Thin and deep Gaussian processes

Daniel Augusto de Souza[1], Alexander Nikitin[2], S. T. John[2], Magnus Ross[3], Mauricio A Álvarez[3], Marc Peter Deisenroth[1], João P. P. Gomes[4], Diego Mesquita[5], César L. C. Mattos[4]

[1]University College London [2]Aalto University [3]University of Manchester [4]Universidade Federal do Ceará [5]Fundação Getulio Vargas

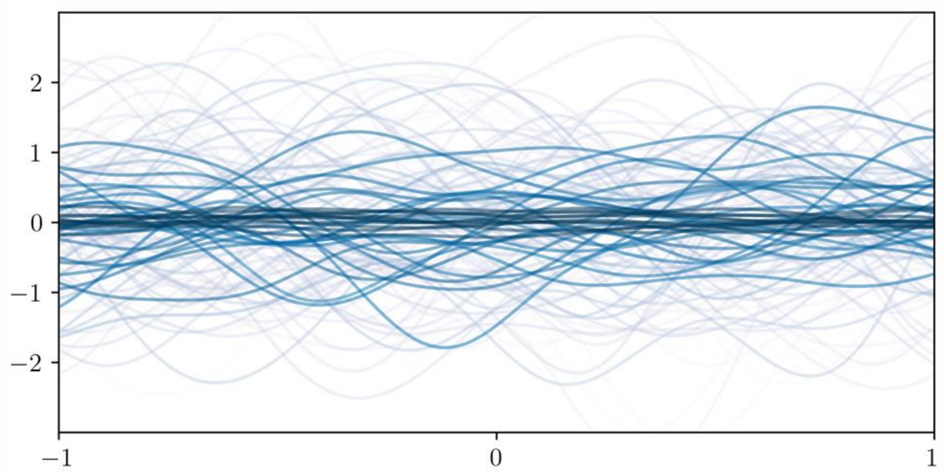
UCL Aalto University MANCHESTER 1824 The University of Manchester UFC FGV

TL;DW

- Current hierarchical Gaussian process models learn one of the following:
  - Latent mappings – reduce dimensionality (DKL/CDGP);
  - Lengthscale fields – easier to interpret (DNSGP);
- We…
  - Propose the thin and deep Gaussian process (TDGP), a new deep GP method that learns both, increasing its interpretability over previous proposals!
  - Show that it has a close relation to the more standard DGP but enables the learning of lengthscales.
  - Demonstrate its performance in synthetic, generic and geospatial datasets.

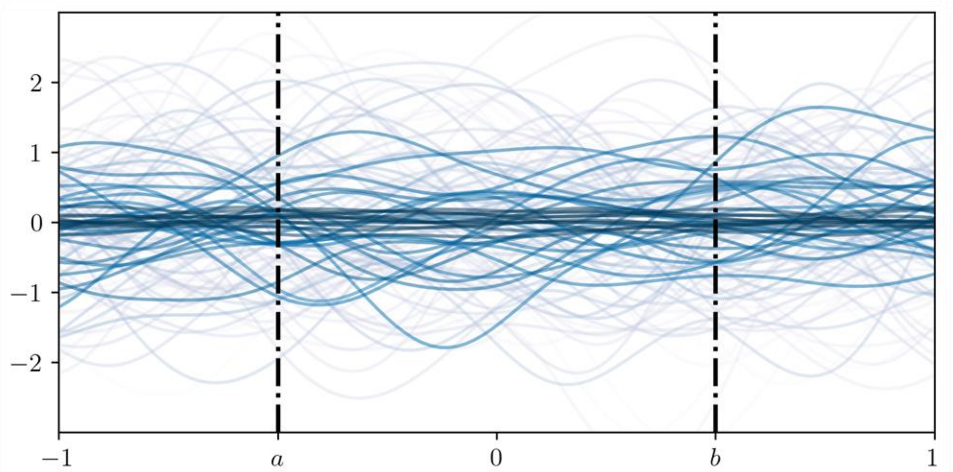
Background

- For a kernel k, a Gaussian process $f(\cdot) \sim \mathcal{GP}(0, k)$ is a distribution over functions.



Background

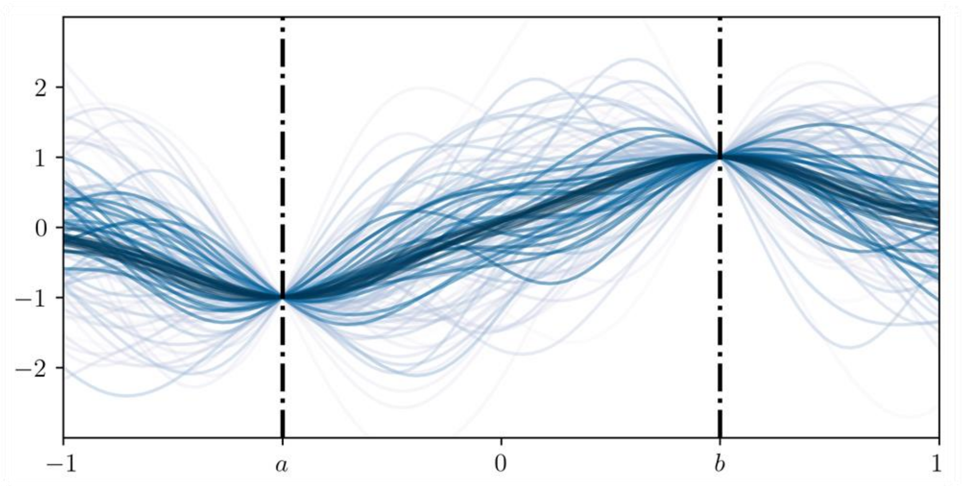
- For a kernel k, a Gaussian process $\mathrm{f}(\cdot) \sim \mathcal{GP}(0, k)$ is a distribution over functions.
- Such that $\mathrm{cov}[\mathrm{f}(a), \mathrm{f}(b)] = k(a, b)$.



Background

- For a kernel k, a Gaussian process $\mathrm{f}(\cdot) \sim \mathcal{GP}(0, k)$ is a distribution over functions.
- Such that $\mathrm{cov}[\mathrm{f}(a), \mathrm{f}(b)] = k(a, b)$.
- Moreover, for training data $(\boldsymbol{X}, \boldsymbol{y})$, the posterior distribution is also a GP:

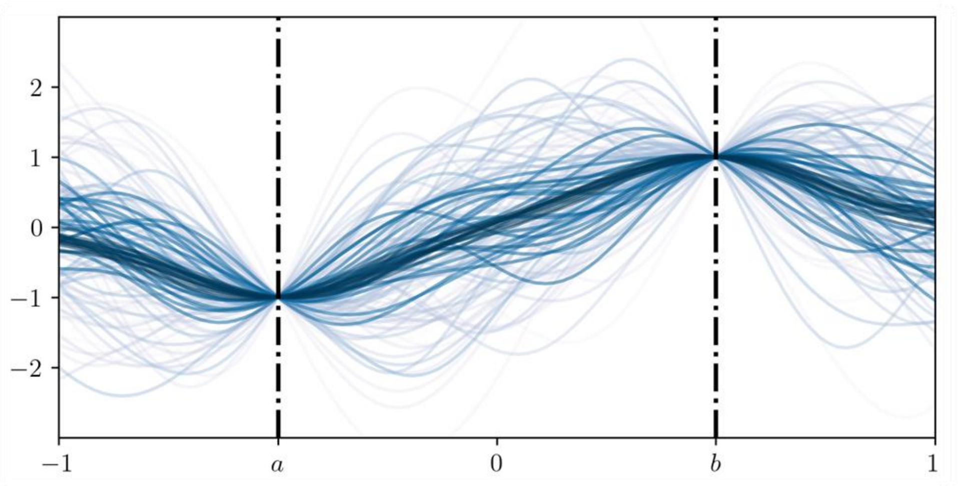
$$\mathrm{f}(\cdot) \mid \mathcal{D} \sim \mathcal{GP}\begin{pmatrix} k(\cdot, \boldsymbol{X}) k(\boldsymbol{X})^{-1} \boldsymbol{y}, \\ k - k(\cdot_1, \boldsymbol{X}) k(\boldsymbol{X})^{-1} k(\boldsymbol{X}, \cdot_2) \end{pmatrix}$$



Background

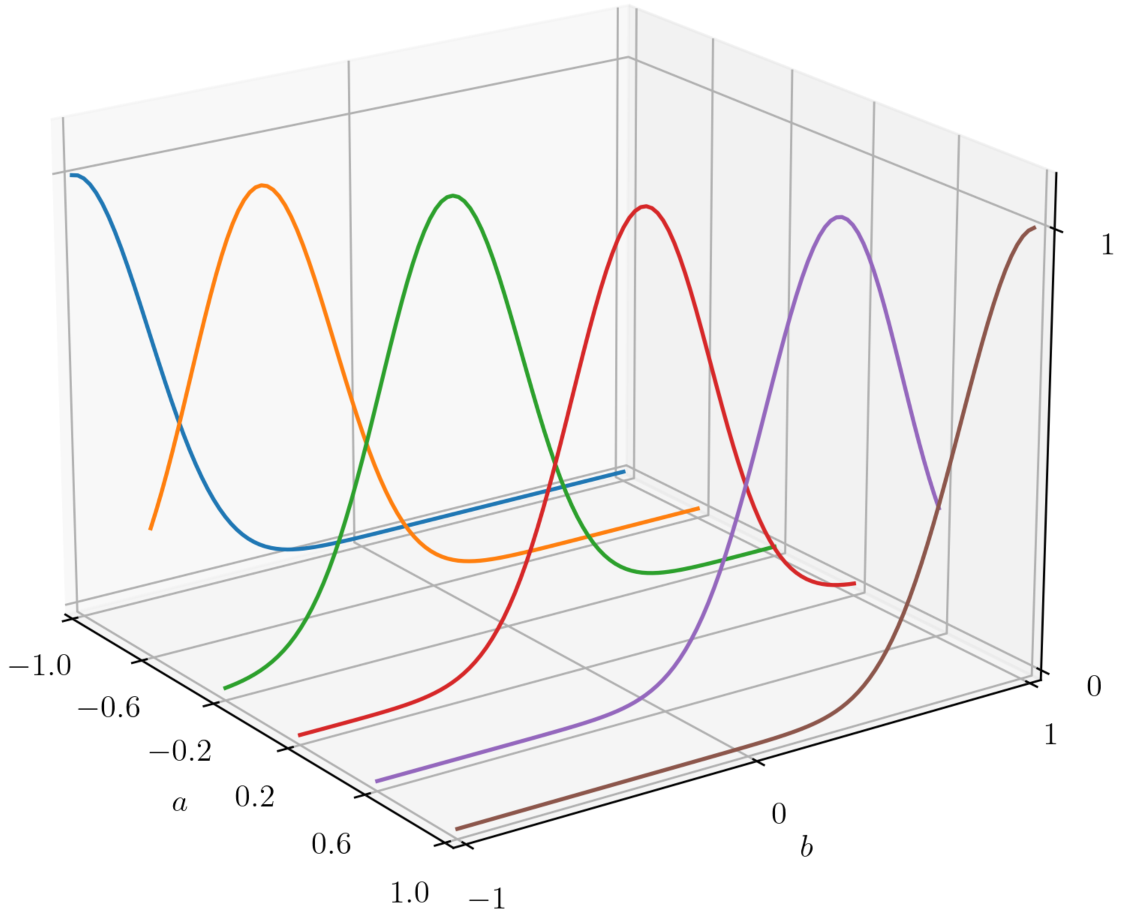
- For a kernel k, a Gaussian process $\mathrm{f}(\cdot) \sim \mathcal{GP}(0, k)$ is a distribution over functions.
- Such that $\mathrm{cov}[\mathrm{f}(a), \mathrm{f}(b)] = k(a, b)$.
- Moreover, for training data $(\boldsymbol{X}, \boldsymbol{y})$, the posterior distribution is also a GP:

$$\mathrm{f}(\cdot) \mid \mathcal{D} \sim \mathcal{GP}\begin{pmatrix} k(\cdot, \boldsymbol{X}) k(\boldsymbol{X})^{-1} \boldsymbol{y}, \\ k - k(\cdot_1, \boldsymbol{X}) k(\boldsymbol{X})^{-1} k(\boldsymbol{X}, \cdot_2) \end{pmatrix}$$



Stationary kernel

$$k(a, b) = k(0, b - a)$$



Isotropic kernel

- A kernel is isotropic stationary with lengthscale $\boldsymbol{\Delta}$ if:

Isotropic kernel

- A kernel is isotropic stationary with lengthscale **Δ** if:

$$k(a,b) = k(a-b,0) \qquad \text{[Stationarity]}$$

Isotropic kernel

- A kernel is isotropic stationary with lengthscale $\mathbf{\Delta}$ if:

$$
\begin{aligned}
k(a,b) &= k(a-b,0) && \text{[Stationarity]} \\
&= \pi_k\big((a-b)^T \mathbf{\Delta}^{-1}(a-b)\big) && \text{[Isotropic]}
\end{aligned}
$$

Isotropic kernel

- A kernel is isotropic stationary with lengthscale **Δ** if:

$$
\begin{aligned}
k(a,b) &= k(a-b,0) && \text{[Stationarity]} \\
&= \pi_k\big((a-b)^T \boldsymbol{\Delta}^{-1}(a-b)\big) && \text{[Isotropic]} \\
&= \pi_k\big((\boldsymbol{W}a-\boldsymbol{W}b)(\boldsymbol{W}a-\boldsymbol{W}b)^T\big)
\end{aligned}
$$

Isotropic kernel

- A kernel is isotropic stationary with lengthscale $\mathbf{\Delta}$ if:

$$
\begin{aligned}
k(a,b) &= k(a-b,0) && \text{[Stationarity]} \\
&= \pi_k\big((a-b)^T \mathbf{\Delta}^{-1}(a-b)\big) && \text{[Isotropic]} \\
&= \pi_k\big((\boldsymbol{W}a - \boldsymbol{W}b)(\boldsymbol{W}a - \boldsymbol{W}b)^T\big)
\end{aligned}
$$

- For example, $k(a,b) = \sigma_f^2 \exp\left[-\frac{1}{2}\sum_i \frac{(a_i - b_i)^2}{\ell_i^2}\right]$, then we have $\pi_k(d^2) = \sigma_f^2 \exp\left[-\frac{1}{2} d^2\right]$ with diagonal $\mathbf{\Delta}_{ii} = \ell_i$.

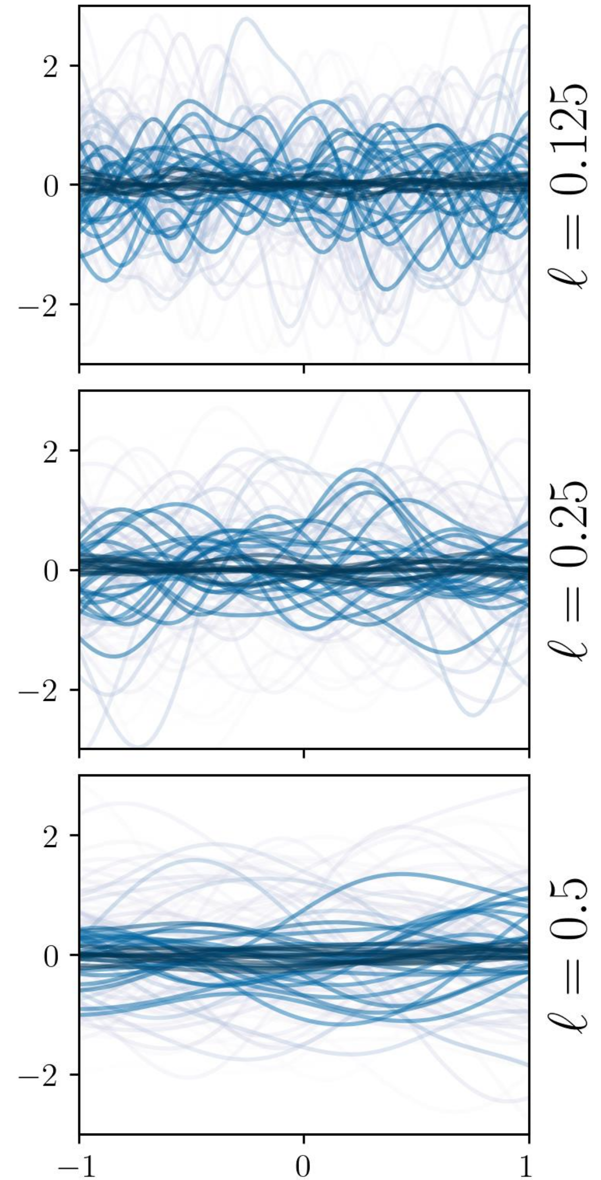
Understanding lengthscales

- Lengthscales control the spatial variance of a Gaussian process;
- For example, with the squared exponential kernel:

$$k_{\mathrm{SE}}(a,b) = \exp\left[-\frac{1}{2}\frac{(a-b)^2}{\ell^2}\right],$$

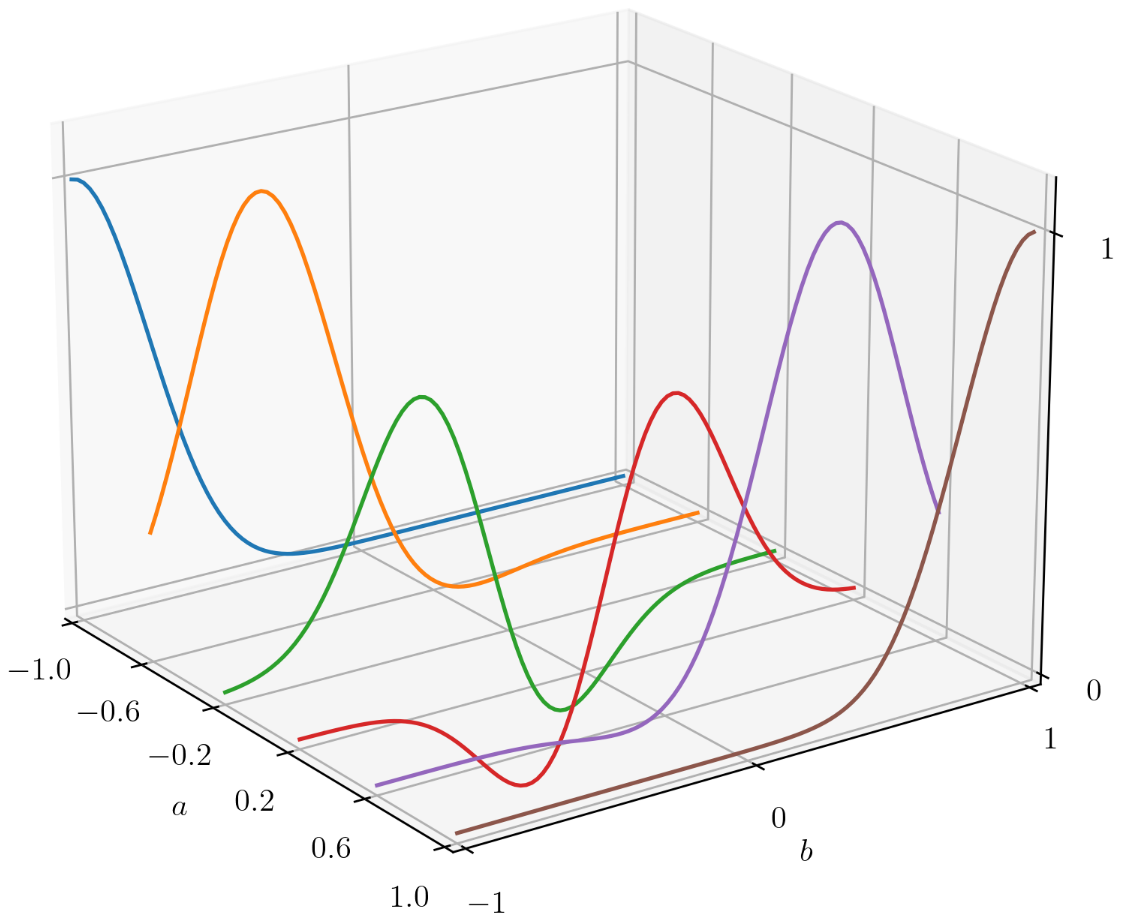
then,

$$\frac{\mathrm{df}}{\mathrm{d}x} \sim \mathcal{GP}\left(0, \frac{1}{\ell^2}\right).$$



Non-stationary kernel

$$k(a,b) \neq k(0, b-a)$$



Non-stationarity from stationary kernels

Compositional kernels

- Let $\tau(\cdot)$ be an arbitrary warping function:

$$k_\tau(a, b) = k\big(\tau(a), \tau(b)\big)$$

Non-stationarity from stationary kernels

Compositional kernels

- Let $\tau(\cdot)$ be an arbitrary warping function:

$$k_\tau(a, b) = k\big(\tau(a), \tau(b)\big)$$

- If $\tau(x) = \ell^{-1} \cdot x$ is a linear function, then k_τ is stationary with lengthscale ℓ.

Non-stationarity from stationary kernels

Compositional kernels

- Let $\tau(\cdot)$ be an arbitrary warping function:
$$k_\tau(a, b) = k\big(\tau(a), \tau(b)\big)$$
- If $\tau(x) = \ell^{-1} \cdot x$ is a linear function, then k_τ is stationary with lengthscale ℓ.
- If $\tau(x)$ is a parametric non-linear function, this corresponds to the deep kernel learning model. [Wilson et al., 2016]

Non-stationarity from stationary kernels

Compositional kernels

- Let $\tau(\cdot)$ be an arbitrary warping function:
$$k_\tau(a,b) = k\big(\tau(a), \tau(b)\big)$$
- If $\tau(x) = \ell^{-1} \cdot x$ is a linear function, then k_τ is stationary with lengthscale ℓ.
- If $\tau(x)$ is a parametric non-linear function, this corresponds to the deep kernel learning model. [Wilson et al., 2016]
- If $\tau(x) \sim \mathcal{GP}(m, k')$, this corresponds to the traditional compositional deep Gaussian process.
[Damianou & Lawrence, 2013; Salimbeni & Deisenroth, 2017a]

Non-stationarity from stationary kernels

Lengthscale mixture kernels

- Let $\mathbf{\Delta}(\cdot)$ be a lengthscale field:

$$\sqrt{\frac{\sqrt{|\boldsymbol{\Delta}(a)|}\sqrt{|\boldsymbol{\Delta}(b)|}}{|\boldsymbol{\Delta}(a)+\boldsymbol{\Delta}(b)|}}\pi_k\left((a-b)^{\mathrm{T}}\left[\frac{\boldsymbol{\Delta}(a)+\boldsymbol{\Delta}(b)}{2}\right]^{-1}(a-b)\right)$$

Non-stationarity from stationary kernels

Lengthscale mixture kernels

- Let $\mathbf{\Delta}(\cdot)$ be a lengthscale field:

$$\sqrt{\frac{\sqrt{|\boldsymbol{\Delta}(a)|}\sqrt{|\boldsymbol{\Delta}(b)|}}{|\boldsymbol{\Delta}(a)+\boldsymbol{\Delta}(b)|}}\,\pi_k\left((a-b)^{\mathrm{T}}\left[\frac{\boldsymbol{\Delta}(a)+\boldsymbol{\Delta}(b)}{2}\right]^{-1}(a-b)\right)$$

- If k is squared exponential, this is the Gibbs' kernel. [Gibbs, 1997]

Non-stationarity from stationary kernels

Lengthscale mixture kernels

- Let $\boldsymbol{\Delta}(\cdot)$ be a lengthscale field:

$$\sqrt{\frac{\sqrt{|\boldsymbol{\Delta}(a)|}\sqrt{|\boldsymbol{\Delta}(b)|}}{|\boldsymbol{\Delta}(a)+\boldsymbol{\Delta}(b)|}}\,\pi_k\left((a-b)^{\mathrm{T}}\left[\frac{\boldsymbol{\Delta}(a)+\boldsymbol{\Delta}(b)}{2}\right]^{-1}(a-b)\right)$$

- If k is squared exponential, this is the Gibbs' kernel. [Gibbs, 1997]

- If, $w\big(\boldsymbol{\Delta}(\cdot)\big) \sim \mathcal{GP}(0, k)$, for a warping function $w(\cdot)$, we obtain a deep non-stationary model.
  [Paciorek & Schervish, 2013; Salimbeni & Deisenroth, 2017b]

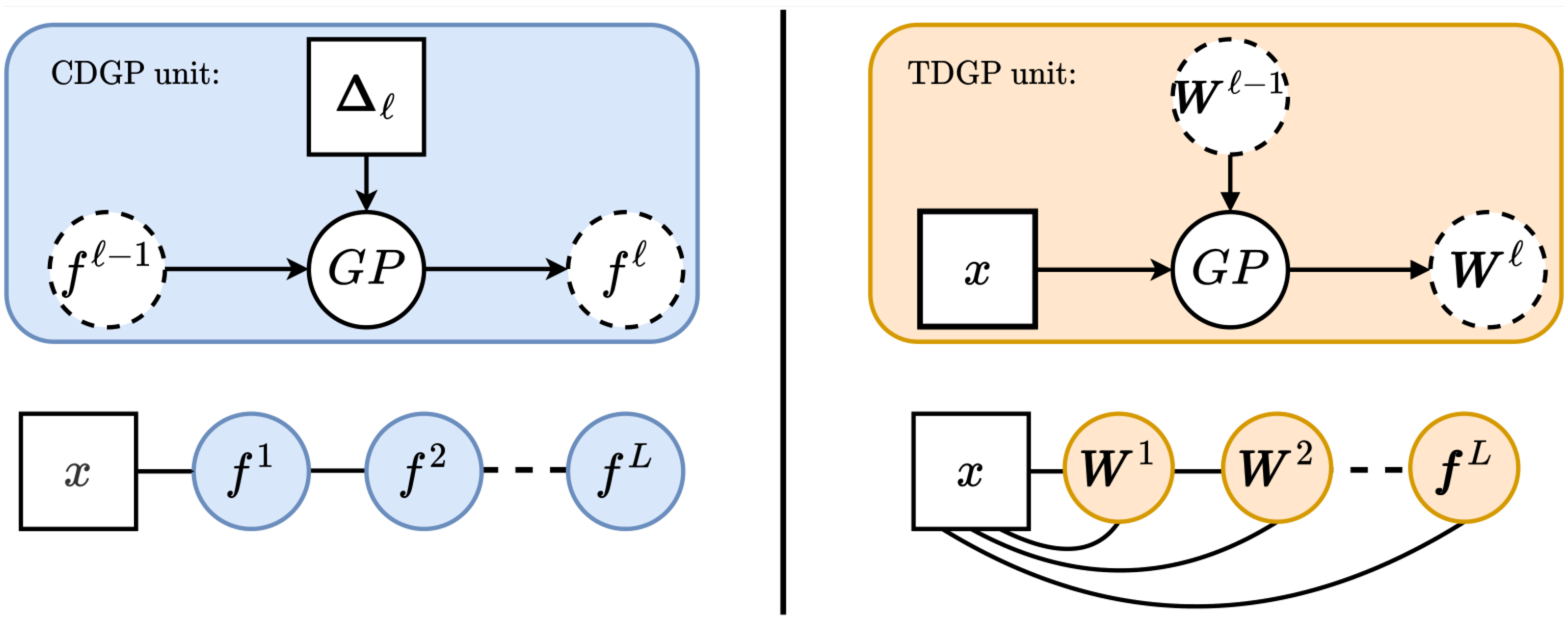
Our proposal

Kernel construction

- We choose a hybrid approach:

$$k(\boldsymbol{W}(a) \cdot a, \boldsymbol{W}(b) \cdot b\ )$$

- Defines a latent space $\tau(x) = \boldsymbol{W}(x) \cdot x$.
- Induces a lengthscale field, $\boldsymbol{\Delta}(x) = [\boldsymbol{W}(x)\boldsymbol{W}(x)^T]^{-\mathbf{1}}$



Our proposal

Variational inference

- As a deep GP, inference must be approximate; Extending the approach of Titsias & Lázaro-Gredilla (2013), our variational distribution for a two-layer model is:

$$p(f \mid u)\mathrm{N}(u \mid \mu_u, \Sigma_u) \prod_{q,d}^{Q,D} p\left(w_{qd} | v_{qd}\right)\mathrm{N}\left(v_{qd} \mid \mu_{v_{qd}}, \Sigma_{v_{qd}}\right)$$

Our proposal

Variational inference

- As a deep GP, inference must be approximate; Extending the approach of Titsias & Lázaro-Gredilla (2013), our variational distribution for a two-layer model is:

$$p(f \mid u)\mathrm{N}(u \mid \mu_u, \Sigma_u) \prod_{q,d}^{Q,D} p\left(w_{qd} | v_{qd}\right)\mathrm{N}\left(v_{qd} \mid \mu_{v_{qd}}, \Sigma_{v_{qd}}\right)$$

- Additionally, to compute the ELBO, the Ψ-statistics need to be computed:

$$[\boldsymbol{\Psi}_1]_{ij} = \int k\left(\boldsymbol{W}(\boldsymbol{x}_i) \cdot \boldsymbol{x}_i, \boldsymbol{z}_j\right) q(\boldsymbol{W}) \, \mathrm{d}\boldsymbol{W}$$

Our proposal

Variational inference

- As a deep GP, inference must be approximate; Extending the approach of Titsias & Lázaro-Gredilla (2013), our variational distribution for a two-layer model is:

$$p(f \mid u)\mathrm{N}(u \mid \mu_u, \Sigma_u) \prod_{q,d}^{Q,D} p(w_{qd} | v_{qd})\mathrm{N}\left(v_{qd} \mid \mu_{v_{qd}}, \Sigma_{v_{qd}}\right)$$

- Additionally, to compute the ELBO, the Ψ-statistics need to be computed:

$$[\boldsymbol{\Psi}_1]_{ij} = \int k\left(\boldsymbol{W}(\boldsymbol{x}_i) \cdot \boldsymbol{x}_i, \boldsymbol{z}_j\right) q(\boldsymbol{W}) \,\mathrm{d}\boldsymbol{W}$$

So, we restrict $k(a, b)$ to the squared exponential kernel and obtain closed form solutions to Ψ-statistics.

Our proposal

Variational inference

- As a deep GP, inference must be approximate; Extending the approach of Titsias & Lázaro-Gredilla (2013), our variational distribution for a two-layer model is:

$$p(f \mid u)\mathrm{N}(u \mid \mu_u, \Sigma_u) \prod_{q,d}^{Q,D} p(w_{qd} | v_{qd}) \mathrm{N}\left(v_{qd} \mid \mu_{v_{qd}}, \Sigma_{v_{qd}}\right)$$

- Additionally, to compute the ELBO, the Ψ-statistics need to be computed:

$$[\boldsymbol{\Psi}_1]_{ij} = \int k\left(\boldsymbol{W}(\boldsymbol{x}_i) \cdot \boldsymbol{x}_i, \boldsymbol{z}_j\right) q(\boldsymbol{W})\, \mathrm{d}\boldsymbol{W}$$

  So, we restrict $k(a, b)$ to the squared exponential kernel and obtain closed form solutions to Ψ-statistics.

  - As an alternative, doubly stochastic inference doesn't require these assumptions.

Bathymetry case study

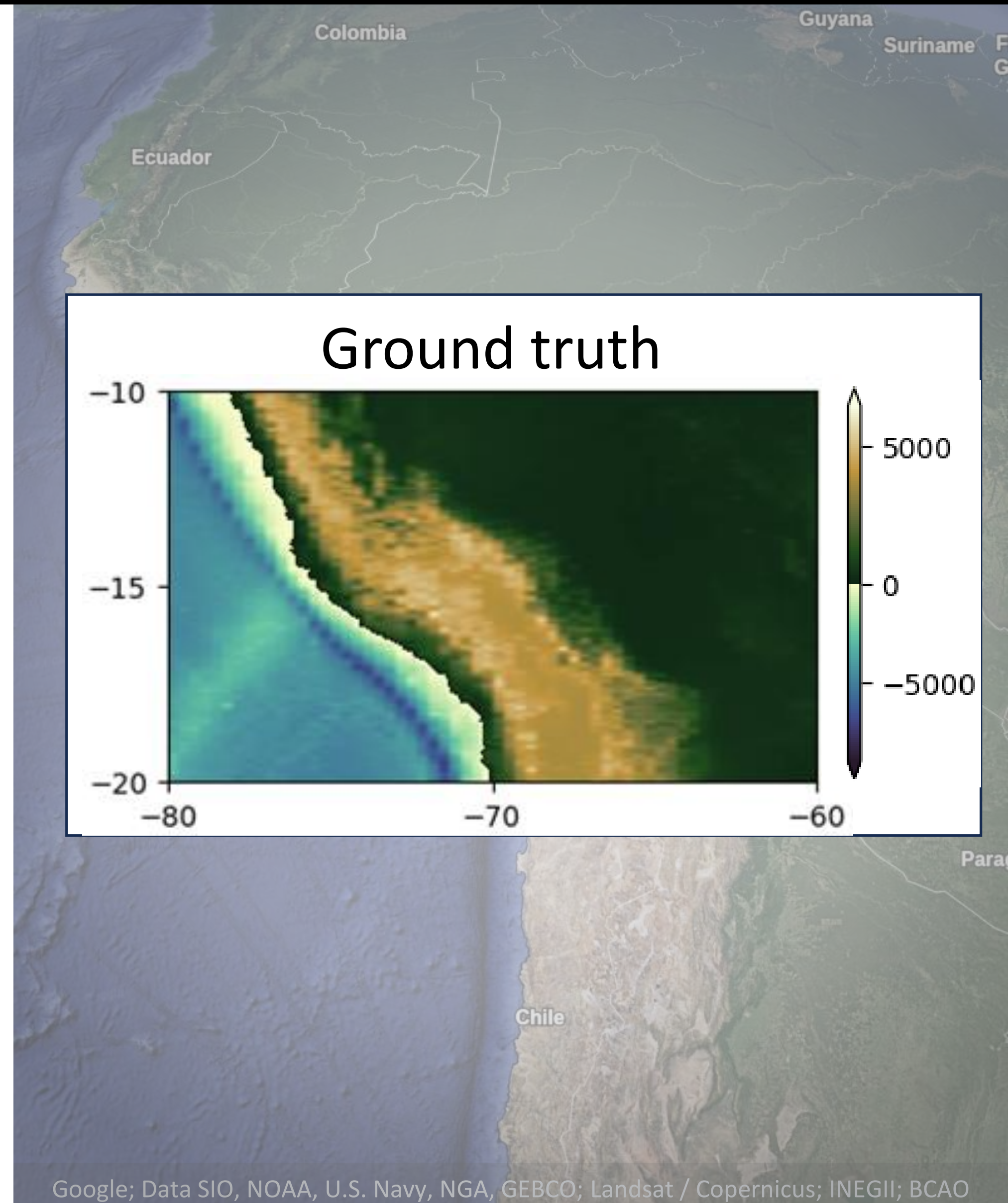
- As a case-study, we also apply TDGP to the GEBCO gridded bathymetry dataset. It contains a global terrain model (elevation data) for ocean and land.

Bathymetry case study

- As a case-study, we also apply TDGP to the GEBCO gridded bathymetry dataset. It contains a global terrain model (elevation data) for ocean and land.
- As an example of a non-stationary task, we selected an especially challenging subset of the data covering the Andes mountain range, ocean, and land.

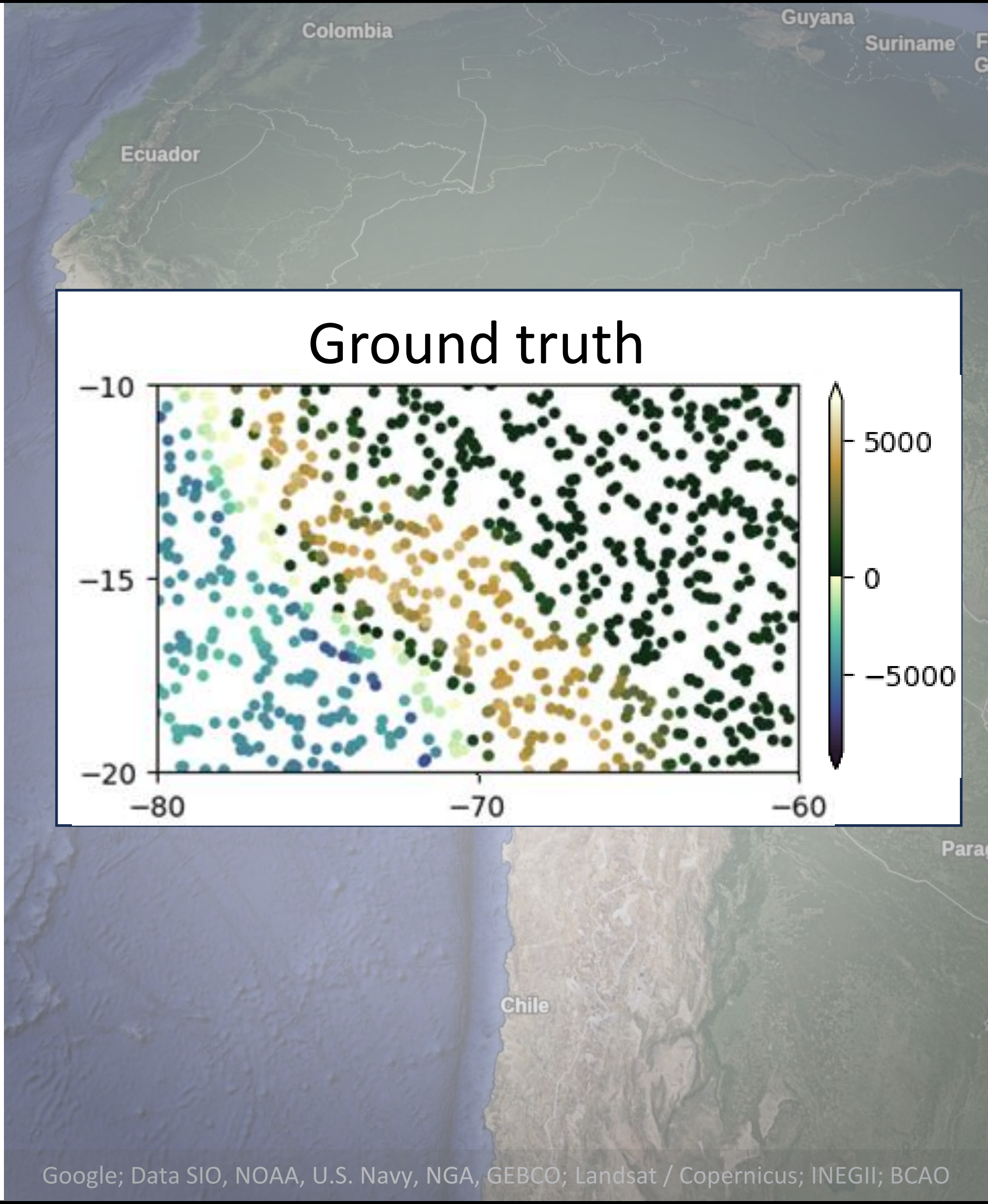
Bathymetry case study

- As a case-study, we also apply TDGP to the GEBCO gridded bathymetry dataset. It contains a global terrain model (elevation data) for ocean and land.
- As an example of a non-stationary task, we selected an especially challenging subset of the data covering the Andes mountain range, ocean, and land.



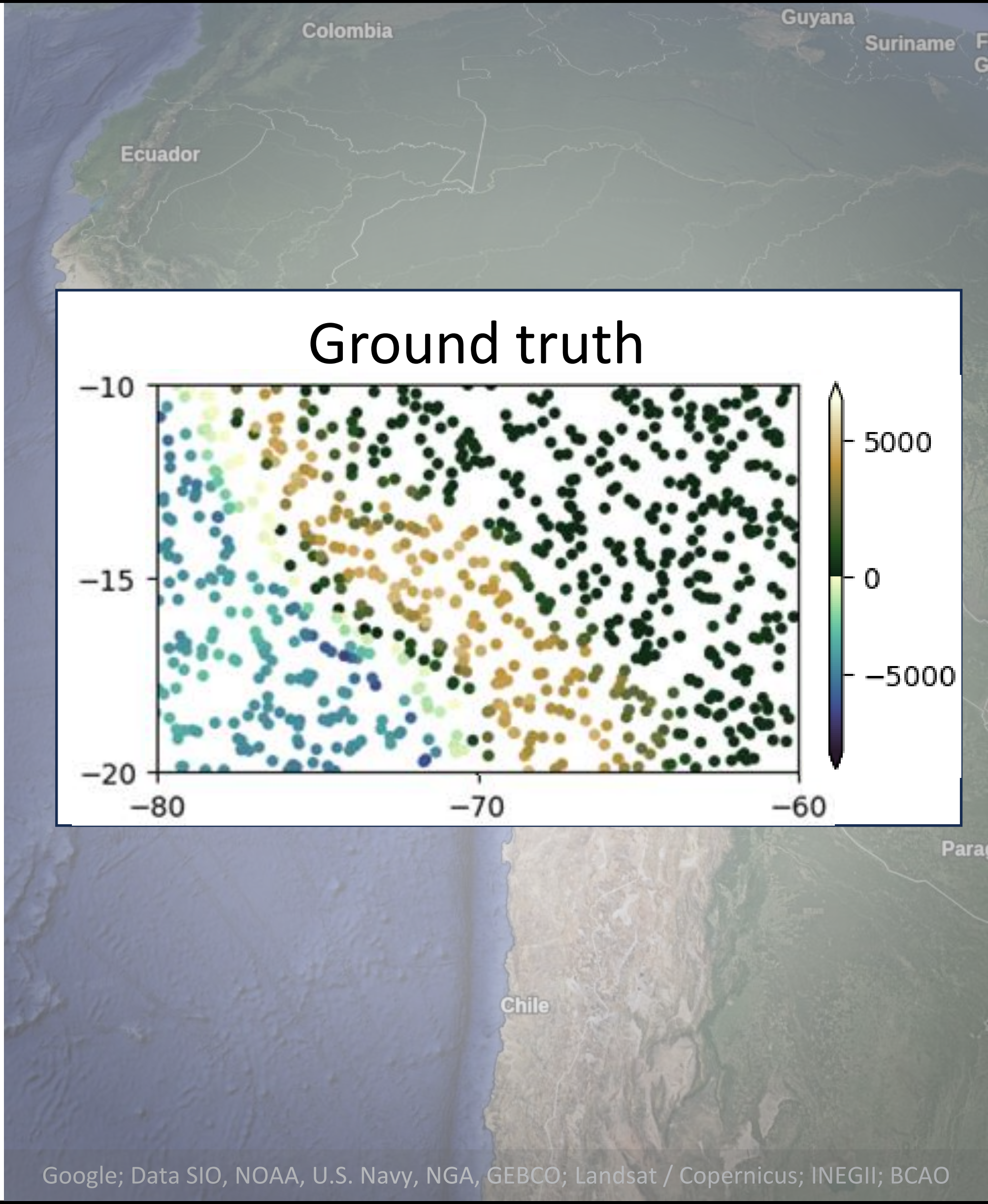
Bathymetry case study

- As a case-study, we also apply TDGP to the GEBCO gridded bathymetry dataset. It contains a global terrain model (elevation data) for ocean and land.
- As an example of a non-stationary task, we selected an especially challenging subset of the data covering the Andes mountain range, ocean, and land.
- This region was subsampled to 1,000 points from this region and compared with the methods via five-fold cross-validation.



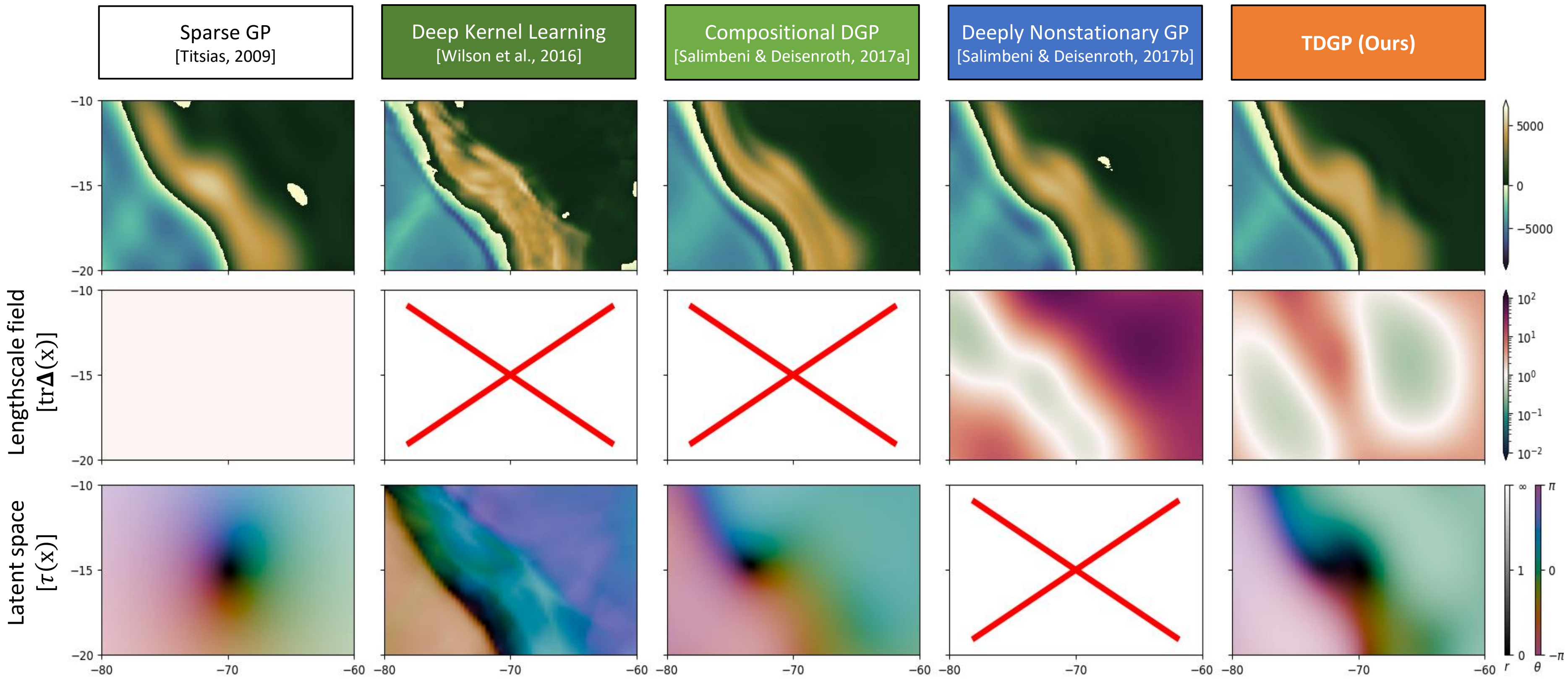
Bathymetry case study

- As a case-study, we also apply TDGP to the GEBCO gridded bathymetry dataset. It contains a global terrain model (elevation data) for ocean and land.
- As an example of a non-stationary task, we selected an especially challenging subset of the data covering the Andes mountain range, ocean, and land.
- This region was subsampled to 1,000 points from this region and compared with the methods via five-fold cross-validation.
- We compare our method against popular inference methods of the previous alternatives.



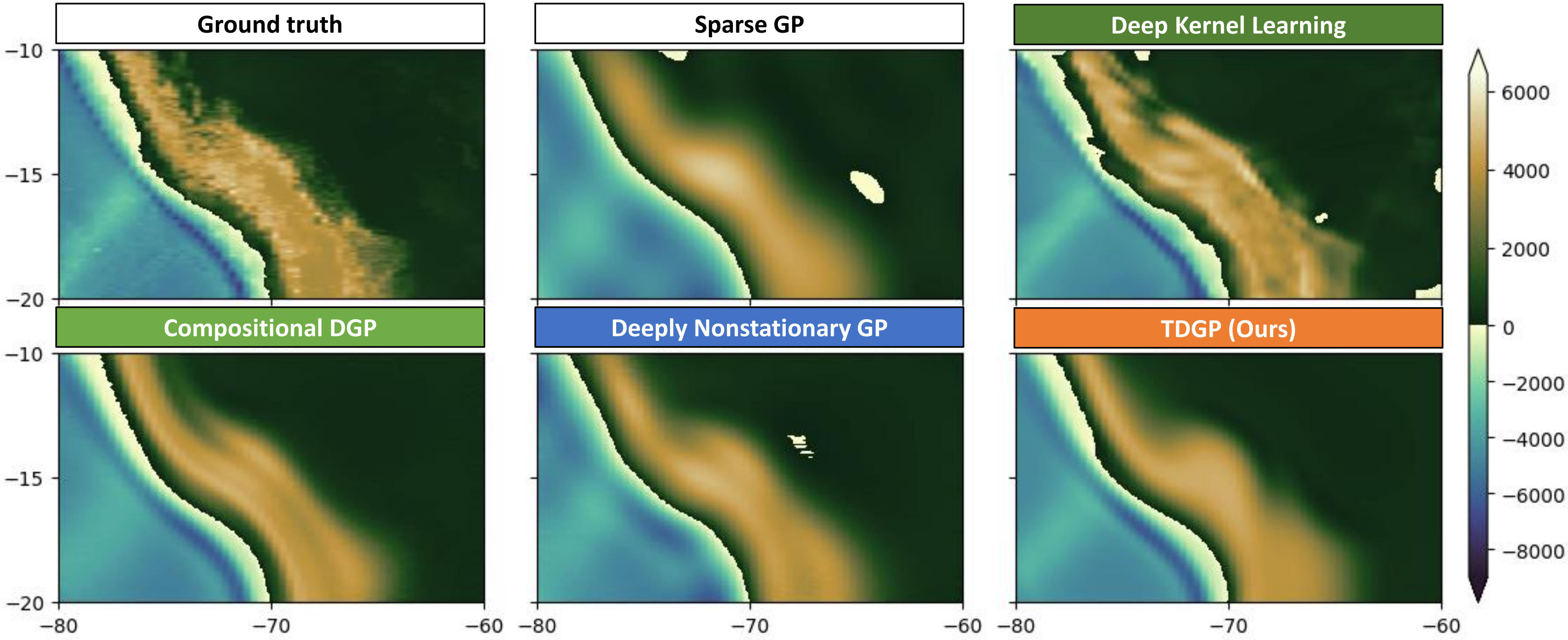
Bathymetry case study

Results



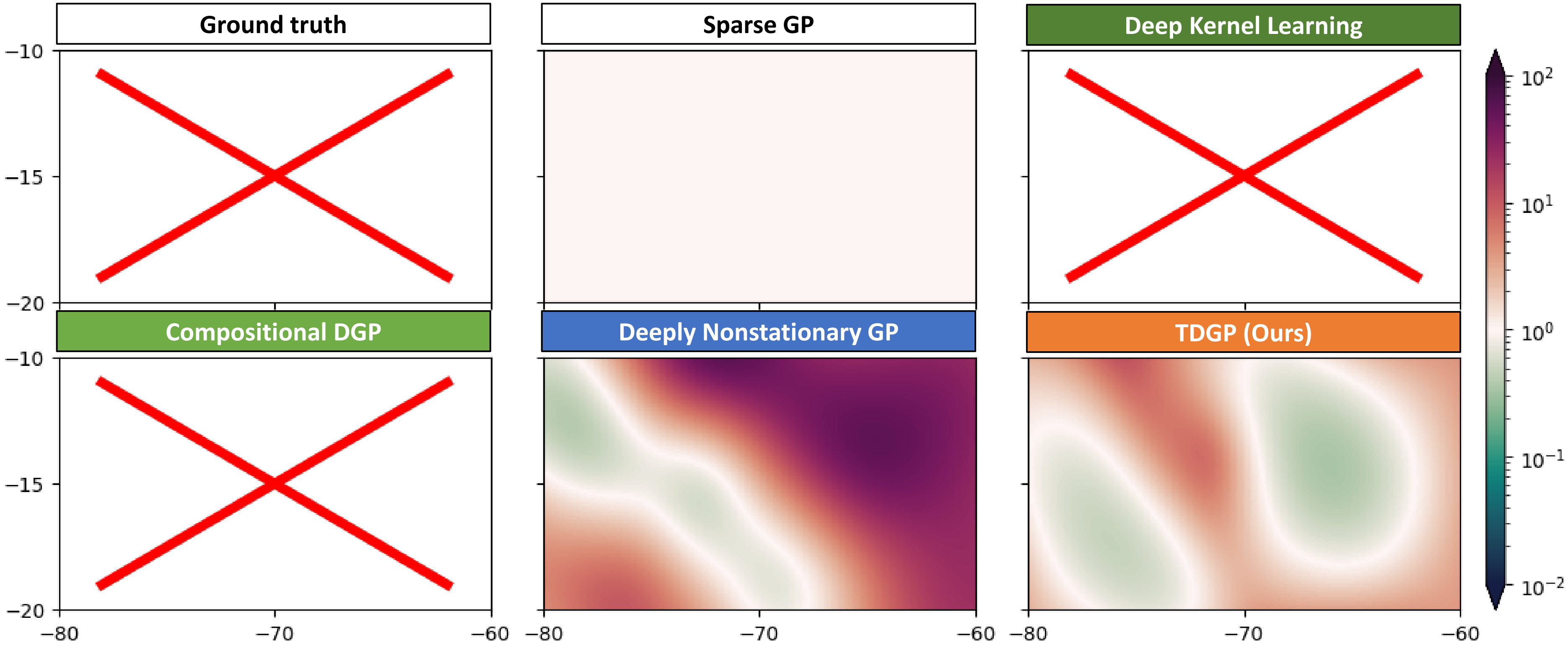
Bathymetry case study

Results – Bathymetry (m)



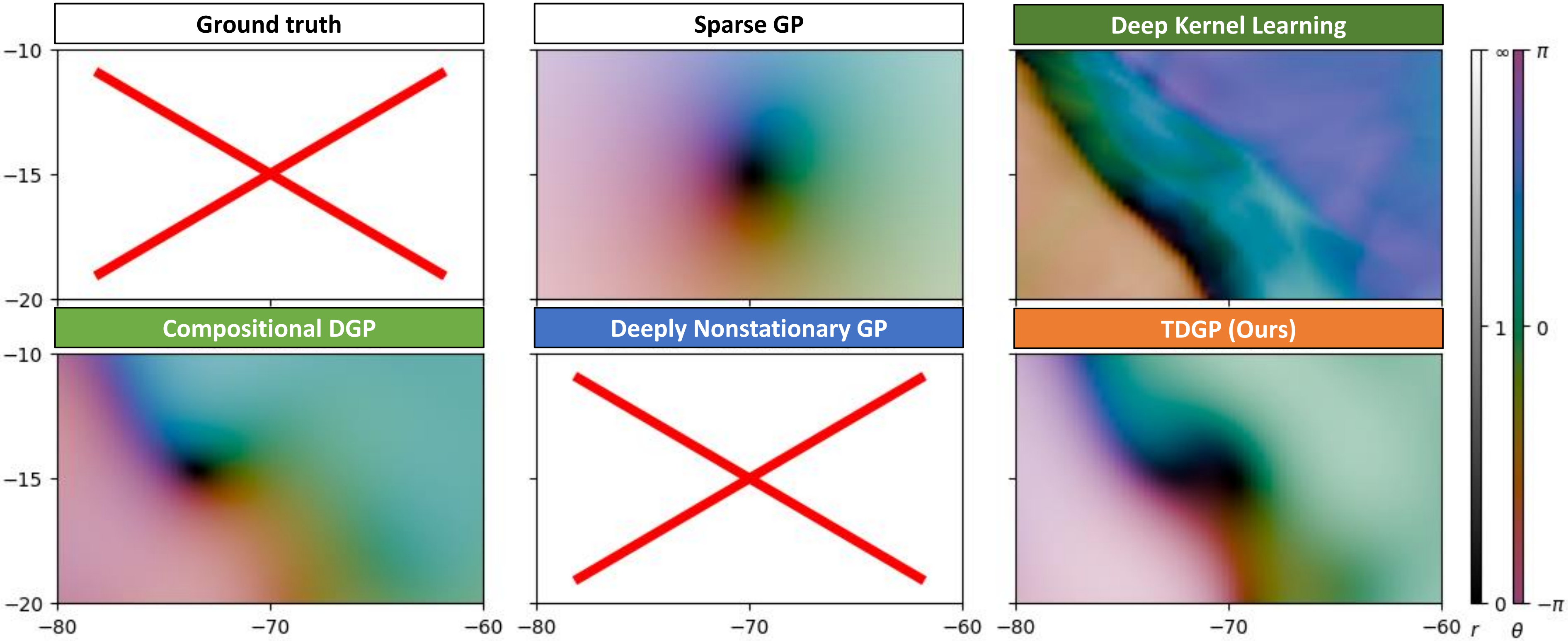
Bathymetry case study

Results – Lengthscale field [$\mathrm{tr}\Delta(\mathrm{x})$]



Bathymetry case study

Results – Latent space $[\boldsymbol{\tau}(\mathrm{x})]$



Bathymetry case study

Results – Test metrics

	NLPD	MRAE
Sparse GP	-0.13 ± 0.09	1.19 ± 0.63
Deep Kernel Learning	3.85 ± 0.92	**0.59 ± 0.31**
Compositional DGP	-0.44 ± 0.12	0.83 ± 0.56
Deeply Nonstationary GP	-0.31 ± 0.12	1.12 ± 0.75
TDGP (Ours)	**-0.53 ± 0.10**	0.66 ± 0.43

References

1. Gibbs, Mark N.
   "Bayesian Gaussian Processes for Regression and Classification" (1997)

2. Paciorek, Christopher J. & Schervish, Mark J.
   "Nonstationary Covariance Functions for Gaussian Process Regression" (2003)

3. Titsias, Michalis K.
   "Variational Learning of Inducing Variables in Sparse Gaussian Processes" (2009)

4. Damianou, Andreas C. & Lawrence, Neil D.
   "Deep Gaussian Processes" (2013)

5. Titsias, Michalis K. & Lázaro-Gredilla, Miguel.
   "Variational Inference for Mahalanobis Distance Metrics in Gaussian Process Regression" (2013)

6. Wilson, Andrew Gordon & Zhiting Hu & Ruslan Salakhutdinov & Eric P. Xing.
   "Stochastic Variational Deep Kernel Learning" (2016)

7. Salimbeni, Hugh & Deisenroth, Marc Peter.
   "Doubly Stochastic Variational Inference for Deep Gaussian Processes" (2017a)

8. Salimbeni, Hugh & Deisenroth, Marc Peter.
   "Deeply Non-Stationary Gaussian Processes" (2017b)

Thank you!