Deconditional Downscaling with Gaussian processes

Siu Lun Chau*, Shahine Bouabid*, Dino Sejdinovic

35th Conference on Neural Information Processing, 2021

Outline

Background on Kernel Embeddings

Deconditional Mean Embedding

Deconditional Downscaling with Gaussian processes

Background on Kernel Embeddings

- Kernel method is any method that endows a generic abstract domain $\mathcal{X}$ with an inner product structure induced by some feature transformation $\varphi : \mathcal{X} \to \mathcal{H}$.

- Kernel method is any method that endows a generic abstract domain $\mathcal{X}$ with an inner product structure induced by some feature transformation $\varphi : \mathcal{X} \to \mathcal{H}$.
- Kernel function is as an inner product of features: any function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for which there exists a Hilbert space $\mathcal{H}$ and a map $\varphi : \mathcal{X} \to \mathcal{H}$ s.t. $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$ for all $x, x' \in \mathcal{X}$.

Kernels and Reproducing Kernel Hilbert Spaces

- Kernel method is any method that endows a generic abstract domain $\mathcal{X}$ with an inner product structure induced by some feature transformation $\varphi : \mathcal{X} \to \mathcal{H}$.
- Kernel function is as an inner product of features: any function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for which there exists a Hilbert space $\mathcal{H}$ and a map $\varphi : \mathcal{X} \to \mathcal{H}$ s.t. $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$ for all $x, x' \in \mathcal{X}$.
- There exists a canonical feature space $\mathcal{H}_k$, called reproducing kernel Hilbert space (RKHS) with canonical feature map $x \mapsto k(\cdot, x)$, where:
  - $\forall x \in \mathcal{X}, k(\cdot, x) \in \mathcal{H}_k$
  - $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}_k, \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} = f(x)$.

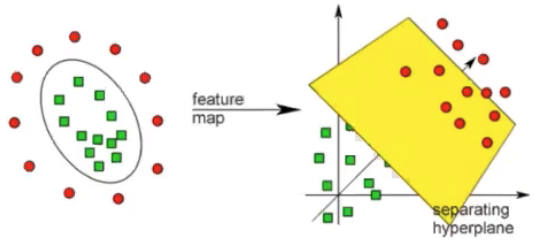
  Thus also $k(x, x') = \langle k(\cdot, x), k(\cdot, x') \rangle_{\mathcal{H}_k}$.

- Kernel method is any method that endows a generic abstract domain $\mathcal{X}$ with an inner product structure induced by some feature transformation $\varphi : \mathcal{X} \to \mathcal{H}$.
- Kernel function is as an inner product of features: any function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for which there exists a Hilbert space $\mathcal{H}$ and a map $\varphi : \mathcal{X} \to \mathcal{H}$ s.t. $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$ for all $x, x' \in \mathcal{X}$.
- There exists a canonical feature space $\mathcal{H}_k$, called reproducing kernel Hilbert space (RKHS) with canonical feature map $x \mapsto k(\cdot, x)$, where:
  - $\forall x \in \mathcal{X}, k(\cdot, x) \in \mathcal{H}_k$
  - $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}_k, \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} = f(x)$.

  Thus also $k(x, x') = \langle k(\cdot, x), k(\cdot, x') \rangle_{\mathcal{H}_k}$.
- Moore-Aronszajn Theorem: every positive semidefinite kernel is the kernel of a unique RKHS.

Kernel Trick and Kernel Mean Trick

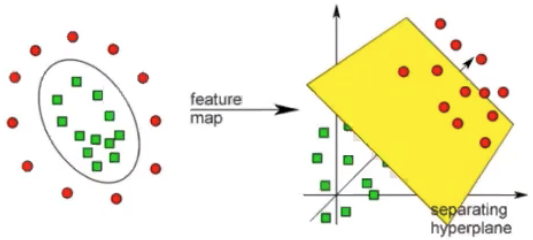
- implicit feature map $x \mapsto k(\cdot, x) \in \mathcal{H}_k$
  replaces $x \mapsto [\phi_1(x), \ldots, \phi_s(x)] \in \mathbb{R}^s$
- $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} = k(x, y)$
  inner products readily available



Cortes & Vapnik, 1995; Schölkopf & Smola, 2001

Kernel Trick and Kernel Mean Trick

- implicit feature map $x \mapsto k(\cdot, x) \in \mathcal{H}_k$
  replaces $x \mapsto [\phi_1(x), \ldots, \phi_s(x)] \in \mathbb{R}^s$
- $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} = k(x, y)$
  inner products readily available

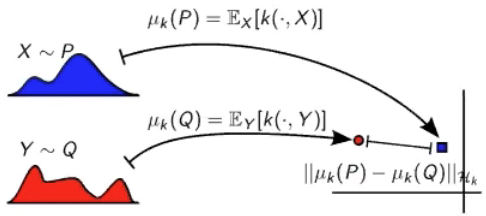


Cortes & Vapnik, 1995; Schölkopf & Smola, 2001

- **RKHS embedding**: implicit feature mean

  Smola et al, 2007; Sriperumbudur et al, 2010; Muandet et al, 2017

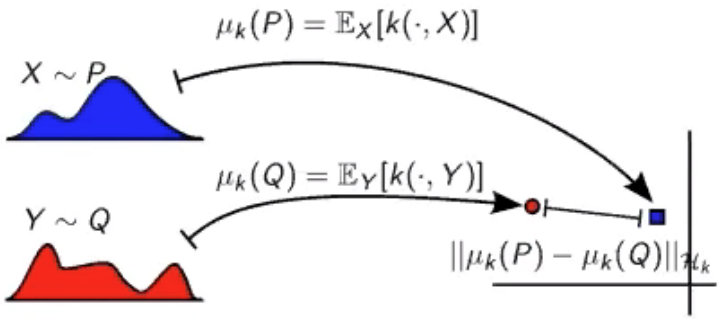
  $P \mapsto \mu_k(P) = \mathbb{E}_{X \sim P} k(\cdot, X) \in \mathcal{H}_k$
  replaces
  $P \mapsto [\mathbb{E}\phi_1(X), \ldots, \mathbb{E}\phi_s(X)] \in \mathbb{R}^s$
- $\langle \mu_k(P), \mu_k(Q) \rangle_{\mathcal{H}_k} = \mathbb{E}_{X \sim P, Y \sim Q} k(X, Y)$
  inner products easy to estimate



Gretton et al, 2005; Gretton et al, 2006; Fukumizu et al, 2007; DS et al, 2013; Muandet et al, 2012; Szabo et al, 2015

Maximum Mean Discrepancy

- Maximum Mean Discrepancy (MMD) Borgwardt et al, 2006; Gretton et al, 2007 between P and Q:



$$\mathrm{MMD}_k(P, Q) = \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k} = \sup_{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1} |\mathbb{E} f(X) - \mathbb{E} f(Y)|$$

- Characteristic kernels: $\mathrm{MMD}_k(P, Q) = 0$ iff $P = Q$ (Gaussian RBF $\exp(-\frac{1}{2\sigma^2}\|x - x'\|_2^2)$, Matérn family, inverse multiquadrics.).
- Can encode structural properties in the data: kernels on non-Euclidean domains, networks, images, text...

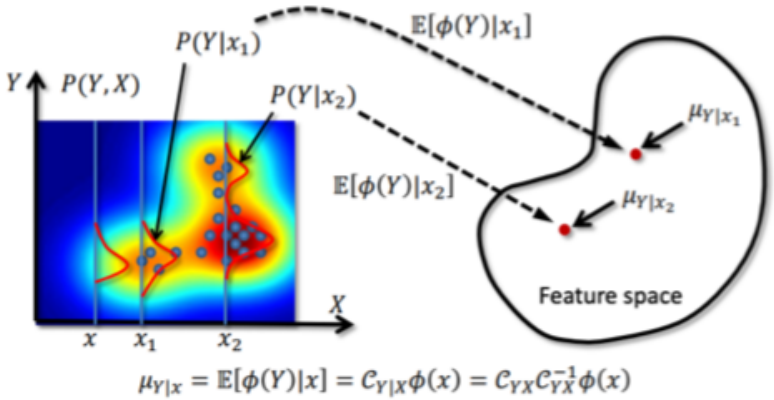
Conditional Mean Embeddings

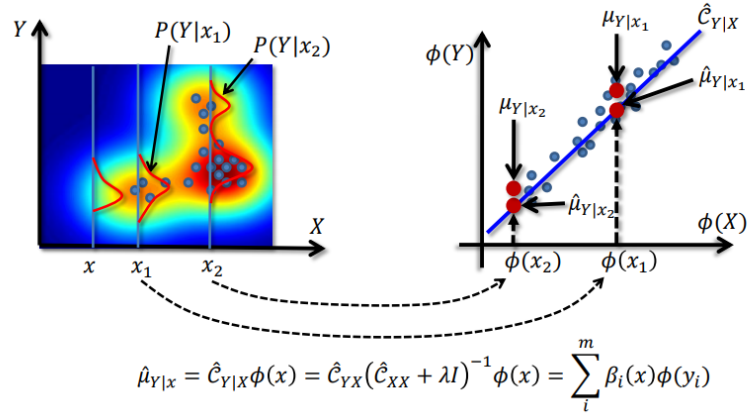
Consider a joint distribution $\mathbb{P}_{XY}$ over rvs (X, Y) taking values in $\mathcal{X} \times \mathcal{Y}$. Given a kernel $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, the conditional mean embedding (CME) of $Y|X = x$ is defined as:

$$\mu_{Y|X=x} := \mathbb{E}[\ell(\cdot, Y)|X = x] = \int_{\mathcal{Y}} \ell(\cdot, y) \, \mathrm{d}\mathbb{P}_{Y|X=x}(y) \in \mathcal{H}_\ell.$$

- Allows to compute $\forall f \in \mathcal{H}_\ell, \mathbb{E}[f(Y)|X = x] = \langle f, \mu_{Y|X=x} \rangle_{\mathcal{H}_\ell}$.

Conditional Mean Embeddings





To model conditional embeddings as functions of x, we associate them with a conditional mean operator (CMO) $C_{Y|X} : \mathcal{H}_k \to \mathcal{H}_\ell$ where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which satisfies

$$\mu_{Y|X=x} = C_{Y|X} k(\cdot, x).$$

This is essentially a feature-to-feature RKHS-valued ridge regression.

GPs and RKHSs: shared mathematical foundations

- The same notion of a positive-definite kernel, but conceptual gaps between communities.
- Orthogonal projection in RKHS $\Leftrightarrow$ Conditioning in GPs.
- 0/1 laws: GP sample paths with infinite-dimensional covariance kernel k lie a.s. outside of $\mathcal{H}_k$. The space of sample paths can be thought of as an "outer shell" of $\mathcal{H}_k$.

Deconditional Mean Embedding

Deconditional Mean Operation

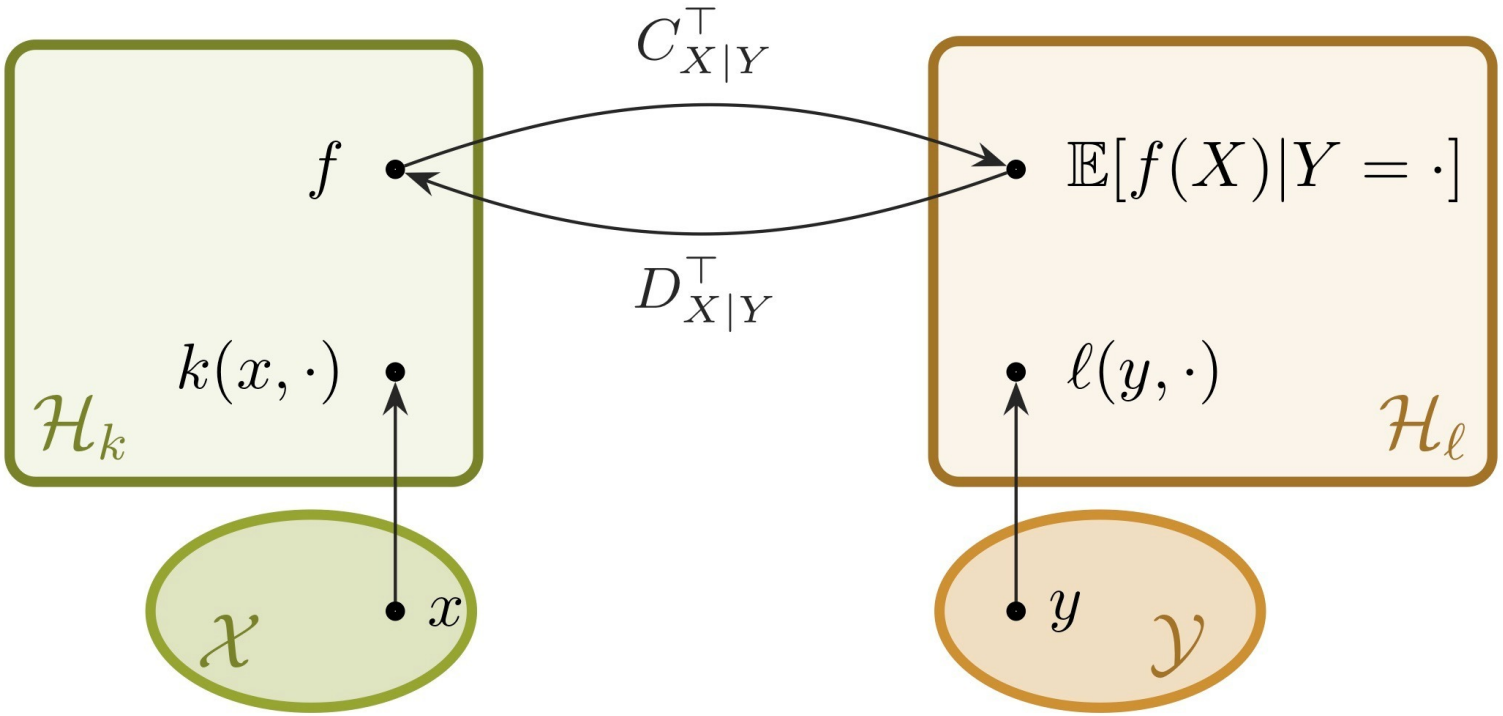
- CMO $C_{Y|X} : \mathcal{H}_k \to \mathcal{H}_\ell$ allows us to reason about the conditional expectation of any $f \in \mathcal{H}_\ell$ in the following way:

$$C_{Y|X}^\top f = \mathbb{E}[f(Y) \mid X = \cdot]$$

- Hsu and Ramos [2019] proposed Deconditional Mean Operator $D_{Y|X} : \mathcal{H}_\ell \to \mathcal{H}_k$ (DMO) as a natural counterpart to CMO.
- Given the conditional mean function, we recover f by setting

$$D_{Y|X}^\top \mathbb{E}[f(Y) \mid X = \cdot] = f$$

Deconditional Mean Operation



Deconditional Mean Operator

- DMO can be rewritten in terms of CMOs and cross-covariance operators:

$$D_{Y|X} = (C_{Y|X} C_{XX})^\top (C_{Y|X} C_{XX} C_{Y|X}^\top)^{-1}$$

  under regularity assumption.
- In our work we further characterise their formulation and show that deconditioning can be seen as a two-staged vector valued kernel ridge regression.

Deconditioning as two-staged regression

- First introduced by Grunewalder et al. [2012], CME has a feature-to-feature regression interpretation:

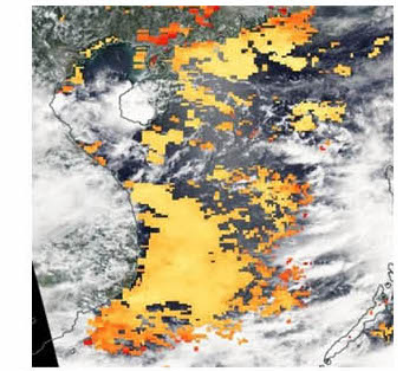
$$\mathcal{E}_c(C) := \frac{1}{N}\sum_{i=1}^{N} ||\ell_{y_i} - C^\top k_{x_i}||^2_{\mathcal{H}_\ell} + \lambda\Omega(C)$$

- Deconditioning has a similar interpretation, thus allowing one to apply regression convergence result to theoretically analyze DMO,

$$\mathcal{E}_d(D) := \frac{1}{M}\sum_{j=1}^{M} ||k_{x_j} - D\hat{C}^\top_{Y|X} k_{x_j}||^2_{\mathcal{H}_k} + \epsilon\Omega(D)$$

Deconditional Downscaling with Gaussian processes

Motivation

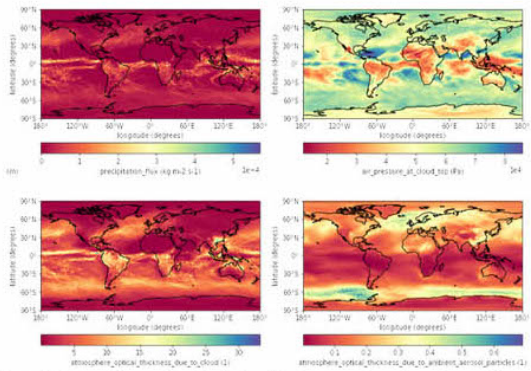


AOD over Vietnamese coast
(MODIS aqua+terra)

Cannot match

Reflectance over Marnes - France
(Landsat-8)

Can match



Can match

CMIP6 cloud fields historical simulations

Problem Setup

Dataset

- We have a dataset of N bags of high-resolution (HR) covariates ${}^{b}\boldsymbol{x}_j := \{x_j^{(1)}, \ldots, x_j^{(n_j)}\}$ paired with a mediating variable y_j

$$\mathcal{D}_1 = \{{}^{b}\boldsymbol{x}_j, y_j\}_{j=1}^{N}, \tag{1}$$

and a second separate dataset of M mediating variables $\tilde{y}_j$ paired with a low-resolution (LR) response $\tilde{z}_j$.

$$\mathcal{D}_2 = \{\tilde{y}_j, \tilde{z}_j\}_{j=1}^{M}. \tag{2}$$

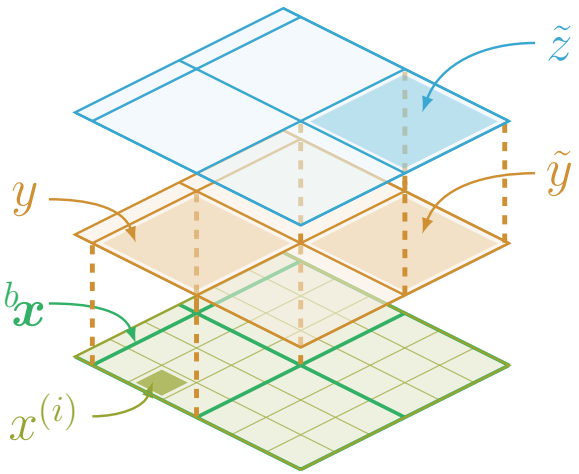


Figure 1: Representation of covariates indirect pairing

Problem Setup

Objective

- Downscale the response $\tilde{z}$ to the HR granularity level of $x^{(i)}$ covariates, i.e. find a function $f : \mathcal{X} \to \mathbb{R}$ which maps between HR covariates and HR responses (not observed!).

How?

- Assume there exists mapping $f : \mathcal{X} \to \mathbb{R}$.
- But only observe "aggregates" $\mathbb{E}_X[f(X)|Y = y]$.
- Recover underlying function f out of aggregated observations.

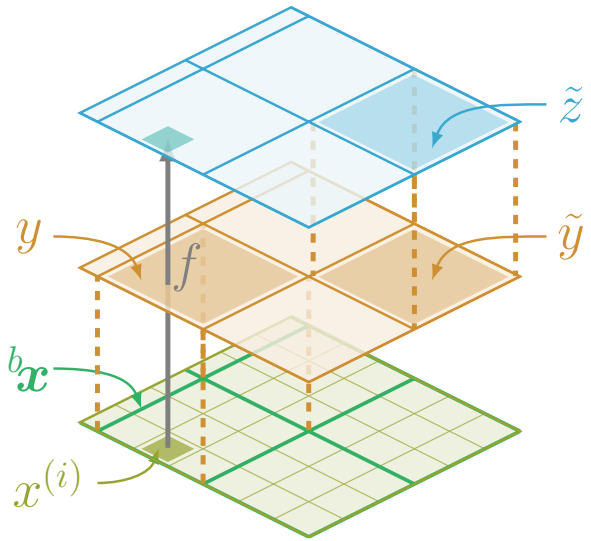


Figure 2: We wish to learn a map from HR covariates to HR estimate of the response.

Deconditional Formulation

Given dataset $\mathcal{D}_2 = \{\tilde{y}_j, \tilde{z}_j\}_{j=1}^{M}$,

Observation Model

- We suppose that the HR response f aggregates into the LR response $\tilde{z}_j$ as

$$\tilde{z}_j = \mathbb{E}_X[f(X)|Y = \tilde{y}_j] + \varepsilon_j \tag{3}$$

with noise $\varepsilon_j \sim \mathcal{N}(0, \sigma^2)$.

Recovering f corresponds to the *deconditioning* problem of Hsu and Ramos [1]:

- Given $g : \mathcal{Y} \to \mathbb{R}$, infer the function $f : \mathcal{X} \to \mathbb{R}$ such that

$$g(y) = \mathbb{E}_X[f(X)|Y = y]. \tag{4}$$

f is called the *deconditional mean of* g w.r.t. $\mathbb{P}_{X|Y}$.

Deconditional Formulation

Given dataset $\mathcal{D}_2 = \{\tilde{y}_j, \tilde{z}_j\}_{j=1}^M$,

Observation Model

- We suppose that the HR response f aggregates into the LR response $\tilde{z}_j$ as

$$\tilde{z}_j = \mathbb{E}_X[f(X)|Y = \tilde{y}_j] + \varepsilon_j \tag{5}$$

with noise $\varepsilon_j \sim \mathcal{N}(0, \sigma^2)$.

Recovering f corresponds to the *deconditioning* problem of Hsu and Ramos [1]:

- Given $g : \mathcal{Y} \to \mathbb{R}$, infer the function $f : \mathcal{X} \to \mathbb{R}$ such that

$$g(y) = \mathbb{E}_X[f(X)|Y = y]. \tag{6}$$

f is called the *deconditional mean of* g w.r.t. $\mathbb{P}_{X|Y}$.

Bayesian formulation for f and g

Conditional Mean Process

- By placing a GP prior on $f \sim \mathrm{GP}(m, k)$, we can represent the LR field as:

$$g(y) = \mathbb{E}_X[f(X)|Y = y] = \int_{\mathcal{X}} f(x) \mathrm{d}\mathbb{P}_{X|Y=y}(x) \sim \mathrm{GP}(\nu, q)$$

where by linearity of the integral,

$$\begin{aligned} \nu(y) &= \mathbb{E}_X[m(X)|Y = y] \\ q(y, y') &= \mathbb{E}_{X,X'}[k(X, X')|Y = y, Y' = y'] = \langle \mu_{X|Y=y}, \mu_{X|Y=y'} \rangle_k. \end{aligned}$$

- ν and q are estimated via conditional mean operator $C_{X|Y}$ using $\mathcal{D}_1 = \{{}^b\boldsymbol{x}_j, y_j\}_{j=1}^N$ and a second kernel $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.

Joint normality between LR and HR field

- $f \sim \mathrm{GP}(m, k)$ and $g \sim \mathrm{GP}(\nu, q)$
- The latent HR field $f(x)$ and the observed noisy LR field $\tilde{z} = g(\tilde{y}) + \epsilon$ are both normally distributed:

$$\begin{bmatrix} f(x) \\ \tilde{z} \end{bmatrix} \mid \tilde{y} \sim \mathcal{N}\left(\begin{bmatrix} m(x) \\ \nu(\tilde{y}) \end{bmatrix}, \begin{bmatrix} k(x,x) & \mathrm{Cov}(f(x), g(\tilde{y})) \\ \mathrm{Cov}(g(\tilde{y}), f(x)) & q(\tilde{y}, \tilde{y}) + \sigma^2 \end{bmatrix} \right)$$

Joint normality between LR and HR field

- $f \sim \mathrm{GP}(m, k)$ and $g \sim \mathrm{GP}(\nu, q)$
- The latent HR field $f(x)$ and the observed noisy LR field $\tilde{z} = g(\tilde{y}) + \epsilon$ are both normally distributed:

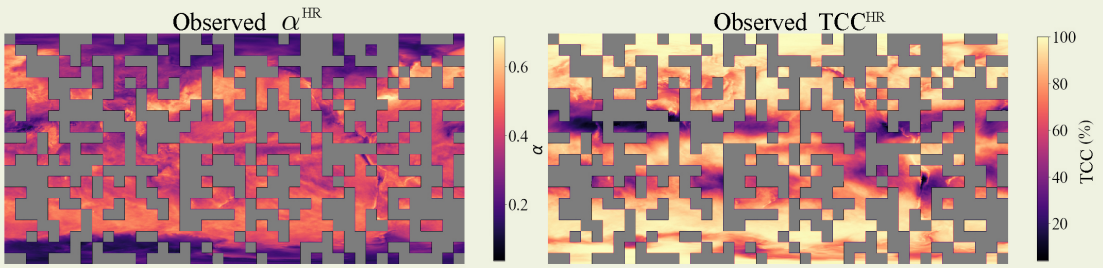
$$\begin{bmatrix} f(x) \\ \tilde{z} \end{bmatrix} \mid \tilde{y} \sim \mathcal{N}\left(\begin{bmatrix} m(x) \\ \nu(\tilde{y}) \end{bmatrix}, \begin{bmatrix} k(x,x) & \mathrm{Cov}(f(x), g(\tilde{y})) \\ \mathrm{Cov}(g(\tilde{y}), f(x)) & q(\tilde{y},\tilde{y}) + \sigma^2 \end{bmatrix} \right)$$

- Allows to directly obtain *deconditional posterior* $f|\tilde{z} \sim \mathrm{GP}(m_d, k_d)$ from $\mathcal{D}_2$ with:

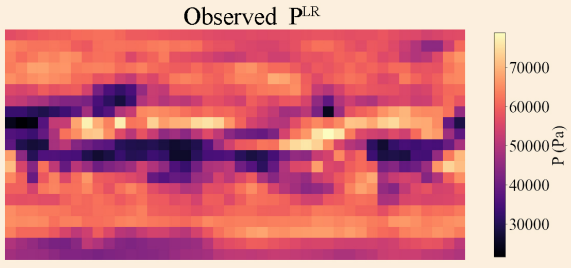
$$\hat{m}_{\mathrm{d}}(x) = m(x) + k(x, \mathbf{x})\mathbf{A}(\hat{\mathbf{Q}} + \sigma^2 \mathbf{I}_M)^{-1}(\tilde{\mathbf{z}} - \nu(\tilde{\mathbf{y}}))$$
$$\hat{k}_{\mathrm{d}}(x, x') = k(x, x') - k(x, \mathbf{x})\mathbf{A}(\hat{\mathbf{Q}} + \sigma^2 \mathbf{I}_M)^{-1}\mathbf{A}^\top k(\mathbf{x},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

where $\mathbf{A} := (\ell(\mathbf{y}, \mathbf{y}) + N\lambda \mathbf{I}_N)^{-1}\ell(\mathbf{y}, \tilde{\mathbf{y}})$ with $\lambda > 0$, $\hat{\mathbf{Q}} := \hat{q}(\tilde{\mathbf{y}}, \tilde{\mathbf{y}})$.

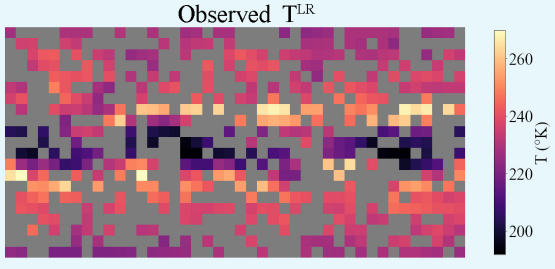
$\{\boldsymbol{x}_j\}_{j=1}^{N}$



$\{y_j\}_{j=1}^{N} \cup \{\tilde{y}_j\}_{j=1}^{B-N}$

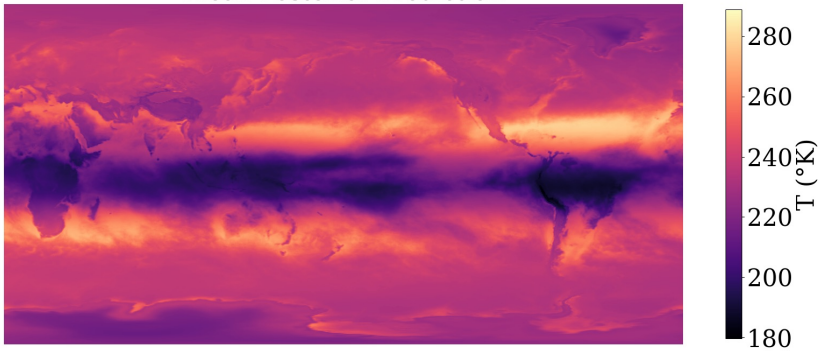


$\{\tilde{z}_j\}_{j=1}^{B-N}$

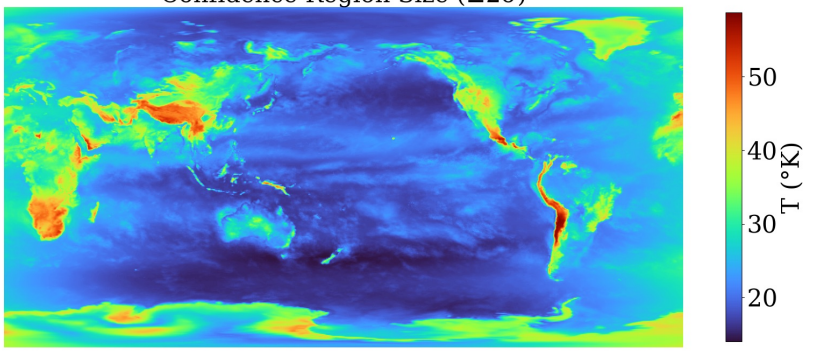


Mediated Downscaling of Atmospheric Temperatures

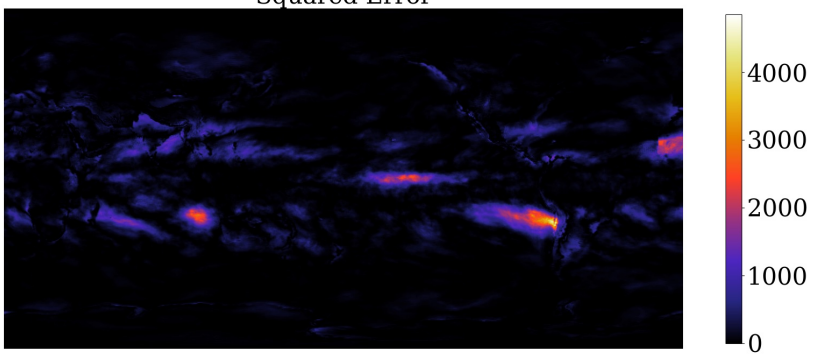
Mean Posterior Prediction



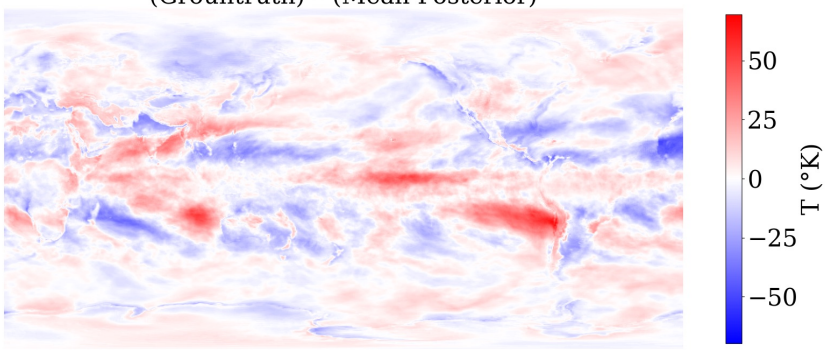
Confidence Region Size (±2σ)

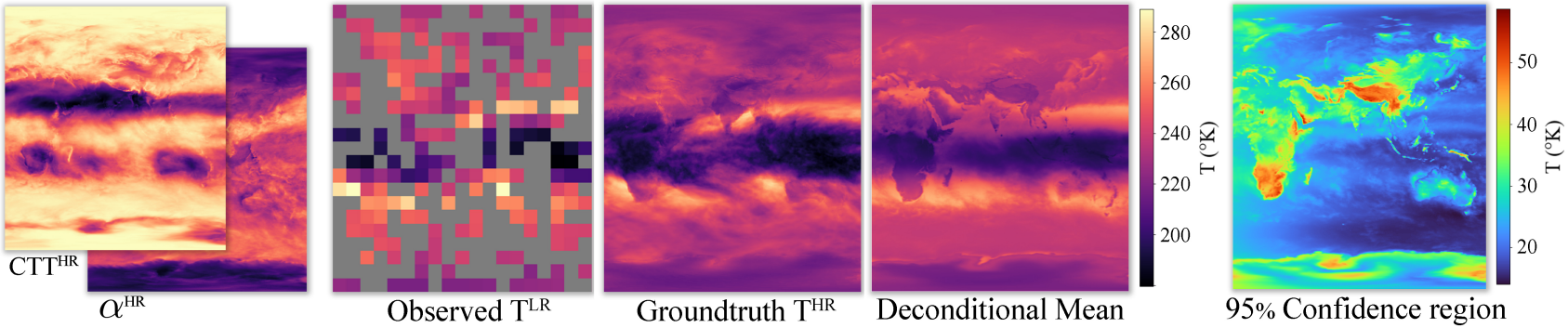


Squared Error



(Grountruth) – (Mean Posterior)





Model	RMSE ↓	MAE ↓	Corr. ↑	SSIM ↑
Krigging	8.02±0.28	5.55±0.17	0.831±0.012	**0.212**±0.011
VBAgg [2]	8.25±0.15	5.82±0.11	0.821±0.006	0.182±0.004
Our method	**7.40**±0.25	**5.34**±0.22	**0.848**±0.011	**0.212**±0.013

Table 1: Downscaling similarity scores of posterior mean against HR groundtruth; reports 1 s.d.

References

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

[2] 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.

Usual CMO estimation

$$
\begin{aligned}
C_{X|Y} &= C_{XY}(C_{YY} + \lambda \,\mathrm{Id}_{\mathcal{H}_\ell})^{-1} \\
&= \mathbb{E}_{XY}[k(\cdot, X) \otimes \ell(Y, \cdot)](\mathbb{E}_Y[\ell(\cdot, Y) \otimes \ell(Y, \cdot)] + \lambda \,\mathrm{Id}_{\mathcal{H}_\ell})^{-1} \\
&\approx \frac{1}{N} k(\cdot, \mathbf{x}) \ell(\mathbf{y}, \cdot) \left(\frac{1}{N} \ell(\cdot, \mathbf{y}) \ell(\mathbf{y}, \cdot) + \lambda \,\mathrm{Id}_{\mathcal{H}_\ell} \right)^{-1} \\
&= k(\cdot, \mathbf{x}) \big(\ell(\mathbf{y}, \cdot) \ell(\cdot, \mathbf{y}) + N\lambda \mathbf{I}_n \big)^{-1} \ell(\mathbf{y}, \cdot) \\
&= k(\cdot, \mathbf{x}) \left(\mathbf{L}_{\mathbf{yy}} + n\lambda \mathbf{I}_n \right)^{-1} \ell(\mathbf{y}, \cdot)
\end{aligned}
$$

where $k(\cdot, \mathbf{x}) = [k(\cdot, x_1), \ldots, k(\cdot, x_N)]$, $\ell(\cdot, \mathbf{y}) = [\ell(\cdot, y_1), \ldots, \ell(\cdot, y_N)]$.

Bagged CMO estimation

$$
\begin{aligned}
C_{X|Y} &= C_{XY}(C_{YY} + \lambda \operatorname{Id}_{\mathcal{H}_\ell})^{-1} \\
&= \mathbb{E}_{XY}[k(\cdot, X) \otimes \ell(Y, \cdot)](C_{YY} + \lambda \operatorname{Id}_{\mathcal{H}_\ell})^{-1} \\
&= \mathbb{E}_{Y}[\mathbb{E}_{X}[k(\cdot, X)|Y] \otimes \ell(Y, \cdot)](C_{YY} + \lambda \operatorname{Id}_{\mathcal{H}_\ell})^{-1} \\
&= \mathbb{E}_{Y}[\mu_{X|Y} \otimes \ell(Y, \cdot)](C_{YY} + \lambda \operatorname{Id}_{\mathcal{H}_\ell})^{-1} \\
&\approx \frac{1}{N} \hat{\boldsymbol{\mu}}_{X|\mathbf{y}} \ell(\mathbf{y}, \cdot) \left(\frac{1}{N} \ell(\cdot, \mathbf{y}) \ell(\mathbf{y}, \cdot) + \lambda \operatorname{Id}_{\mathcal{H}_\ell} \right)^{-1} \\
&= \hat{\boldsymbol{\mu}}_{X|\mathbf{y}} \left(\mathbf{L}_{\mathbf{yy}} + N\lambda \mathbf{I}_n \right)^{-1} \ell(\mathbf{y}, \cdot) \\
&:=^{S} \hat{C}_{X|Y}
\end{aligned}
$$

where $\hat{\boldsymbol{\mu}}_{X|\mathbf{y}} = [\hat{\mu}_{X|Y=y_1}, \ldots, \hat{\mu}_{X|Y=y_N}]$ and $\hat{\mu}_{X|Y=y_j} = \frac{1}{n_j} \sum_{i=1}^{n_j} k(\cdot, x_j^{(i)})$.

Notation Abuses

$$k(\cdot, \mathbf{x})\ell(\mathbf{y}, \cdot) = \sum_{i=1}^{N} k(\cdot, x_i) \otimes \ell(\cdot, y_i) \in \mathcal{H}_k \otimes \mathcal{H}_\ell$$

$$\begin{aligned}
\ell(\mathbf{y}, \cdot)\ell(\cdot, \mathbf{y}) &= [\langle \ell(\cdot, y_i), \ell(\cdot, y_j) \rangle_{\mathcal{H}_\ell}]_{1 \leq i,j \leq N} \\
&= [\ell(y_i, y_j)]_{1 \leq i,j \leq N} \\
&= \mathbf{L}_{\mathbf{yy}} \in \mathbb{R}^{N \times N}
\end{aligned}$$