *DeepConvSet*. Given the convolved functional condition $\psi(\mathbf{r}_S, \mathbf{z})$ and the target inputs \mathbf{x}_T , a decoder *DeepConvSet* maps the condition to the prediction distribution. We use a new space \mathbf{x}_S instead of the original space \mathbf{x}_C to obtain the

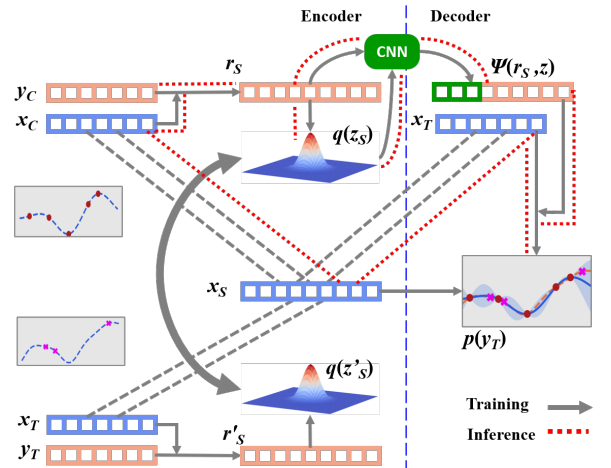


Fig. 3. The model structure for GBCoNP

variational inference of the intractable prior $p(\mathbf{z})$. Thus, (6) can be optimized with another differentiable network ($\psi(\cdot)$ omitted for better clarification):

$$p(\mathbf{y}_{\mathcal{T}}|\mathbf{x}_{\mathcal{T}}, \mathbf{x}_{\mathcal{C}}, \mathbf{y}_{\mathcal{C}}) := \int p(\mathbf{y}_{\mathcal{T}}|\mathbf{x}_{\mathcal{T}}, \mathbf{r}_{\mathcal{S}}, \mathbf{z})q(\mathbf{z}|\mathbf{r}_{\mathcal{S}}) dz \quad (7)$$

According to [6], the joint distribution of $p(\mathbf{y}_{\mathcal{T}})$ must suffice exchangeability to be a stochastic process:

$$\begin{aligned} & p(\mathbf{y}_{\mathcal{T}_1}, \mathbf{y}_{\mathcal{T}}, \dots, \mathbf{y}_{\mathcal{T}_n}|\mathbf{x}_{\mathcal{T}_1}, \dots, \mathbf{x}_{\mathcal{T}_n}, \mathbf{x}_{\mathcal{C}_1}, \dots, \mathbf{x}_{\mathcal{C}_m}, \mathbf{y}_{\mathcal{C}_1}, \dots, \mathbf{y}_{\mathcal{C}_m}) \\ &= p(\mathbf{y}_{\mathcal{T}_{\pi(1)}}, \dots, \mathbf{y}_{\mathcal{T}_{\pi(n)}}|\mathbf{x}_{\mathcal{T}_{\pi(1)}}, \dots, \mathbf{x}_{\mathcal{T}_{\pi(n)}}, \\ & \quad \mathbf{x}_{\mathcal{C}_{\pi'(1)}}, \dots, \mathbf{x}_{\mathcal{C}_{\pi'(m)}}, \mathbf{y}_{\mathcal{C}_{\pi'(1)}}, \dots, \mathbf{y}_{\mathcal{C}_{\pi'(m)}}) \end{aligned} \quad (8)$$

where π and π' are permutations of the target index set $\{1, \dots, n\}$ and the context index set $\{1, \dots, m\}$, meaning the permutation of the context and the target points cannot change the prediction outcome. The previous latent NPs need an **aggregation module** (mean, sum, etc.) on context and target space to ensure $q(\mathbf{z}|\mathbf{x}_{\mathcal{T}}, \mathbf{y}_{\mathcal{T}})$ and $q(\mathbf{z}|\mathbf{x}_{\mathcal{C}}, \mathbf{y}_{\mathcal{C}})$ are permutation invariant and have the identical dimensions for divergence.

Similar to attention, the *DeepConvSet* already suffices the exchangeability by calculating the inner product between the key ($\mathbf{x}_{\mathcal{C}}$ or $\mathbf{x}_{\mathcal{T}}$) and query $\mathbf{x}_{\mathcal{S}}$. This operation is insensitive to the input order and thus the context and target are projected to the same space \mathcal{S} regardless of the order. Presumably, $\mathbf{r}_{\mathcal{S}}$ can be directly passed through an MLP for \mathbf{z} without aggregation. However, we discovered that the aggregation on \mathbf{z} can mitigate the coherence deficiency in the samples, a drawback caused by the independent prediction assumption in NPFs [8]. It maybe attributed to the diminishing effect on the divergence between $q(\mathbf{z}|\mathbf{r}_{\mathcal{S}}) := q(\mathbf{z}|\mathbf{x}_{\mathcal{S}}, \mathbf{x}_{\mathcal{C}}, \mathbf{y}_{\mathcal{C}})$ and $q(\mathbf{z}|\mathbf{r}'_{\mathcal{S}}) := q(\mathbf{z}|\mathbf{x}_{\mathcal{S}}, \mathbf{x}_{\mathcal{T}}, \mathbf{y}_{\mathcal{T}})$ after the fusion.

An observation that supports the aggregation is the non-translational invariance of the pre-aggregated \mathbf{z} . Normally, $\psi(\cdot)$ in (6) comprises several cascaded convolutional nets; when we directly convolve on $\mathbf{r}_{\mathcal{S}}$ concatenated with the pre-aggregated \mathbf{z} , the local pattern near a context point (e.g., a fluctuation) is propagated to the entire space \mathcal{S} . Besides, similar patterns are obtained when \mathbf{z} is derived from an intermediate convolution state of $\psi(\mathbf{r}_{\mathcal{S}})$ and get aggregated afterwards, meaning $\mathbf{r}_{\mathcal{S}}$ is convolved by a subset of the cascaded nets. Both cases above imply that local patterns on \mathbf{z} cannot be as transferable to the whole space as $\mathbf{r}_{\mathcal{S}}$ can. Therefore, we add an aggregation module to \mathbf{z} to ensure every point in the space \mathcal{S} gets the same latent prior.

Adopting the variation inference, the objective function for GBCoNP now becomes (9):

Prop 1. Evidence-Lower-BOund for GBCoNP:

$$\log p(\mathbf{y}_{\mathcal{T}}|\mathbf{x}_{\mathcal{T}}, \mathbf{x}_{\mathcal{C}}, \mathbf{y}_{\mathcal{C}}) \geq \mathbb{E}_{q(\mathbf{z}|\mathbf{r}'_{\mathcal{S}})} \log p(\mathbf{y}_{\mathcal{T}}|\mathbf{x}_{\mathcal{T}}, \mathbf{r}_{\mathcal{S}}, \mathbf{z}) - \mathcal{KL}(q(\mathbf{z}|\mathbf{r}'_{\mathcal{S}})||q(\mathbf{z}|\mathbf{r}_{\mathcal{S}})) \quad (9)$$

Proof. The conditional probability of the prediction can be built on the marginalization of a joint distribution with a latent

variable \mathbf{z} .

$$\begin{aligned} \log p(\mathbf{y}_{\mathcal{T}}|\mathbf{x}_{\mathcal{T}}, \mathbf{x}_{\mathcal{C}}, \mathbf{y}_{\mathcal{C}}) &= \log \int p(\mathbf{y}_{\mathcal{T}}, \mathbf{z}|\mathbf{x}_{\mathcal{T}}, \mathbf{r}_{\mathcal{S}}) dz \\ &= \log \int p(\mathbf{y}_{\mathcal{T}}|\mathbf{x}_{\mathcal{T}}, \mathbf{r}_{\mathcal{S}}, \mathbf{z})p(\mathbf{z}) dz \\ &= \log \int \frac{p(\mathbf{y}_{\mathcal{T}}|\mathbf{x}_{\mathcal{T}}, \mathbf{r}_{\mathcal{S}}, \mathbf{z})p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x}_{\mathcal{T}}, \mathbf{y}_{\mathcal{T}}, \mathbf{x}_{\mathcal{S}})} q(\mathbf{z}|\mathbf{x}_{\mathcal{T}}, \mathbf{y}_{\mathcal{T}}, \mathbf{x}_{\mathcal{S}}) dz \\ &= \log \mathbb{E}_{q(\mathbf{z}|\mathbf{r}'_{\mathcal{S}})} \frac{p(\mathbf{y}_{\mathcal{T}}|\mathbf{x}_{\mathcal{T}}, \mathbf{r}_{\mathcal{S}}, \mathbf{z})p(\mathbf{z})}{q(\mathbf{z}|\mathbf{r}'_{\mathcal{S}})} \end{aligned} \quad (10)$$

Since the prior \mathbf{z} is intractable, it is replaced by the conditional posterior $q(\mathbf{z}|\mathbf{r}_{\mathcal{S}})$ on the space \mathcal{S} . Using Jensens Inequality and the concave log function, we get the lower bound of (10):

$$\begin{aligned} & \geq \mathbb{E}_{q(\mathbf{z}|\mathbf{r}'_{\mathcal{S}})} \log p(\mathbf{y}_{\mathcal{T}}|\mathbf{x}_{\mathcal{T}}, \mathbf{r}_{\mathcal{S}}, \mathbf{z}) + \mathbb{E}_{q(\mathbf{z}|\mathbf{r}'_{\mathcal{S}})} \log \frac{q(\mathbf{z}|\mathbf{r}_{\mathcal{S}})}{q(\mathbf{z}|\mathbf{r}'_{\mathcal{S}})} \\ &= \mathbb{E}_{q(\mathbf{z}|\mathbf{r}'_{\mathcal{S}})} \log p(\mathbf{y}_{\mathcal{T}}|\mathbf{x}_{\mathcal{T}}, \mathbf{r}_{\mathcal{S}}, \mathbf{z}) - \mathcal{KL}(q(\mathbf{z}|\mathbf{r}'_{\mathcal{S}})||q(\mathbf{z}|\mathbf{r}_{\mathcal{S}})) \end{aligned} \quad (11)$$

Computational complexity. An attention layer takes $\mathcal{O}(nmd)$, where n and m are the key and query lengths, and d implies the weight dimension [10]. A convolutional layer costs $\mathcal{O}(nkdf)$, where k and f refer to the kernel size and channel depth. Therefore, two *DeepConvSets* in encoder and decoder cost $\mathcal{O}(msd)$ and $\mathcal{O}(nsd)$, where m and n refer to the context and target set sizes, and s refers to the grid length of the discretization ($s \gg m, n$). The computationally expensive part $\psi(\cdot)$ costs $\mathcal{O}(skdf)$ for convolutions on the \mathcal{S} space. Compared with ConvCNP, the latent path in GBCoNP brought in two MLP modules: the latent MLP from $\mathbf{r}_{\mathcal{S}}$ to \mathbf{z} and the merger MLP (which compress the concatenation of $(\mathbf{r}_{\mathcal{S}}, \mathbf{z})$ to a low dimension before convolution). Both modules cost $\mathcal{O}(s)$. Such modifications preserve the total number of parameters in convolution and are thus affordable.

III. EXPERIMENTS

We evaluate our proposed model on three groups of datasets covering 1D, 2D and spatial temporal scenarios across a broad range of domains. All the models are built with Python 3.6.8 and Pytorch 1.4.0. Two TITAN RTX GPUs are used for training. Wall-clock time for training one epoch of GBCoNP and ConvCNP is shown in Appendix B.

A. 1D Datasets

In 1D scenarios, each task deals with a series of context points. The objective is to predict the values and uncertainty for the unseen target set (The datasets are detailed in Appendix A). For each task, x values are normalized to $[-1, 1]$ and y values are standard normalized before training. The number of context points range within $\mathcal{U}(1, 50)$, and the target set comprises all the data points in the task. Each dataset is trained for 100 epochs and tested for 6 runs.

Table I shows the log-likelihood of the latent neural process families along with their conditional members on 1D datasets. The poor results of a neural network (NN) that abandons the context set $y_{\mathcal{T}} = f(x_{\mathcal{T}})$ reflect that non-NP models are probably unsuitable for the meta-setting. ConvCNP and ANP achieve the best baseline log-likelihood in their category; their

predictions are compared further with the proposed GBCoNP in Fig 4.

Almost all the ground truths lie in the predicted uncertainties for the three methods (except for the periodic kernel), validating that NP families are capable of modeling uncertainty. The highest uncertainty is normally achieved at the furthest point away from all the context points. The latent NP members generate more diverse samples when compared with their counterpart conditional members at the expense of slight performance drop. This may be attributed to the inference of the intractable latent variables z using context data—the resulting distributional gap $\mathcal{KL}(q(z|r'_S)||q(z|r_S))$ causes information loss in exchange for global uncertainty. GBCoNP and ConvCNP produce smoother mean values whereas ANP achieves more diverse samples on real-world datasets. The most fluctuated predictive means tend to occur in the middle of two context points (see RBF, Matern, and HousePricing in Fig 4), where the target points are sensitive to both context points. By virtue of their convolution filters, GBCoNP and ConvCNP are able to mitigate those fluctuations.

With regard to global uncertainty **causal-effect analysis**, we present the values of μ_z and σ_z in (4) after training (Table II). Considering that z is a high dimensional distribution and each task batch presents a pair of (μ_z, σ_z) , we therefore averaged the results on both the dimension and batch levels. Table II indicates that the global uncertainty depends largely on model representation capacity and the intra-task diversity. Model representation capacity refers to constraints a model impose on the global uncertainty. For instance, NP constrains equally loose for all the points, ANP only constrains stricter to the points closer to the context while GBCoNP constrain almost all the points differently with the convolution on S . As the constraints grow (NP < ANP < GBCoNP), the global uncertainty decreases (σ_z : NP > ANP > GBCoNP). Intra-task diversity implies the meta-setting characteristic of the dataset – how many possibilities there are in this dataset given the same context. For example, although Periodic and SmartMeter are both seasonal datasets, the possible target sets, given a certain context set for Periodic, are not unique; therefore, GBCoNP can predict seasonal means with different amplitudes. In contrast, in SmartMeter, there is only one possible target set corresponding to a context set; causing a smaller σ_z in this case (< Periodic).

To **manipulate global uncertainty** for increasing intra-task diversity, we reduce the latent dimension of z from 128 to 4 and display μ_{y_T} with different priors in Fig 5. The variance bound in the prior(4) is relaxed from σ_z to $40 \sigma_z$ for Stock50 and Periodic to amplify the effects. The results show that a group of different functions can be sampled meanwhile fitting well with context data. As highlighted in Fig 5 (a), some dimension controls the trend of the curve after a context point (up/down) while some others control the amplitude of the trend. Fig 5 (b) shows the controlling factors become the amplitude and the phase of a wave, and the resulting positions of context data in the prediction curves shift from a “crest” to a “trough”.

B. 2D Datasets

MNIST, SVHN, CelebA32. In 2D scenarios, we aim to inpaint the whole image with uncertainties given a set of context pixels, and we select three image datasets (MNIST, SVHN, CelebA32) for the task. For convolutional-based methods, the context values (x) are built using a binary mask $\in \mathbb{R}^{W \times H}$, where $\mathcal{U}(0, 0.3(W \times H))$ pixels are unveiled. For non-convolutional baselines, the mask is transformed into a list of 2d inputs representing the relative pixel locations (pixel index/image size). Y values are either 1d or 3d implying the pixel intensity. Each datasets is trained for 50 epochs and tested for 6 runs. Table III shows the log-likelihood of the latent and the conditional neural process families on the 2D datasets. GBCoNP achieves the SOTA performance among the latent members. ConvCNP performs the best conditional result on MNIST, and ACNP outperforms on SVHN and CelebA32.

Fig 6 shows the predictive means and variances on the three datasets. Overall, all the three methods give reasonable results, even with only 5% of the total pixels. The local uncertainty occurs on the strokes of a digit for MNIST and SVHN, while in CelebA, the variances lie in the profile of a face, including face shape, hair, eyes, noise and mouth. The variance on SVHN and CelebA plummets as the amount of context data increases. With 30% of unveiled pixels, all the models can recover the whole image with little local uncertainty. GBCoNP and ConvCNP can generate smoother predictive means (see Fig 6 MNIST digits “2” and “5”, and the plate numbers “42” and “25” in SVHN). ANP achieves more coherent local uncertainties regarding variances while GBCoNP and ConvCNP tend to reduce the variances around context points to zero.

As shown in Table II, when the intra-task diversity is large enough, the global uncertainties achieved by several methods are quite similar (NP \approx ANP \approx GBCoNP in MNIST). For instance, given a small set of pixels indicating a number “6”, there are plenty of possibilities in this dataset to finish the inpainting, therefore producing a large uncertainty σ_z . To manipulate this global uncertainty, we generate images with respect to different priors in Fig 7. Similar to 1D, we relax the prior standard deviation to $12 \sigma_z$. For MNIST, the global uncertainty in ANP determines the thickness and extension of a stroke, e.g., a “1” can be transformed to a “9” then to a “4” if the upper stroke of “1” is gradually extended and thickened. Besides, extension towards different directions can result in variants of orientations of a digit (see digit “6” in Fig 7). The results of GBCoNP comply with this extension pattern yet with limited variation due to the model representation capacity as discussed in 1D. For CelebA32, global uncertainty depicts the background and appearance. Changing values across the prior can transit the background color from completely black to white with different shades of brightness. Appearance variations range from the size of the eyes, noses, face shape, to the hair volume and color.

TABLE I
LOGLIKELIHOOD ON 1D DATASETS (MEAN \pm STD)

Model Name	Dataset Name					
	RBF	Periodic	Matrn-3/2	Stock50	SmartMeter	HousePricing
NN	-1.42 \pm 0.00	-1.42 \pm 0.00	-1.42 \pm 0.00	-1.40 \pm 4E-3	-1.42 \pm 7E-5	-1.16 \pm 4E-3
CNP [11]	-0.73 \pm 0.03	-1.27 \pm 2E-3	-0.89 \pm 8E-3	-0.53 \pm 0.05	-0.43 \pm 0.03	0.33 \pm 0.08
ACNP	0.74 \pm 0.12	-1.08 \pm 7E-3	0.23 \pm 0.03	-0.11 \pm 0.11	-0.33 \pm 0.05	1.07 \pm 0.15
ConvCNP [9]	1.19 \pm 0.15	0.71 \pm 0.03	0.41 \pm 0.04	0.03 \pm 0.16	0.09 \pm 0.07	1.25 \pm 0.18
NP [6]	-1.08 \pm 0.03	-1.30 \pm 2E-3	-1.04 \pm 0.01	-0.79 \pm 0.08	-0.47 \pm 0.02	-1.20 \pm 0.39
ANP [7]	0.74 \pm 0.12	-1.06 \pm 8E-3	0.19 \pm 0.03	-0.19 \pm 0.15	-0.33 \pm 0.05	0.80 \pm 0.37
GBCoNP	1.24 \pm 0.14	0.25 \pm 0.04	0.37 \pm 0.03	0.03 \pm 0.13	0.02 \pm 0.05	1.11 \pm 0.24

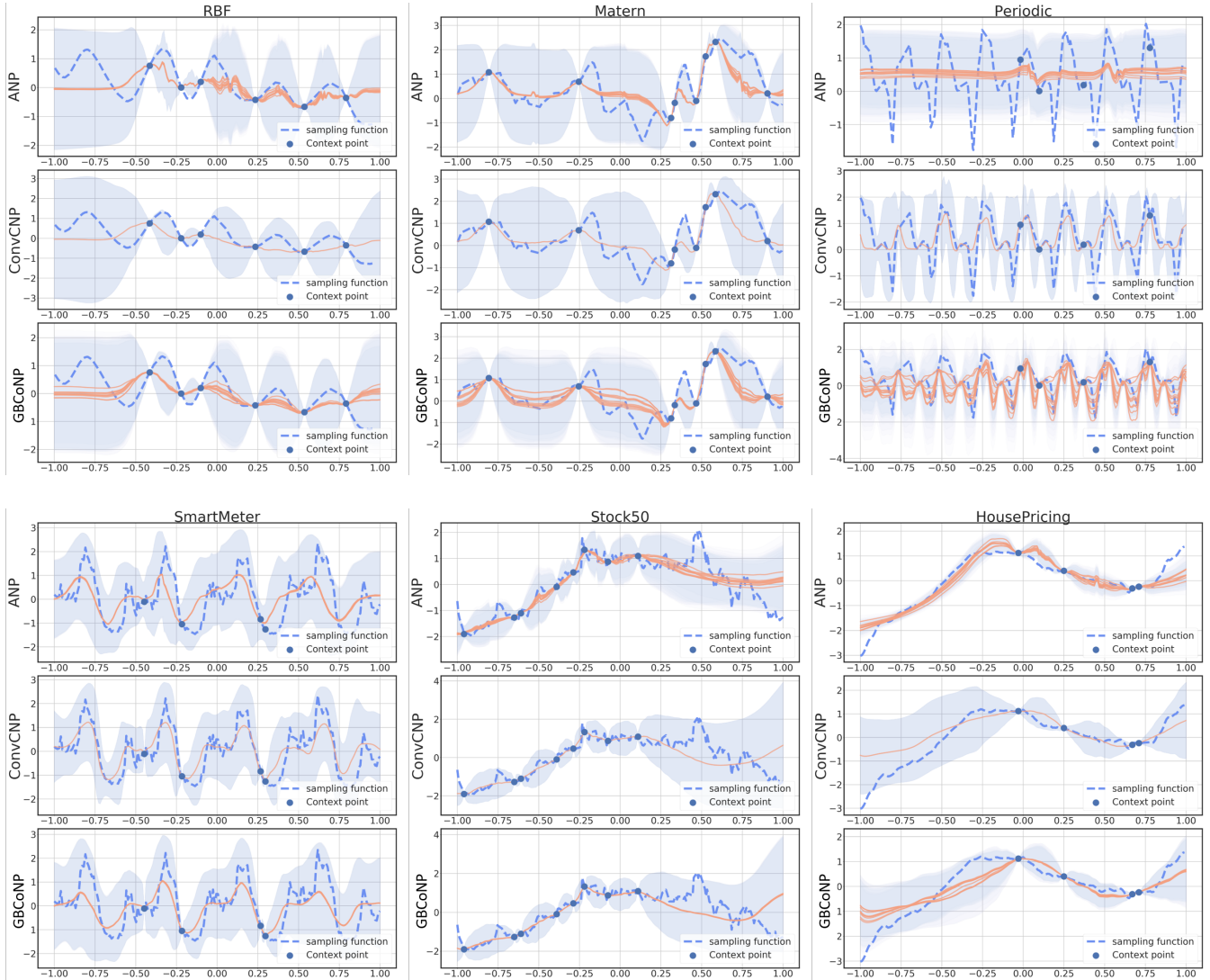


Fig. 4. Model predictions with uncertainty for ANP, ConvCNP and GBCoNP on 6 datasets. For latent based models, 10 different global latent values are sampled and displayed. Each Solid line represents one sample of the predicted mean. The shaded areas, $\mu \pm 2\sigma$, indicate 95% confidence intervals. Ground truths are showed in dotted lines, with the context data highlighted in big dots.

C. Spatial-Temporal Dataset

COVID¹. In spatial temporal scenarios, we use this dataset that contains daily total confirmed coronavirus cases in all

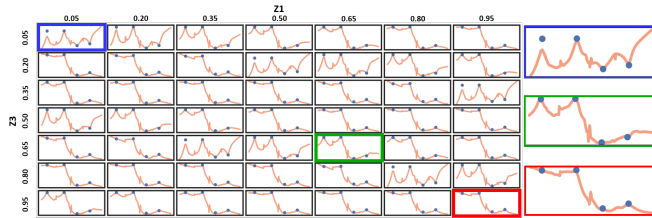
counties of US from 21/01/2020 to 20/04/2021. We use a window of 14 days with daily total confirmed cases sampled in each task to build temporal features that could depict the trend. Each element is normalized by $\log(x - x_{min} + 1)$, where x_{min} suggests the lowest cases therein. Since three points are sufficient to describe the curve, i.e., the last 7 days, the last

¹<https://www.kaggle.com/fireballbyedimynnmm/us-counties-covid-19-dataset>

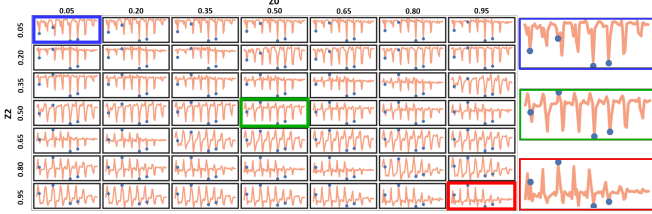
TABLE II
GLOBAL UNCERTAINTY FOR LATENT NEURAL PROCESS FAMILIES

Model	NP		ANP		GBCoNP	
	μ_z (e-3)	σ_z	μ_z (e-3)	σ_z	μ_z (e-3)	σ_z
RBF	-7.25	0.48	-0.48	0.20	0.29	0.19
Periodic	-1.26	0.52	7.21	0.57	-3.12	0.21
Matrn-3/2	-0.61	0.35	-0.80	0.35	0.12	0.19
Stock50	0.74	0.29	5.93	0.36	-0.05	0.11
SmartMeter	-2.67	0.45	-0.21	0.28	10.66	0.11
HousePricing	3.76	0.46	6.58	0.25	4.53	0.18
MNIST	-1.00	0.32	55.01	0.34	-1.77	0.29
SVHN	-12.08	0.49	-8.61	0.13	35.76	0.11
CelebA32	18.30	0.18	2.83	0.14	50.04	0.13
Covid	-2.25	0.52	-10.56	0.13	-55.88	0.44

^a ϵ in (4) for 1D, 2D datasets are 1, and 5% total pixels respectively. Covid datasets remains providing all previous records as a context set.



(a) ANP on Stock50



(b) GBCoNP on Periodic

Fig. 5. Manipulation of global uncertainty for sample generation. A grid of 7×7 latent variables z is sampled to generate predictive means μ_{y_T} . The sampling distribution $\mathcal{N}(\mu_z, \sigma_z^2)$ is obtained from the dotted context data. Sampling values vary from the 5th to the 95th percentile values.

3 days and today, according to our observation, these context points are used for predicting the logged relative growth 7 days after. To fully utilize spatial information, the data for each county are projected on a geographic map (shown in the first row Fig 8), where spatial relationships between counties are displayed while preserving temporal information with color intensity.

The whole map is segmented into a grid of 104 of 40×40 cells and a model only processes one cell at a time to minimize the computational cost. For convolutional methods, context values x is a binary mask $\in \mathbb{R}^{T \times W \times H}$, where only the last time channel is set to zero. For non-convolutional baselines, the mask is transformed into a list of 3d inputs with the relative spatial and temporal locations where Y values correspond to the log-relative growth. The dataset is trained for 100 epochs and tested for 6 runs.

The last column of table III compares the log-likelihood metrics among NP families. GBCoNP and ACNP achieve

TABLE III
LOGLIKELIHOOD ON 2D AND SPATIAL-TEMPORAL DATASETS (MEAN \pm STD)

Model Name	Dataset Name			
	MNIST	SVHN	CelebA32	Covid
NN	1.15 ± 0	0.06 ± 0	-0.06 ± 0	-3.03 ± 0.07
CNP [11]	2.03 ± 0.03	$1.26 \pm 7E-3$	0.97 ± 0.03	-1.91 ± 0.29
ACNP	2.04 ± 0.03	2.48 ± 0.02	1.89 ± 0.10	0.53 ± 0.58
ConvCNP	2.68 ± 0.07	1.84 ± 0.28	1.87 ± 0.11	0.34 ± 0.41
NP [6]	-8.87 ± 2.09	-1.34 ± 0.41	0.24 ± 0.23	-1.50 ± 0.42
ANP [7]	1.36 ± 0.19	2.48 ± 0.03	1.66 ± 0.16	0.67 ± 0.69
GBCoNP	2.72 ± 0.06	2.57 ± 0.03	1.82 ± 0.15	0.83 ± 0.10

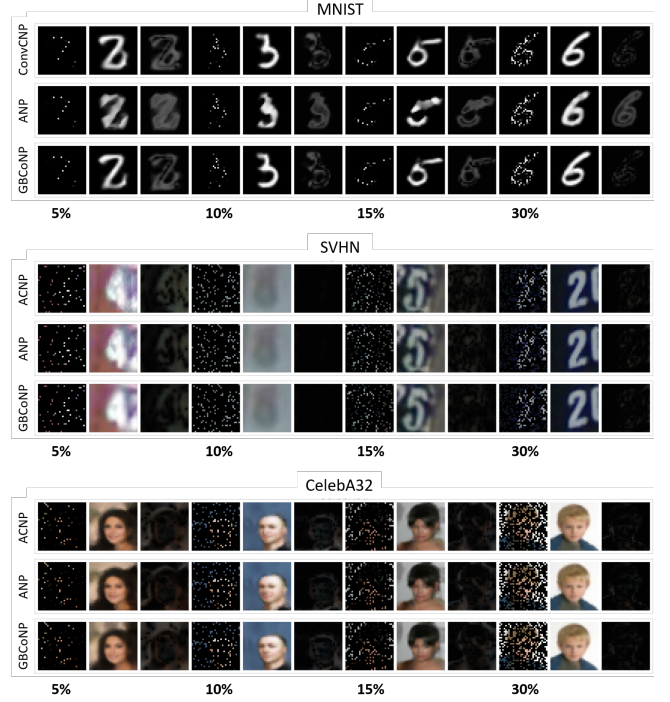


Fig. 6. Model predictions with uncertainties on MNIST, SVHN and CelebA32. For each task, four samples are presented with differed proportions of the unveiled context pixels $\in \{5\%, 10\%, 15\%, 30\%\}$. Each sample gives three consecutive images: masked images, predictive means, and standard deviations.

the SOTA performance in their categories, and GBCoNP outperforms ConvCNP in stability. Most models have high log-likelihood variances due to random sampling of the cell. In contrast with non-convolutional neural processes that could only handle a maximum of $40 \times 40 \times 3 = 4800$ (one cell in the grid) context points each time, GBCoNP is the only one that can process the whole grid (104 cells) by convolutions which results in a more stabilized metric. Interestingly, the locations with the low mean values tend to have high uncertainty (regions with whiter means have redder variances), as shown in Fig 8(b) and (c)—the model believes that these “safer” regions surrounded by high risk neighbourhoods have higher tendency to be infected in the future. Besides, GBCoNP yields more precise mean values (purple line in Fig 8(c)) with



Fig. 7. Manipulation of the global uncertainty for ANP and GBCoNP on MNIST and CelebA32. A grid of 10×10 different latent variables z is sampled to generate predictive means. Four bounding boxes in the grid are enlarged at the bottom line for clarity. The first element is the masked image, i.e., context data. The other three are diverse samplings for comparison.

higher standard deviations compared with ANP (see Fig 8(b)). Our supplementary experiments on global uncertainty reveal that the COVID dataset has the highest intra-task diversity compared with other datasets, as this covid spread trend has the most diverse possibilities. Manipulation on different priors show the standard deviation change of a region from fairly low risk to high risk.

IV. RELATED WORKS

Neural Processes. To enhance uncertainty modelling with Gaussian Process(GPs), previous efforts [12] [13] focus on representing deep kernels in GPs before Neural Processes(NPs) were proposed. Although they inherited the powerful feature representation from neural networks(NNs), they still suffer from computationally expensive matrix inversion $\mathcal{O}(n^3)$. NPs novelly represent stochastic processes with neural networks under two constraints: exchangeability and consistency, which frees the model from inversion with fully back-propagation and induces the whole neural processes family [14]. There are two major branches in the family: conditional NPs(CNP) [11] and latent NPs. Others members add different inductive biases on context and target relationships; some famous ones include Attentive NPs(ANPs) [7], Sequential NPs(SNPs) [15], and Convolutional CNPs (ConvCNPs) [11]. A recent work, ConvNP [8], adds latent path to ConvCNP; but instead of using variational inference, it adopts Monte Carlo estimation for z and focuses on local coherence. Inspired by this, our work further elaborates on the global uncertainty and the related effect.

Generative model with priors. Generative models help increase intra-diversity of data distributions and mitigate few shot learning issues. They generally generate synthetic samples with priors over model condition. For deep learning-based time series generation, [16] applies LSTMs to capture context information and fill the missing values. [17] uses adversarial nets to generate sequence with temporal dynamics across time. In image inpainting with priors, conditional VAEs [18] and GANs [19], [20] manipulate the semantic features in an image (e.g., "smile", "eyeglasses", "age"). However, very few existing studies have meta-learning settings for a distribution of sampling functions, even fewer have considered model uncertainty. [21] employs a meta-agnostic structure for reinforcement learning but uses GPs rather than NPs.

V. CONCLUSION

In this paper, we answered three important questions about the global uncertainty in latent neural processes: How to formalize global uncertainty? What causes and affects global uncertainty? How to manipulate the global uncertainty for data generation? We define the global uncertainty as a prior of z from a latent functional distribution given a small set of context data. We discovered that global uncertainty is affected by the model representation capacity and the data intra-task diversity. Our manipulation of the global uncertainty not only achieves generating the desired samples to tackle few-shot learning, but also enables the probability evaluation on the functional priors.

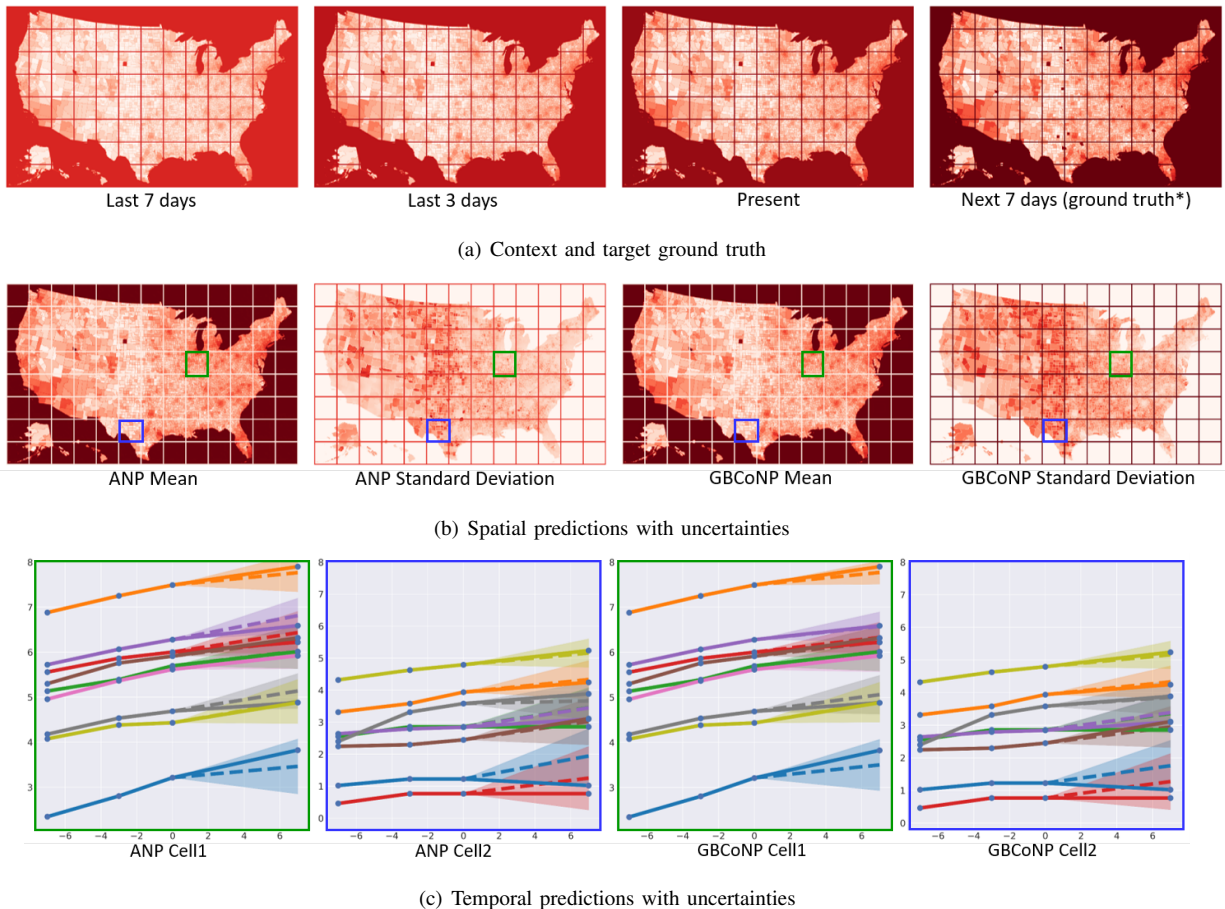


Fig. 8. Model predictions with uncertainties on COVID (sample date: 01-April-21). (a): Context data include the last 7, 3 days and the present day. The shades of color represent the relative growth compared with the data in the last 14 days. Saturation of the colors in context data are adjusted in the plot to increase contrast for better visualization. (b) Spatial predictions with uncertainties for ANP and GBCoNP. Each model processes one cell in the grid at a time. (c) Temporal predictions on two 40×40 cells. Nine pixels are presented in each cell. The ground truth are depicted as solid lines. Predictions are shown in dotted lines with the shaded area indicating uncertainty.

APPENDIX A DATASET DETAILS

For 1d datasets, We use synthetic Gaussian processes with 3 kernels to sample values. The kernel functions are as follows:

- **RBF**: $\mathcal{K}(x, x') = e^{-\frac{1}{2}(\frac{x-x'}{0.2})^2}$
- **Periodic**: $\mathcal{K}(x, x') = e^{-2(\frac{\sin 2\pi(|x-x'|)}{0.5})^2}$
- **Matern** $-\frac{3}{2}$: $\mathcal{K}(x, x') = (1 + 5\sqrt{3}|x - x'|)e^{-5\sqrt{3}|x-x'|}$

where the x values lie within $x \in [-2, 2]$, and the function values y are sampled from $y \sim \mathcal{GP}(0, \mathcal{K}(x, x'))$. Training, validating, and testing data sizes are set to 50,000, 10,000 and 5,000.

Stock50² contains daily trading volumes of 50 stocks from National Stock Exchange India. In each sampling task, 200 days of data are sampled starting from a random date between 11/2016 to 11/2017. X values represent the days count from the first day, and y values represent the corresponding volume weighted average price. Training, validating and testing sizes of the stocks are 36, 10, and 4, respectively.

²<https://www.kaggle.com/rohanrao/nifty50-stock-market-data>

SmartMeter³ includes half-hourly average energy consumption readings from 5,567 London households during 03/12/2011 – 28/02/2014 [22]. For each task, 100 hours of data are sampled with a random timestamp. X values refer to the relative time gaps measured in days, and y values refer to the consumption in kWh/half-hour. The proportions of the training, validating and testing sizes are 7:2:1 based on time line.

HousePricing⁴ comprises monthly average house prices in 9,473 American cities from 01/1996 to 03/2020. One hundred months of data are sampled per sampling task. X values are the time differences measured in months, and values y are prices in dollars. The proportions of the training, validating and testing sets are 7:2:1 based on total number of cities index.

³<https://www.kaggle.com/jeanmidev/smart-meters-in-london/version/11>

⁴https://www.kaggle.com/paultimothymooney/zillow-house-price-data?select=City_Zhvi_AllHomes.csv

APPENDIX B
TRAINING DETAILS

The running time costs of GBCoNP and ConvCNP are compared in Table IV. The total number of epochs for 1D, 2D and Covid datasets is 100, 50, 100 respectively.

TABLE IV
WALL-CLOCK TIME COSTS IN SECONDS (MEAN \pm STD) ON 5 EPOCHS

Model Dataset	GBCoNP		ConvCNP	
	mean	std	mean	std
RBF	17.53	0.11	7.60	0.06
Periodic	17.59	0.07	7.48	0.11
Matrn-3/2	17.54	0.10	7.49	0.14
Stock50	33.42	0.14	20.86	0.14
SmartMeter	50.87	0.10	38.49	0.41
HousePricing	41.61	0.09	21.97	0.24
MNIST	618.30	6.04	188.60	0.78
SVHN	763.44	7.14	234.04	0.36
CelebA32	2322.96	3.53	854.94	2.32
Covid	93.46	3.53	82.05	0.18

REFERENCES

- [1] R. Sawhney, S. Agarwal, A. Wadhwa, and R. R. Shah, "Spatiotemporal hypergraph convolution network for stock movement forecasting," in *2020 IEEE International Conference on Data Mining (ICDM)*, 2020, pp. 482–491.
- [2] Y. Lin, Y.-Y. Chiang, M. Franklin, S. P. Eckel, and J. L. Ambite, "Building autocorrelation-aware representations for fine-scale spatiotemporal prediction," in *2020 IEEE International Conference on Data Mining (ICDM)*, 2020, pp. 352–361.
- [3] X. Chen, L. Yao, and Y. Zhang, "Residual attention u-net for automated multi-class segmentation of covid-19 chest ct images," *arXiv preprint arXiv:2004.05645*, 2020.
- [4] A. Zeroual, F. Harrou, A. Dairi, and Y. Sun, "Deep learning methods for forecasting covid-19 time-series data: A comparative study," *Chaos, Solitons & Fractals*, vol. 140, p. 110121, 2020.
- [5] Z. Qian, A. M. Alaa, and M. van der Schaar, "When and how to lift the lockdown? global covid-19 scenario analysis and policy assessment using compartmental gaussian processes," *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [6] M. Garnelo, J. Schwarz, D. Rosenbaum, F. Viola, D. J. Rezende, S. Eslami, and Y. W. Teh, "Neural processes," *arXiv preprint arXiv:1807.01622*, 2018.
- [7] H. Kim, A. Mnih, J. Schwarz, M. Garnelo, A. Eslami, D. Rosenbaum, O. Vinyals, and Y. W. Teh, "Attentive neural processes," in *International Conference on Learning Representations*, 2018.
- [8] A. Foong, W. Bruinsma, J. Gordon, Y. Dubois, J. Requeima, and R. Turner, "Meta-learning stationary stochastic process prediction with convolutional neural processes," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [9] J. Gordon, W. P. Bruinsma, A. Y. Foong, J. Requeima, Y. Dubois, and R. E. Turner, "Convolutional conditional neural processes," in *International Conference on Learning Representations*, 2019.
- [10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.
- [11] M. Garnelo, D. Rosenbaum, C. Maddison, T. Ramalho, D. Saxton, M. Shanahan, Y. W. Teh, D. Rezende, and S. A. Eslami, "Conditional neural processes," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1704–1713.
- [12] A. Garriga-Alonso, C. E. Rasmussen, and L. Aitchison, "Deep convolutional networks as shallow gaussian processes," in *International Conference on Learning Representations*, 2018.
- [13] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing, "Deep kernel learning," in *Artificial Intelligence and Statistics*, 2016, pp. 370–378.
- [14] Y. Dubois, J. Gordon, and A. Y. Foong, "Neural process family," <http://yanndubs.github.io/Neural-Process-Family/>, September 2020.
- [15] G. Singh, J. Yoon, Y. Son, and S. Ahn, "Sequential neural processes," in *Advances in Neural Information Processing Systems*, 2019, pp. 10254–10264.
- [16] X. Tang, H. Yao, Y. Sun, C. Aggarwal, P. Mitra, and S. Wang, "Joint modeling of local and global temporal dynamics for multivariate time series forecasting with missing values," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 5956–5963.
- [17] J. Yoon, D. Jarrett, and M. van der Schaar, "Time-series generative adversarial networks," *Advances in Neural Information Processing Systems*, 2019.
- [18] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9446–9454.
- [19] J. Gu, Y. Shen, and B. Zhou, "Image processing using multi-code gan prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3012–3021.
- [20] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of gans for semantic face editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9243–9252.
- [21] M. Xu, W. Ding, J. Zhu, Z. LIU, B. Chen, and D. Zhao, "Task-agnostic online reinforcement learning with an infinite mixture of gaussian processes," *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [22] 3springs, "Neural processes for sequential data," 2019, available at: <https://github.com/3springs/attentive-neural-processes>.