Deconditional Downscaling with Gaussian Processes

Anonymous Author(s)
Affiliation
Address
email

Abstract

Refining low-resolution (LR) spatial fields with high-resolution (HR) information is challenging as the diversity of spatial datasets often prevents direct matching of observations. Yet, when LR samples are modeled as aggregate conditional means of HR samples with respect to a mediating variable that is globally observed, the recovery of the underlying fine-grained field can be framed as taking an "inverse" of the conditional expectation, namely a *deconditioning problem*. In this work, we introduce *conditional mean process* (CMP), a new class of Gaussian Processes describing conditional means. By treating CMPs as inter-domain features of the underlying field, a posterior for the latent field can be established as a solution to the deconditioning problem. Furthermore, we show that this solution can be viewed as a two-staged vector-valued kernel ridge regressor and show that it has a minimax optimal convergence rate under mild assumptions. Lastly, we demonstrate its proficiency in a synthetic and a real-world atmospheric field downscaling problem, showing substantial improvements over existing methods.

1 Introduction

Spatial observations often operate at limited resolution due to practical constraints. For example, remote sensing atmosphere products [1, 2, 3, 4] provide measurement of atmospheric properties such as cloud top temperatures and optical thickness, but only at a low resolution. This hinders local scale analysis of variables playing a critical role in our understanding of the anthropogenic impact on climate.

By postulating an underlying detailed spatial field that aggregates into the coarse observations, it becomes possible to frame the inference of high resolution samples as weakly-supervised learning [5] given coarse aggregated targets. This problem, also known as *statistical downscaling* or *spatial disaggregation*, has been studied in the literature in various forms, notably giving it a probabilistic treatment [6, 7, 8, 9, 10, 11], in which Gaussian processes (GP) [12] or deeper combinations of GPs [13, 14] are typically used in conjunction with a sparse variational formulation [15] to recover the underlying latent field at high resolution.

However, these models require access to bags of high resolution (HR) covariates that are paired with aggregate targets, which in practice might be infeasible. For example, distinct satellite trajectories, instrument limitations or variability of atmospheric conditions often lead to coarse and fine grained covariates that are unmatched spatially and temporally. Climate simulations [16, 17, 18], on the other hand, provide a comprehensive low resolution (LR) coverage of meteorological variables that can be matched to both coarse and high resolution covariates. This motivates us to introduce a less restrictive mediated formulation of statistical downscaling that only requires indirect matching between low and high resolution covariates.

Formally, let ${}^b\boldsymbol{x} = \{x^{(1)}, \ldots, x^{(n)}\} \subset \mathcal{X}$ be a general notation for bags of HR covariates, $f : \mathcal{X} \to \mathbb{R}$ the field of interest we wish to recover and $\tilde{z}$ observations from this field but at LR. We suppose that ${}^b\boldsymbol{x}$

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

and $\tilde{z}$ are unmatched, but that there exists LR covariates $y, \tilde{y} \in \mathcal{Y}$, such that y is jointly observed with ${}^b\boldsymbol{x}$ and likewise $\tilde{y}$ with $\tilde{z}$ as illustrated in Figure 1. Then y and $\tilde{y}$ can be used to mediate a two-staged learning process as follows: **(i)** Model the aggregate observation process by $\tilde{z} = \mathbb{E}[f(X)|Y = \tilde{y}] + \varepsilon$ with some noise ε ; **(ii)** Separately learn the conditional expectation operator $y \mapsto \mathbb{E}[f(X)|Y = y]$ using $\{{}^b\boldsymbol{x}, y\}$.

The task of recovering f under observation model $\tilde{z} = \mathbb{E}[f(X)|Y = \tilde{y}] + \varepsilon$ is referred to as *deconditioning*. Motivated by applications in likelihood-free inference and task-transfer regression, Hsu and Ramos [19] first studied the deconditioning problem through the lens of reproducing kernel Hilbert space (RKHS) and introduced the framework of *Deconditional Mean Embeddings* (DME) as its solution. In this work, motivated by the application to mediated statistical downscaling, we extend deconditioning to a multi-resolution setup, which is not considered by Hsu and Ramos [19]. Placing a GP prior on the sought field f, we first characterise the stochastic process resulting from taking the conditional expectations of f, which we term the *Conditional Mean Process* (CMP). By exploiting joint normality of the sought field and its CMP, we can obtain the posterior distribution of f as a principled Bayesian solution to the downscaling task on indirectly matched data. The posterior mean under our model elegantly recovers the DME-based estimator of Hsu and Ramos [19] and the posterior covariance reflects confidence on the deconditioning quality. While Hsu and Ramos [19] also consider a Bayesian interpretation of DME, they do this through the lens of task-transformed regression and do not link f to its CMP. Rather, the likelihoods they use imply $\tilde{z} = f(x)$ which does not reflect a different granularity level of the aggregate response $\tilde{z}$. Our approach also admits a tractable variational inference approximation method, enabling scalability to large datasets. In addition, we study theoretical properties of Deconditional Mean Operator (DMO) associated to DME, by establishing it as a two-staged vector-valued regressor with a natural reconstruction loss. This perspective allows us to leverage vector-valued regression theory and obtain novel convergence rate results for the DMO estimator. Under mild assumptions, we obtain conditions under which this rate is a minimax optimal in terms of statistical-computational efficiency. Our contributions are summarized as follows:

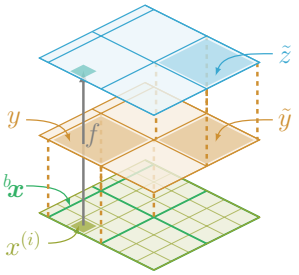

Figure 1: The LR response $\tilde{z}$ (blue) and the bag HR covariates ${}^b\boldsymbol{x}$ (green) are unmatched. The downscaling is mediated through bag-level LR covariates y and $\tilde{y}$ (orange).

- We propose a Bayesian formulation of the mediated statistical downscaling problem, establish its posterior mean as a DME-based estimate and its posterior covariance as a gauge of the deconditioning quality.
- We demonstrate that the DMO estimate minimises a two-staged vector-valued regression and derive its convergence rate under mild assumptions, with conditions for minimax optimality.
- We benchmark our model against existing methods for spatial disaggregation tasks, on both synthetic and real-world multi-resolution atmospheric fields data, and show improved performance

2 Background Materials

2.1 Notations

Let X, Y be a pair of random variables taking values in non-empty sets $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be positive definite kernels. The closure of the span of their canonical feature maps $k_x := k(x, \cdot)$ and $\ell_y := \ell(y, \cdot)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ respectively induces RKHS $\mathcal{H}_k \subseteq \mathbb{R}^{\mathcal{X}}$ and $\mathcal{H}_\ell \subseteq \mathbb{R}^{\mathcal{Y}}$ endowed with inner products $\langle \cdot, \cdot \rangle_k$ and $\langle \cdot, \cdot \rangle_\ell$.

We observe realizations ${}^b\mathcal{D}_1 = \{{}^b\boldsymbol{x}_j, y_j\}_{j=1}^N$ of bags ${}^b\boldsymbol{x}_j = \{x_j^{(i)}\}_{i=1}^{n_j}$ from conditional distribution $\mathbb{P}_{X|Y=y_j}$, with bag-level covariates y_j sampled from $\mathbb{P}_Y$. We concatenate them into vectors $\mathbf{x} := \begin{bmatrix} {}^b\boldsymbol{x}_1 & \dots & {}^b\boldsymbol{x}_N \end{bmatrix}^\top$ and $\mathbf{y} := \begin{bmatrix} y_1 & \dots & y_N \end{bmatrix}^\top$. For simplicity, our exposition will initially use the notation without bagging of the observations, i.e. where $\mathcal{D}_1 = \{x_j, y_j\}_{j=1}^N$. We will come back to a bagged dataset formalism in Section 3.3, which corresponds to our motivating application of statistical downscaling.

With an abuse of notation, we define feature matrices by stacking feature maps along columns as $\mathbf{\Phi_x} := [k_{x_1} \quad \dots \quad k_{x_N}]$ and $\mathbf{\Psi_y} := [\ell_{y_1} \quad \dots \quad \ell_{y_N}]$ and we denote Gram matrices as $\mathbf{K_{xx}} = \mathbf{\Phi_x}^\top \mathbf{\Phi_x} = [k(x_i, x_j)]_{1 \leq i,j \leq N}$ and $\mathbf{L_{yy}} = \mathbf{\Psi_y}^\top \mathbf{\Psi_y} = [\ell(y_i, y_j)]_{1 \leq i,j \leq N}$. The notation abuse $(\cdot)^\top(\cdot)$ is a shorthand for the elementwise RKHS inner products when it is clear from the context.

Let Z denote the real-valued random variable stemming from the noisy conditional expectation of some unknown latent function $f : \mathcal{X} \to \mathbb{R}$, as $Z = \mathbb{E}[f(X)|Y] + \varepsilon$. We suppose one observes another set of realizations $\mathcal{D}_2 = \{\tilde{y}_j, \tilde{z}_j\}_{j=1}^M$ from $\mathbb{P}_{YZ}$, which is sampled independently from $\mathcal{D}_1$. Likewise, we stack observations into vectors $\tilde{\mathbf{y}} := [\tilde{y}_1 \quad \dots \quad \tilde{y}_M]^\top$, $\tilde{\mathbf{z}} := [\tilde{z}_1 \quad \dots \quad \tilde{z}_M]^\top$ and define feature map $\mathbf{\Psi_{\tilde{y}}} := [\ell_{\tilde{y}_1} \quad \dots \quad \ell_{\tilde{y}_M}]$.

2.2 Conditional and Deconditional Kernel Mean Embeddings

Marginal and Joint Mean Embeddings: Kernel mean embeddings of distributions provide a powerful framework for representing and manipulating distributions without specifying their parametric form [20, 21]. The marginal mean embedding of measure $\mathbb{P}_X$ is defined as $\mu_X := \mathbb{E}[k_X] \in \mathcal{H}_k$ and corresponds to the Riesz representer of expectation functional $f \mapsto \mathbb{E}[f(X)]$. It can hence be used to evaluate expectations $\mathbb{E}[f(X)] = \langle f, \mu_X \rangle_k$. If the mapping $\mathbb{P}_X \mapsto \mu_X$ is injective, the kernel k is said to be characteristic [22], a property satisfied for the Gaussian and Matérn kernels on $\mathbb{R}^d$ [23]. In practice, Monte Carlo estimator $\hat{\mu}_X := \frac{1}{N}\sum_{i=1}^N k_{x_i}$ provides an unbiased estimate of μ_X [24].

Extending this rationale to embeddings of joint distributions, we define $C_{YY} := \mathbb{E}[\ell_Y \otimes \ell_Y] \in \mathcal{H}_\ell \otimes \mathcal{H}_\ell$ and $C_{XY} := \mathbb{E}[k_X \otimes \ell_Y] \in \mathcal{H}_k \otimes \mathcal{H}_\ell$, which can be identified with the cross-covariance operators between Hilbert spaces $C_{YY} : \mathcal{H}_\ell \to \mathcal{H}_\ell$ and $C_{XY} : \mathcal{H}_\ell \to \mathcal{H}_k$. They correspond to the Riesz representers of bilinear forms $(g, g') \mapsto \mathrm{Cov}(g(Y), g'(Y)) = \langle g, C_{YY} g' \rangle_\ell$ and $(f, g) \mapsto \mathrm{Cov}(f(X), g(Y)) = \langle f, C_{XY} g \rangle_k$. As above, empirical estimates are obtained as $\hat{C}_{YY} = \frac{1}{N}\mathbf{\Psi_y}\mathbf{\Psi_y}^\top = \frac{1}{N}\sum_{i=1}^N \ell_{y_i} \otimes \ell_{y_i}$ and $\hat{C}_{XY} = \frac{1}{N}\mathbf{\Phi_x}\mathbf{\Psi_y}^\top = \frac{1}{N}\sum_{i=1}^N k_{x_i} \otimes \ell_{y_i}$.

Conditional Mean Embeddings: Similarly, one can introduce RKHS embeddings for conditional distributions referred to as *Conditional Mean Embeddings* (CME). The CME of conditional probability measure $\mathbb{P}_{X|Y=y}$ is defined as $\mu_{X|Y=y} := \mathbb{E}[k_X|Y = y] \in \mathcal{H}_k$. As introduced by Fukumizu et al. [22], it is common to formulate conditioning in terms of a Hilbert space operator $C_{X|Y} : \mathcal{H}_\ell \to \mathcal{H}_k$ called the *Conditional Mean Operator* (CMO). $C_{X|Y}$ satisfies by definition $C_{X|Y}\ell_y = \mu_{X|Y=y}$ and $C_{X|Y}^\top f = \mathbb{E}[f(X)|Y = \cdot], \forall f \in \mathcal{H}_k$, where $C_{X|Y}^\top$ denotes the adjoint of $C_{X|Y}$. Plus, the CMO admits expression $C_{X|Y} = C_{XY}C_{YY}^{-1}$, provided $\ell_y \in \mathrm{Range}(C_{YY})$, $\forall y \in \mathcal{Y}$ [20, 22]. Song et al. [25] show that a nonparametric empirical form of the CMO writes

$$\hat{C}_{X|Y} = \mathbf{\Phi_x}(\mathbf{L_{yy}} + N\lambda\mathbf{I}_N)^{-1}\mathbf{\Psi_y}^\top, \tag{1}$$

where $\lambda > 0$ is some regularisation ensuring the empirical operator is globally defined and bounded.

As observed by Grünewälder et al. [26], since $C_{X|Y}$ defines a mapping from $\mathcal{H}_\ell$ to $\mathcal{H}_k$, it can be interpreted as the solution to a vector-valued regression problem. This perspective enables derivation of probabilistic convergence bounds on the empirical CMO estimator [26, 27].

Deconditional Mean Embeddings: Introduced by Hsu and Ramos [19] as a new class of embeddings, *Deconditional Mean Embeddings* (DME) are natural counterparts of CMEs. While CME $\mu_{X|Y=y} \in \mathcal{H}_k$ allows to take the conditional expectation of any $f \in \mathcal{H}_k$ through inner product $\mathbb{E}[f(X)|Y = y] = \langle f, \mu_{X|Y=y} \rangle_k$, the DME denoted $\mu_{X=x|Y} \in \mathcal{H}_\ell$ solves the inverse problem[1] and allows to recover the initial function of which the conditional expectation was taken, through inner product $\langle \mathbb{E}[f(X)|Y = \cdot], \mu_{X=x|Y} \rangle_\ell = f(x)$.

The associated Hilbert space operator, the *Deconditional Mean Operator* (DMO), is thus defined as the operator $D_{X|Y} : \mathcal{H}_k \to \mathcal{H}_\ell$ such that $D_{X|Y}^\top \mathbb{E}[f(X)|Y = \cdot] = f, \forall f \in \mathcal{H}_k$. It admits an expression in terms of CMO and cross-covariance operators $D_{X|Y} = (C_{X|Y}C_{YY})^\top (C_{X|Y}C_{YY}(C_{X|Y})^\top)^{-1}$ provided $\ell_y \in \mathrm{Range}(C_{YY})$ and $k_x \in \mathrm{Range}(C_{X|Y}C_{YY}C_{X|Y}^\top), \forall y \in \mathcal{Y}$ and $\forall x \in \mathcal{X}$. A nonparametric empirical estimate of the DMO using datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ as described above, is given by $\hat{D}_{X|Y} = \mathbf{\Psi_{\tilde{y}}}\left(\mathbf{A}^\top\mathbf{K_{xx}}\mathbf{A} + M\epsilon\mathbf{I}_M\right)^{-1}\mathbf{A}^\top\mathbf{\Phi_x}^\top$ where $\epsilon > 0$ is a regularisation term and

[1] we adopt slightly unusual notation $\mu_{X=x|Y}$ from Hsu and Ramos [19] which is meant to contrast the usual conditioning $\mu_{X|Y=y}$

$\mathbf{A} := (\mathbf{L_{yy}} + N\lambda\mathbf{I})^{-1}\mathbf{L_{y\tilde{y}}}$ can be interpreted as an aggregation operator. Applying the DMO to expected responses $\tilde{\mathbf{z}}$, Hsu and Ramos [19] are able to recover an estimate of f as

$$\hat{f}(x) = k(x, \mathbf{x})\mathbf{A}\left(\mathbf{A}^\top \mathbf{K_{xx}}\mathbf{A} + M\epsilon\mathbf{I}_M\right)^{-1}\tilde{\mathbf{z}}. \tag{2}$$

Note that since separate samples $\tilde{\mathbf{y}}$ can be used to estimate C_{YY}, this naturally fits a mediating variables setup where $\mathbf{x}$ and the conditional means $\tilde{\mathbf{z}}$ are not jointly observed.

3 Methodology

In this section, we introduce *Conditional Mean Process* (CMP), a stochastic process stemming from the conditional expectation of a GP. We provide a characterisation of the CMP and show that the corresponding posterior mean of its integrand is a DME-based estimator. We also derive in Appendix B a variational formulation of our model that scales to large datasets and demonstrate its performance in Section 5. In what follows, $\mathcal{X}$ is a measurable space, $\mathcal{Y}$ a Borel space and feature maps k_x and ℓ_y are Borel-measurable functions for any $x \in \mathcal{X}$, $y \in \mathcal{Y}$. All proofs and derivations of the paper are included in the appendix.

3.1 Conditional Mean Process

Bayesian quadrature [12, 28, 29] is based on the observation that the integral of a GP with respect to some marginal measure is a Gaussian random variable. We start by probing the implications of integrating with respect to conditional distribution $\mathbb{P}_{X|Y=y}$ and considering such integrals as functions of the conditioning variable y. This gives rise to the notion of conditional mean processes.

Definition 3.1 (Conditional Mean Process)**.** Let $f \sim \mathcal{GP}(m, k)$ with integrable sample paths, i.e. $\int_{\mathcal{X}} |f| \,\mathrm{d}\mathbb{P}_X < \infty$ a.s. The CMP induced by f with respect to $\mathbb{P}_{X|Y}$ is defined as the stochastic process $\{g(y) : y \in \mathcal{Y}\}$ given by $g(y) = \int_{\mathcal{X}} f(x) \,\mathrm{d}\mathbb{P}_{X|Y=y}(x)$.

By linearity of the integral, it is clear that $g(y)$ is a Gaussian random variable for each $y \in \mathcal{Y}$. The sample paths integrability requirement ensures g is well-defined almost everywhere. The following result characterizes CMP as a GP on $\mathcal{Y}$.

Proposition 3.2 (CMP characterization)**.** *Suppose* $\mathbb{E}[|m(X)|] < \infty$ *and* $\mathbb{E}[\|k_X\|_k] < \infty$ *and let* $(X', Y') \sim \mathbb{P}_{XY}$*. Then* g *is a Gaussian process* $g \sim \mathcal{GP}(\nu, q)$ *a.s. , specified by*

$$\nu(y) = \mathbb{E}[m(X)|Y = y] \qquad q(y, y') = \mathbb{E}[k(X, X')|Y = y, Y' = y'] \tag{3}$$

$\forall y, y' \in \mathcal{Y}$*. If we further assume that* k *is characteristic, then* $q(y, y') = \langle \mu_{X|Y=y}, \mu_{X|Y=y'} \rangle_k$*.*

Intuitively, the CMP can be understood as a GP on the conditional means where its covariance $q(y, y')$ is induced by the similarity between the CMEs at y and y'. Resorting to the kernel ℓ defined on $\mathcal{Y}$, we can reexpress the covariance using Hilbert space operators as $q(y, y') = \langle C_{X|Y}\ell_y, C_{X|Y}\ell_{y'} \rangle_k$. A natural nonparametric estimate of the CMP covariance thus comes using the CMO estimator from (1) as $\hat{q}(y, y') = \ell(y, \mathbf{y})\left(\mathbf{L_{yy}} + N\lambda\mathbf{I}_N\right)^{-1}\mathbf{K_{xx}}\left(\mathbf{L_{yy}} + N\lambda\mathbf{I}_N\right)^{-1}\ell(\mathbf{y}, y')$. When $m \in \mathcal{H}_k$, the mean function can be written as $\nu(y) = \langle \mu_{X|Y=y}, m \rangle_k$ for which we can also use empirical estimate $\hat{\nu}(y) = \ell(y, \mathbf{y})\left(\mathbf{L_{yy}} + N\lambda\mathbf{I}_N\right)^{-1}\mathbf{\Phi}_{\mathbf{x}}^\top m$. Finally, one can also quantify the covariance between the CMP g and its integrand f, i.e. $\mathrm{Cov}(f(x), g(y)) = \mathbb{E}[k(x, X)|Y = y]$. Under the same assumptions as Proposition 3.2, this covariance can be expressed using mean embeddings, i.e. $\mathrm{Cov}(f(x), g(y)) = \langle k_x, \mu_{X|Y=y} \rangle_k$ and admits empirical estimate $k(x, \mathbf{x})\left(\mathbf{L_{yy}} + N\lambda\mathbf{I}_N\right)^{-1}\ell(\mathbf{y}, y)$.

3.2 Deconditional Posterior

Given independent observations introduced above, $\mathcal{D}_1 = \{\mathbf{x}, \mathbf{y}\}$ and $\mathcal{D}_2 = \{\tilde{\mathbf{y}}, \tilde{\mathbf{z}}\}$, we may now consider an additive noise model with CMP prior on aggregate observations $\tilde{\mathbf{z}}|\tilde{\mathbf{y}} \sim \mathcal{N}(g(\tilde{\mathbf{y}}), \sigma^2\mathbf{I}_M)$. Let $\mathbf{Q}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}} := q(\tilde{\mathbf{y}}, \tilde{\mathbf{y}})$ be the kernel matrix induced by q on $\tilde{\mathbf{y}}$ and let $\mathbf{\Upsilon} := \mathrm{Cov}(f(\mathbf{x}), \tilde{\mathbf{z}}) = \mathbf{\Phi}_{\mathbf{x}}^\top C_{X|Y}\mathbf{\Psi}_{\tilde{\mathbf{y}}}$ be the cross-covariance between $f(\mathbf{x})$ and $\tilde{\mathbf{z}}$. The joint normality of $f(\mathbf{x})$ and $\tilde{\mathbf{z}}$ gives

$$\begin{bmatrix} f(\mathbf{x}) \\ \tilde{\mathbf{z}} \end{bmatrix} \mid \mathbf{y}, \tilde{\mathbf{y}} \sim \mathcal{N}\left(\begin{bmatrix} m(\mathbf{x}) \\ \nu(\tilde{\mathbf{y}}) \end{bmatrix}, \begin{bmatrix} \mathbf{K_{xx}} & \mathbf{\Upsilon} \\ \mathbf{\Upsilon}^\top & \mathbf{Q}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}} + \sigma^2\mathbf{I}_M \end{bmatrix}\right). \tag{4}$$

Using Gaussian conditioning, we can then readily derive the posterior distribution of the underlying GP field f given the aggregate observations $\tilde{\mathbf{z}}$ corresponding to $\tilde{\mathbf{y}}$. The latter can naturally be

degenerated if observations are paired, i.e. $\mathbf{y} = \tilde{\mathbf{y}}$. This formulation can be seen as an example of the inter-domain GP [30], where we utilise the observed conditional means $\tilde{\mathbf{z}}$ as inter-domain inducing features for inference of f.

Proposition 3.3 (Deconditional Posterior)**.** *Given aggregate observations $\tilde{\mathbf{z}}$ with homoscedastic noise σ^2, the deconditional posterior of f is defined as the Gaussian process $f|\tilde{\mathbf{z}} \sim \mathcal{GP}(m_{\mathrm{d}}, k_{\mathrm{d}})$ where*

$$m_{\mathrm{d}}(x) = m(x) + \boldsymbol{\Upsilon}(\mathbf{Q}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}} + \sigma^2 \mathbf{I}_M)^{-1}(\tilde{\mathbf{z}} - \nu(\tilde{\mathbf{y}})), \tag{5}$$

$$k_{\mathrm{d}}(x, x') = k(x, x') - \boldsymbol{\Upsilon}(\mathbf{Q}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}} + \sigma^2 \mathbf{I}_M)^{-1}\boldsymbol{\Upsilon}^\top. \tag{6}$$

Substituting terms by their empirical forms, we can define a nonparametric estimate of the m_{d} as

$$\hat{m}_{\mathrm{d}}(x) := m(x) + k(x, \mathbf{x})\mathbf{A}(\hat{\mathbf{Q}}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}} + \sigma^2 \mathbf{I}_M)^{-1}(\tilde{\mathbf{z}} - \hat{\nu}(\tilde{\mathbf{y}}))) \tag{7}$$

which, when $m = 0$, reduces to the DME-based estimator in (2) by taking the noise variance $\frac{\sigma^2}{N}$ as the inverse regularization parameter. Hsu and Ramos [19] recover a similar posterior mean expression in their Bayesian interpretation of DME. However, they do not link the distributions of f and its CMP, which leads to much more complicated chained inference derivations combining fully Bayesian and MAP estimates, while we naturally recover it using simple Gaussian conditioning.

Likewise, an empirical estimate of the deconditional covariance is given by

$$\hat{k}_{\mathrm{d}}(x, x') := k(x, x') - k(x, \mathbf{x})\mathbf{A}(\hat{\mathbf{Q}}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}} + \sigma^2 \mathbf{I}_M)^{-1}\mathbf{A}^\top k(\mathbf{x}, x'). \tag{8}$$

Interestingly, the latter can be rearranged to write as the difference between the original kernel and kernel undergoing conditioning and deconditioning steps $\hat{k}_{\mathrm{d}}(x, x') = k(x, x') - \langle k_x, \hat{D}_{X|Y}\hat{C}_{X|Y}k_{x'}\rangle_k$. This can be interpreted as a measure of reconstruction quality, which degenerates in the case of perfect deconditioning, i.e. $\hat{D}_{X|Y}\hat{C}_{X|Y} = \mathrm{Id}_{\mathcal{H}_k}$.

3.3 Deconditioning and multiresolution data

Downscaling application would typically correspond to multiresolution data, with bag dataset ${}^b\mathcal{D}_1 = \{({}^b\boldsymbol{x}_j, y_j)\}_{j=1}^N$ where ${}^b\boldsymbol{x}_j = \{x_j^{(i)}\}_{i=1}^{n_j}$. In this setup, the mismatch in size between vector concatenations $\mathbf{x} = [x_1^{(1)} \ \dots \ x_N^{(n_N)}]^\top$ and $\mathbf{y} = [y_1 \quad \dots \quad y_N]^\top$ prevents from readily applying (1) to estimate the CMO and thus infer the deconditional posterior. There is, however, a straightforward approach to alleviate this: simply replicate bag-level covariates y_j to match bags sizes n_j. Although simple, this method incurs a $\mathcal{O}((\sum_{j=1}^N n_j)^3)$ computational cost due to matrix inversion in (1). Alternatively, since bags ${}^b\boldsymbol{x}_j$ are sampled from conditional distribution $\mathbb{P}_{X|Y=y_j}$, unbiased Monte Carlo estimators of CMEs are given by $\hat{\mu}_{X|Y=y_j} = \frac{1}{n_j}\sum_{i=1}^{n_j} k_{x_j^{(i)}}$. Let $\hat{\mathbf{M}}_{\mathbf{y}} = [\hat{\mu}_{X|Y=y_1} \dots \hat{\mu}_{X|Y=y_N}]^\top$ denote their concatenation along columns. We can then rewrite the cross-covariance operator as $C_{XY} = \mathbb{E}_Y[\mathbb{E}_{X|Y}[k_X] \otimes \ell_Y]$ and hence take $\frac{1}{N}\hat{\mathbf{M}}_{\mathbf{y}}\boldsymbol{\Psi}_{\mathbf{y}}^\top$ as an estimate for C_{XY}. Substituting empirical forms into $C_{X|Y} = C_{X|Y}C_{YY}^{-1}$ and applying Woodbury identity, we obtain an alternative CMO estimator that only requires inversion of a $N \times N$ matrix. We call it *Conditional Mean Shrinkage Operator* and define it as

$${}^S\hat{C}_{X|Y} := \hat{\mathbf{M}}_{\mathbf{y}}(\mathbf{L}_{\mathbf{yy}} + \lambda N \mathbf{I}_N)^{-1}\boldsymbol{\Psi}_{\mathbf{y}}^\top. \tag{9}$$

This estimator can be seen as a generalisation of the Kernel Mean Shrinkage Estimator [31] to the conditional case. We provide in Appendix C modifications of (7) and (8) including this estimator for the computation of the deconditional posterior.

4 Deconditioning as a regression

In Section 3.2, we obtain a DMO-based estimate for the posterior mean of $f|\tilde{\mathbf{z}}$. When the estimate gets closer to the exact operator, the uncertainty collapses and the Bayesian view meets the frequentist. It is however unclear how the empirical operators effectively converge in finite data size. Adopting an alternative perspective, we now demonstrate that the DMO estimate can be obtained as the minimiser of a two-staged vector-valued regression. This frequentist turn enables us to leverage rich theory of vector-valued regression and establish under mild assumptions a convergence rate on the

DMO estimator, with conditions to fulfill minimax optimality in terms of statistical-computational efficiency. In the following, we briefly review CMO's vector-valued regression viewpoint and construct an analogous regression problem for DMO. We refer the reader to [32] for a comprehensive overview of vector-valued RKHS theory.

Stage 1: Regressing the Conditional Mean Operator: As first introduced by Grünewälder et al. [26] and generalized to infinite dimensional spaces by Singh et al. [27], estimating $C_{X|Y}^\top$ is equivalent to solving a vector-valued kernel ridge regression problem in the hypothesis space of Hilbert-Schmidt operators from $\mathcal{H}_k$ to $\mathcal{H}_\ell$, denoted as $\mathrm{HS}(\mathcal{H}_k, \mathcal{H}_\ell)$. Specifically, we may consider the operator-valued kernel defined over $\mathcal{H}_k$ as $\Gamma(f, f') := \langle f, f\rangle_k \, \mathrm{Id}_{\mathcal{H}_\ell}$. We denote $\mathcal{H}_\Gamma$ the $\mathcal{H}_\ell$-valued RKHS spanned by Γ with norm $\|\cdot\|_\Gamma$, which can be identified to $\mathrm{HS}(\mathcal{H}_k, \mathcal{H}_\ell)$. Singh et al. [27] frame CMO regression as the minimisation surrogate discrepancy $\mathcal{E}_\mathrm{c}(C) := \mathbb{E}\left[\|k_X - C^\top \ell_Y\|_k^2\right]$, to which they substitute an empirical regularised version restricted to $\mathcal{H}_\Gamma$ given by $\hat{\mathcal{E}}_\mathrm{c}(C) := \frac{1}{N}\sum_{i=1}^N \|k_{x_i} - C^\top \ell_{y_i}\|_k^2 + \lambda \|C\|_\Gamma^2$. This $\mathcal{H}_k$-valued kernel ridge regression problem admits a closed-form minimiser which shares the same empirical form as the CMO, i.e. $\hat{C}_{X|Y}^\top = \arg\min_{C\in\mathcal{H}_\Gamma} \hat{\mathcal{E}}_\mathrm{c}(C)$ [26, 27].

Stage 2 : Regressing the Deconditional Mean Operator: The DMO on the other hand is defined as the operator $D_{X|Y} : \mathcal{H}_k \to \mathcal{H}_\ell$ such that $\forall f \in \mathcal{H}_k$, $D_{X|Y}^\top C_{X|Y}^\top f = f$. Since deconditioning corresponds to finding a pseudo-inverse to the CMO, it is natural to consider a reconstruction objective $\mathcal{E}_\mathrm{d}(D) := \mathbb{E}\left[\|\ell_Y - DC_{X|Y}\ell_Y\|_\ell^2\right]$. Introducing a novel characterization of the DMO, we propose to minimise this objective in the hypothesis space of Hilbert-Schmidt operators $\mathrm{HS}(\mathcal{H}_k, \mathcal{H}_\ell)$ which identifies to $\mathcal{H}_\Gamma$. As per above, we denote $\hat{C}_{X|Y}$ the empirical CMO learnt in Stage 1, and we substitute the loss with an empirical regularised formulation on $\mathcal{H}_\Gamma$

$$\hat{\mathcal{E}}_\mathrm{d}(D) := \frac{1}{M}\sum_{j=1}^M \|\ell_{\tilde{y}_j} - D\hat{C}_{X|Y}\ell_{\tilde{y}_j}\|_\ell^2 + \epsilon\|D\|_\Gamma^2 \qquad D \in \mathcal{H}_\Gamma \qquad \epsilon > 0 \tag{10}$$

Proposition 4.1 (Empirical DMO as vector-valued regressor)**.** *The minimiser of the empirical reconstruction risk is the empirical DMO, i.e.* $\hat{D}_{X|Y} = \arg\min_{D\in\mathcal{H}_\Gamma} \hat{\mathcal{E}}_\mathrm{d}(D)$

Since it requires to estimate the CMO first, minimising (10) can be viewed as a two-staged vector value regression problem.

Convergence results: Following the footsteps of [27, 33], this perspective enables us to state the performance of the DMO estimate in terms of asymptotic convergence of the objective $\mathcal{E}_\mathrm{d}$. As in Caponnetto and De Vito [34], we must restrict the class of probability measure for $\mathbb{P}_{XY}$ and $\mathbb{P}_Y$ to ensure uniform convergence even when $\mathcal{H}_k$ is infinite dimensional. The family of distribution considered is a general class of priors that does not assume parametric distributions and is parametrized by two variables: $b > 1$ controls the effective input dimension and $c \in]1, 2]$ controls functional smoothness. Mild regularity assumptions are also placed on the original spaces $\mathcal{X}, \mathcal{Y}$, their corresponding RKHS $\mathcal{H}_k, \mathcal{H}_\ell$ and the vector-valued RKHS $\mathcal{H}_\Gamma$. We discuss these assumptions in details in Appendix D. Importantly, while $\mathcal{H}_k$ can be infinite dimensional, we nonetheless have to assume the RKHS $\mathcal{H}_\ell$ is finite dimensional. In further research, we will relax this assumption.

Theorem 4.2 (Empirical DMO Convergence Rate)**.** *Denote* $D_{\mathbb{P}_Y} = \arg\min_{D\in\mathcal{H}_\Gamma} \mathcal{E}_\mathrm{d}(D)$. *Assume assumptions stated in Appendix D are satisfied. In particular, let* (b, c) *and* $(0, c')$ *be the parameters of the restricted class of distribution for* $\mathbb{P}_Y$ *and* $\mathbb{P}_{XY}$ *respectively and let* $\iota \in]0, 1]$ *be the Hölder continuity exponent in* $\mathcal{H}_\Gamma$. *Then, if we choose* $\lambda = N^{-\frac{1}{c'+1}}$, $N = M^{\frac{a(c'+1)}{\iota(c'-1)}}$ *where* $a > 0$, *we have the following result,*

- *If* $a \leq \frac{b(c+1)}{bc+1}$, *then* $\mathcal{E}_\mathrm{d}(\hat{D}_{X|Y}) - \mathcal{E}_\mathrm{d}(D_{\mathbb{P}_Y}) = \mathcal{O}(M^{\frac{-ac}{c+1}})$ *with* $\epsilon = M^{\frac{-a}{c+1}}$
- *If* $a \geq \frac{b(c+1)}{bc+1}$, *then* $\mathcal{E}_\mathrm{d}(\hat{D}_{X|Y}) - \mathcal{E}_\mathrm{d}(D_{\mathbb{P}_Y}) = \mathcal{O}(M^{\frac{-bc}{bc+1}})$ *with* $\epsilon = M^{\frac{-b}{bc+1}}$

This theorem underlines a trade-off between the computational and statistical efficiency with respect to the datasets cardinalities $N = |\mathcal{D}_1|$, $M = |\mathcal{D}_2|$ and the problem difficulty b, c, c'. For $a \leq \frac{b(c+1)}{bc+1}$, smaller a means less samples from $\mathcal{D}_1$ at fixed M and thus computational savings. But it also hampers convergence, resulting in reduced statistical efficiency. At $a = \frac{b(c+1)}{bc+1} < 2$, convergence rate is a minimax computational-statistical efficiency optimal, i.e. convergence rate is optimal with smallest

possible M. We note that at this optimal, $N > M$ and which means less samples are required from $\mathcal{D}_2$. $a \geq \frac{b(c+1)}{bc+1}$ does not improve the convergence rate but only increases the size of $\mathcal{D}_1$ and hence the computational cost it bears.

5 Deconditional Downscaling Experiments

We demonstrate and evaluate our CMP-based downscaling approaches on both synthetic experiments and a challenging atmospheric temperature field downscaling problem with unmatched multi-resolution data. We denote the exact CMP deconditional posterior from Section 3 as CMP, the CMP using with efficient shrinkage CMO estimation as S-CMP and the variational formulation as VARCMP. They are compared against VBAGG [7] — which we describe below — and a GP regression [12] baseline (GPR) modified to take bags centroids as the input. Experiments are implemented in *PyTorch* [35, 36], all code and datasets are made available [37] and computational details are provided in Appendix E.

Variational Aggregate Learning: VBAGG is introduced by Law et al. [7] as a variational aggregate learning framework to disaggregate exponential family models, with emphasis on the Poisson family. We consider its application to the Gaussian family, which models the relationship between aggregate targets z_j and bag covariates $\{x_j^{(i)}\}_i$ by bag-wise averaging of a GP prior on the function of interest. In fact, the Gaussian VBAGG corresponds exactly to a special case of CMP on matched data, where the bag covariates are simply one hot encoded indices with kernel $\ell(j, j') = \delta(j, j')$ where δ is the Kronecker delta. However, VBAGG cannot handle unmatched data as bag indices do not instill the smoothness that is used for mediation. For fair analysis, we compare variational methods VARCMP and VBAGG together, and exact methods CMP/S-CMP to an exact version of VBAGG, which we implement and refer to as BAGG-GP.

5.1 Swiss Roll

The *scikit-learn* [38] swiss roll manifold sampling function allows to generate a 3D manifold of points $x \in \mathbb{R}^3$ mapped with their position along the manifold $t \in \mathbb{R}$. Our objective will be to recover t for each point x by only observing t at an aggregate level. In the first experiment, we compare our model to existing weakly supervised spatial disaggregation methods when all high-resolution covariates are matched with a coarse-level aggregate target. We then proceed to withdraw this requirement in a companion experiment.

5.1.1 Direct matching

Experimental Setup: Inspired by the experimental setup from Law et al. [7], we regularly split space along height $B-1$ times as depicted in Figure 2 and group together manifold points within each height level, hence mixing together points with very different positions on the manifold. We obtain bags of samples $\{({}^b\boldsymbol{x}_j, \boldsymbol{t}_j)\}_{j=1}^B$ where the jth bag contains n_j points ${}^b\boldsymbol{x}_j = \{x_j^{(i)}\}_{i=1}^{n_j}$ and their corresponding targets $\boldsymbol{t}_j = \{t_j^{(i)}\}_{i=1}^{n_j}$. We then construct bag aggregate targets by taking noisy bag targets average $z_j := \frac{1}{n_j}\sum_{i=1}^{n_j} t_j^{(i)} + \varepsilon_j$, where $\varepsilon_j \sim \mathcal{N}(0, \sigma^2)$. We thus obtain matched weakly supervised bag learning dataset $\mathcal{D}^\circ = \{({}^b\boldsymbol{x}_j, z_j)\}_{j=1}^B$. Since each bag corresponds to a height-level, the center altitude of each height split $y_j \in \mathbb{R}$ is a natural candidate bag-level covariate that informs on relative positions of the bags. We can augment the above dataset as $\mathcal{D} = \{({}^b\boldsymbol{x}_j, y_j, z_j)\}_{j=1}^B$. Using these bag datasets, we now wish to downscale aggregate targets z_j to recover the unobserved manifold locations $\{\boldsymbol{t}_j\}_{j=1}^B$ and be able to query the target at any previously unseen input x.

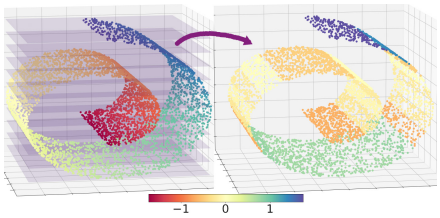

Figure 2: **Step 1:** Split space regularly along height. **Step 2:** Group points into height-level bags. **Step 3:** Average points targets into bag-level aggregate targets.

Models: We use a zero-mean prior on f and choose a Gaussian kernel for k and ℓ. Inducing points location is initialized with K-means++ procedure for VARCMP and VBAGG such that they spread evenly across the manifold. For exact methods, kernel hyperparameters and noise variance σ^2 are learnt on $\mathcal{D}$ by optimising the marginal likelihood. For VARCMP, they are learnt jointly with variational distribution parameters by maximising an evidence lower bound objective. While

Table 1: RMSE of the swissroll experiment for models trained over directly and indirectly matched datasets ; scores averaged over 20 seeds and 1 s.d is reported ; * indicates our proposed methods.

Matching	CMP*	S-CMP*	VARCMP*	BAGG-GP	VBAGG	GPR
Direct	$0.33_{\pm 0.06}$	$0.25_{\pm 0.04}$	$\mathbf{0.18}_{\pm 0.04}$	$0.60_{\pm 0.01}$	$0.22_{\pm 0.04}$	$0.70_{\pm 0.05}$
Indirect	$\mathbf{0.80}_{\pm 0.14}$	$1.05_{\pm 0.04}$	$0.87_{\pm 0.07}$	$1.13_{\pm 0.11}$	$1.46_{\pm 0.34}$	$1.04_{\pm 0.05}$

CMP-based methods can leverage bag-level covariates y_j, baselines are restricted to learn from $\mathcal{D}^\circ$. Adam optimiser [39] is used in all experiments.

Results: We test models against unobserved groundtruth $\{\boldsymbol{t}_j\}_{j=1}^B$ by evaluating the root mean square error (RMSE) to the posterior mean. Table 1 shows that CMP, S-CMP and VARCMP outperform their corresponding counterparts i.e. BAGG-GP and VBAGG, with statistical significance confirmed by a Wilcoxon signed-rank test in Appendix E. Most notably, this shows that the additional knowledge on bag-level dependence instilled by ℓ is reflected even in a setting where each bag is matched with an aggregate target.

5.1.2 Indirect matching

Experimental Setup: We now impose indirect matching through mediating variable y_j. We randomly select $N = \lfloor \frac{B}{2} \rfloor$ bags which we consider to be the N first ones without loss of generality and split $\mathcal{D}$ into $\mathcal{D}_1 = \{({}^b\boldsymbol{x}_j, y_j)\}_{j=1}^N$ and $\mathcal{D}_2 = \{(\tilde{y}_j, \tilde{z}_j)\}_{j=1}^{B-N} = \{(y_{N+j}, z_{N+j})\}_{j=1}^{B-N}$, such that no pair of covariates bag ${}^b\boldsymbol{x}_j$ and aggregate target $\tilde{z}_j$ are jointly observed.

Models: CMP-based methods are naturally able to learn from this setting and are trained by independently drawing samples from $\mathcal{D}_1$ and $\mathcal{D}_2$. Baseline methods however require bags of covariates to be matched with an aggregate bag target. To remedy this, we place a separate prior $g \sim \mathcal{GP}(0, \ell)$ and fit regression model $\tilde{z}_j = g(\tilde{y}_j) + \varepsilon_j$ over $\mathcal{D}_2$. We then use the predictive posterior mean to augment the first dataset as $\mathcal{D}_1' = \left\{\left({}^b\boldsymbol{x}_j, \mathbb{E}[g(y_j)|\mathcal{D}_2]\right)\right\}_{j=1}^N$. This dataset can then be used to train BAGG-GP, VBAGG and GPR.

Results: For comparison, we use the same evaluation as in the direct matching experiment. Table 1 underlines an anticipated drop in RMSE for all models, but we observe that BAGG-GP and VBAGG suffer most from the mediated matching of the dataset while CMP and VARCMP report best scores by a substantial margin. This highlights how using a separate prior on g to mediate $\mathcal{D}_1$ and $\mathcal{D}_2$ turns out to be suboptimal in contrast to using the prior naturally implied by CMP. While it is more computationally efficient than CMP, we observe a relative drop in performance for S-CMP.

5.2 Mediated downscaling of atmospheric temperature

Given the large diversity of sources and formats of remote sensing and model data, expecting directly matched observations is often unrealistic [40]. For example, two distinct satellite products will often provide low and high resolution imagery that can be matched neither spatially nor temporally [1, 2, 3, 4]. Climate simulations [16, 17, 18] on the other hand provide a comprehensive coarse resolution coverage of meteorological variables that can serve as a mediating dataset. For the purpose of demonstration, we create an experimental setup inspired by this problem using Coupled Model Intercomparison Project Phase 6 (CMIP6) [16] simulation data. This grants us access to groundtruth high-resolution covariates to facilitate model evaluation.

Experimental Setup: We collect monthly mean 2D atmospheric fields simulation from CMIP6 data [41, 42] for the following variables: air temperature at cloud top (T), mean cloud top pressure (P), total cloud cover (TCC) and cloud albedo (α). First, we collocate TCC and α onto a HR latitude-longitude grid of size 360×720 to obtain fine grained fields (latitude^{HR}, longitude^{HR}, altitude^{HR}, TCC^{HR}, α^{HR}) augmented with a static HR surface altitude field. Then we collocate P and T onto a LR grid of size 21×42 to obtain coarse grained fields (latitude^{LR}, longitude^{LR}, P^{LR}, T^{LR}). We denote by B the number of low resolution pixels.

Our objective is to disaggregate T^{LR} to the HR fields granularity. We assimilate the jth coarse temperature pixel to an aggregate target $z_j := \text{T}_j^{\text{LR}}$ corresponding to bag j. Each bag includes HR covariates ${}^b\boldsymbol{x}_j = \{x_j^{(i)}\}_{i=1}^{n_j} := \{(\text{latitude}_j^{\text{HR}(i)}, \text{longitude}_j^{\text{HR}(i)}, \text{altitude}_j^{\text{HR}(i)}, \text{TCC}_j^{\text{HR}(i)}, \alpha_j^{\text{HR}(i)})\}_{i=1}^{n_j}$. To emulate unmatched observations, we randomly select $N = \lfloor \frac{B}{2} \rfloor$ of the bags $\{{}^b\boldsymbol{x}_j\}_{j=1}^N$ and keep

Figure 3: **Left:** High-resolution atmoshperic covariates used for prediction; **Center-Left:** Observed low-resolution temperature field, grey pixels are unobserved; **Center** Unobserved high-resolution groundtruth temperature field; **Center-Right:** VARCMP deconditional posterior mean; **Right** 95% confidence region size on prediction; temperature values are in Kelvin.

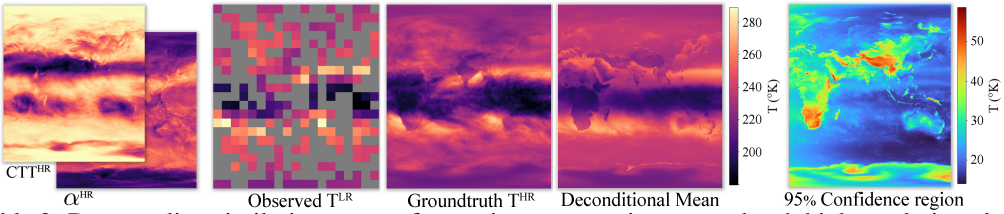

Table 2: Downscaling similarity scores of posterior mean against groundtruth high resolution cloud top temperature field ; averaged over 10 seeds; we report 1 s.d. ; "↓": lower is better ; "↑": higher is better.

Model	RMSE ↓	MAE ↓	Corr. ↑	SSIM ↑
VARGPR	8.02±0.28	5.55±0.17	0.831±0.012	**0.212**±0.011
VBAGG	8.25±0.15	5.82±0.11	0.821±0.006	0.182±0.004
VARCMP	**7.40**±0.25	**5.34**±0.22	**0.848**±0.011	**0.212**±0.013

the opposite half of LR observations $\{z_{N+j}\}_{j=1}^{B-N}$, such that there is no single aggregate bag target that corresponds to one of the available bags. Finally, we choose the pressure field P^{LR} as the mediating variable. We hence compose a third party low resolution field of bag-level covariates $y_j := (\mathrm{latitude}_j^{\mathrm{LR}}, \mathrm{longitude}_j^{\mathrm{LR}}, \mathrm{P}_j^{\mathrm{LR}})$ which can separately be matched with both above sets to obtain datasets $\mathcal{D}_1 = \{(\boldsymbol{x}_j, y_j)\}_{j=1}^{N}$ and $\mathcal{D}_2 = \{(\tilde{y}_j, \tilde{z}_j)\}_{j=1}^{B-N} = \{(y_{N+j}, z_{N+j})\}_{j=1}^{B-N}$.

Models Setup: We only consider variational methods to scale to large number of pixels. VARCMP is naturally able to learn from indirectly matched data. We use a Matérn-1.5 kernel for rough spatial covariates (latitude, longitude) and a Gaussian kernel for atmospheric covariates (P, TCC, α) and surface altitude. k and ℓ are both taken as sums of Matérn and Gaussian kernels, and their hyperparameters are learnt along with noise variance during training. A high-resolution noise term is also introduced, with details provided in Appendix E. Inducing points locations are uniformly initialized across the HR grid. We replace GPR with an inducing point variational counterpart VARGPR [15]. Since baseline methods require a matched dataset, we proceed as with the unmatched swiss roll experiment and fit a GP regression model g with kernel ℓ on $\mathcal{D}_2$ and then use its predictive posterior mean to obtain pseudo-targets for the bags of HR covariates from $\mathcal{D}_1$.

Results: Performance is evaluated by comparing downscaling deconditional posterior mean against original high resolution field T^{HR} available in CMIP6 data [42], which we emphasise is never observed. We use random Fourier features [43] approximation of kernel k to scale kernel evaluation to the HR covariates grid during testing. As reported in Table 2, VARCMP substantially outperforms both baselines with statistical significance provided in Appendix E. Figure 3 shows the reconstructed image with VARCMP, plots for other methods are included in the Appendix E. The model resolves statistical patterns from HR covariates into coarse resolution temperature pixels, henceforth reconstructing a faithful HR version of the temperature field.

6 Discussion

We introduced a scalable Bayesian solution to the mediated statistical downscaling problem, which handles unmatched multi-resolution data. The proposed approach combines Gaussian Processes with the framework of deconditioning using RKHSs and recovers previous approaches as its special cases. We provided convergence rates for the associated deconditioning operator. Finally, we demonstrated the advantages over spatial disaggregation baselines in synthetic data and in a challenging atmospheric fields downscaling problem.

In future work, exploring theoretical guarantees of the computationally efficient shrinkage formulation in a multi-resolution setting and relaxing finite dimensionality assumptions for the convergence rate will have fruitful practical and theoretical implications. Further directions also open up to quantify uncertainty over the deconditional posterior since it is computed using empirical estimates of the CMP covariance. This may be problematic if the mediating variable undergoes covariate shift between the two datasets.

References

[1] L. A. Remer, Y. J. Kaufman, D. Tanré, S. Mattoo, D. A. Chu, J. V. Martins, R.-R. Li, C. Ichoku, R. C. Levy, R. G. Kleidman, T. F. Eck, E. Vermote, and B. N. Holben. The MODIS Aerosol Algorithm, Products, and Validation. *Journal of the Atmospheric Sciences*, 2005.

[2] S. Platnick, M.D. King, S.A. Ackerman, W.P. Menzel, B.A. Baum, J.C. Riedi, and R.A. Frey. The MODIS cloud products: algorithms and examples from Terra. *IEEE Transactions on Geoscience and Remote Sensing*, 2003.

[3] Graeme L Stephens, Deborah G Vane, Ronald J Boain, Gerald G Mace, Kenneth Sassen, Zhien Wang, Anthony J Illingworth, Ewan J O'connor, William B Rossow, Stephen L Durden, et al. The CloudSat mission and the A-train: A new dimension of space-based observations of clouds and precipitation. *Bulletin of the American Meteorological Society*, 2002.

[4] William L. Barnes, Thomas S. Pagano, and Vincent V. Salomonson. Prelaunch characteristics of the Moderate Resolution Imaging Spectroradiometer (MODIS) on EOS-AM1. *IEEE Transactions on Geoscience and Remote Sensing*, 1998.

[5] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 2017.

[6] Yivan Zhang, Nontawat Charoenphakdee, Zhenguo Wu, and Masashi Sugiyama. Learning from aggregate observations. In *Advances in Neural Information Processing Systems*, 2020.

[7] Leon Ho Chung Law, Dino Sejdinovic, Ewan Cameron, Tim C.D. Lucas, Seth Flaxman, Katherine Battle, and Kenji Fukumizu. Variational learning on aggregate outputs with Gaussian processes. In *Advances in Neural Information Processing Systems*, 2018.

[8] Oliver Hamelijnck, Theodoros Damoulas, Kangrui Wang, and Mark A. Girolami. Multi-resolution multi-task Gaussian processes. In *Advances in Neural Information Processing Systems*, 2019.

[9] Fariba Yousefi, Michael Thomas Smith, and Mauricio A. Álvarez. Multi-task learning for aggregated data using Gaussian processes. In *Advances in Neural Information Processing Systems*, 2019.

[10] Yusuke Tanaka, Toshiyuki Tanaka, Tomoharu Iwata, Takeshi Kurashima, Maya Okawa, Yasunori Akagi, and Hiroyuki Toda. Spatially aggregated Gaussian processes with multivariate areal outputs. *Advances in Neural Information Processing Systems*, 2019.

[11] Arto Klam Ville Tanskanen , Krista Longi. Non-Linearities in Gaussian Processes with Integral Observations. *IEEE international Workshop on Machine Learning for Signal*, 2020.

[12] C Rasmussen and C Williams. Gaussian Processes for Machine Learning, 2005.

[13] Andrew Gordon Wilson, David A. Knowles, and Zoubin Ghahramani. Gaussian process regression networks. *Proceedings of the 29th International Conference on Machine Learning, ICML*, 2012.

[14] Andreas C. Damianou and Neil D. Lawrence. Deep Gaussian Processes. *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2013.

[15] Michalis K. Titsias. Variational learning of inducing variables in sparse gaussian processes. In *AISTATS*, 2009.

[16] Veronika Eyring, Sandrine Bony, Gerald A Meehl, Catherine A Senior, Bjorn Stevens, Ronald J Stouffer, and Karl E Taylor. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 2016.

[17] Gregory M. Flato. Earth system models: an overview. *WIREs Climate Change*, 2011.

[18] Marko Scholze, J. Icarus Allen, William J. Collins, Sarah E. Cornell, Chris Huntingford, Manoj M. Joshi, Jason A. Lowe, Robin S. Smith, and Oliver Wild. *Earth system models*. Cambridge University Press, 2012.

[19] Kelvin Hsu and Fabio Ramos. Bayesian Deconditional Kernel Mean Embeddings. Proceedings of Machine Learning Research. PMLR, 2019.

[20] Le Song, Kenji Fukumizu, and Arthur Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 2013.

[21] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *arXiv preprint arXiv:1605.09522*, 2016.

[22] Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 2004.

[23] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems*, 2008.

[24] Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 2012.

[25] Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.

[26] Steffen Grünewälder, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimilano Pontil. Conditional Mean Embeddings as Regressors. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 2012.

[27] Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems*, 2019.

[28] FM Larkin. Gaussian measure in Hilbert space and applications in numerical analysis. *The Rocky Mountain Journal of Mathematics*, 1972.

[29] François-Xavier Briol, Chris J Oates, Mark Girolami, Michael A Osborne, Dino Sejdinovic, et al. Probabilistic integration: A role in statistical computation? *Statistical Science*, 2019.

[30] Tim GJ Rudner, Dino Sejdinovic, and Yarin Gal. Inter-domain deep Gaussian processes. In *International Conference on Machine Learning*, 2020.

[31] Krikamol Muandet, Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, and Bernhard Schölkopf. Kernel mean shrinkage estimators. *Journal of Machine Learning Research*, 2016.

[32] Vern Paulsen and Mrinal Raghupathi. *An introduction to the theory of reproducing kernel Hilbert spaces*. Cambridge university press, 2016.

[33] Zoltán Szabó, Bharath K. Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 2016.

[34] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 2007.

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*. 2019.

[36] Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. Gpytorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. In *Advances in Neural Information Processing Systems*, 2018.

[37] Anonymous Author(s). `https://anonymous.4open.science/r/deconditional-downscaling`, 2021.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.

[39] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR*, 2015.

[40] D. Watson-Parris, N. Schutgens, N. Cook, Z. Kipling, P. Kershaw, E. Gryspeerdt, B. Lawrence, and P. Stier. Community Intercomparison Suite (CIS) v1.4.0: a tool for intercomparing models and observations. *Geoscientific Model Development*, 2016.

[41] Malcolm Roberts. MOHC HadGEM3-GC31-HM model output prepared for CMIP6 HighResMIP hist-1950. *Earth System Grid Federation*, 2018 v20180730. doi: 10.22033/ESGF/CMIP6.6040.

[42] Aurore Voldoire. CNRM-CERFACS CNRM-CM6-1-HR model output prepared for CMIP6 HighResMIP hist-1950. *Earth System Grid Federation*, 2019 v20190221. doi: 10.22033/ESGF/CMIP6.4040.

[43] Ali Rahimi, Benjamin Recht, et al. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, 2007.

[44] Olav Kallenberg. *Foundations of Modern Probability*. Springer, 2002.

[45] S. Saitoh and Yoshihiro Sawano. *Theory of Reproducing Kernels and Applications*. Springer, 2016.

[46] Charles A Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural computation*, 2005.

[47] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Publishing Company, Incorporated, 2008.

[48] Houman Owhadi and Clint Scovel. Conditioning Gaussian measure on Hilbert space. *arXiv preprint arXiv:1506.04208*, 2015.

[49] Ilja Klebanov, Ingmar Schuster, and TJ Sullivan. A rigorous theory of conditional mean embeddings. *SIAM Journal on Mathematics of Data Science*, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See below.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Societal Impact

In this paper, we consider a general framework for statistical downscaling which handles unmatched multi-resolution data.

Enhancing our understanding of climate atmospheric phenomenons by analyzing them at a finer scale could critically benefit decision making in planning adaptation and mitigation policies against climate change. Atmospheric observations are however difficult to collect, and obtaining both matched and high-resolution data is unrealistic. We believe our method can alleviate this problem by allowing us to combine unmatched measurements at different resolutions.

A potential negative impact of downscaling approaches more generally resides in applications where improving images resolution do not serve the common interest. For example, surveillance and privacy intruding image enhancement technologies can be misused and stand as a threat to our society.