Learned wave-based imaging – variational inference at scale

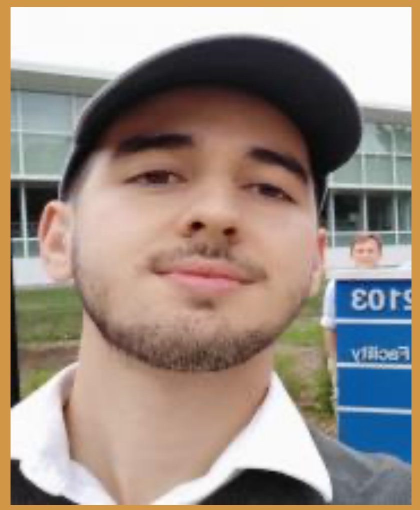
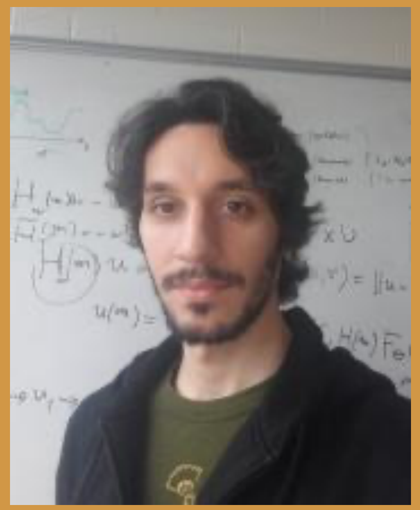
Felix J. Herrmann, June 29, 2021

Delft University of Technology

SLIM
Georgia Institute of Technology

Learned wave-based imaging – variational inference at scale

Ali Siahkoohi, Rafael Orozco, Gabrio Rizutti*, Philipp Witte, and Mathias Louboutin**

* now at Utrecht University
** now at Microsoft Research

SLIM
Georgia Institute of Technology

SLIM

Research program

computational imaging @SLIM

Interested in

- solving imaging & monitoring problems that involve physics (PDEs, waves)
- tackling challenges that come w/ 3D
- seismic & medical imaging

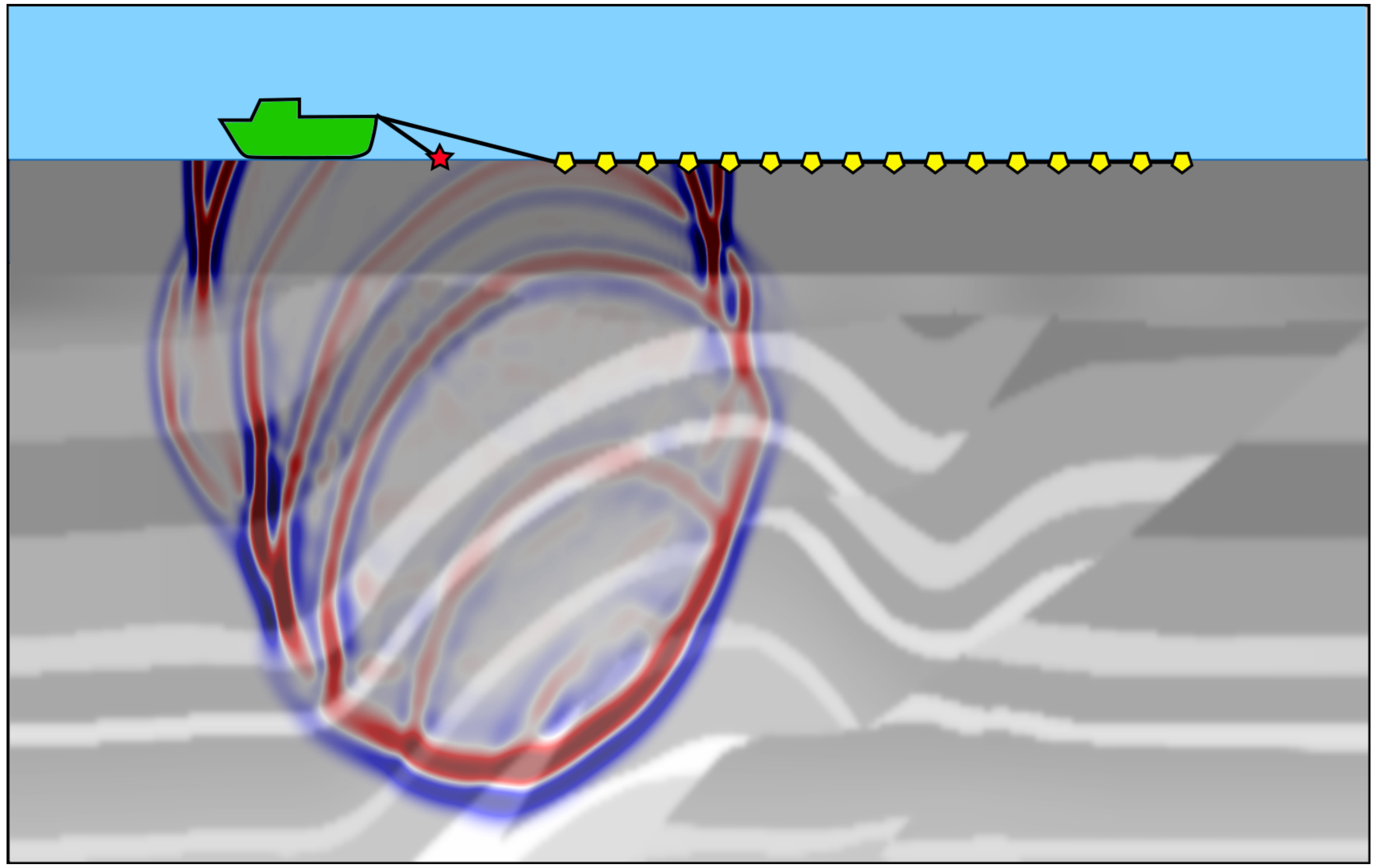
Use techniques from

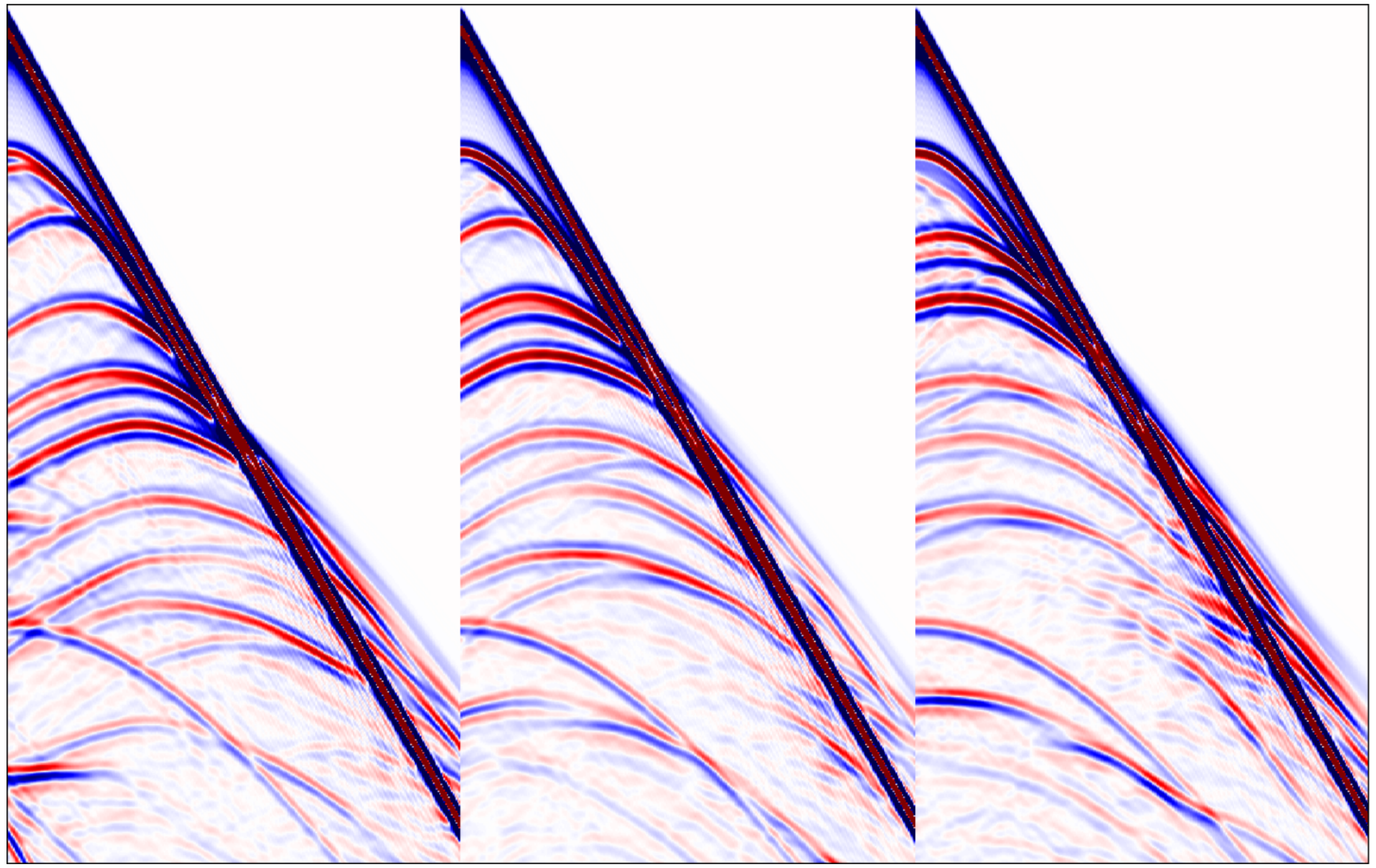
- high performance computing
- randomized linear algebra & stochastic optimization
- machine learning & Bayesian inference

Seismic inverse problems

Estimate unknown subsurface properties from seismic (surface) data:

- image geological structures/discontinuities
- estimate physical rock properties (wavespeed, density, etc.)
- turn around times order 1 year

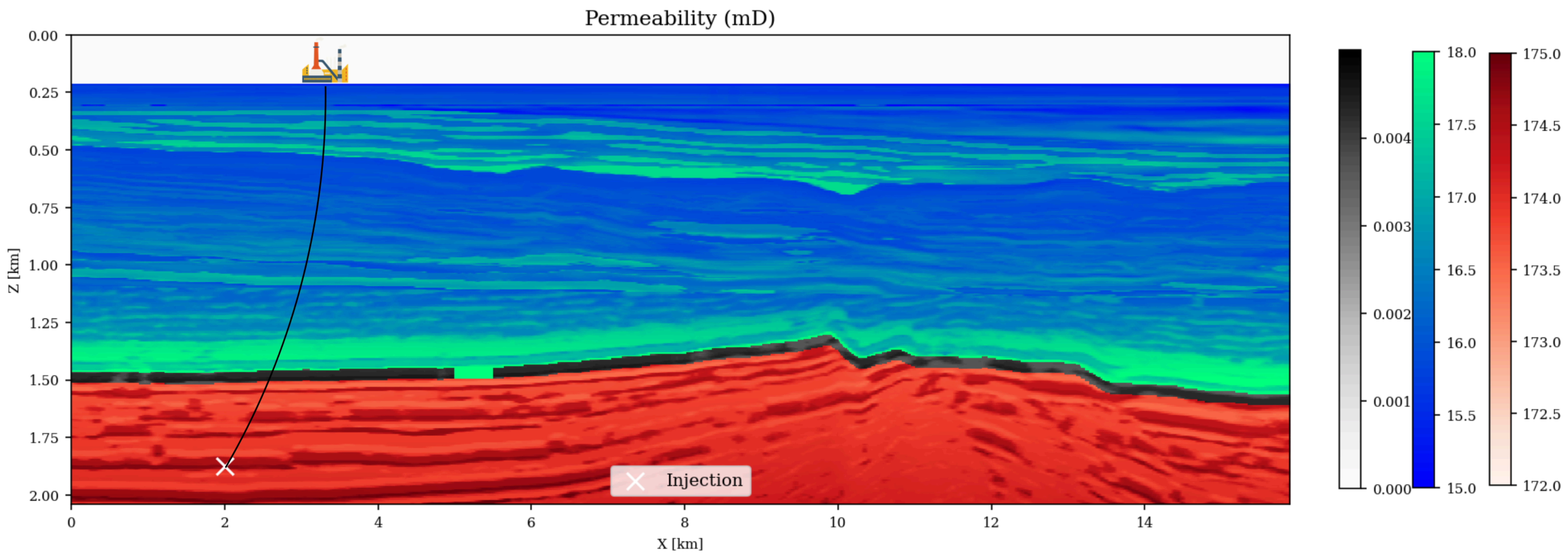


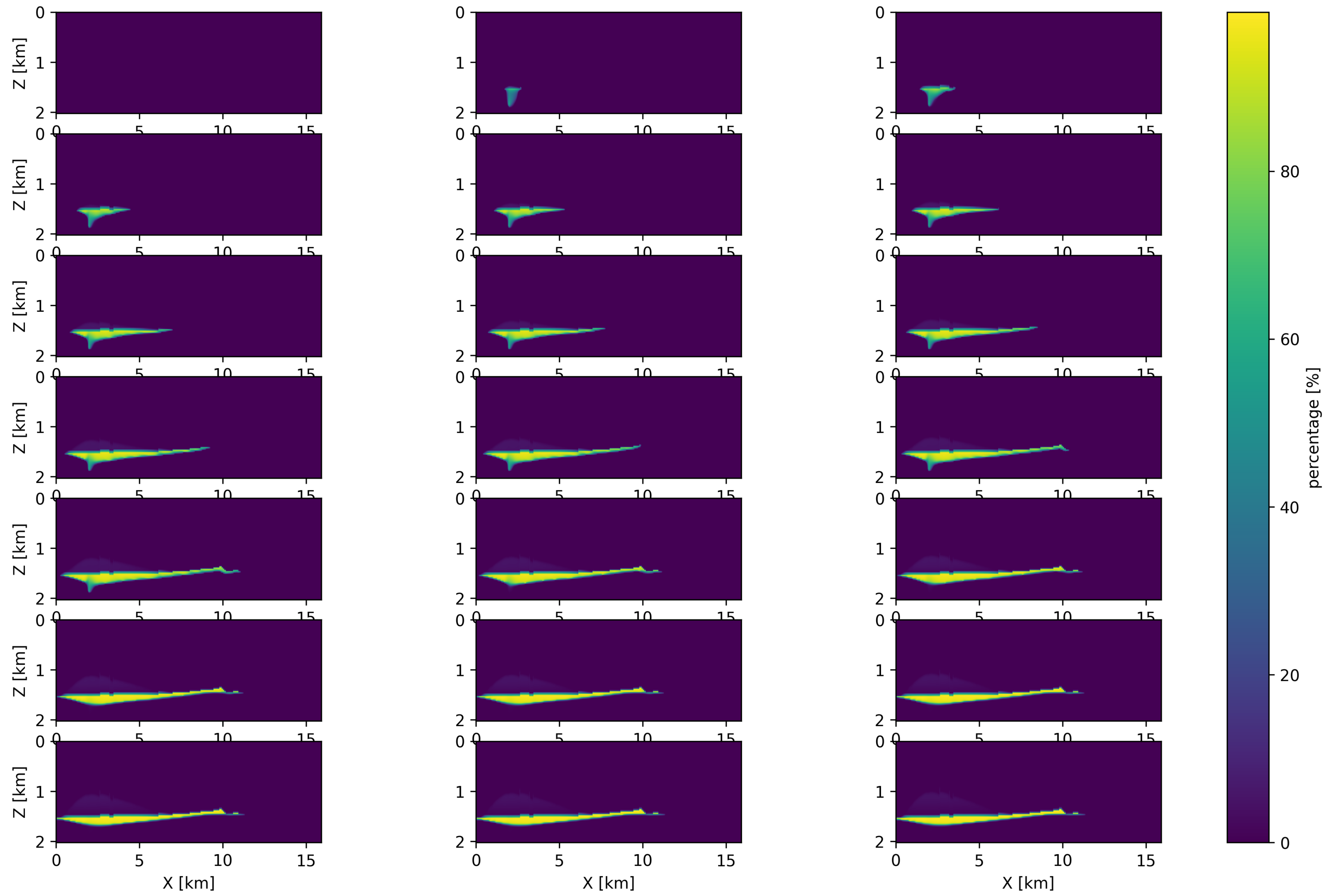


Li, Dongzhuo, et al. "Coupled Time–Lapse Full–Waveform Inversion for Subsurface Flow Problems Using Intrusive Automatic Differentiation." *Water Resources Research* 56.8 (2020): e2019WR027032.
Li, Zongyi, et al. "Fourier neural operator for parametric partial differential equations." arXiv preprint arXiv:2010.08895 (2020).
Kolster, Clea, et al. "The impact of time-varying CO2 injection rate on large scale storage in the UK Bunter Sandstone." International Journal of Greenhouse Gas Control 68 (2018): 77-85.

Monitoring CO2 plumes

https://kailaix.github.io/ADCME.jl/dev/apps_ad/



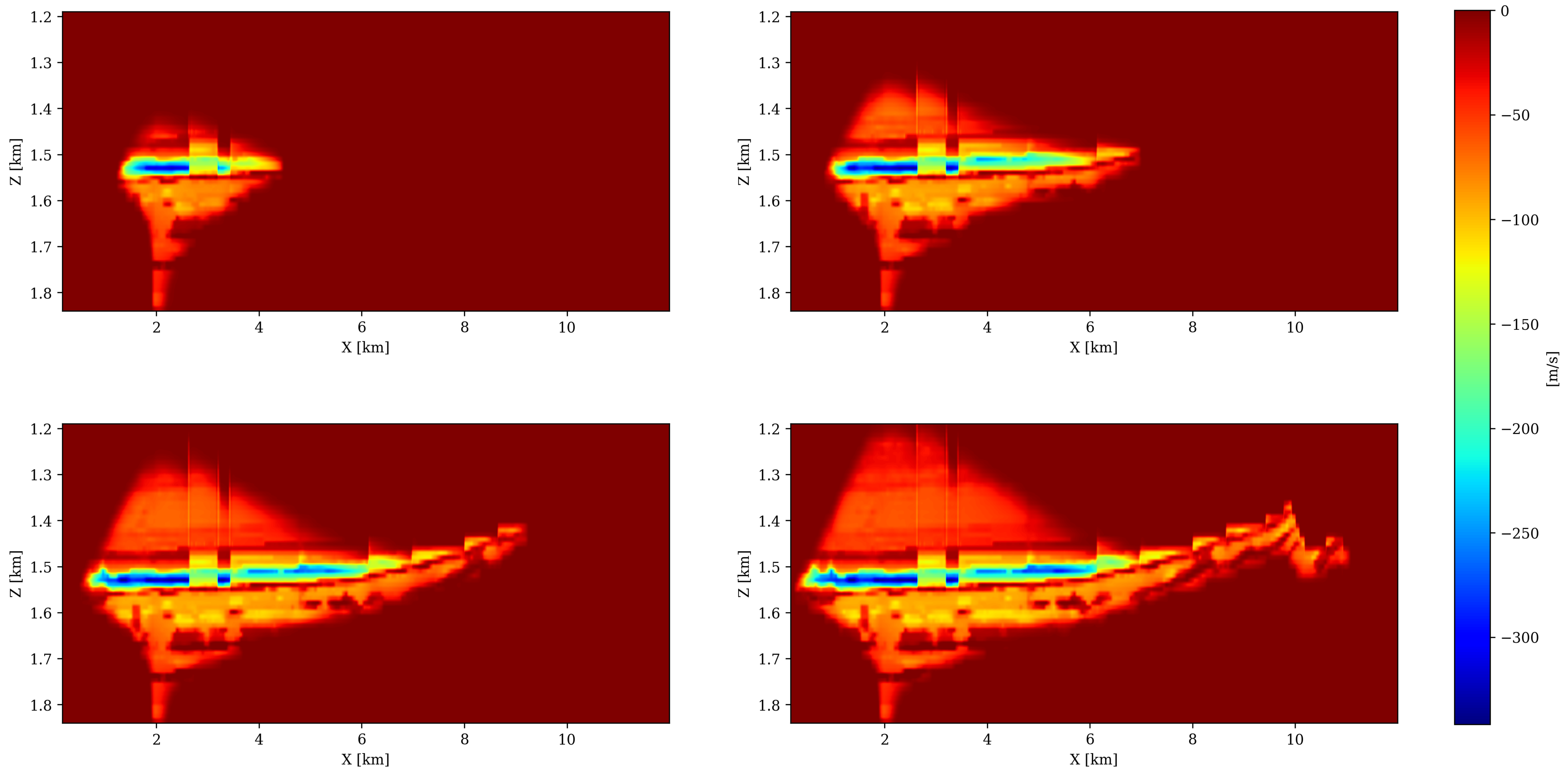


Simulations 2 phase flow of CO_2

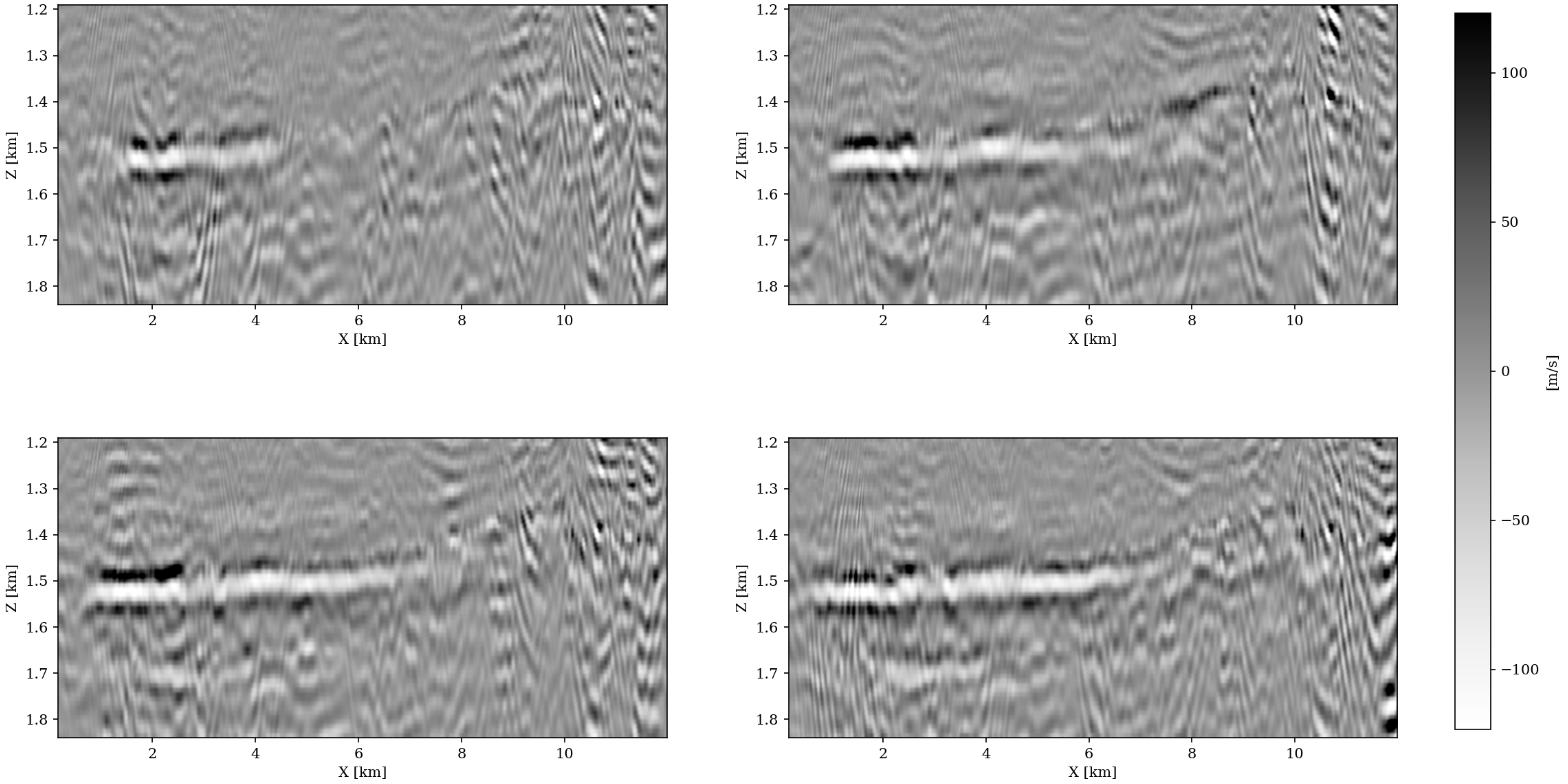
- monitor w/ time-lapse seismic imaging
- optimize sequestration & lower cost & risk

Li, Dongzhuo, et al. "Coupled Time−Lapse Full−Waveform Inversion for Subsurface Flow Problems Using Intrusive Automatic Differentiation." *Water Resources Research* 56.8 (2020): e2019WR027032.

Converted to wave speeds



Joint recovery model

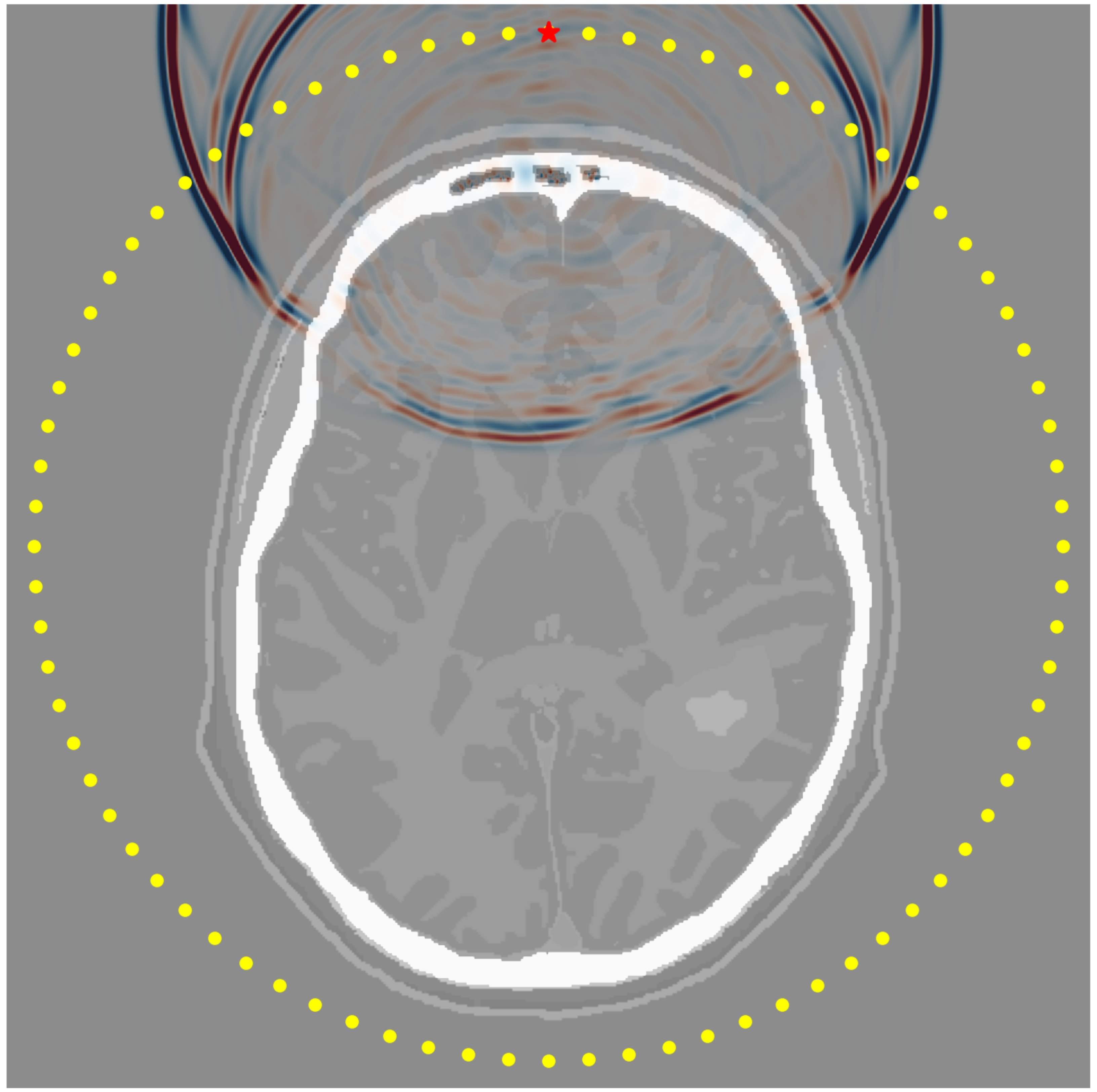


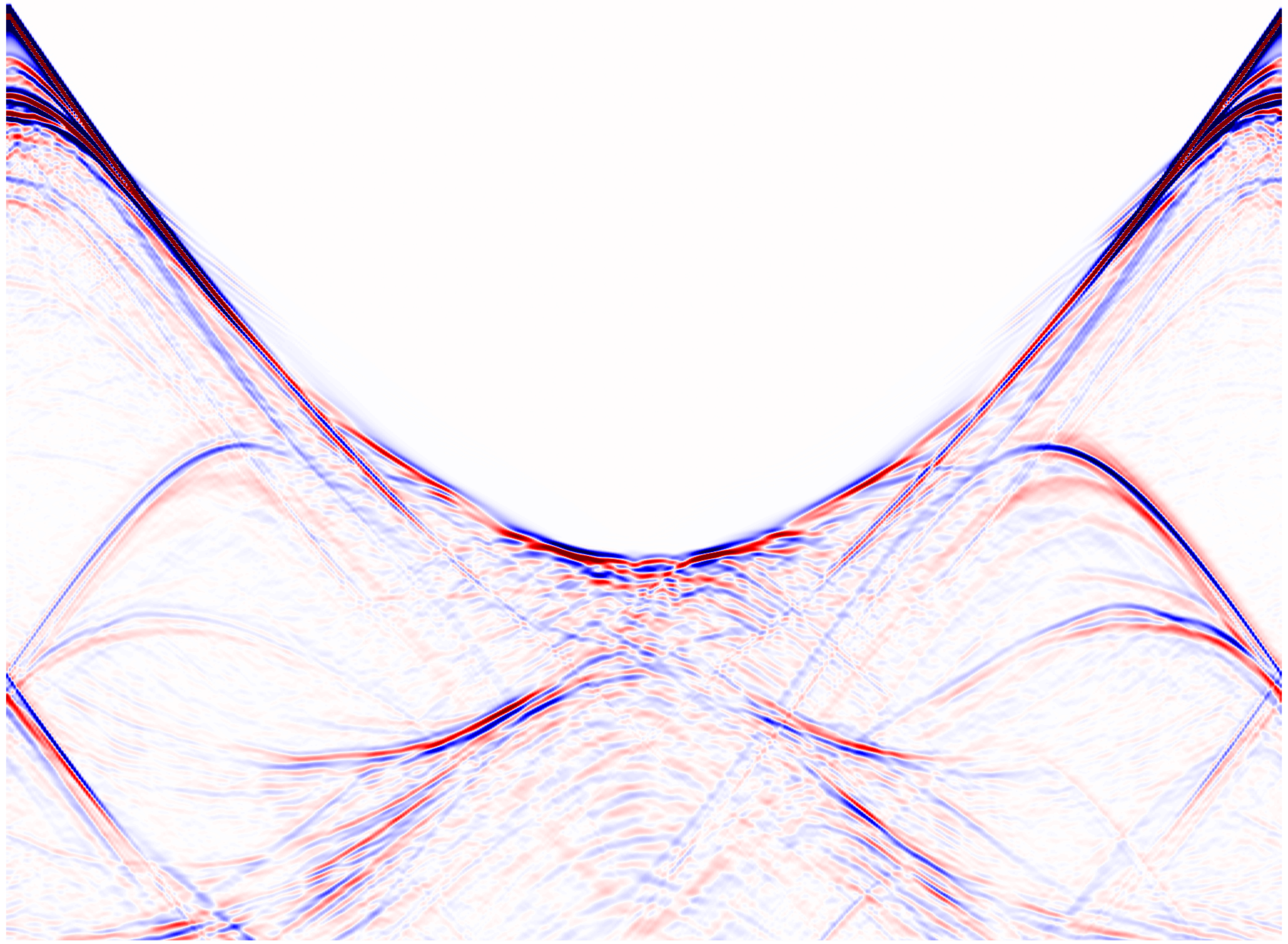
[1] Iacono, Maria Ida, et al. "MIDA: a multimodal imaging-based detailed anatomical model of the human head and neck." *PloS one* 10.4 (2015)
[2] Guasch, Lluís, et al. "Full-waveform inversion imaging of the human brain." *NPJ digital medicine* 3.1 (2020): 1-12.

Medical inverse problems

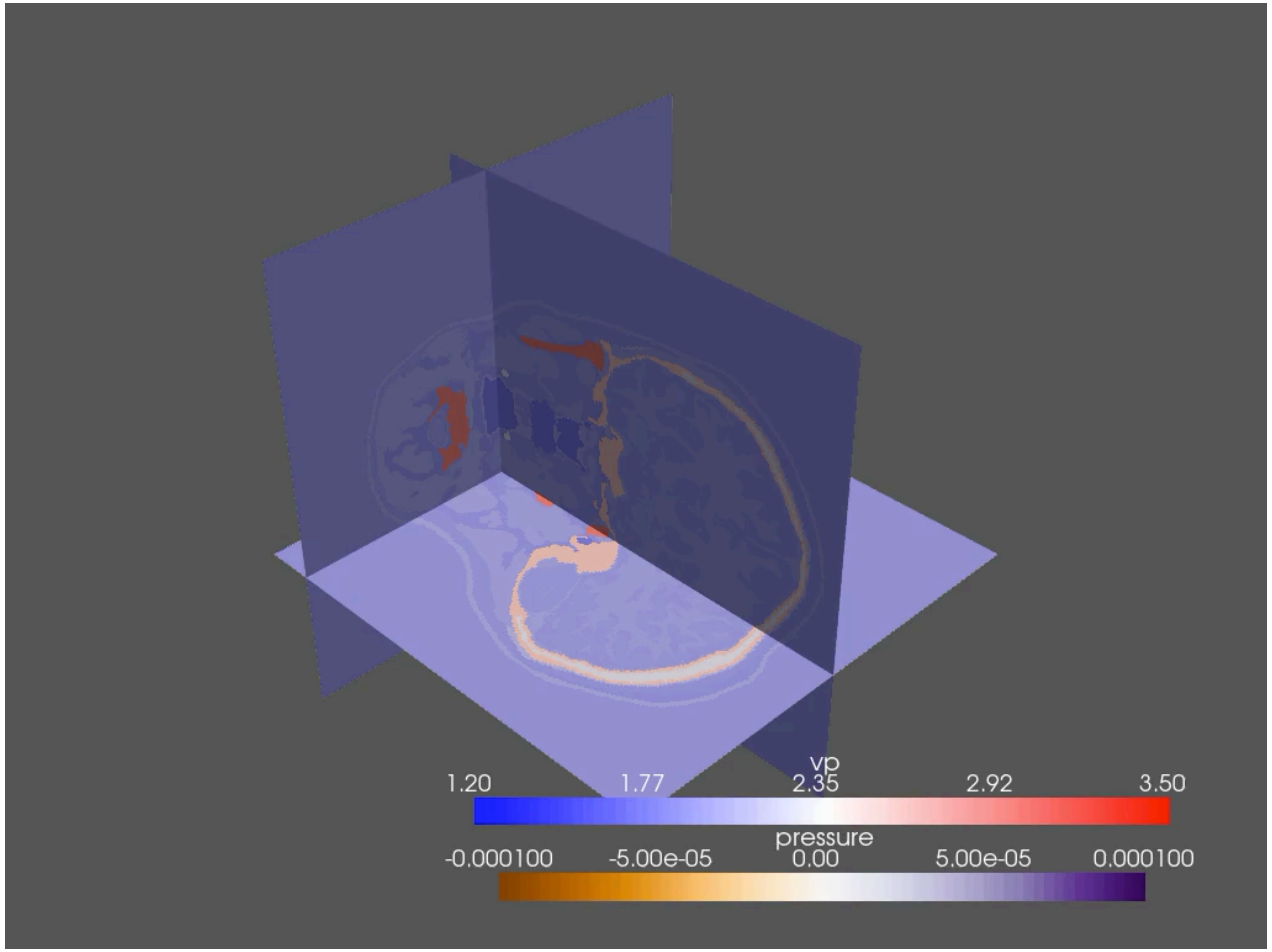
Estimate unknown tissue properties from (omnidirectional) ultrasound data:

- image healthy & malignant tissue
- estimate physical tissue properties (wavespeed, density, etc.)
- turn around times seconds to minutes

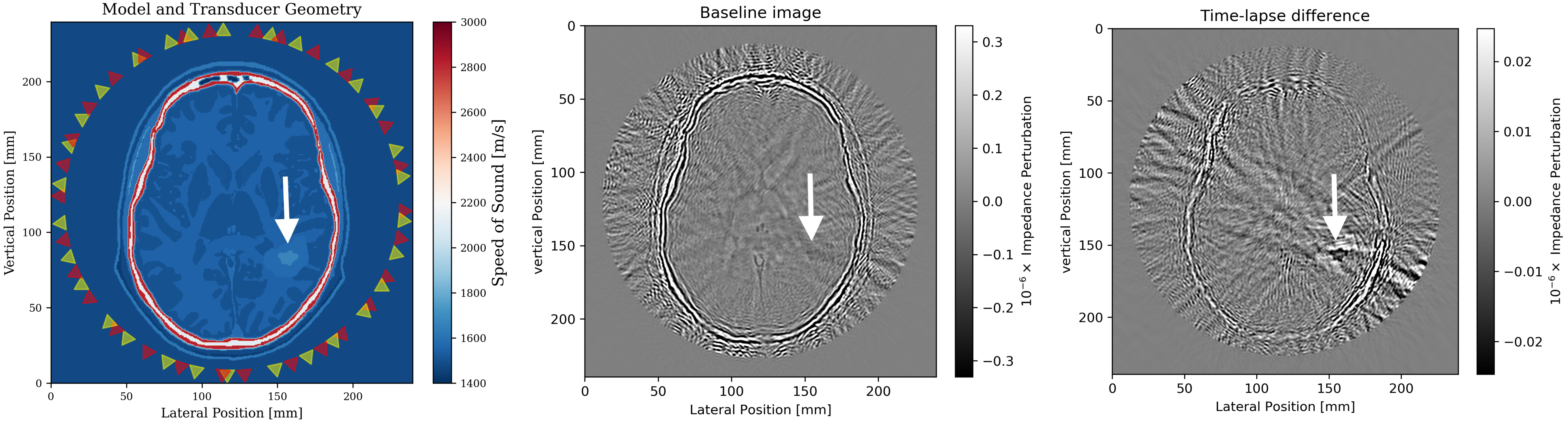




3D wave simulation



Monitoring thermal ablation



Forward problem

The *forward* problem:

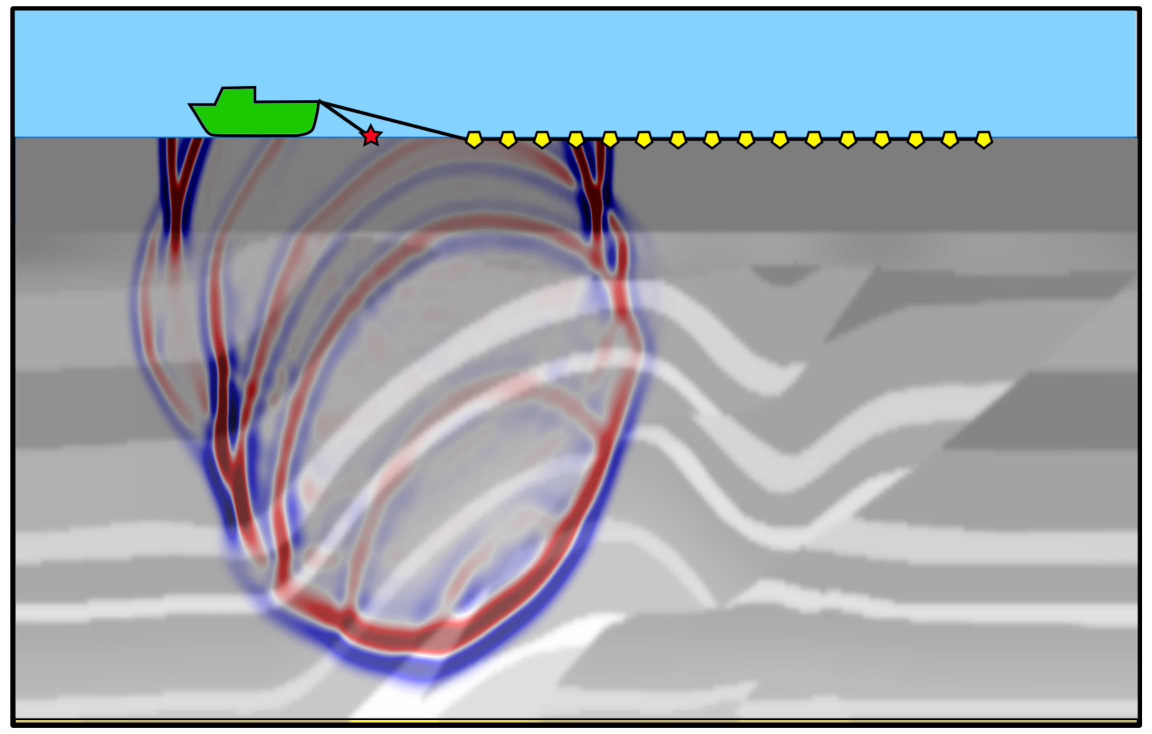
$$\mathbf{F}[\mathbf{m}] \cdot \mathbf{q} = \mathbf{P}_r \underbrace{\mathbf{A}[\mathbf{m}]^{-1} \mathbf{P}_s^\top \mathbf{q}}_{\mathbf{u}}$$

Discretized acoustic wave equation:

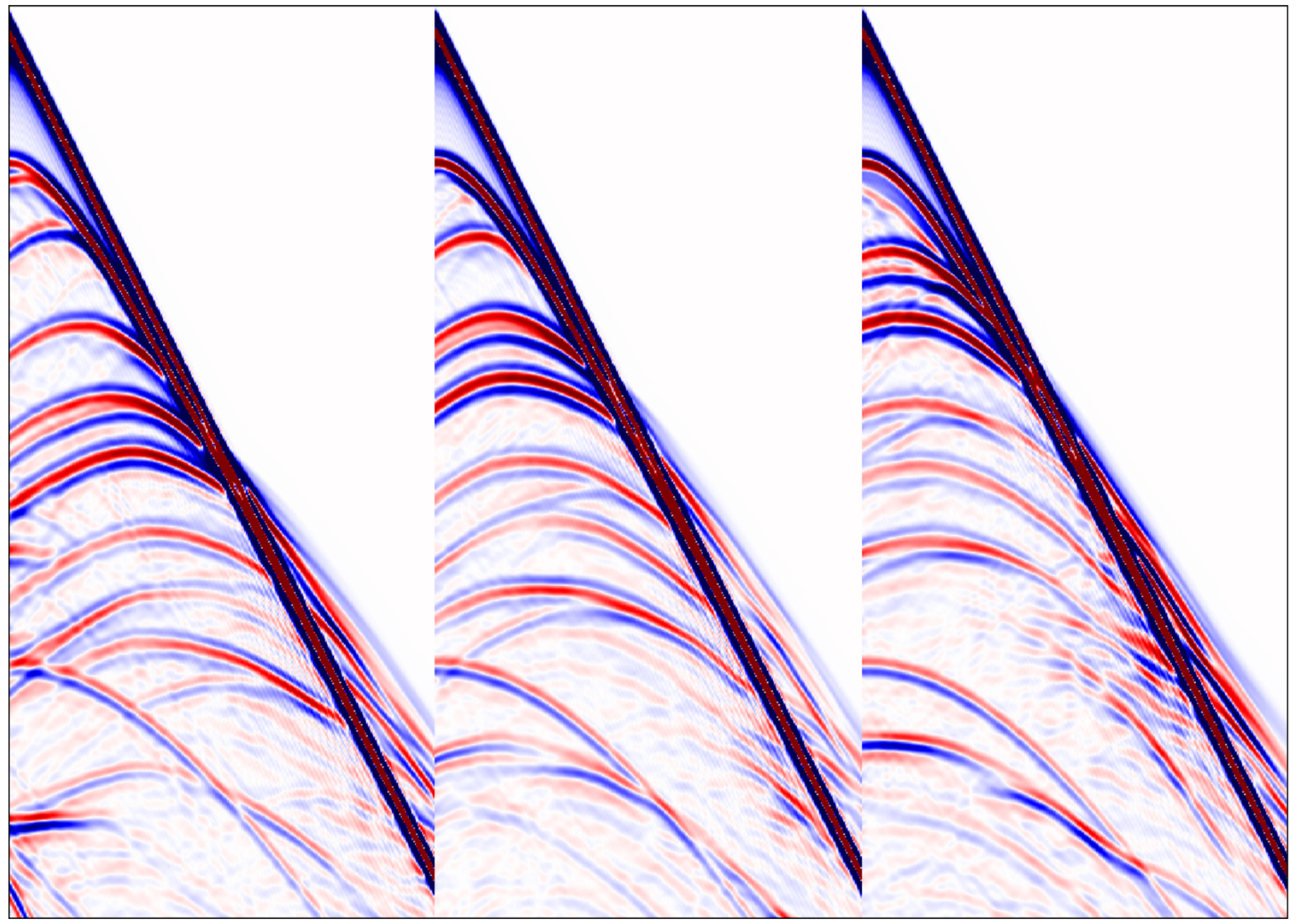
$$\mathbf{A}[\mathbf{m}] = \mathbf{m} \odot \frac{\partial^2}{\partial t^2} - \nabla^2$$

Solve via finite-difference time-stepping:

$$\mathbf{u}^{n+1} = [2 + \frac{\Delta t^2}{\mathbf{m}} \odot \mathbf{L}]\mathbf{u}^n - \mathbf{u}^{n-1} + \frac{\Delta t^2}{\mathbf{m}} \odot \mathbf{P}_s^\top \mathbf{q}^{n+1}$$



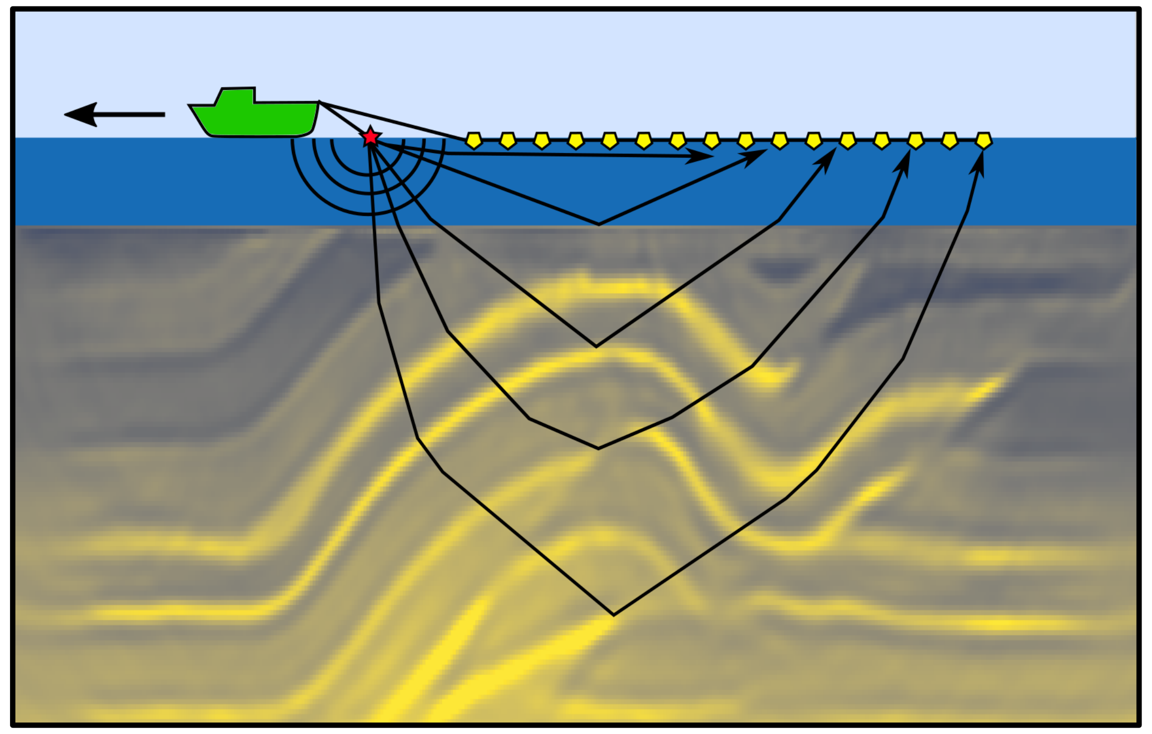
$\mathbf{F}[\mathbf{m}] \cdot \mathbf{q}$



Inverse problem

The *inverse* problem:

$$\min_{\mathbf{m}} \Phi(\mathbf{m}) = \frac{1}{2}\sum_{i=1}^{n_s} \|\mathbf{d}_i - \mathbf{F}[\mathbf{m}]\mathbf{q}_i\|_2^2$$



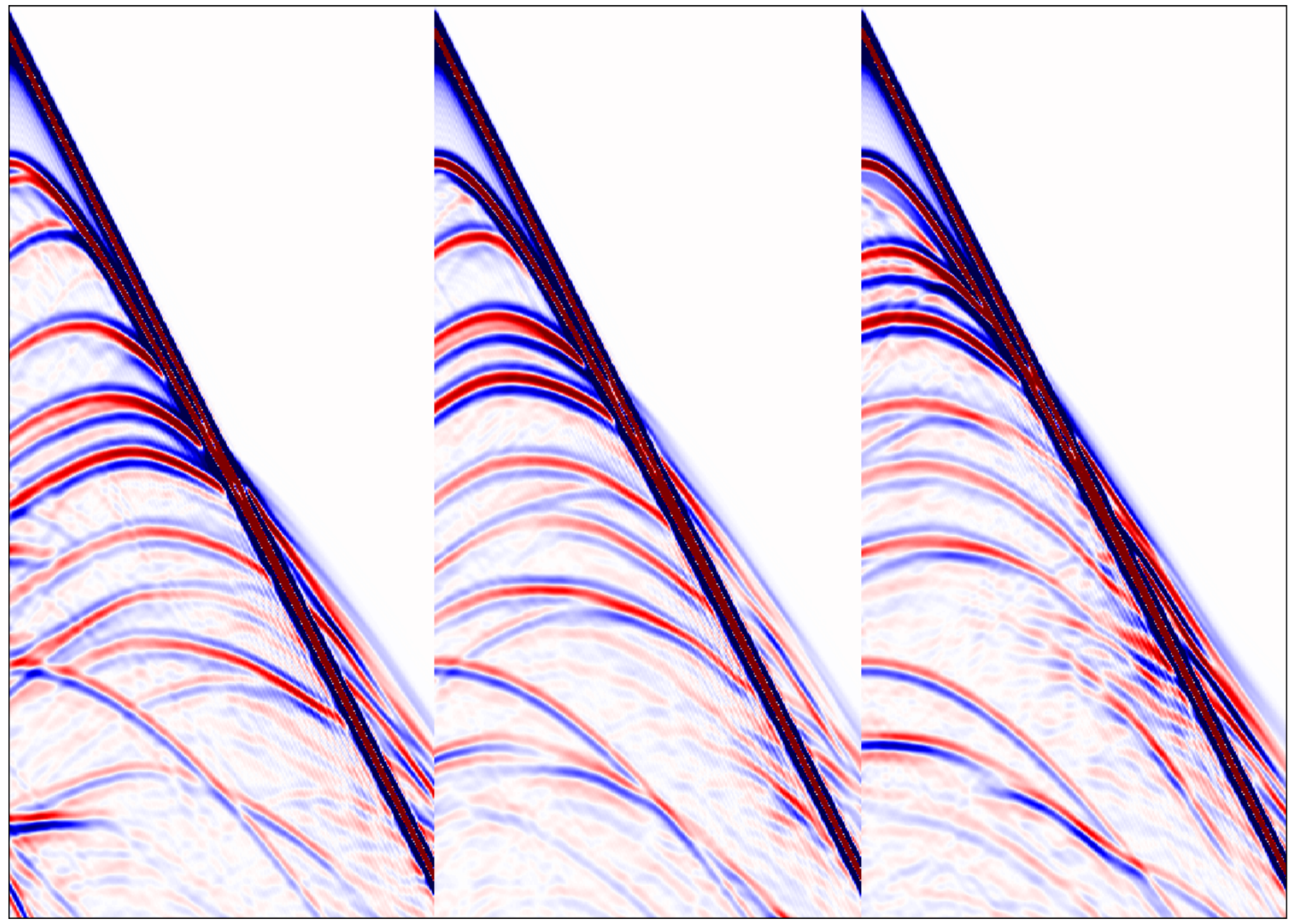
$$\min_{\mathbf{m}} \Phi(\mathbf{m})$$

Sensitivities w.r.t. model parameters: $\dfrac{\partial \mathbf{F}[\mathbf{m}]\mathbf{q}}{\partial \mathbf{m}}$

$$\mathbf{J} = -\mathbf{P}_r \mathbf{A}[\mathbf{m}]^{-1} \mathrm{diag}\left(\frac{\partial \mathbf{A}[\mathbf{m}]}{\partial \mathbf{m}} \mathbf{A}[\mathbf{m}]^{-1} \mathbf{P}_s^\top \mathbf{q}\right)$$

Gradient of objective function via backpropagation:

$$\mathbf{g} = \sum_{i=1}^{n_s} \mathbf{J}^\top (\mathbf{F}[\mathbf{m}]\mathbf{q}_i - \mathbf{d}_i)$$



Challenges
risk mitigation

Increased calls for systematic assessment of uncertainty

- propagate errors in observed data to tasks on the image (e.g. automatic horizon tracking)
- Bayesian Variational Inference–i.e., *distribution* learning

Hurdles:

- Markov Chain Monte Carlo methods for Uncertainty Quantification are too expensive
- tackle complexity numerical PDE solves w/ abstractions & automatic code generation
- data-driven Variational Inference calls for integration physics & machine learning
- access to training data

The Julia Devito Inversion framework (JUDI)

SLIM

Julia

linear opeartors, data containers, IO

↓ parallel modeling function

parallelization: distribute sources, data

↓ serial modeling function

interface to Devito (Python)

↓

Python

Devito: symbolic definition of PDE

↓ automatic code generation and JIT compilation

C

solve PDE w/ OpenMP parallelism

```julia
g = J'*(Pr*A_inv*Ps'*q - d_obs)
```

```python
# Symbolic wavefield
u = TimeFunction(name='u', space_order=16)

# Acoustic wave equation
pde = model.m * u.dt2 - u.laplace
```

```c
float (*rec_coords)[2] = (Float (*[2]) rec_coords_vec;
{
  struct timeval start_kernel, endkernel;
    gettimeofday(&start_kernel, NULL);
    int t0;
    int t1;
    int t2;
    ;
    for (int i3 = 0; i3 < 3; i3+=1)
    {
      flops->kernel += 2.000000;
      {
        {
          t0 = (i3)%(3);
```

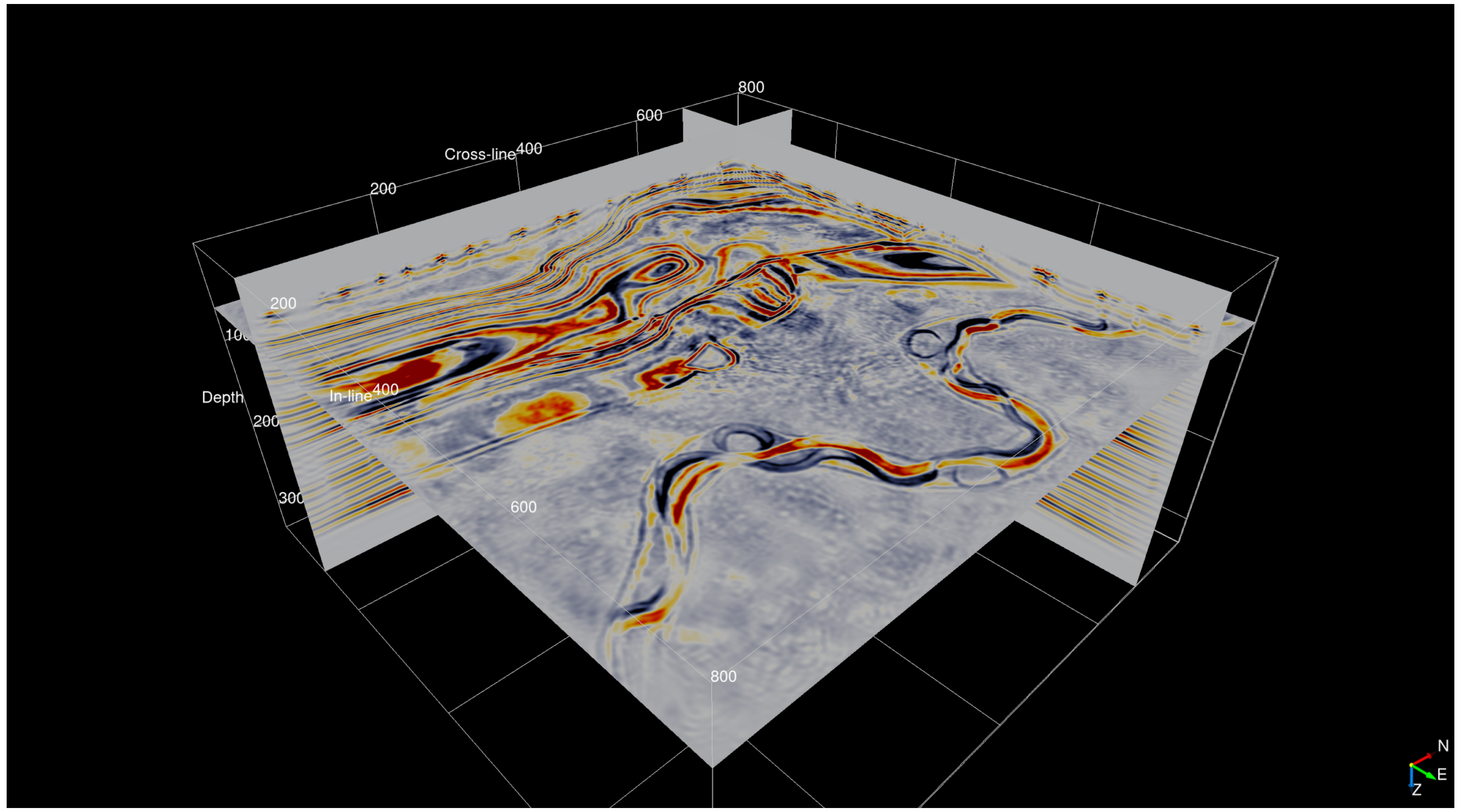
Witte, Philipp A., et al. "An event-driven approach to serverless seismic imaging in the cloud." *IEEE Transactions on Parallel and Distributed Systems* 31.9 (2020): 2032-2049.

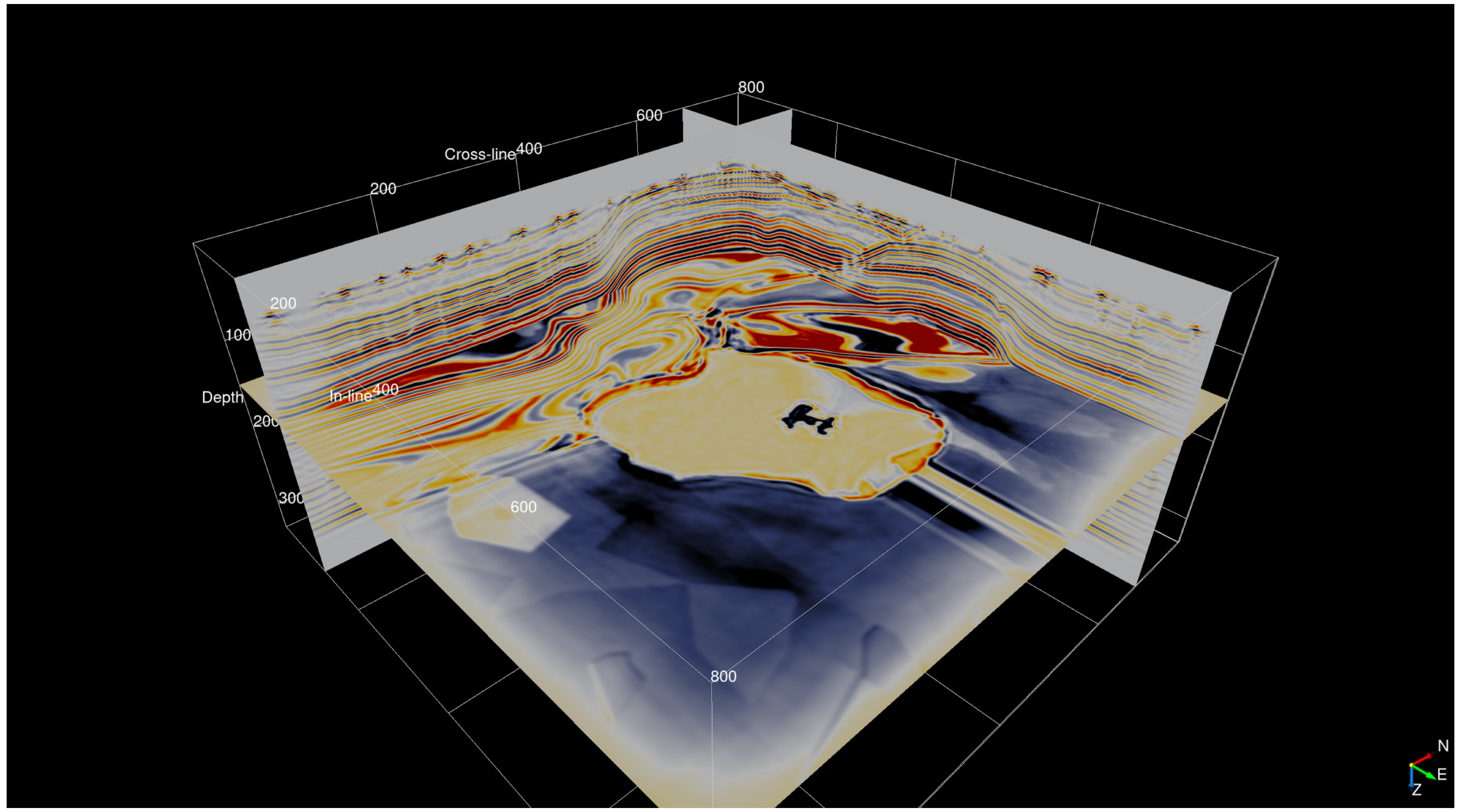
SLIM

Serverless seismic imaging on Azure

"Small" 3D Imaging case study w/ Devito DSL + JUDI

- Data set: 1,500 source locations (~**2.1 TB** data)
- Model: 10 x 10 x 3.325 km (**270 million** unknowns)
- PDE: tilted transversely isotropic (TTI) wave equation, **3,500** time steps
- **Cost: < 10,000$** on 100 E64/E64s instances (2 VMs per gradient with MPI)
- **Peak performance: 140 TFLOP/s**
- **cost single image is comparable to training a large NN**





Example 2: Deep Learning

JUDI operators W/ deep learning packages (e.g. Flux.jl):

```
network(x) = W2*(J*(W1*x + b1)) + b2
# Or:
network(x) = conv2(𝓕(conv1(x)))
loss(x, y) = Flux.mse(network(x), y)
```

Once again, abstractions pay off:

- No need to backpropagate through solvers using automatic differentiation (AD)
- Backpropagation through JUDI operators

```
@adjoint *(J::judiJacobian, x) = *(J, x), Δ -> (nothing, transpose(J) * Δ)
```

Code Repositories – slimgroup

General purpose examples

- JUDI.jl – the Julia Devito Inversion framework
- SetIntersectionProjection.jl – projections onto intersections & sums of sets
- InvertibleNetworks.jl – building blocks for invertible neural networks
- JOLI.jl – serial and distributed linear operators in Julia

Specialized examples

- FastApproximateInference.jl – variational inference for inverse problems
- XConv – Julia/Python code for memory efficient CNNs
- TimeProbeSeismic.jl – low memory WE based inversion

Problem setup

Find $\mathbf{x}$ such that:

$$\mathbf{y} = F(\mathbf{x}) + \boldsymbol{\epsilon}$$

observed data $\mathbf{y}$

(non-Gaussian) noise $\boldsymbol{\epsilon}$

expensive (nonlinear) forward operator F

Stochastic gradient Langevin dynamics

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \frac{\alpha_k}{2}\mathbf{M}_k \nabla_{\mathbf{w}} \left(\frac{n_s}{2\sigma^2} \left\| \mathbf{d}_i - \mathbf{J}[\mathbf{m}_0, \mathbf{q}_i] g(\mathbf{z}, \mathbf{w}_k) \right\|_2^2 + \frac{\lambda^2}{2} \left\| \mathbf{w}_k \right\|_2^2 \right) + \boldsymbol{\xi}_k,$$

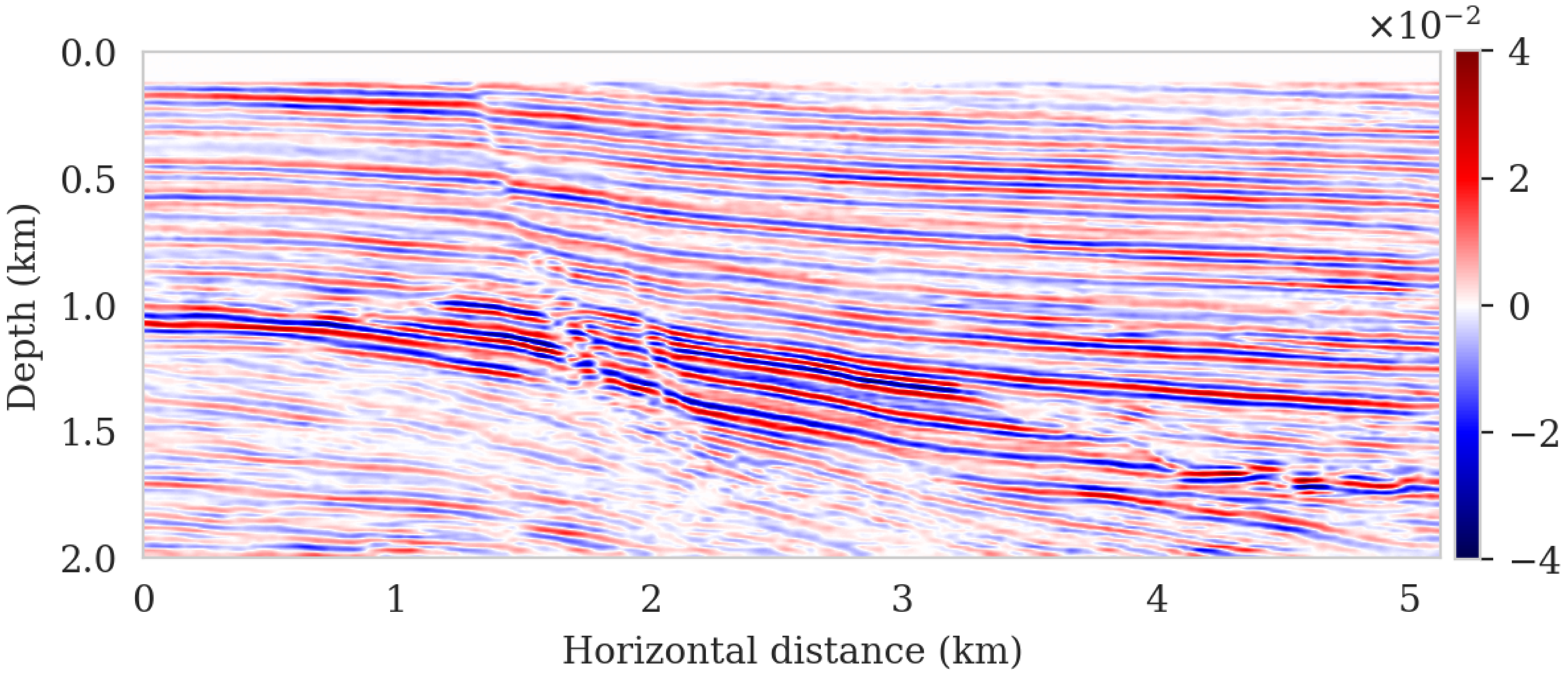
$$\boldsymbol{\xi}_k \sim \mathrm{N}(\mathbf{0}, \alpha_k \mathbf{M}_k)$$

sampling the posterior distribution

requires at least 10k iterations

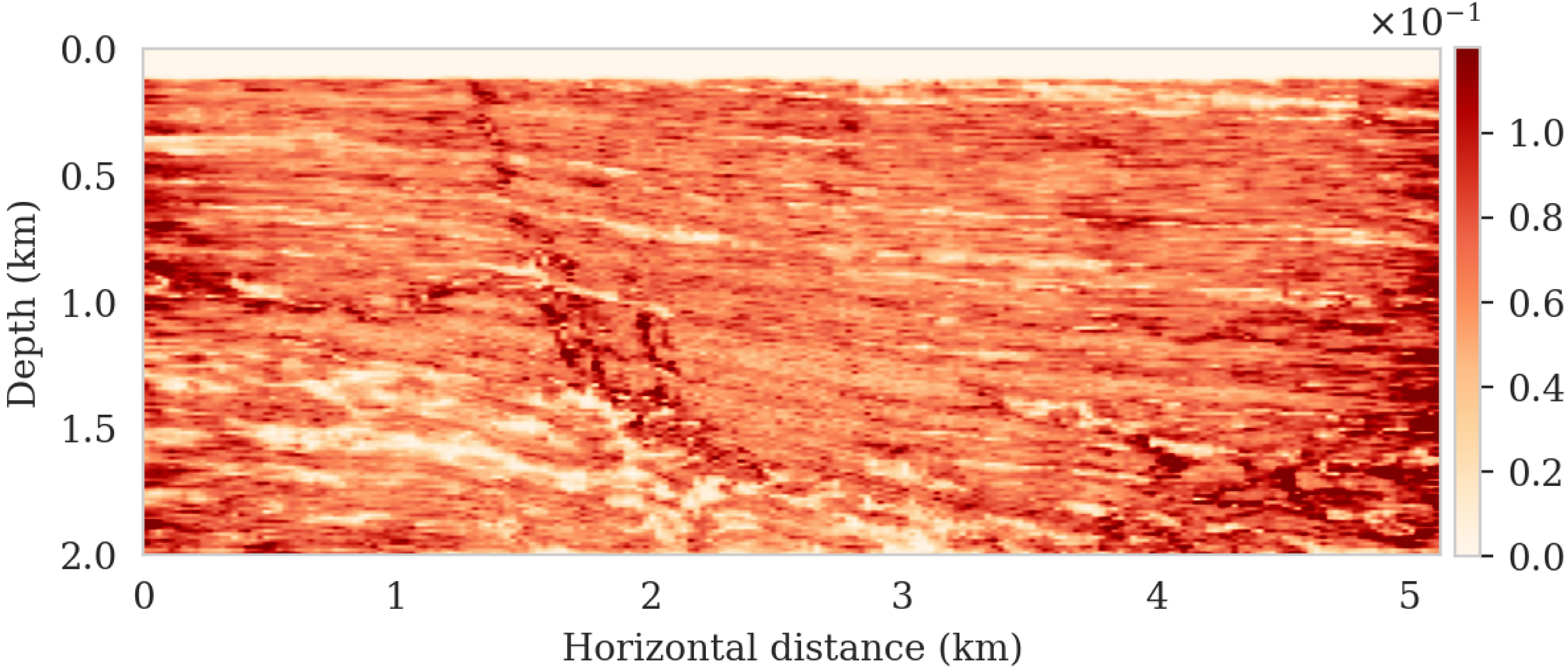
Max Welling and Yee Whye Teh. "Bayesian Learning via Stochastic Gradient Langevin Dynamics". In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML'11. 2011, pp. 681–688. DOI: 10.5555/3104482.3104568.

Ali Siahkoohi, Gabrio Rizzuti, and Felix J. Herrmann. "Uncertainty quantification in imaging and automatic horizon tracking—a Bayesian deep-prior based approach". In: *SEG Technical Program Expanded Abstracts 2020*. Sept. 2020. URL: https://arxiv.org/pdf/2004.00227.pdf.



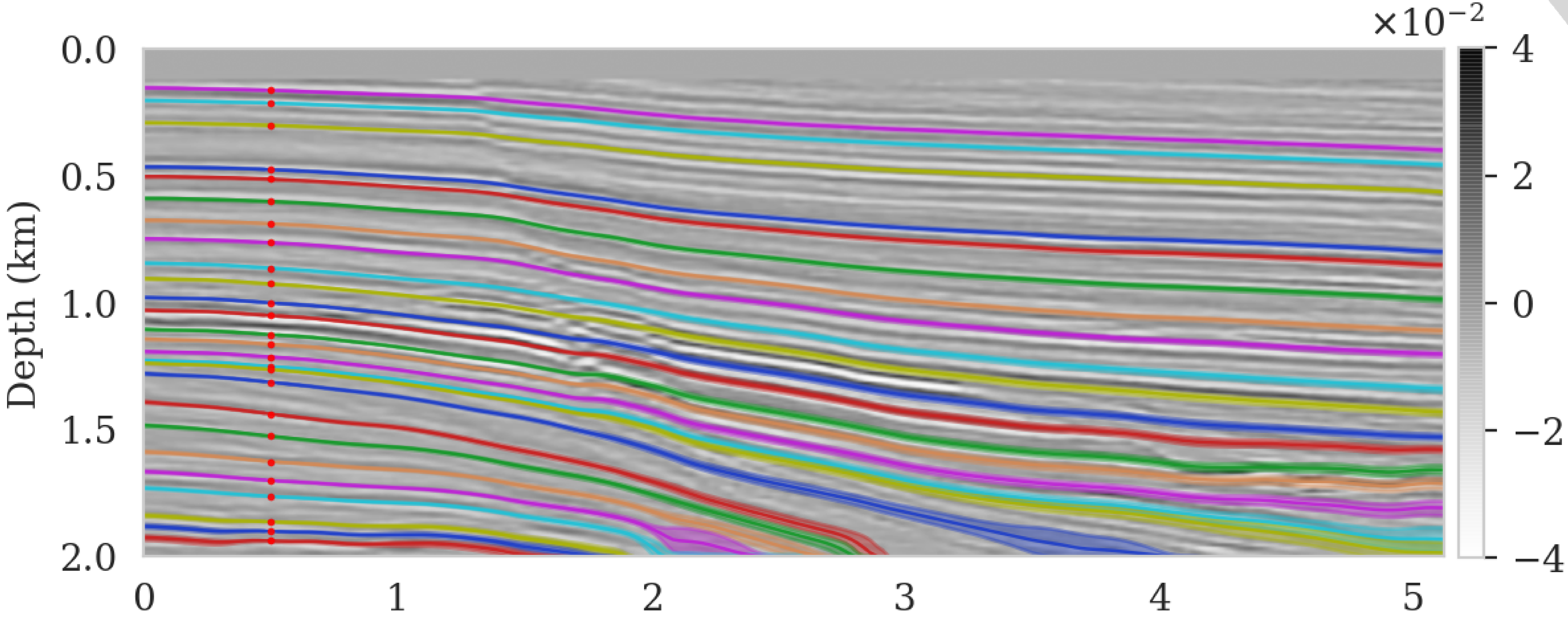
Conditional mean estimate

Ali Siahkoohi, Gabrio Rizzuti, and Felix J. Herrmann. "Uncertainty quantification in imaging and automatic horizon tracking—a Bayesian deep-prior based approach". In: *SEG Technical Program Expanded Abstracts 2020*. Sept. 2020. URL: https://arxiv.org/pdf/2004.00227.pdf.



Pointwise standard deviation, normalized by the envelope of conditional mean

Ali Siahkoohi, Gabrio Rizzuti, and Felix J. Herrmann. "Uncertainty quantification in imaging and automatic horizon tracking—a Bayesian deep-prior based approach". In: *SEG Technical Program Expanded Abstracts 2020*. Sept. 2020. URL: https://arxiv.org/pdf/2004.00227.pdf.



Uncertainty in horizon tracking due to uncertainties in imaging

Xinming Wu and Sergey Fomel. "Least-squares horizons with local slopes and multi-grid correlations". In: *GEOPHYSICS* 83 (4 2018), pp. IM29–IM40. DOI: 10.1190/geo2017-0830.1. URL: https://doi.org/10.1190/geo2017-0830.1.

Xinming Wu and Sergey Fomel. *xinwucwp/mhe*. 2018. URL: https://github.com/xinwucwp/mhe.

Ali Siahkoohi, Gabrio Rizzuti, and Felix J. Herrmann. "Uncertainty quantification in imaging and automatic horizon tracking—a Bayesian deep-prior based approach". In: *SEG Technical Program Expanded Abstracts 2020*. Sept. 2020. URL: https://arxiv.org/pdf/2004.00227.pdf.

Inverse problems and Bayesian inference

Represent the solution as a distribution over the model space

i.e., posterior distribution

Albert Tarantola. *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005. ISBN: 978-0-89871-572-9. DOI: 10.1137/1.9780898717921.

Bayes' rule

$$p(\mathbf{x} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})$$

posterior distribution $p(\mathbf{x} \mid \mathbf{y})$

likelihood function $p(\mathbf{y} \mid \mathbf{x})$

prior distribution $p(\mathbf{x})$

Albert Tarantola. *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005. ISBN: 978-0-89871-572-9. DOI: 10.1137/1.9780898717921.

Maximum likelihood estimator

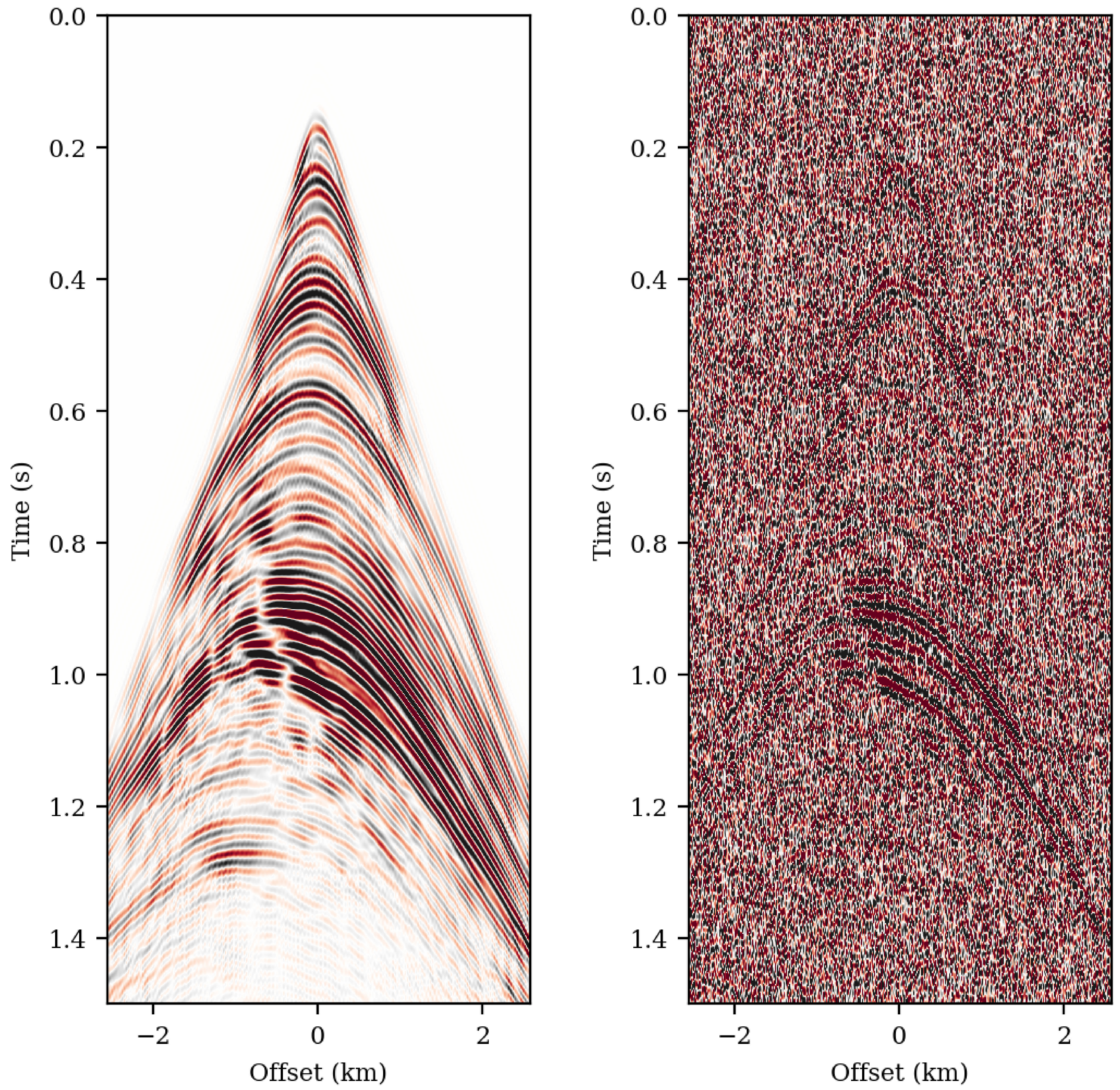
$$\arg\max_{\mathbf{x}} p\left(\mathbf{y} \mid \mathbf{x}\right) = \arg\min_{\mathbf{x}} \frac{1}{2\sigma^2} \left\|\mathbf{y} - F(\mathbf{x})\right\|_2^2$$

no high-dimensional integration/sampling

no prior/regularization

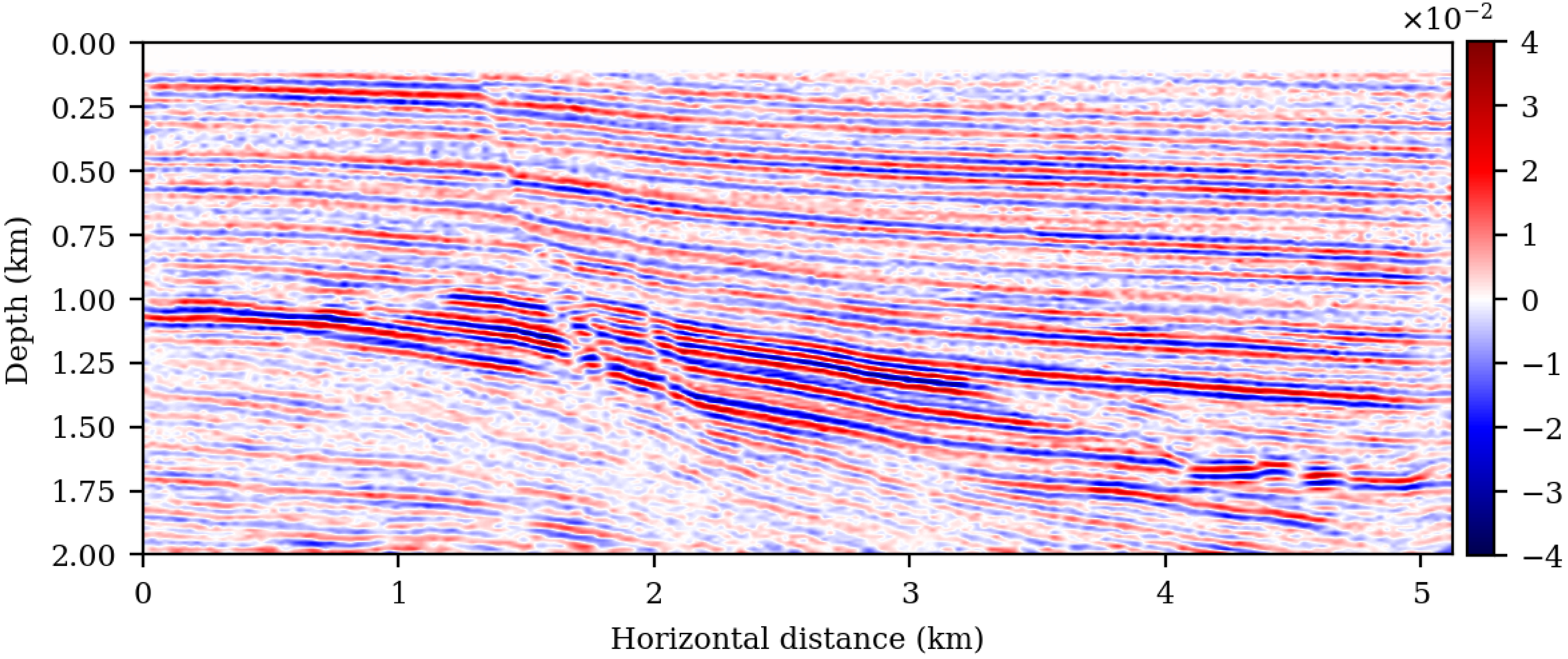
Ozan Öktem. "Bayesian inversion for tomography through machine learning". Mathematics of Imaging Workshop. 2019. URL: https://imaging-in-paris.github.io/semester2019/slides/w3/Oktem.pdf.

Noisy data



Noise-free (left) and noisy (right) linearized data—SNR: -8.75 dB

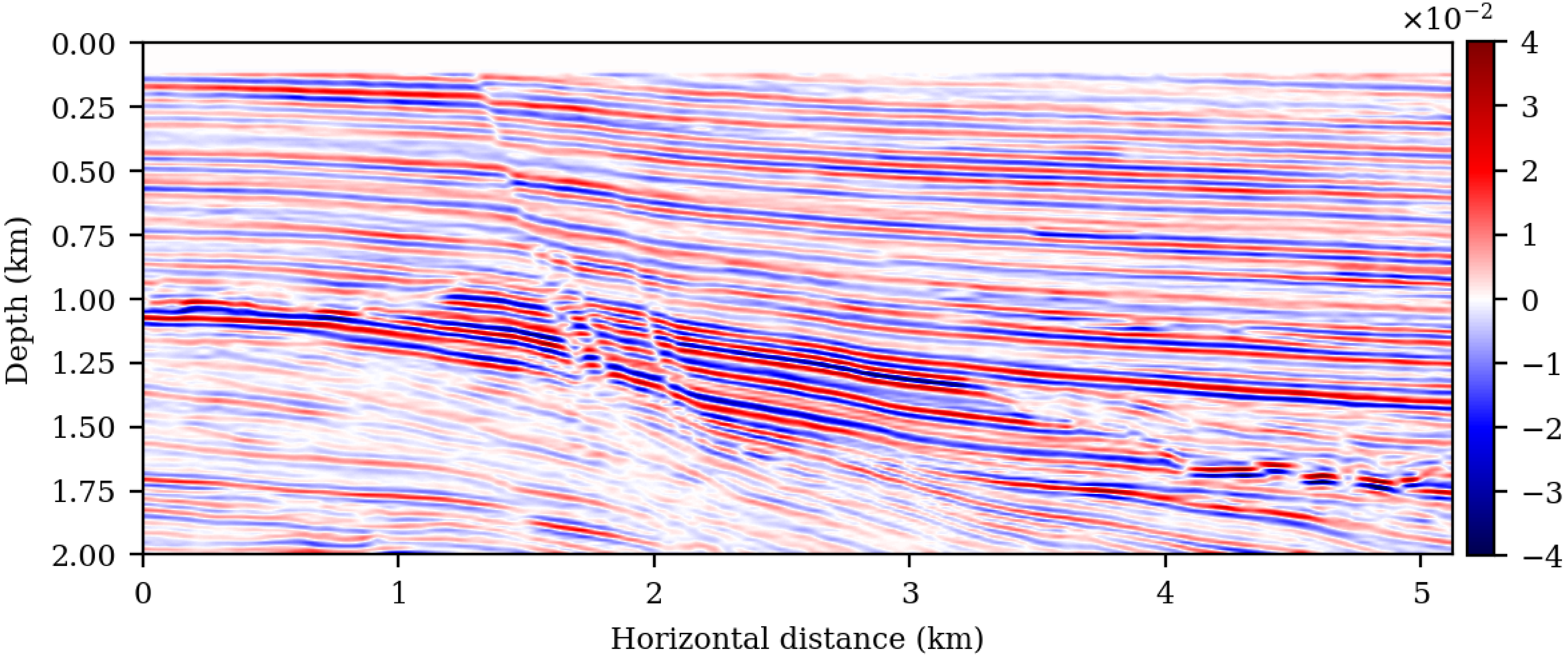
Maximum likelihood



Estimation with no regularization/prior

SNR 8.25 dB

True reflectivity



$\delta\mathbf{m}$—"true" reflectivity model obtain from Parihaka dataset

Maximum a posteriori estimate

$$\arg\max_{\mathbf{x}} \; p\left(\mathbf{x} \mid \mathbf{y}\right) = \arg\min_{\mathbf{x}} \left[-\log p\left(\mathbf{y} \mid \mathbf{x}\right) - \log p\left(\mathbf{x}\right) \right]$$

$$= \arg\min_{\mathbf{x}} \left[\frac{1}{2\sigma^2} \left\| \mathbf{y} - F(\mathbf{x}) \right\|_2^2 - \log p\left(\mathbf{x}\right) \right]$$

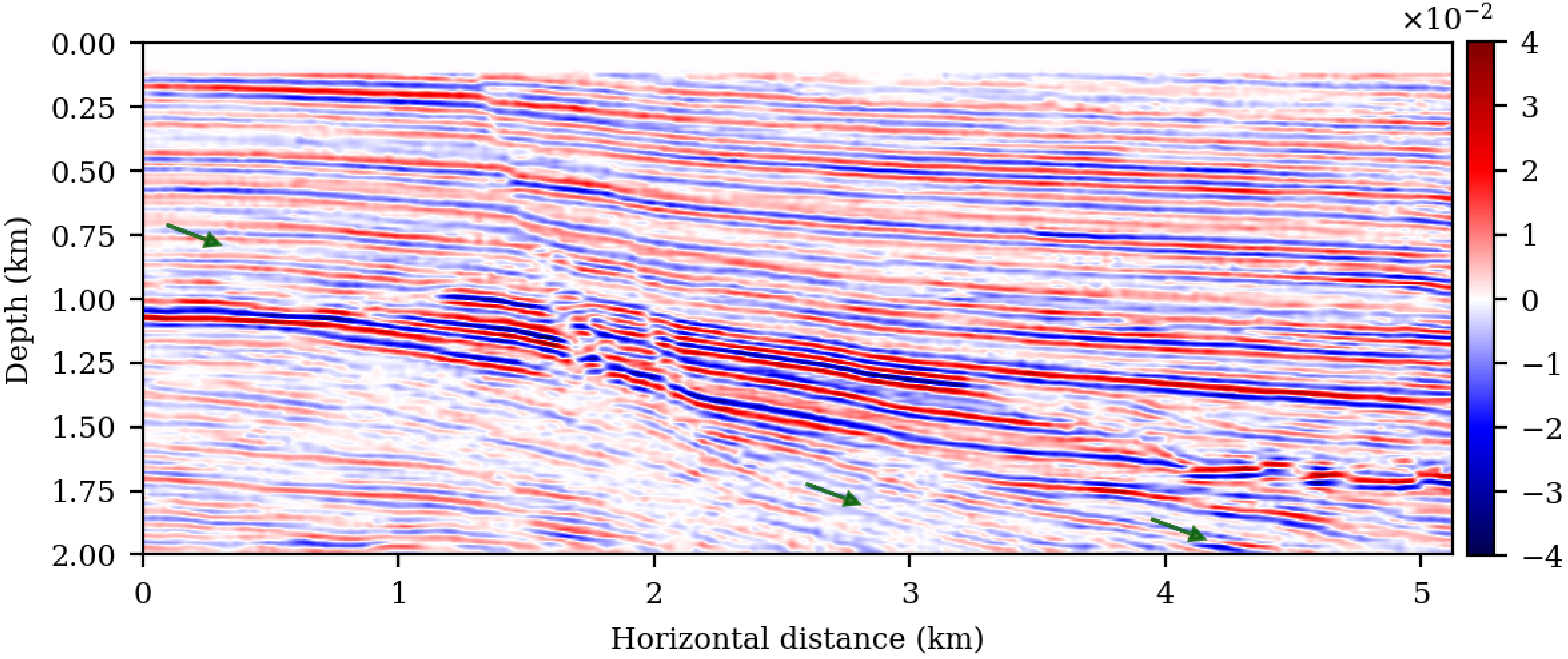
no high-dimensional integration/sampling

uses prior information

requires choosing prior

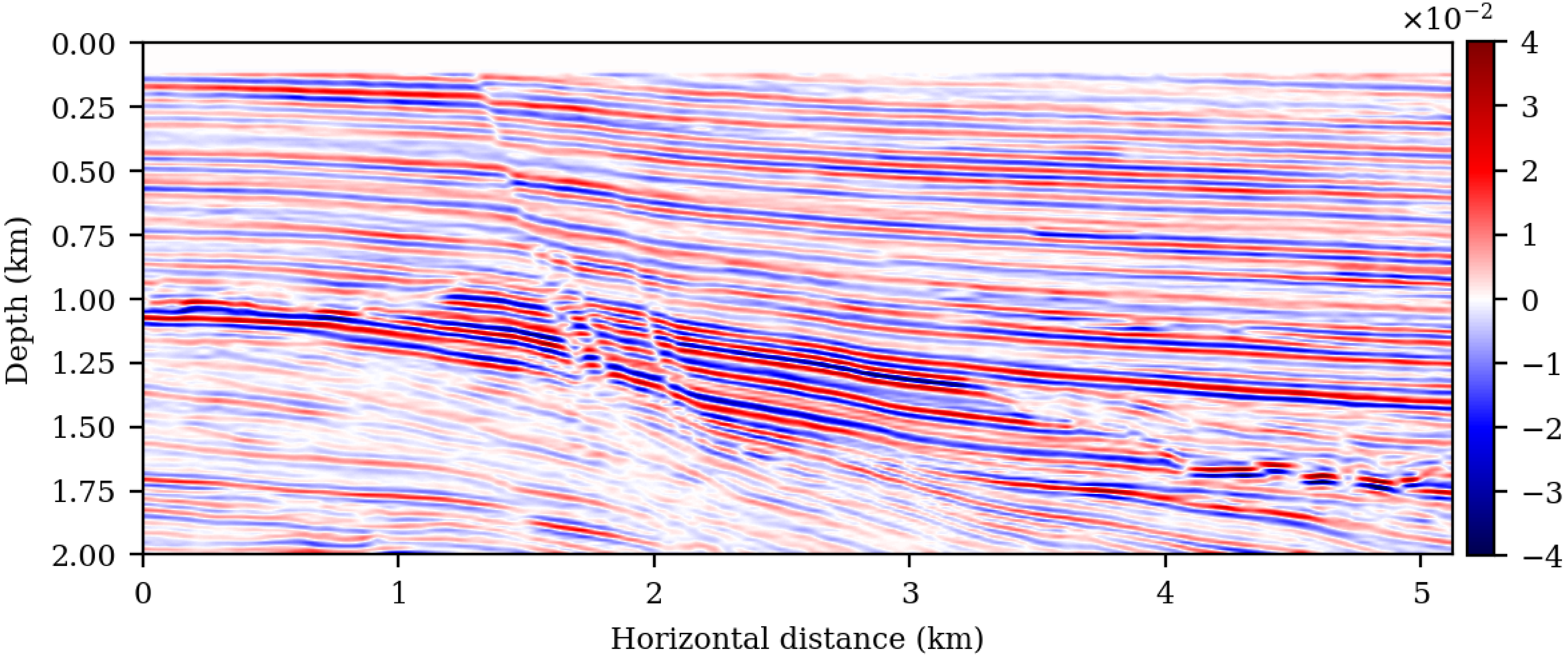
Ozan Öktem. "Bayesian inversion for tomography through machine learning". Mathematics of Imaging Workshop. 2019. URL: https://imaging-in-paris.github.io/semester2019/slides/w3/Oktem.pdf.

Maximum a posteriori estimate (MAP)



MAP estimate—SNR: 8.77 dB

True reflectivity



$\delta\mathbf{m}$—"true" reflectivity model obtain from Parihaka dataset

Conditional mean estimate

$$\mathbb{E}[\mathbf{x} \mid \mathbf{y}] = \int \mathbf{x}\, p(\mathbf{x} \mid \mathbf{y})\, \mathrm{d}\mathbf{x}$$

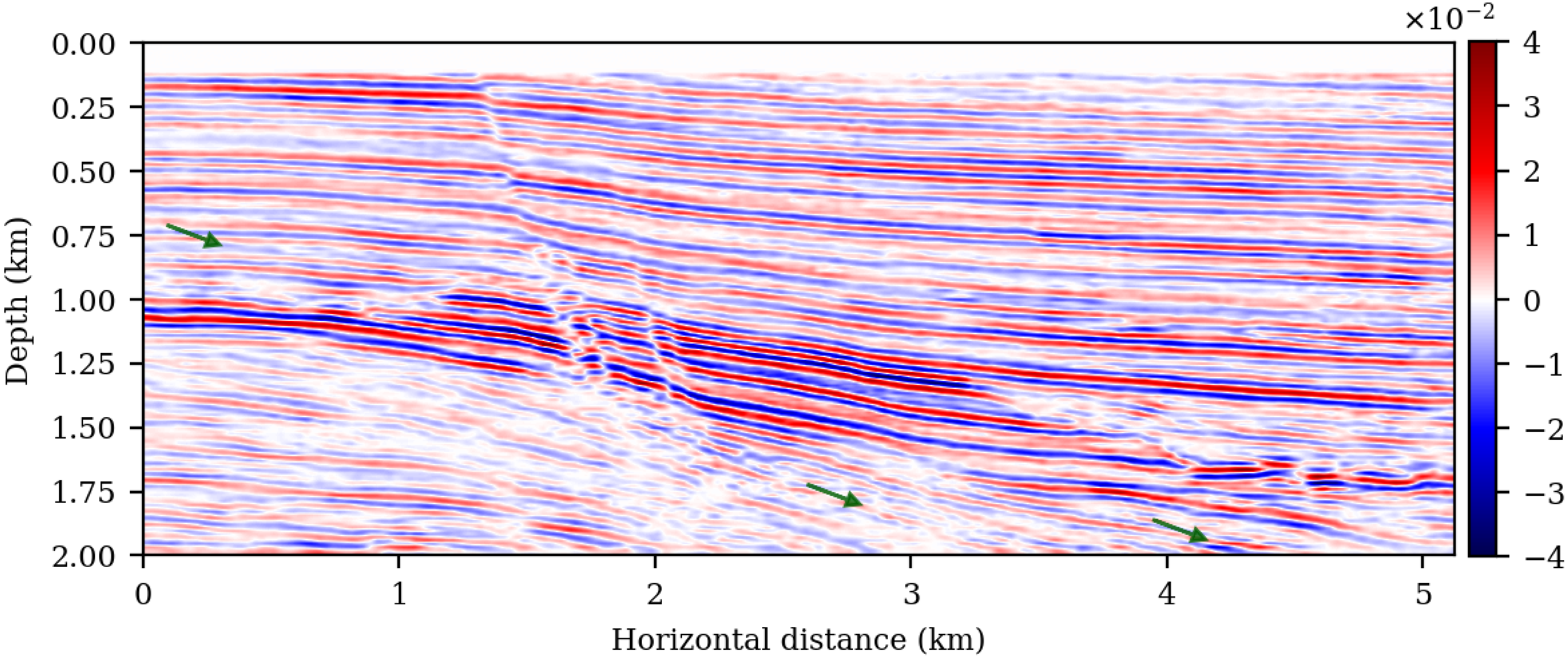
uses prior information

high-dimensional integration/sampling

requires choosing prior

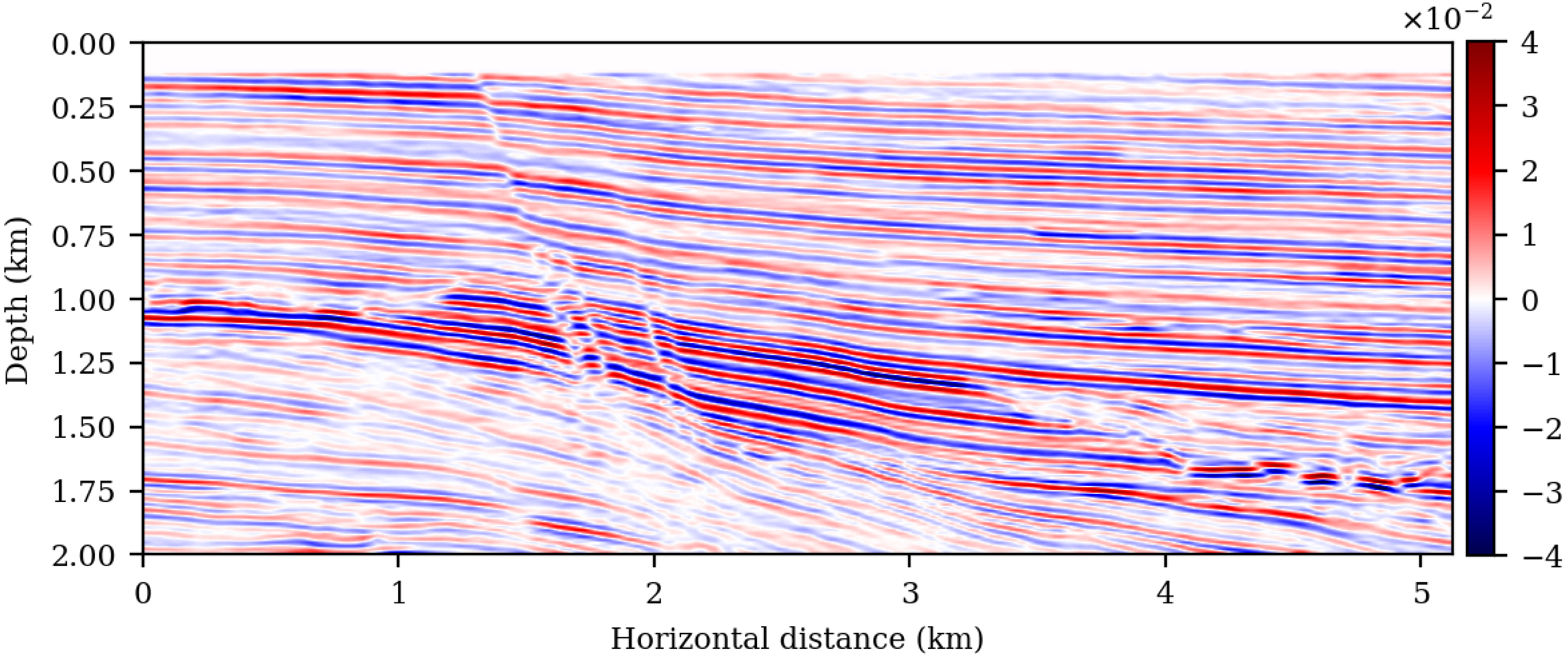
Ozan Öktem. "Bayesian inversion for tomography through machine learning". Mathematics of Imaging Workshop. 2019. URL: https://imaging-in-paris.github.io/semester2019/slides/w3/Oktem.pdf.

Conditional mean estimate



Conditional mean estimate—SNR: 9.66 dB

True reflectivity



$\delta\mathbf{m}$—"true" reflectivity model obtain from Parihaka dataset

Sampling the posterior

Markov chain Monte Carlo (MCMC) sampling

guarantees of producing (asymptotically) exact samples

high-dimensional integration/sampling

costs associated with the forward operator

requires choosing a prior distribution

Christian Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Science & Business Media, 2004. DOI: 10.1007/978-1-4757-4145-2.

Prior art: MCMC

Alberto Malinverno and Victoria A Briggs. "Expanded uncertainty quantification in inverse problems: Hierarchical Bayes and empirical Bayes". In: *GEOPHYSICS* 69.4 (2004), pp. 1005–1016. DOI: 10.1190/1.1778243.

Alberto Malinverno and Robert L Parker. "Two ways to quantify uncertainty in geophysical inverse problems". In: *GEOPHYSICS* 71.3 (2006), W15–W27. DOI: 10.1190/1.2194516.

James Martin, Lucas C. Wilcox, Carsten Burstedde, and Omar Ghattas. "A Stochastic Newton MCMC Method for Large-Scale Statistical Inverse Problems with Application to Seismic Inversion". In: *SIAM Journal on Scientific Computing* 34.3 (2012), A1460–A1487. eprint: http://epubs.siam.org/doi/pdf/10.1137/110845598. URL: http://epubs.siam.org/doi/abs/10.1137/110845598.

Anandaroop Ray, Sam Kaplan, John Washbourne, and Uwe Albertin. "Low frequency full waveform seismic inversion within a tree based Bayesian framework". In: *Geophysical Journal International* 212.1 (Oct. 2017), pp. 522–542. ISSN: 0956-540X. DOI: 10.1093/gji/ggx428. eprint: https://academic.oup.com/gji/article-pdf/212/1/522/21782947/ggx428.pdf. URL: https://doi.org/10.1093/gji/ggx428.

Zhilong Fang, Curt Da Silva, Rachel Kuske, and Felix J. Herrmann. "Uncertainty quantification for inverse problems with weak partial-differential-equation constraints". In: *GEOPHYSICS* 83.6 (2018), R629–R647. DOI: 10.1190/geo2017-0824.1.

Georgia K Stuart, Susan E Minkoff, and Felipe Pereira. "A two-stage Markov chain Monte Carlo method for seismic inversion and uncertainty quantification". In: *GEOPHYSICS* 84.6 (Nov. 2019), R1003–R1020. DOI: 10.1190/geo2018-0893.1.

Zeyu Zhao and Mrinal K Sen. "A gradient based MCMC method for FWI and uncertainty analysis". In: *SEG Technical Program Expanded Abstracts 2019*. Aug. 2019, pp. 1465–1469. DOI: 10.1190/segam2019-3216560.1.

M Kotsi, A Malcolm, and G Ely. "Uncertainty quantification in time-lapse seismic imaging: a full-waveform approach". In: *Geophysical Journal International* 222.2 (May 2020), pp. 1245–1263. ISSN: 0956-540X. DOI: 10.1093/gji/ggaa245.

Andrew Curtis and Anthony Lomax. "Prior information, sampling distributions, and the curse of dimensionality". In: *Geophysics* 66.2 (2001), pp. 372–378.

Felix J. Herrmann, Ali Siahkoohi, and Gabrio Rizzuti. "Learned imaging with constraints and uncertainty quantification". In: *Neural Information Processing Systems (NeurIPS) 2019 Deep Inverse Workshop*. Dec. 2019. URL: https://arxiv.org/pdf/1909.06473.pdf.

Variational Inference

Approximate distribution p by minimizing the Kullback-Leibner (KL) divergence

$$\mathbb{D}_{\mathsf{KL}}\left(p \mid\mid p_\theta\right) = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{p_\theta(\mathbf{x})} \, \mathbf{dx}$$

$$= \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[-\log p_\theta(\mathbf{x}) + \log p(\mathbf{x})\right],$$

parametric distribution p_θ

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. "An Introduction to Variational Methods for Graphical Models". In: *Machine Learning* 37.2 (1999), pp. 183–233. DOI: 10.1023/A:1007665907178.

Variational Inference

Objective of variational inference is to solve

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\arg\min}\, \mathbb{D}_{\mathsf{KL}}\left(p \,||\, p_{\theta}\right)$$

p_{θ} designed to be easily sampled

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. "An Introduction to Variational Methods for Graphical Models". In: *Machine Learning* 37.2 (1999), pp. 183–233. DOI: 10.1023/A:1007665907178.

Sampling the posterior

Variational inference

approximate the posterior with a parametric and easy-to-sample distribution

sampling is turned into an optimization problem

known to scale better than MCMC in high-dimensional problems

requires choosing a prior distribution

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. "Variational inference: A review for statisticians". In: *Journal of the American statistical Association* 112.518 (2017), pp. 859–877.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. "An Introduction to Variational Methods for Graphical Models". In: *Machine Learning* 37.2 (1999), pp. 183–233. DOI: 10.1023/A:1007665907178.

Generative models

probabilistic models to characterize an unknown distribution

Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006.

Generative models

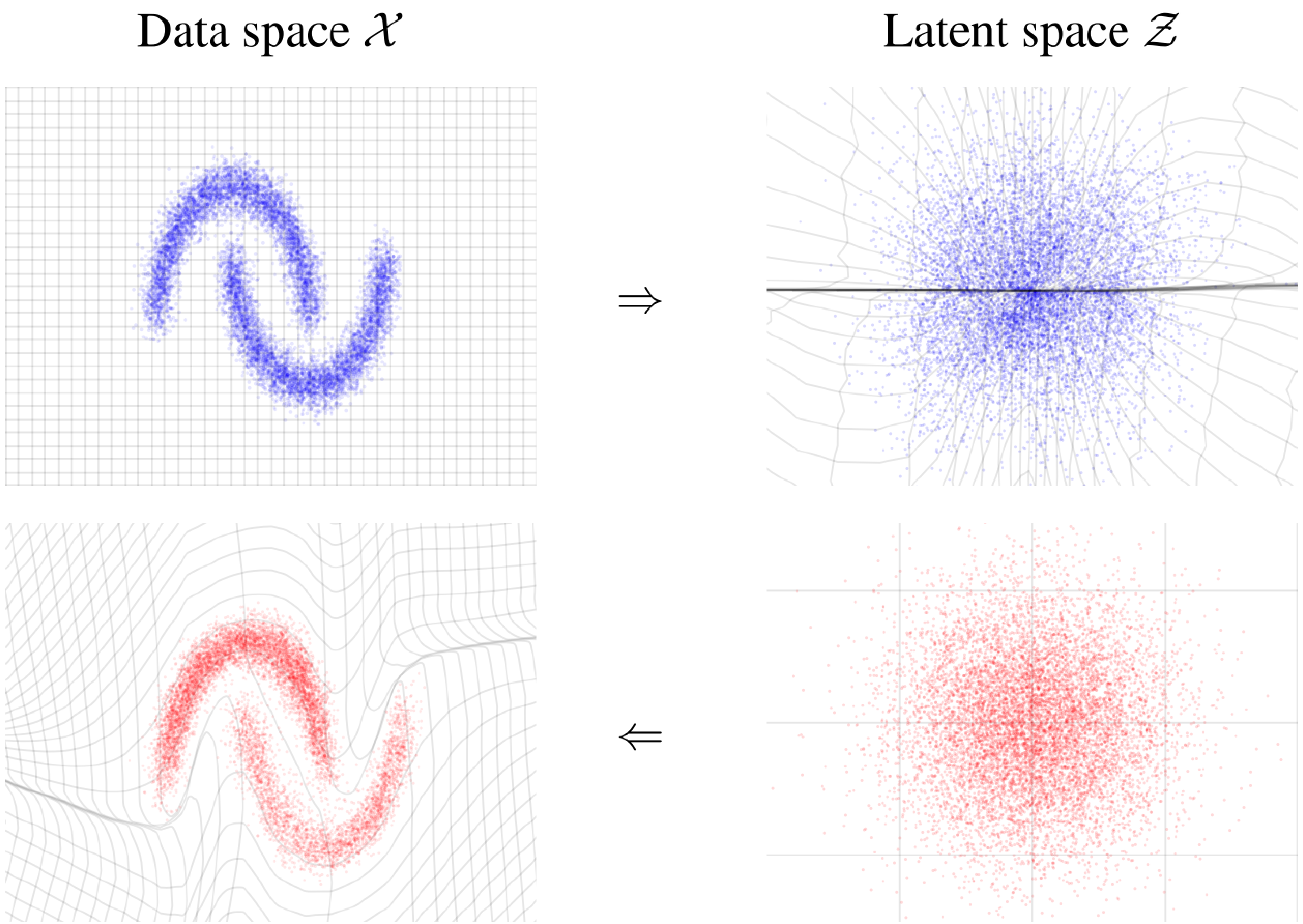
$$p_\theta(\mathbf{x}) \approx p(\mathbf{x}), \quad \mathbf{x} \in \mathcal{X}$$

high dimensional unknown density $p(\mathbf{x})$

generative model $p_\theta(\mathbf{x})$, parameterized by $\boldsymbol{\theta}$

$\boldsymbol{\theta}$ estimated via available samples $\{\mathbf{x}_i\}_{i=1}^n \sim p(\mathbf{x})$

Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006.



Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using Real NVP". In: *International Conference on Learning Representations, ICLR*. 2016. URL: http://arxiv.org/abs/1605.08803.

Inference with NFs

Sampling from $p(\mathbf{x})$

$$G_\theta^{-1}(\mathbf{z}) \sim p(\mathbf{x}), \quad \mathbf{z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}),$$

and fast density estimation via the change-of-variable formula,

$$p_\theta(\mathbf{x}) = p_z(G_\theta(\mathbf{x})) \left| \det \nabla_x G_\theta(\mathbf{x}) \right|$$

Cédric Villani. *Optimal transport: old and new.* Springer-Verlag Berlin Heidelberg, 2009. DOI: 10.1007/978-3-540-71050-9.

Training NFs

$$\boldsymbol{\theta}^* = \arg\min_{\theta} \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})} \Big[-\log p_z(G_\theta(\mathbf{x})) - \log \Big| \det \nabla_x G_\theta(\mathbf{x}^{(i)}) \Big| \Big]$$

$$\approx \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \Big[\frac{1}{2} \big\| G_\theta(\mathbf{x}_i) \big\|_2^2 - \log \Big| \det \nabla_x G_\theta(\mathbf{x}_i) \Big| \Big].$$

expectation estimated via available samples $\{\mathbf{x}_i\}_{i=1}^{N} \sim p(\mathbf{x})$

Posterior inference with NFs

sampling the posterior distribution directly via NFs

Purely data-driven approach

SLIM

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{2} \left\| G_{\theta}(\mathbf{y}_i, \mathbf{x}_i) \right\|^2 - \log \left| \det \nabla_{y,x} G_{\theta}(\mathbf{y}_i, \mathbf{x}_i) \right| \right]$$

$$G_{\theta}(\mathbf{y}, \mathbf{x}) = \begin{bmatrix} G_{\theta_y}(\mathbf{y}) \\ G_{\theta_x}(\mathbf{y}, \mathbf{x}) \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_y \\ \boldsymbol{\theta}_x \end{bmatrix}$$

conditional NF $G_{\theta} : \mathcal{Y} \times \mathcal{X} \to \mathcal{Z}_y \times \mathcal{Z}_x$

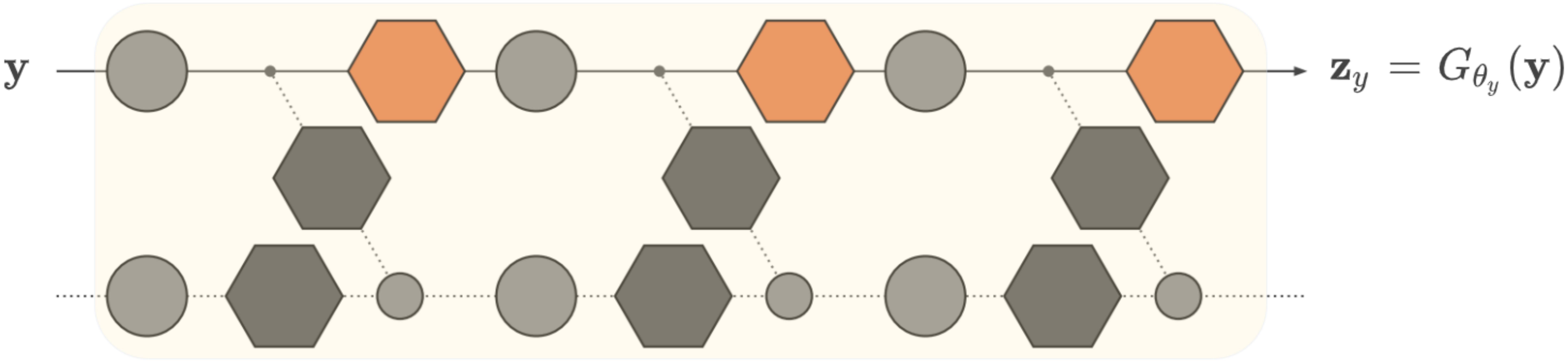
expectation estimated via training pairs $\{\mathbf{y}_i, \mathbf{x}_i\}_{i=1}^{n} \sim p_{y,x}(\mathbf{y}, \mathbf{x})$

Jakob Kruse, Gianluca Detommaso, Robert Scheichl, and Ullrich Köthe. "HINT: Hierarchical Invertible Neural Transport for Density Estimation and Bayesian Inference". In: *Proceedings of AAAI-2021* (2021). URL: https://arxiv.org/pdf/1905.10687.pdf.

Ali Siahkoohi, Gabrio Rizzuti, Mathias Louboutin, Philipp Witte, and Felix J. Herrmann. "Preconditioned training of normalizing flows for variational inference in inverse problems". In: *3rd Symposium on Advances in Approximate Bayesian Inference*. Jan. 2021. URL: https://openreview.net/pdf?id=P9m1sMaNQ8T.

Ali Siahkoohi and Felix J. Herrmann. "Learning by example: fast reliability-aware seismic imaging with normalizing flows". Apr. 2021. URL: https://arxiv.org/pdf/2104.06255.pdf.

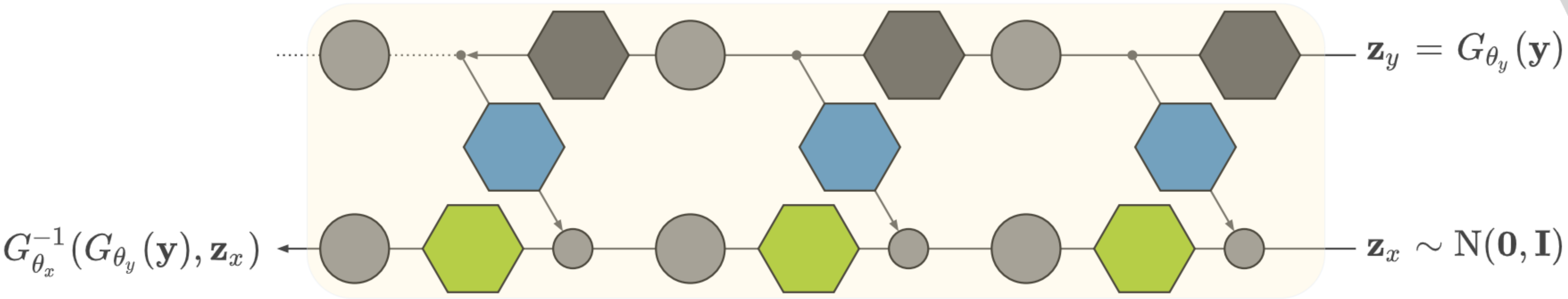
Inference with the data-driven approach



$$\mathbf{z}_y = G_{\theta_y}(\mathbf{y})$$

Jakob Kruse, Gianluca Detommaso, Robert Scheichl, and Ullrich Köthe. "HINT: Hierarchical Invertible Neural Transport for Density Estimation and Bayesian Inference". In: *Proceedings of AAAI-2021* (2021). URL: https://arxiv.org/pdf/1905.10687.pdf.

Inference with the data-driven approach



$$G_{\theta_x}^{-1}(G_{\theta_y}(\mathbf{y}), \mathbf{z}_x) \sim p(\mathbf{x} \mid \mathbf{y}), \quad \mathbf{z}_x \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$$

Jakob Kruse, Gianluca Detommaso, Robert Scheichl, and Ullrich Köthe. "HINT: Hierarchical Invertible Neural Transport for Density Estimation and Bayesian Inference". In: *Proceedings of AAAI-2021* (2021). URL: `https://arxiv.org/pdf/1905.10687.pdf`.

Seismic imaging example

Training dataset for the data-driven approach

Training pairs:

$$\{\mathbf{y}_i, \mathbf{x}_i\}_{i=1}^{n} \sim p(\mathbf{y}, \mathbf{x})$$

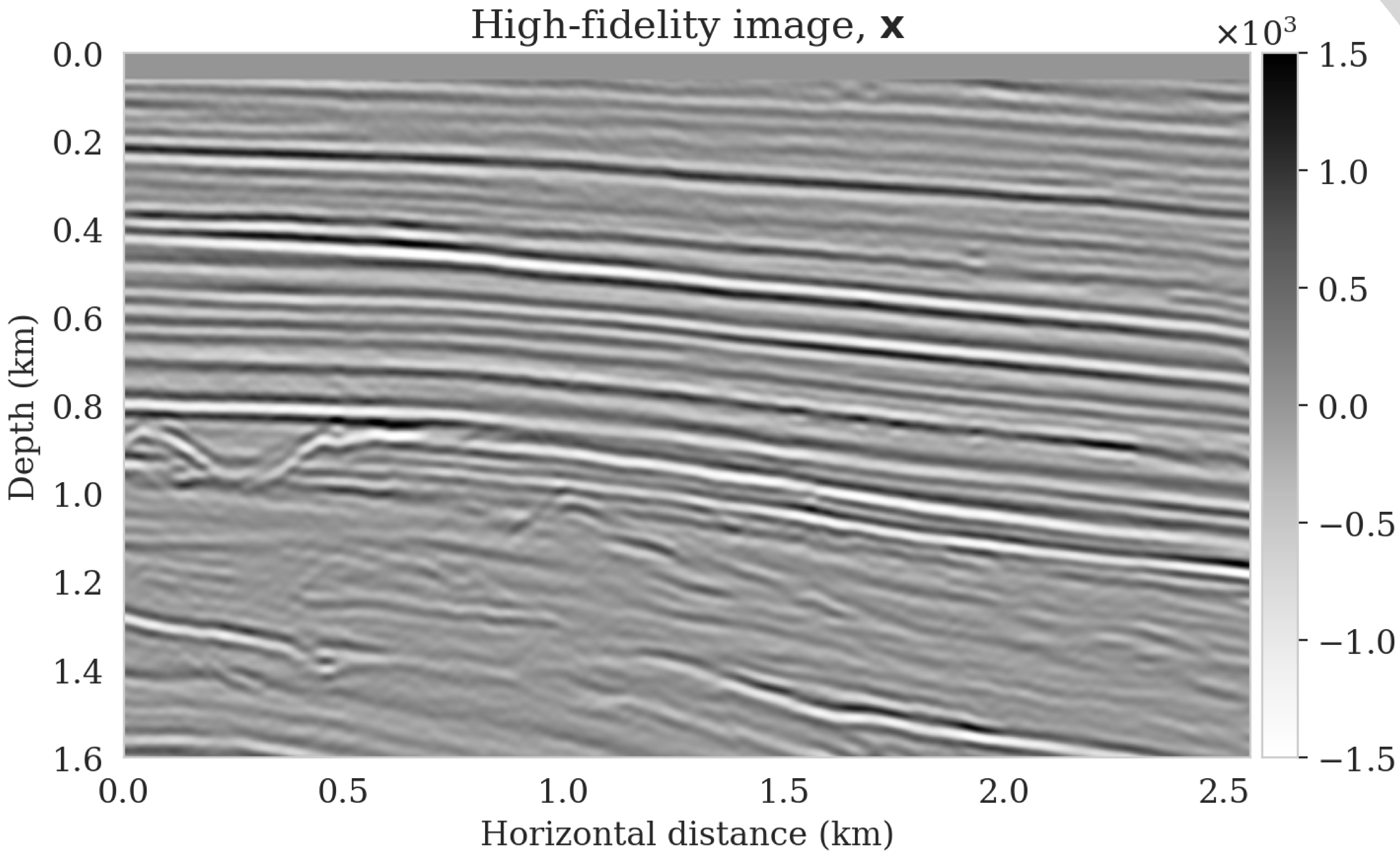
where

$$(\mathbf{y}_i, \mathbf{x}_i) = \left(\mathbf{J}^\top (\mathbf{J}\delta\mathbf{m}_i + \epsilon_i), \delta\mathbf{m}_i\right)$$

WesternGeco. *Parihaka 3D PSTM Final Processing Report*. Tech. rep. New Zealand Petroleum Report 4582. 2012. URL: https://wiki.seg.org/wiki/Parihaka-3D.

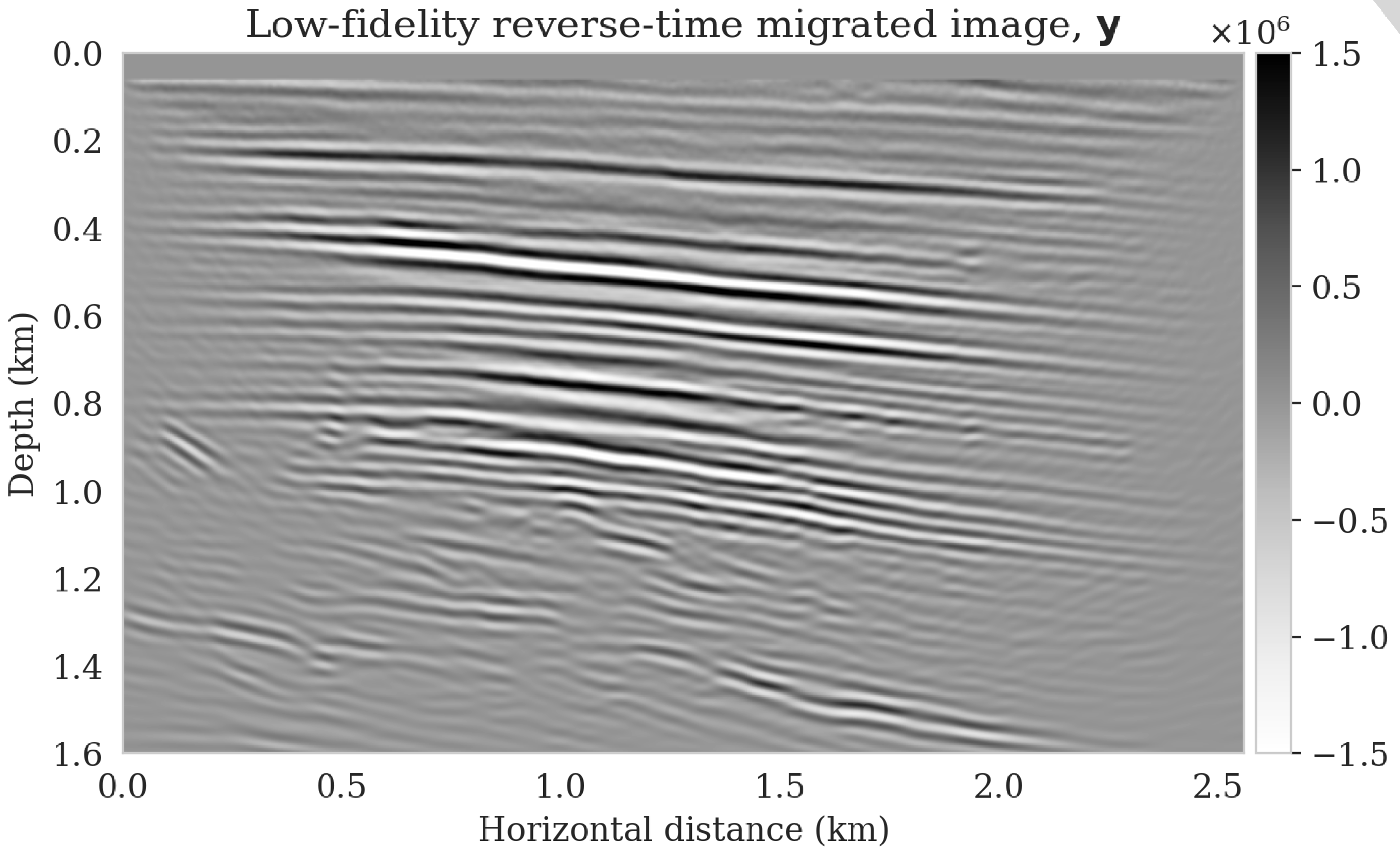
Some data pairs from the training set

WesternGeco. *Parihaka 3D PSTM Final Processing Report*. Tech. rep. New Zealand Petroleum Report 4582. 2012. URL: https://wiki.seg.org/wiki/Parihaka-3D.



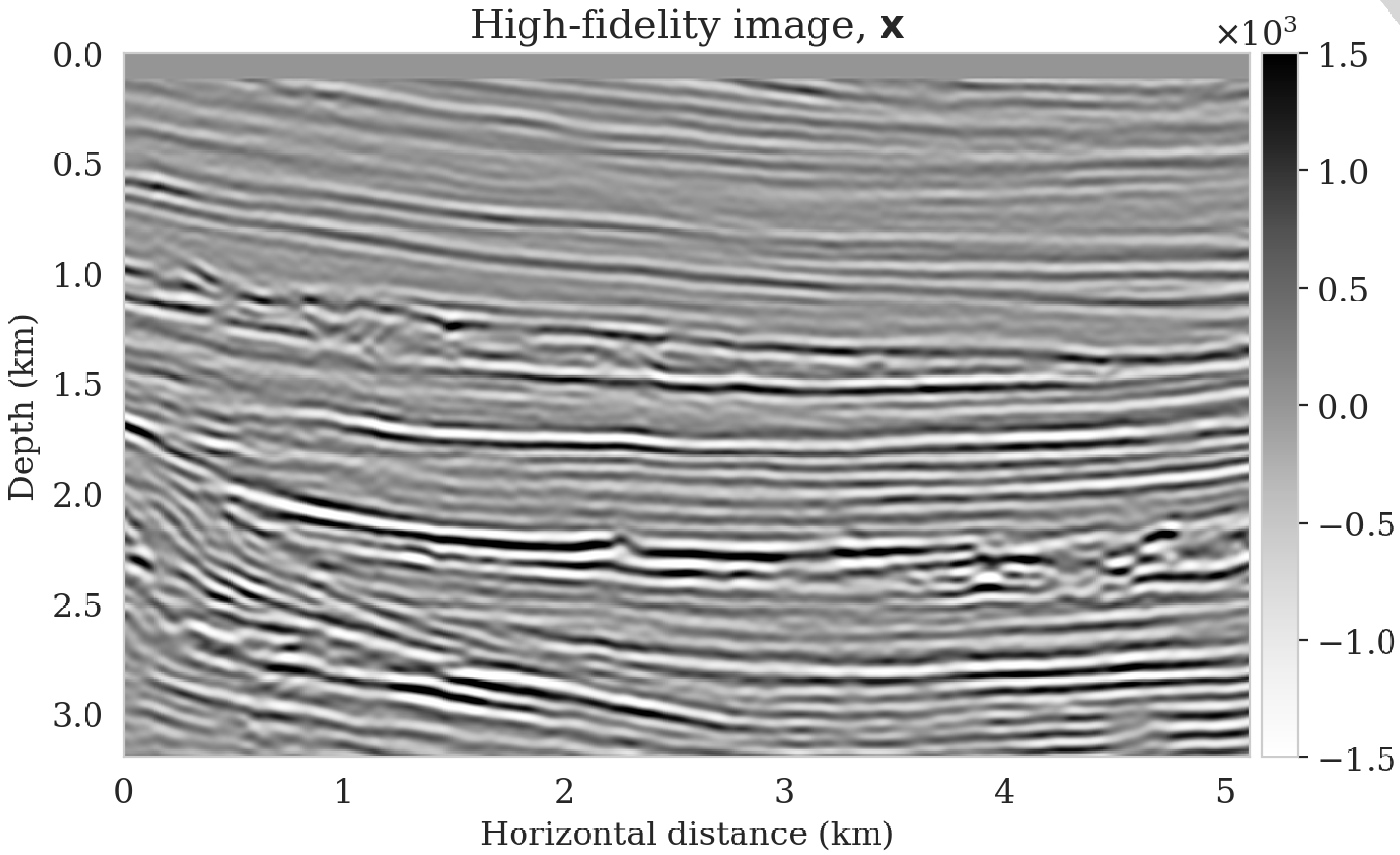
High-fidelity image, x
Depth (km)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Horizontal distance (km)
0.0
0.5
1.0
1.5
2.0
2.5
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

SLIM

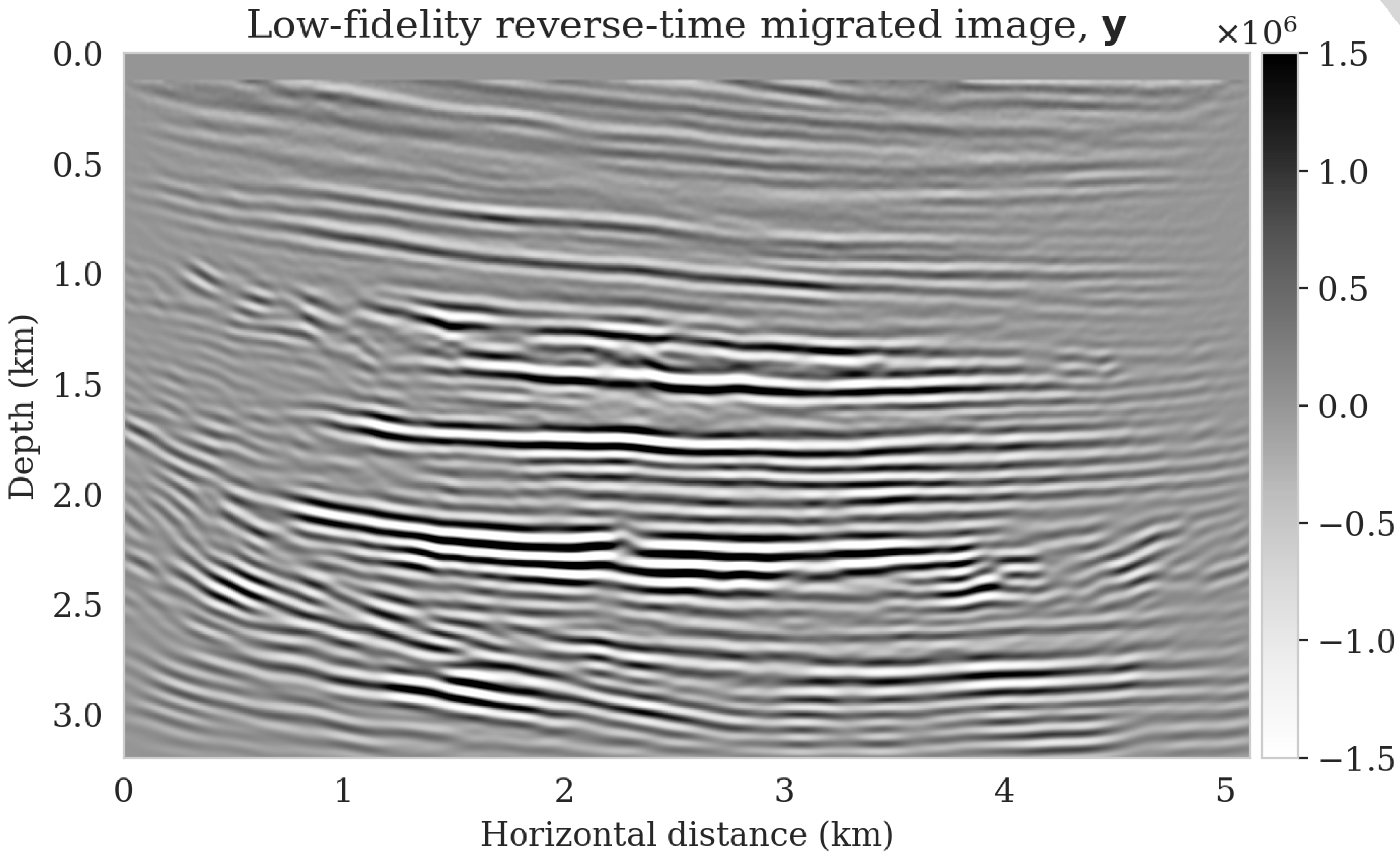


Low-fidelity reverse-time migrated image, **y**
Depth (km)
Horizontal distance (km)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
0.0
0.5
1.0
1.5
2.0
2.5
$\times 10^6$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

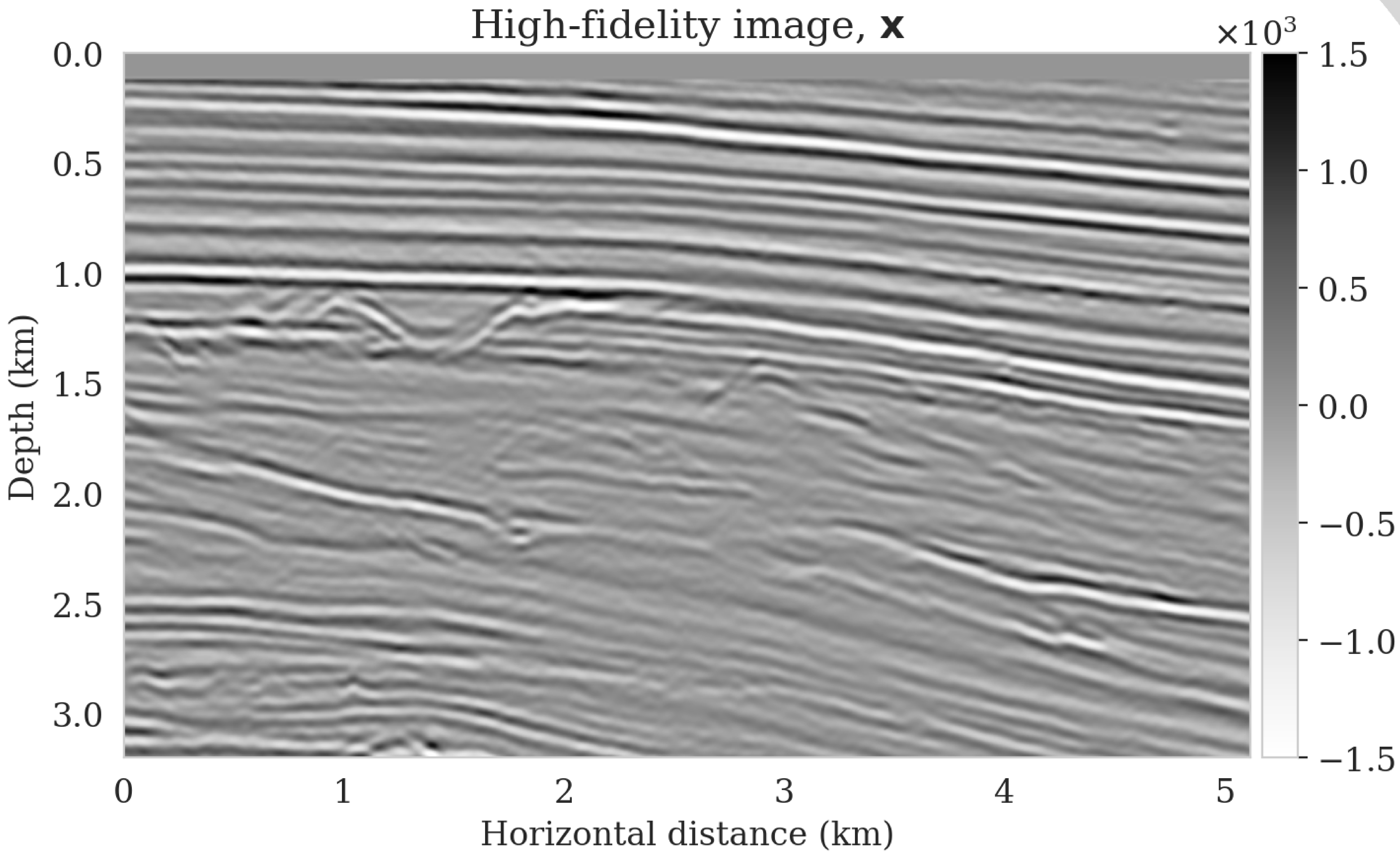
SLIM



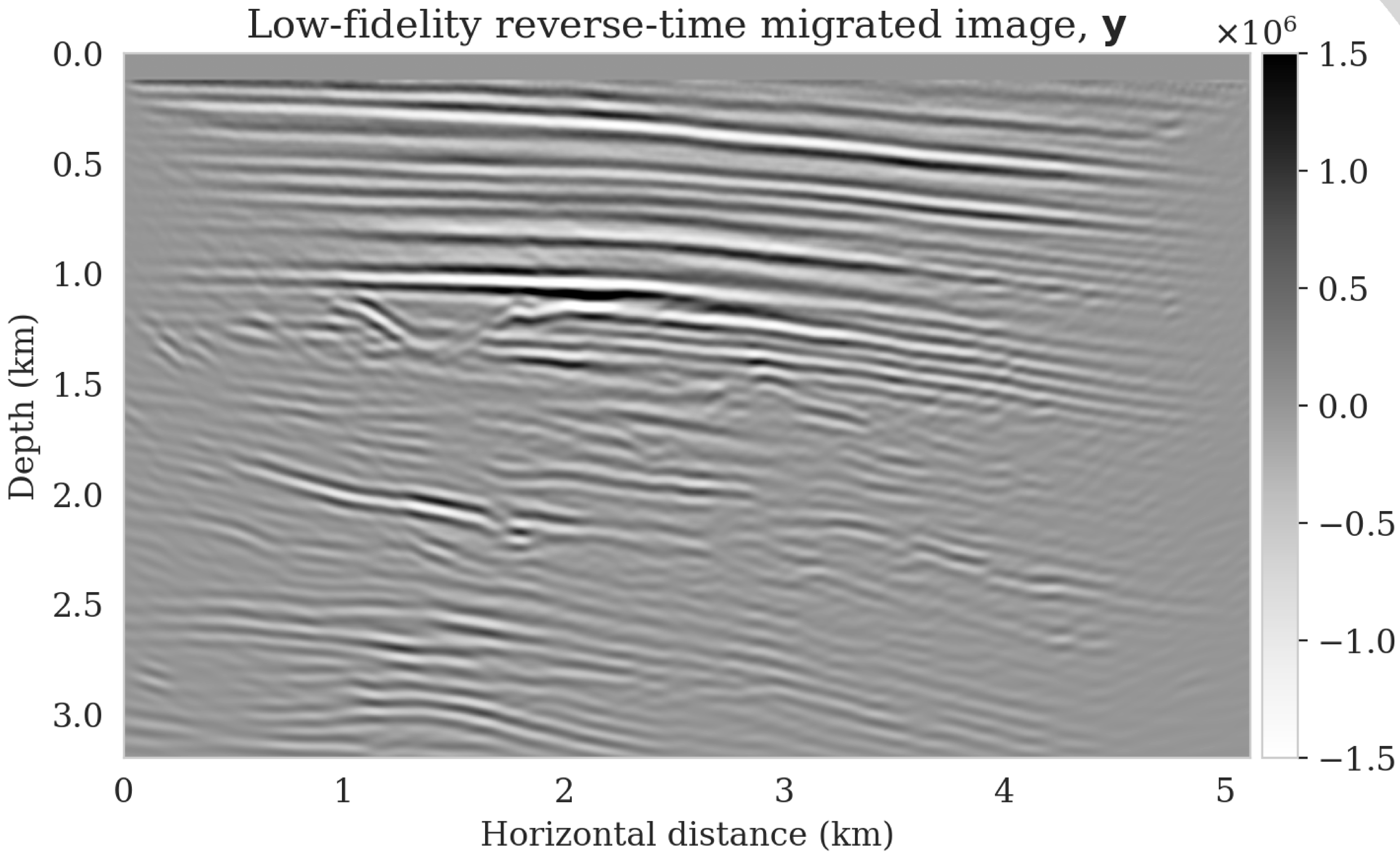
High-fidelity image, **x**
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5



Low-fidelity reverse-time migrated image, **y**
Depth (km)
Horizontal distance (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
×10^6
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

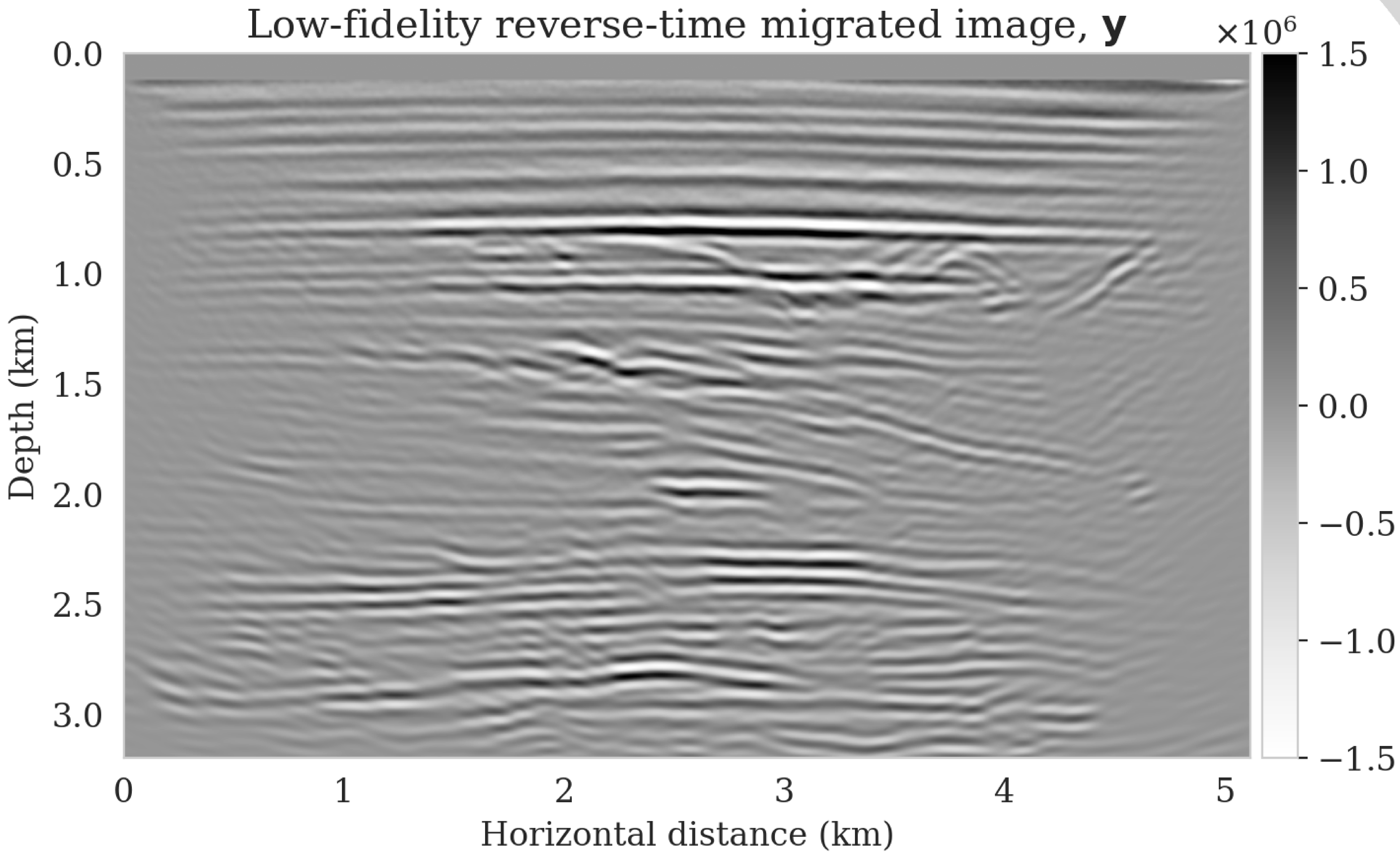


High-fidelity image, **x**
×10³
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
SLIM



Low-fidelity reverse-time migrated image, **y**
Depth (km)
Horizontal distance (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
×10^6
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

Bayesian inference w/ the data-driven approach

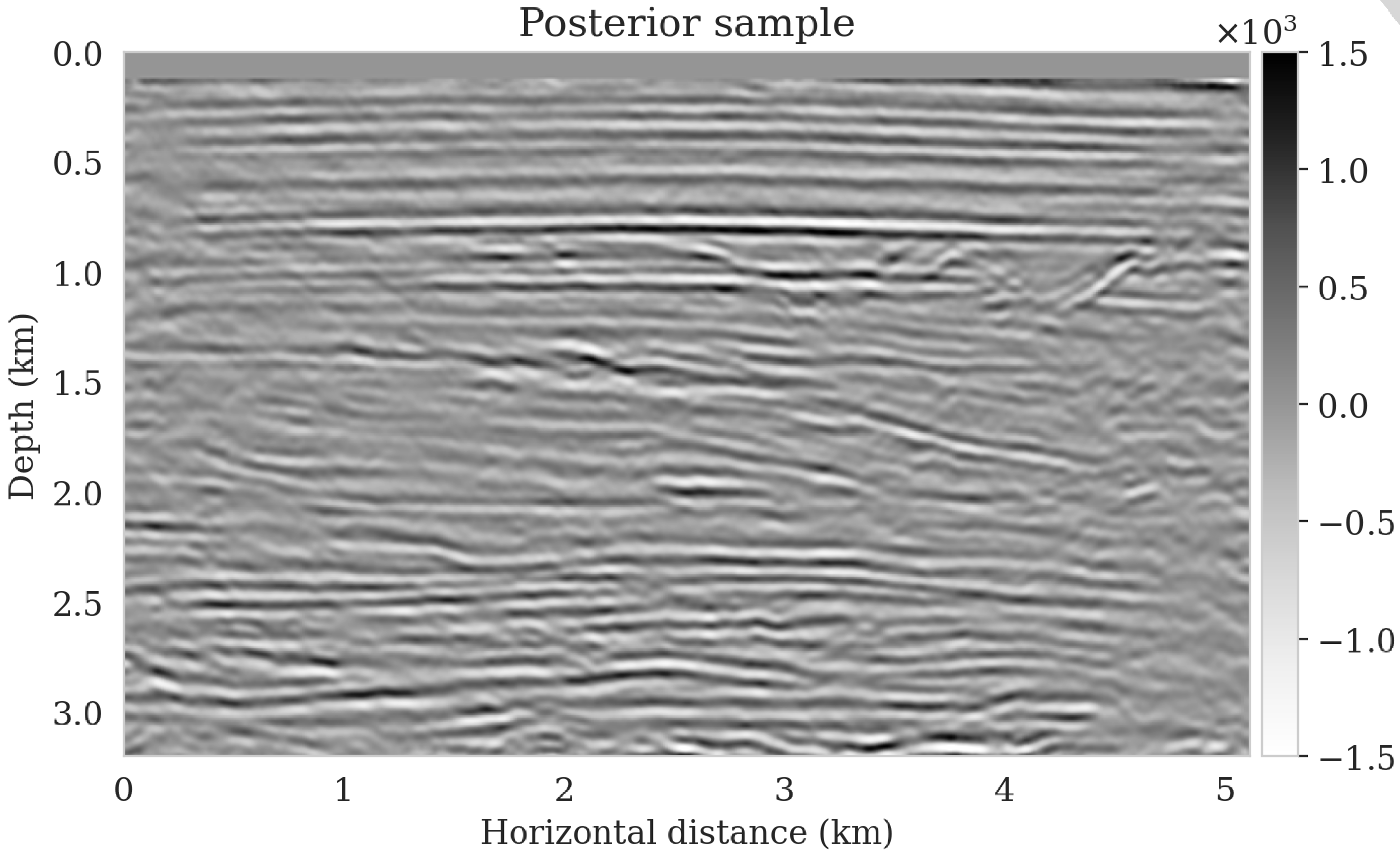


Low-fidelity reverse-time migrated image, **y**
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
$\times 10^6$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

SLIM

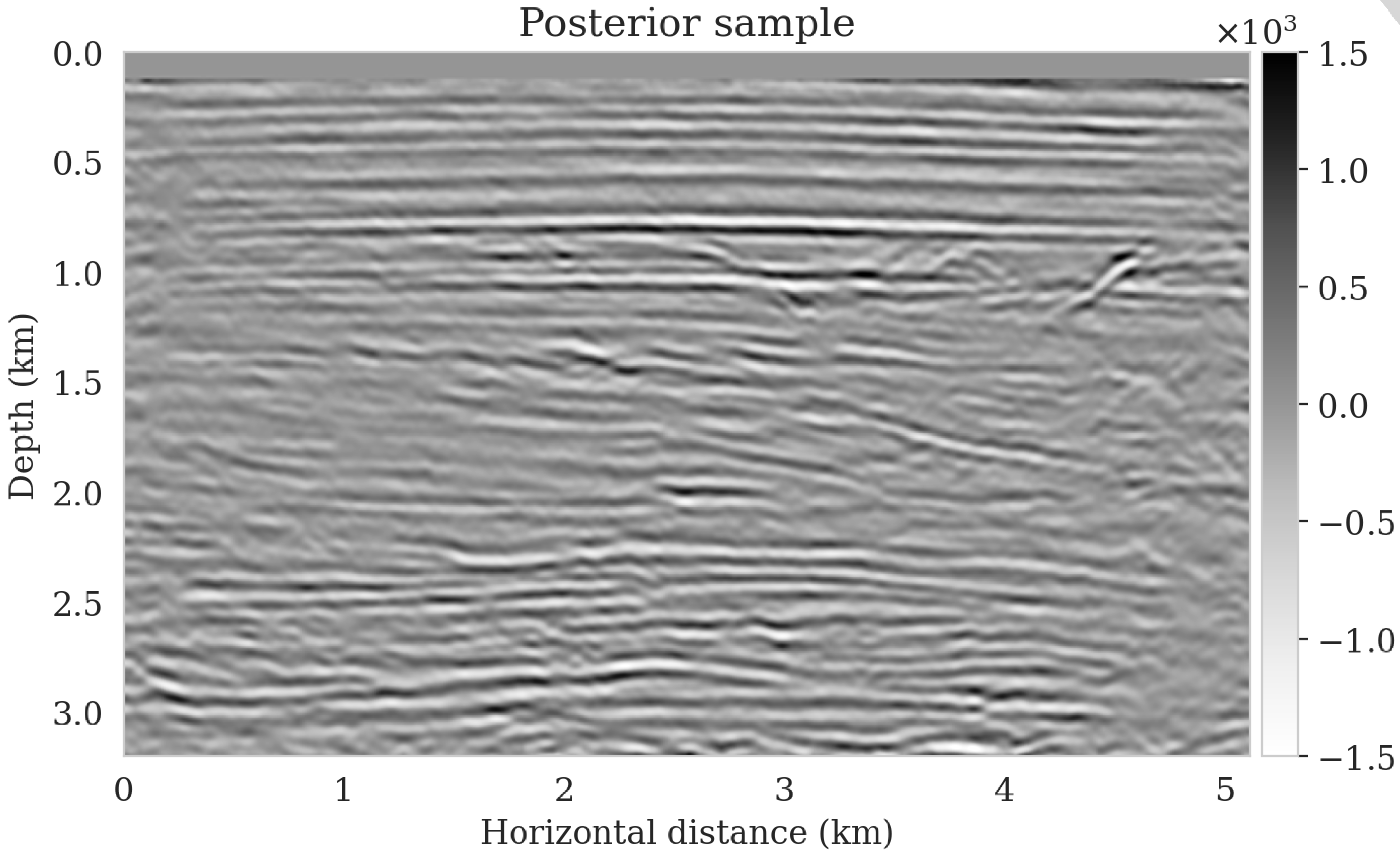
Samples from the posterior

$$G_{\theta_x}^{-1}\left(G_{\theta_y}(\mathbf{y}), \mathbf{z}\right) \sim p\left(\mathbf{x} \mid \mathbf{y}\right), \quad \mathbf{z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$$

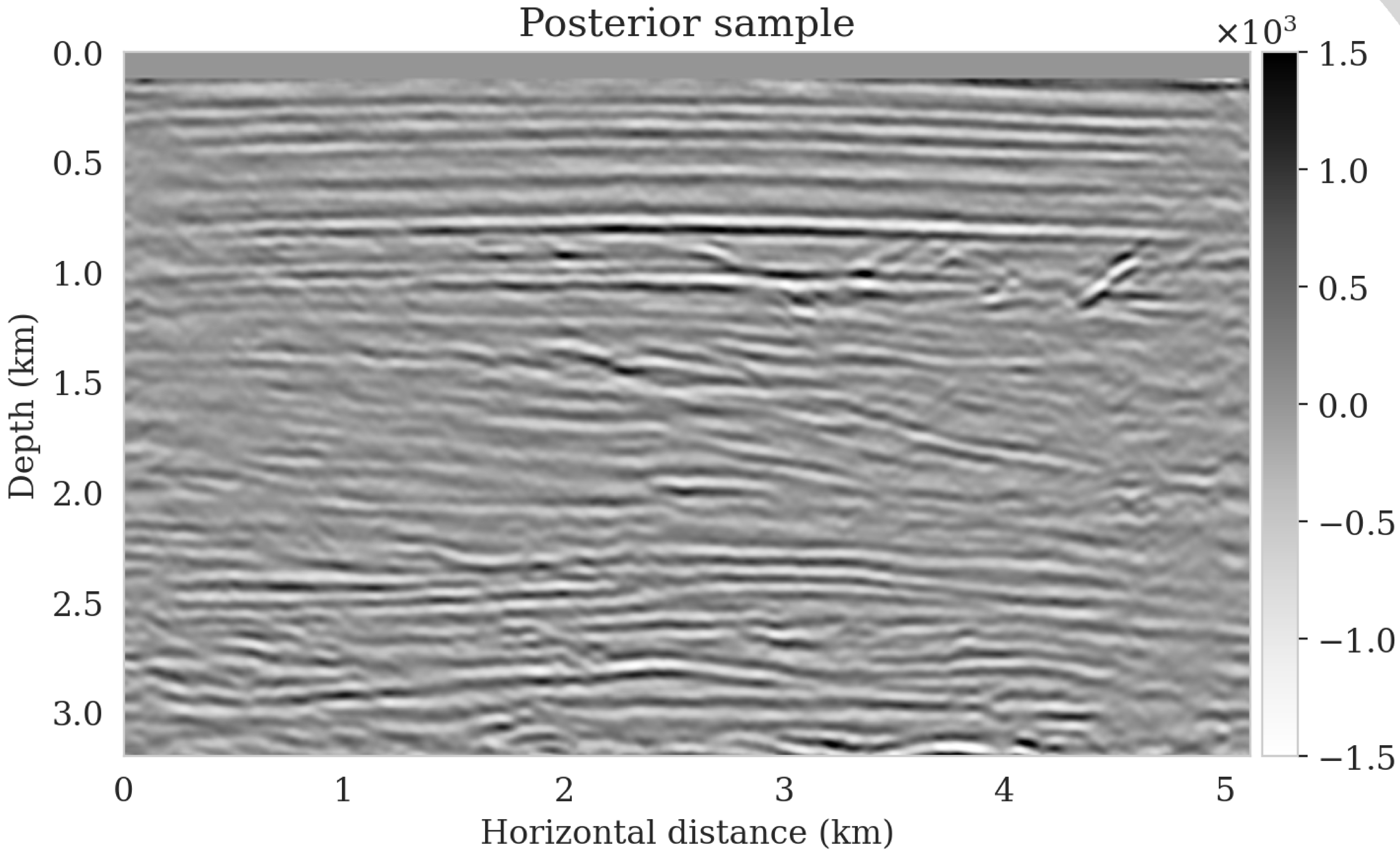


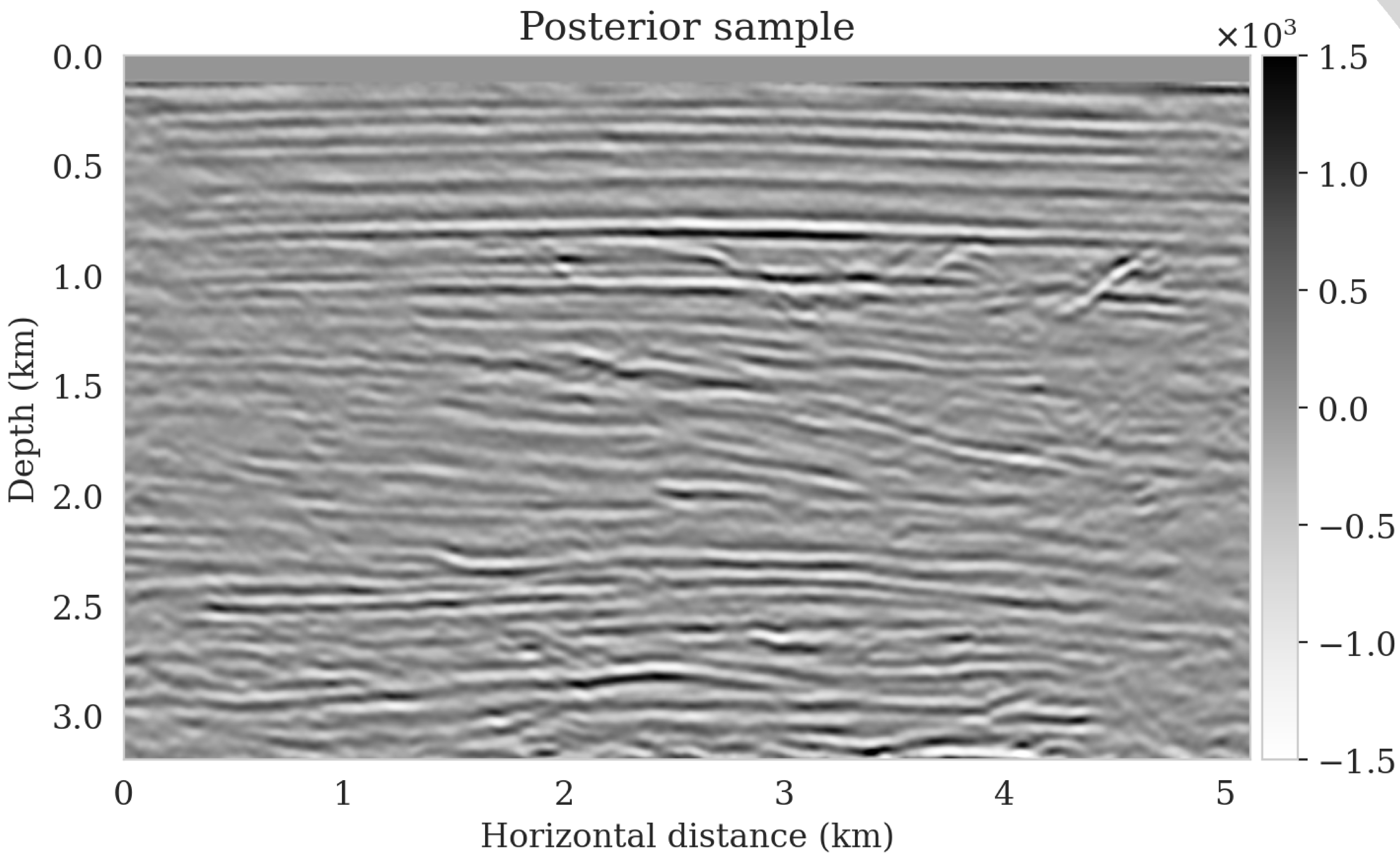
Posterior sample
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
$\times 10^3$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

SLIM



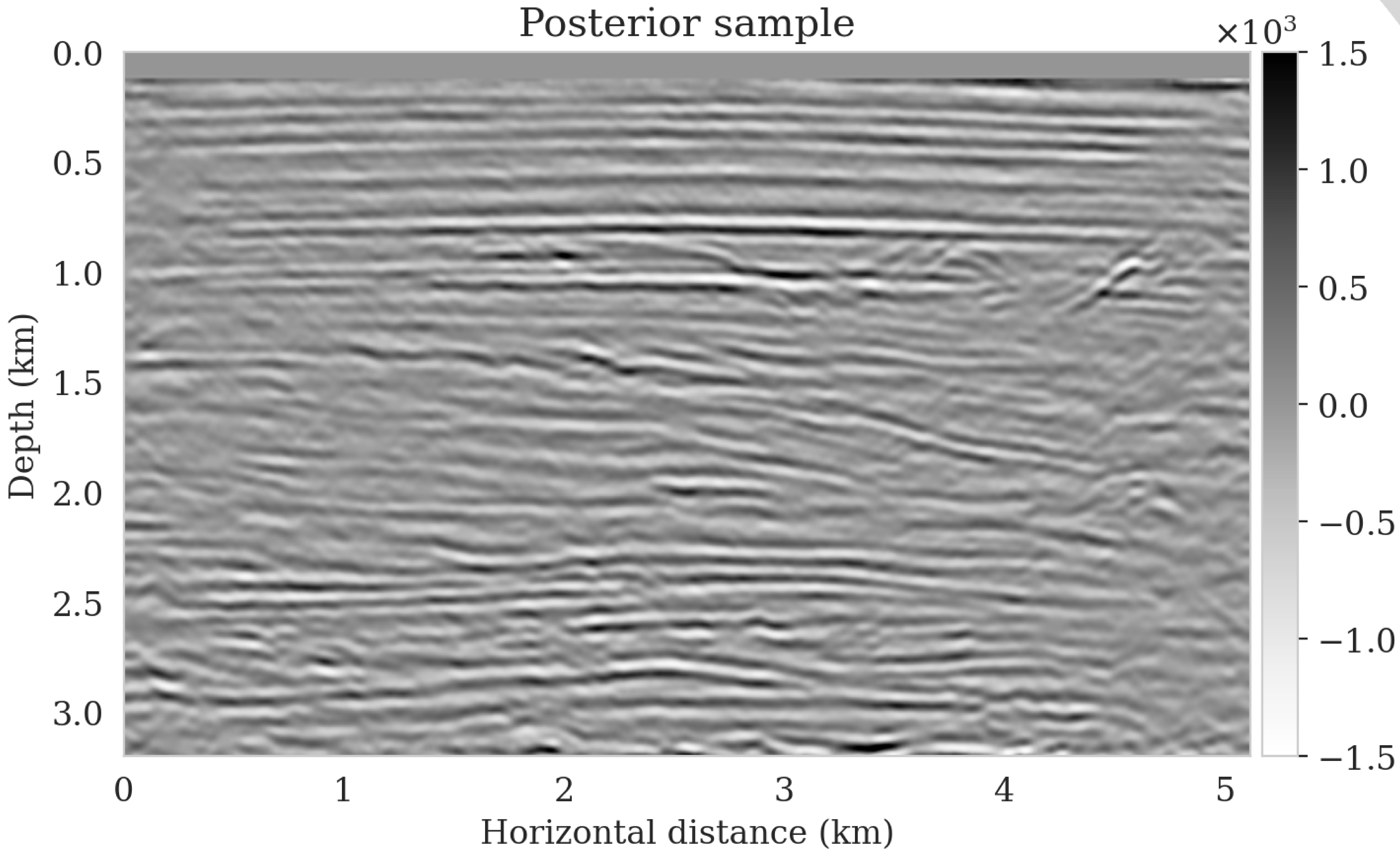
Posterior sample
×10^3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)


Posterior sample
×10^3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Horizontal distance (km)
0
1
2
3
4
5
SLIM

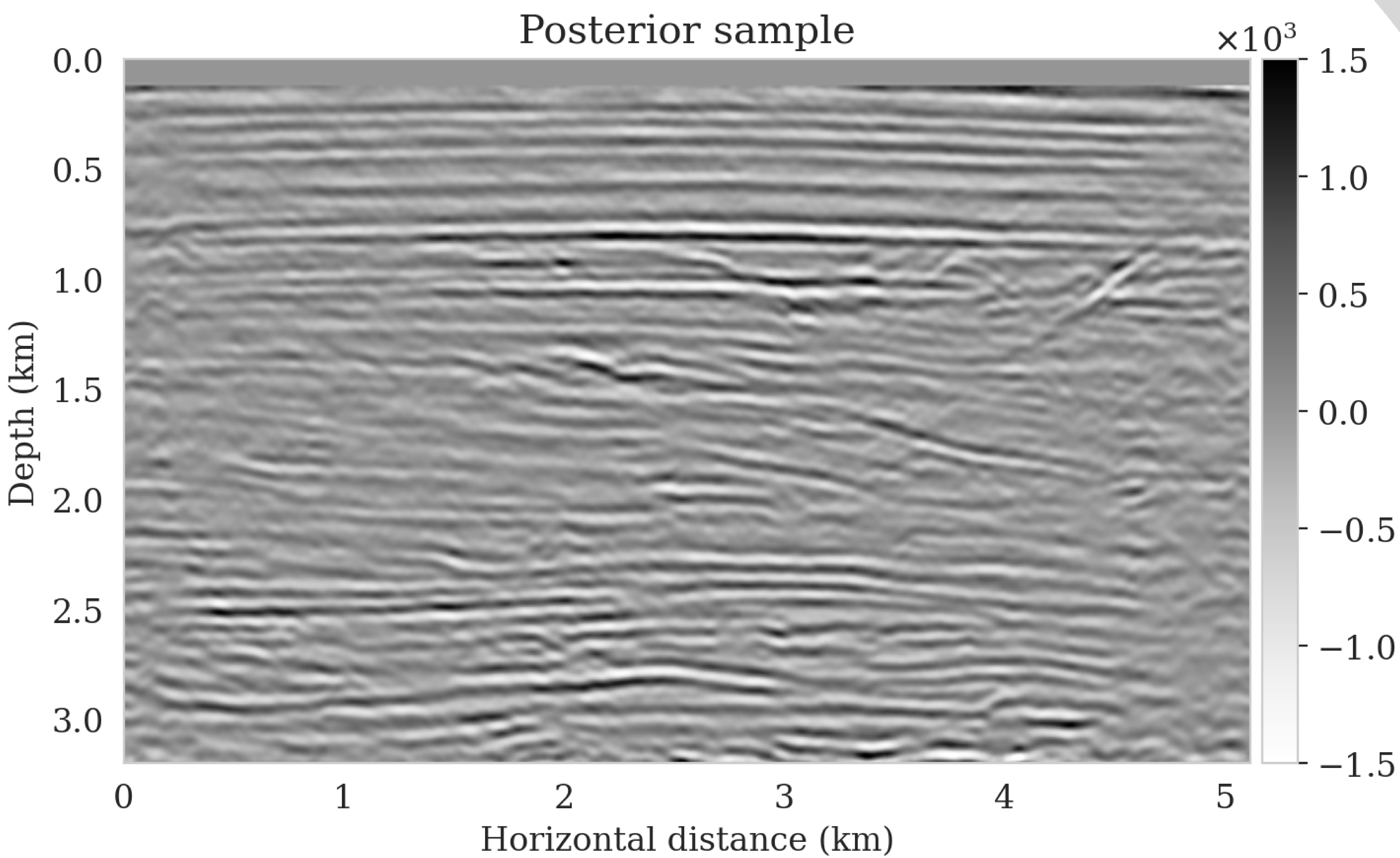


Posterior sample
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

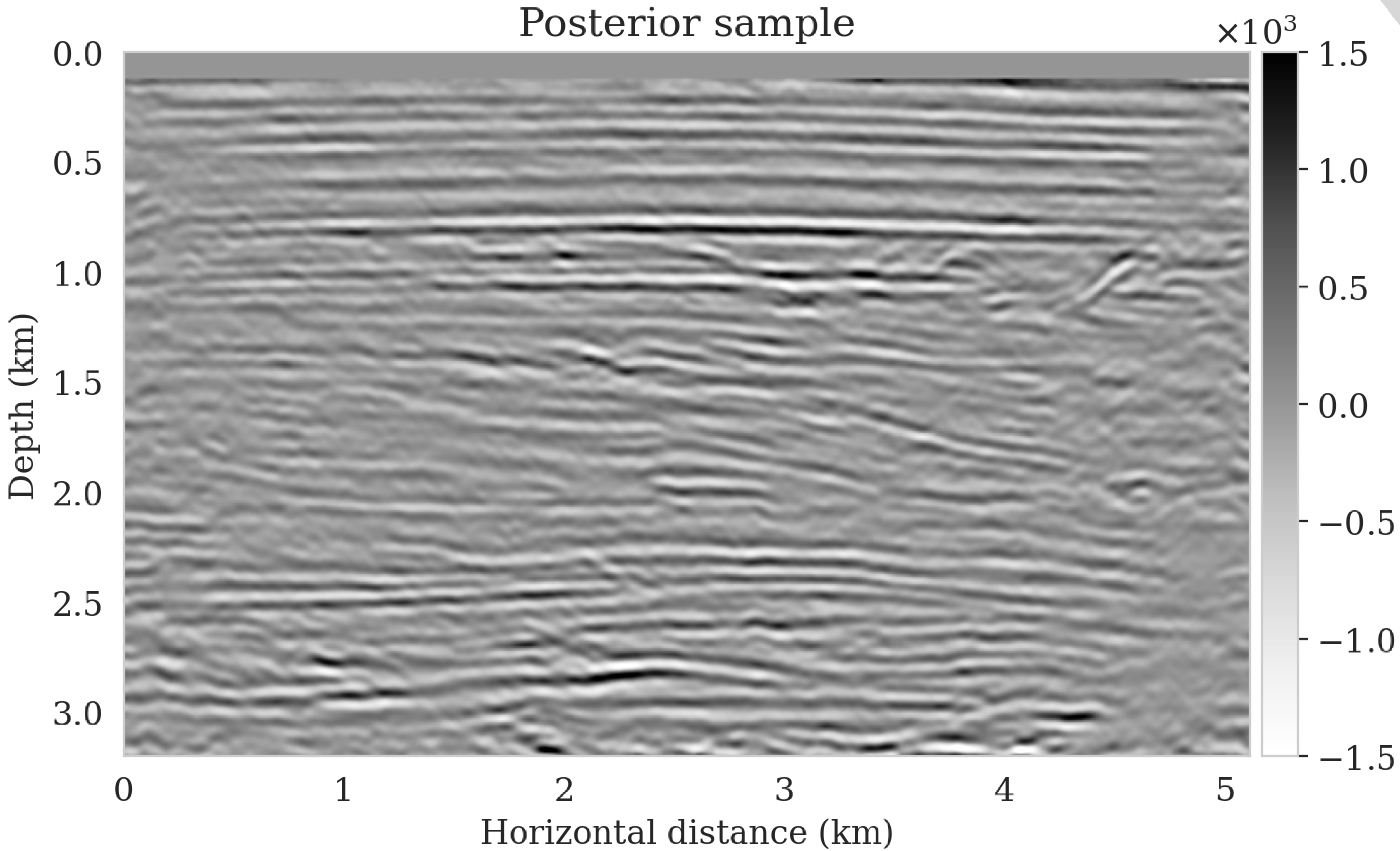
SLIM

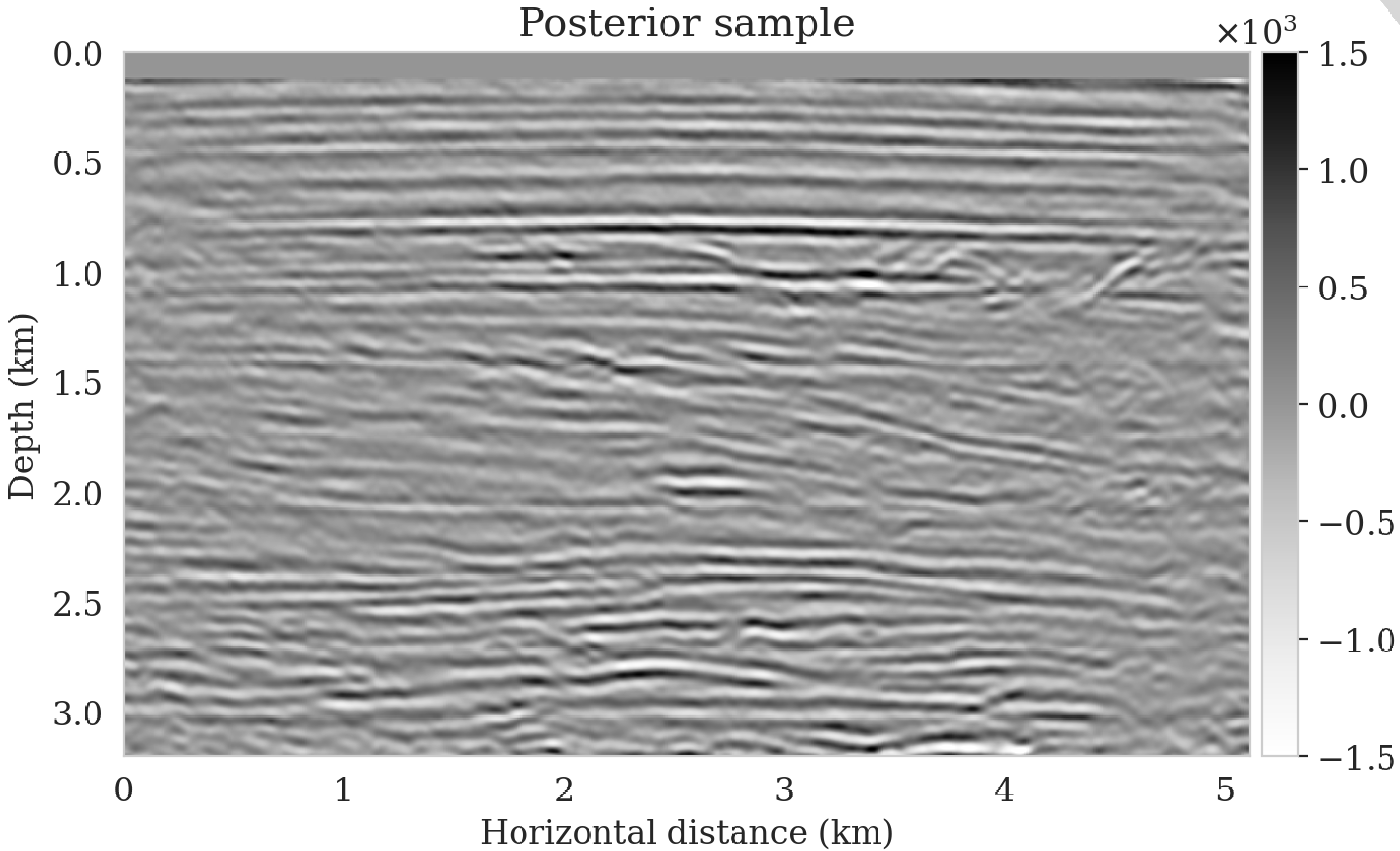

Posterior sample
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
$\times 10^3$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

SLIM

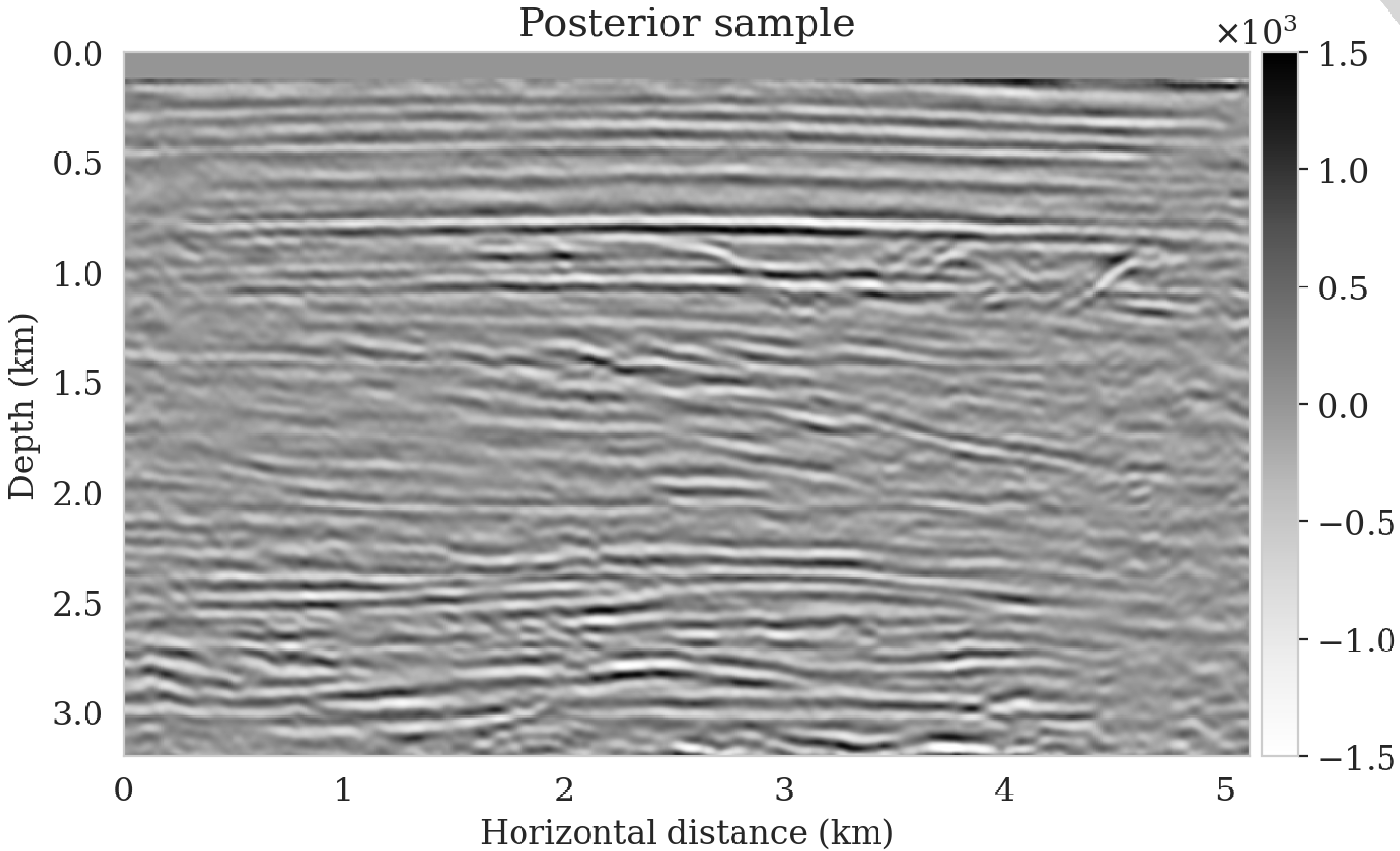


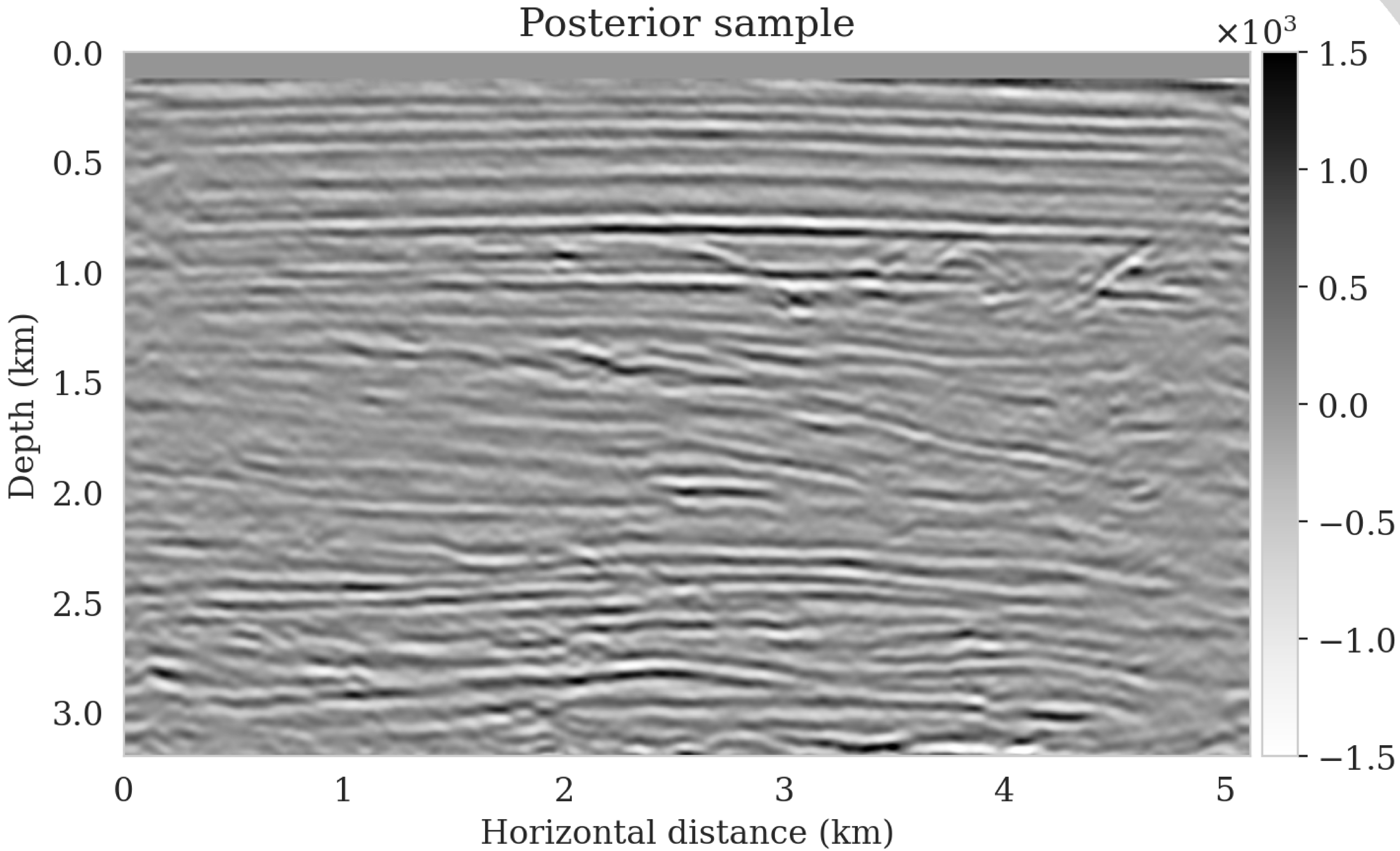
Posterior sample
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


Posterior sample
Depth (km)
Horizontal distance (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
SLIM



Posterior sample
Depth (km)
Horizontal distance (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

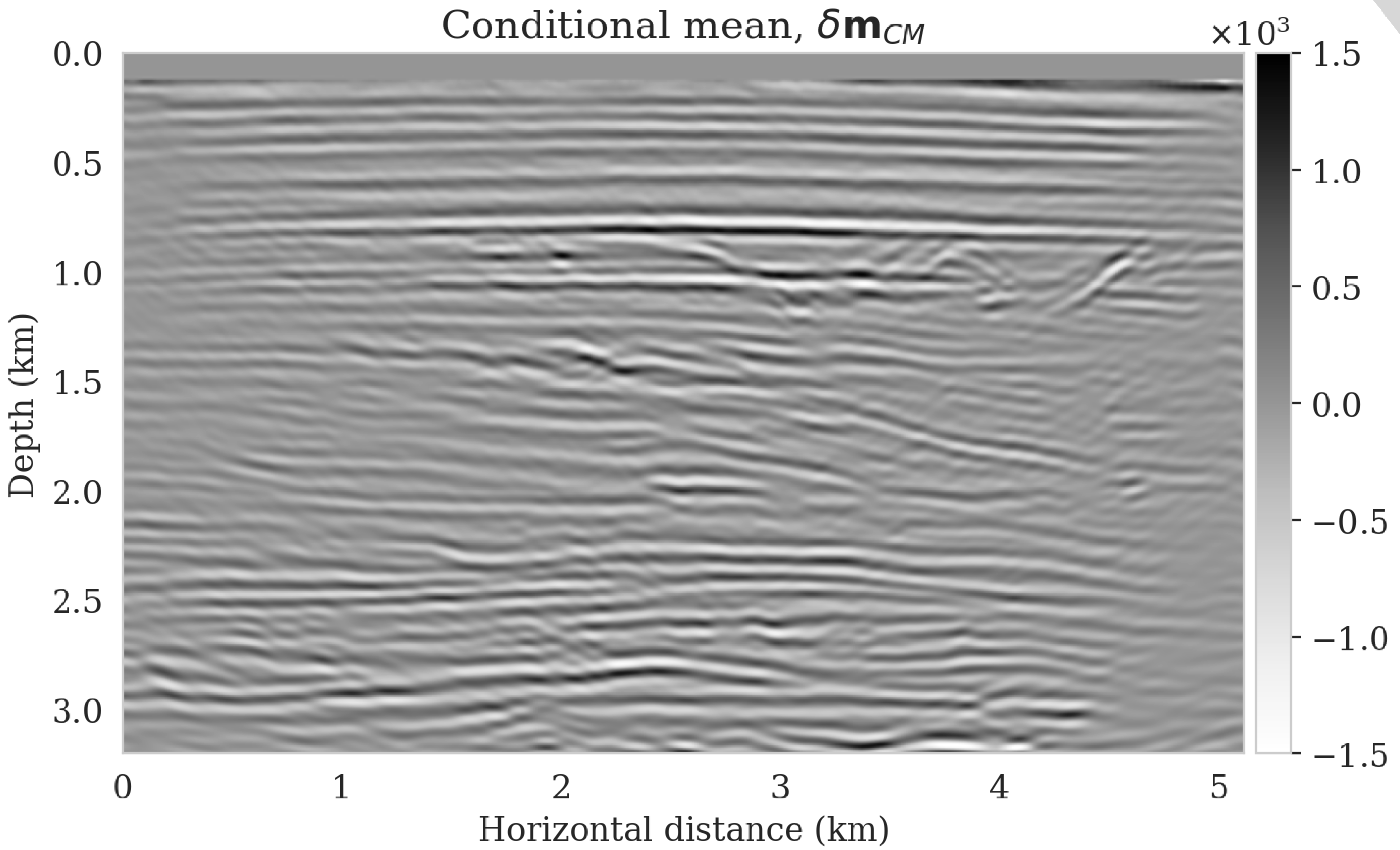

Posterior sample
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5



Posterior sample
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

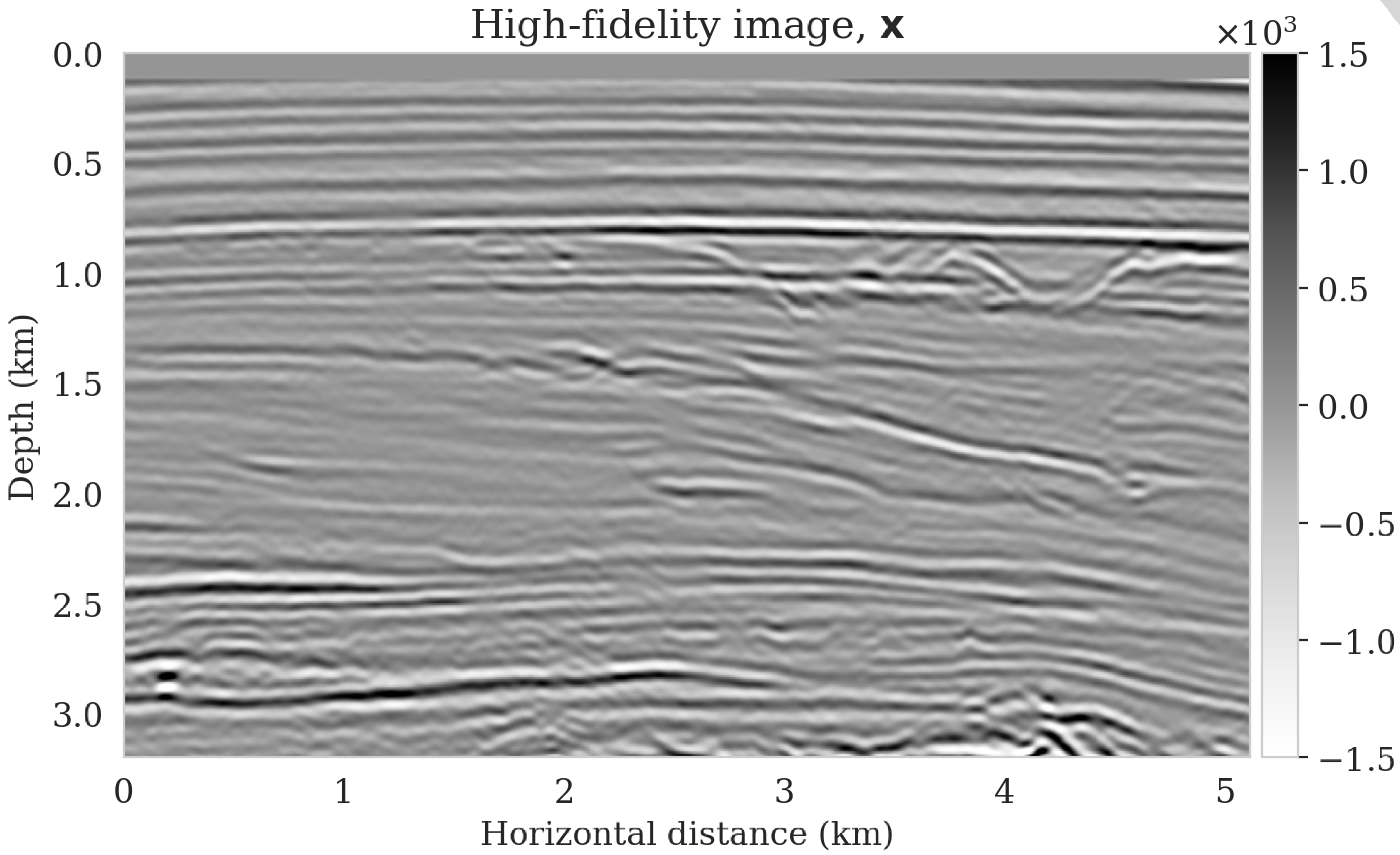
Conditional mean estimate

$$\mathbb{E}\left[\mathbf{x} \mid \mathbf{y}\right] = \int \mathbf{x} \, p\left(\mathbf{x} \mid \mathbf{y}\right) \mathrm{d}\mathbf{x}$$

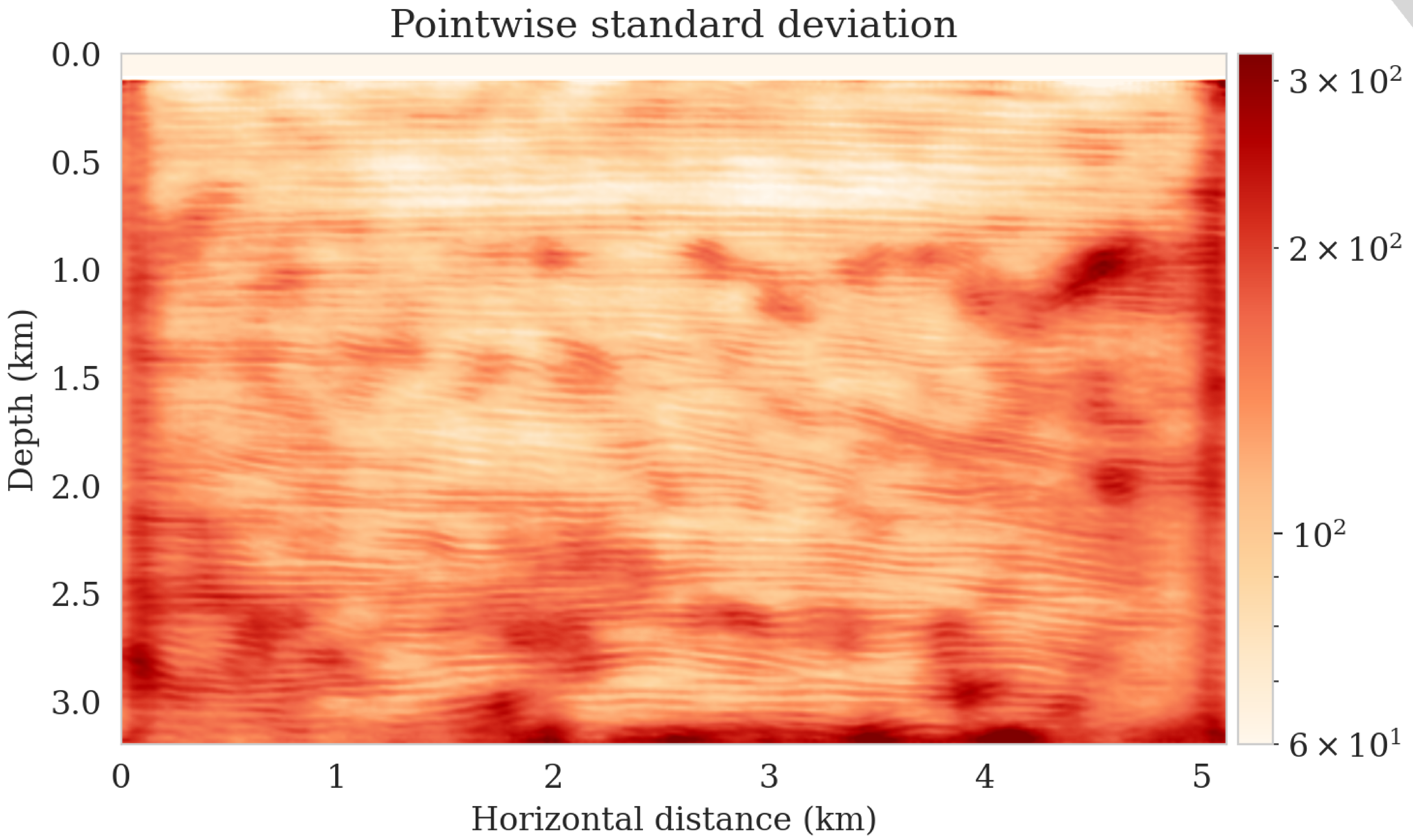


Conditional mean, $\delta \mathbf{m}_{CM}$
$\times 10^3$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)

SLIM



High-fidelity image, x
×10^3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)

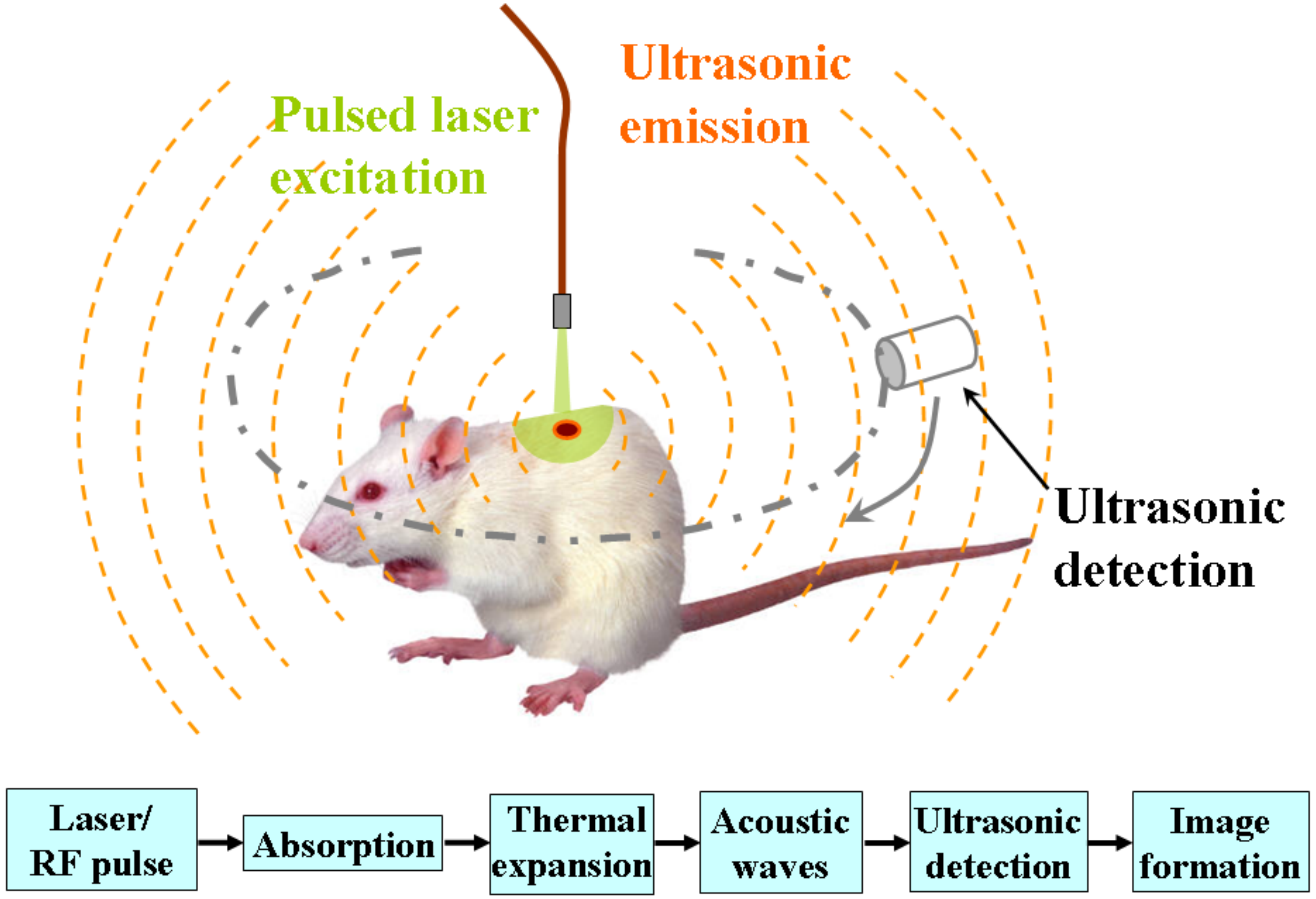


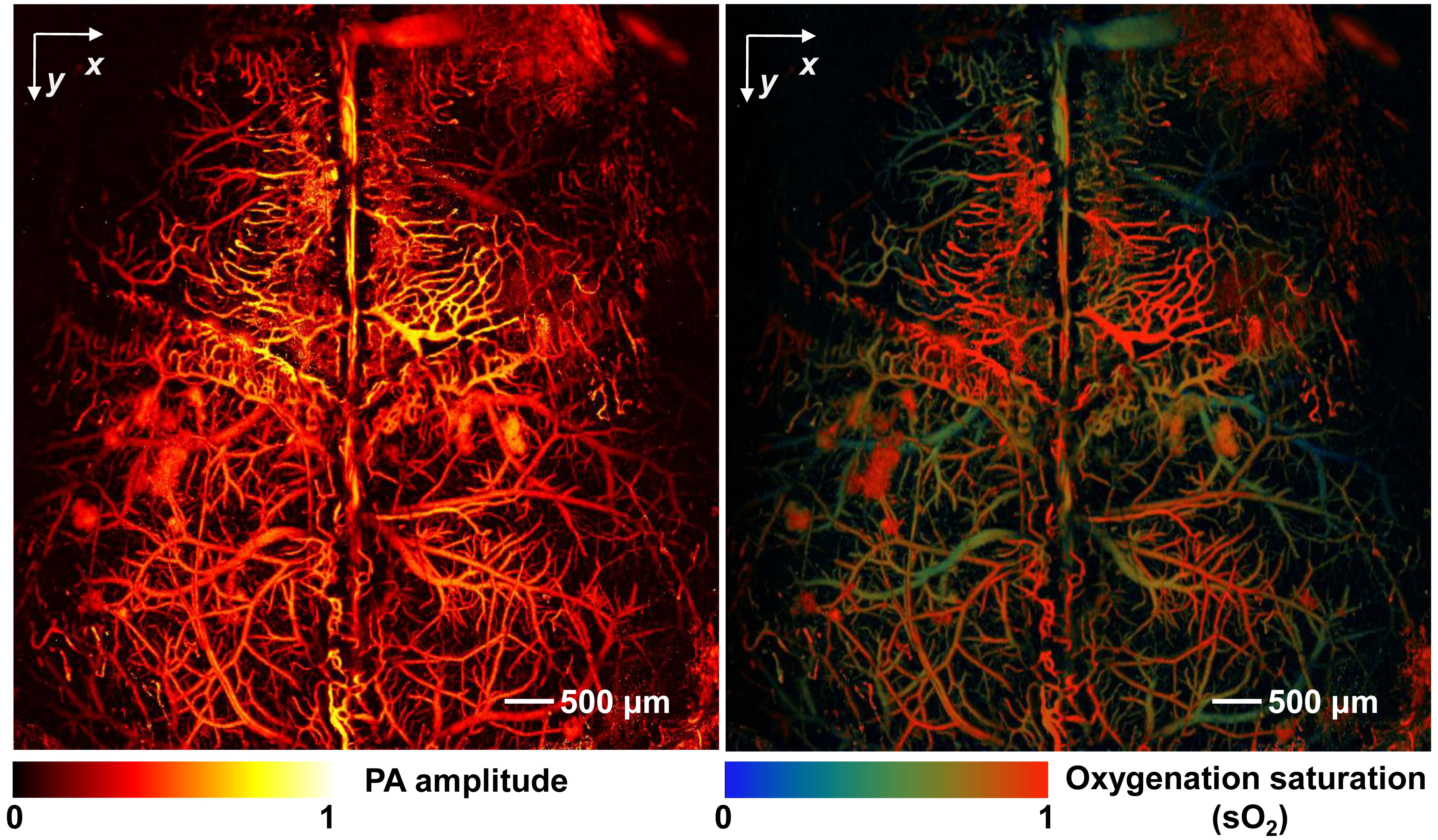
Pointwise standard deviation
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
3×10^2
2×10^2
10^2
6×10^1

Medical imaging example

Junjie Yao and Lihong Wang, WUSTL
Image reproduced from © 2007 User:Bme591wikiproject, Schematic illustration of photoacoustic imaging

Photoacoustic imaging





Photoacoustic operator A

Solve initial value PDE for acoustic pressure $\mathbf{u}$

$$\mathbf{m}\frac{\partial}{\partial t^2}\mathbf{u}(x,t) - \Delta\mathbf{u}(x,t) = 0$$

$$\mathbf{u}(x,0) = \mathbf{p}(x)$$

$$\frac{\partial}{\partial t}\mathbf{u}(x,0) = 0$$

Restrict field $\mathbf{u}$, to limited view, subsampled receivers

Training dataset for the data-driven approach

Training pairs

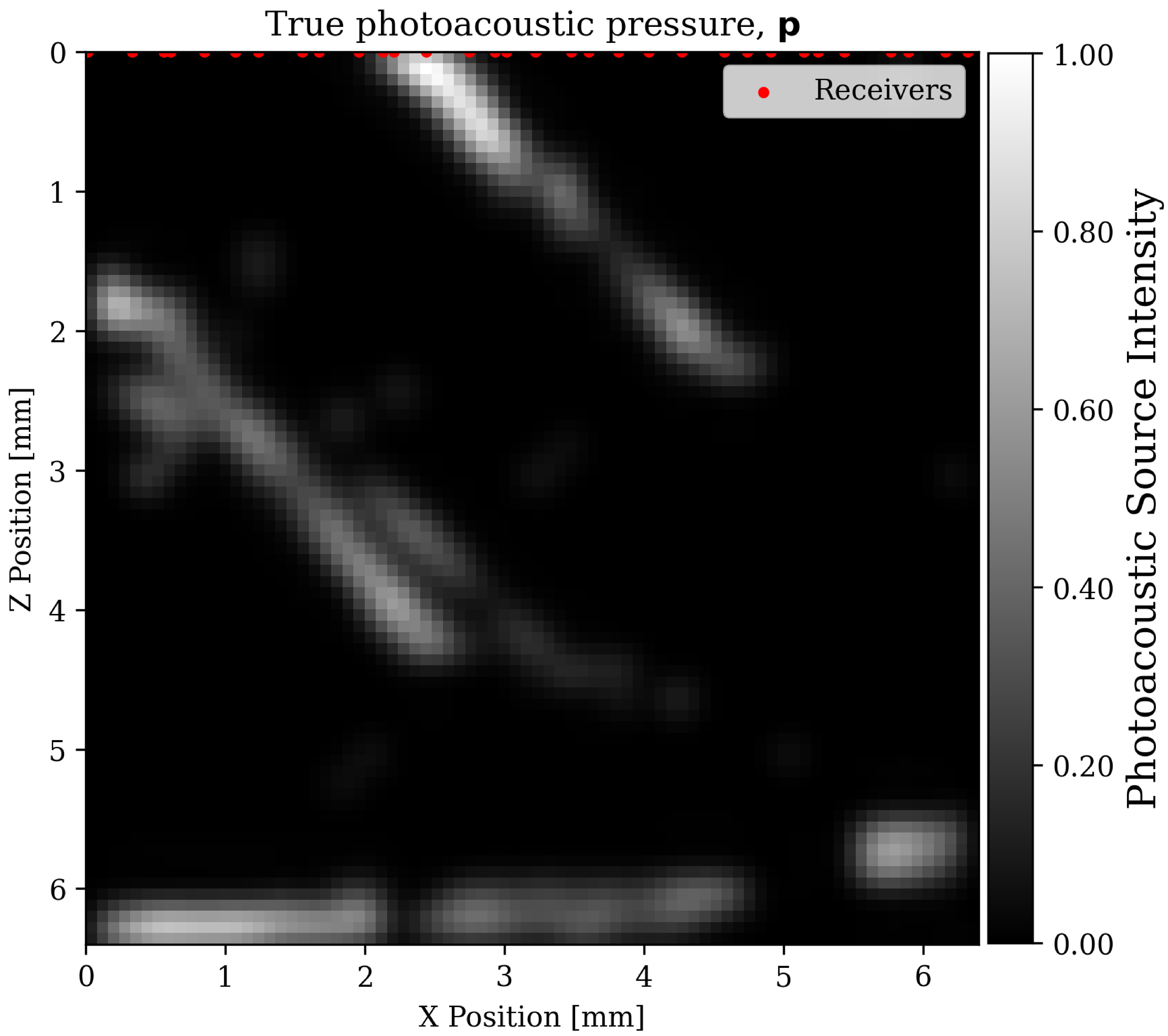
$$\{\mathbf{y}_i, \mathbf{x}_i\}_{i=1}^{n} \sim p(\mathbf{y}, \mathbf{x})$$

where

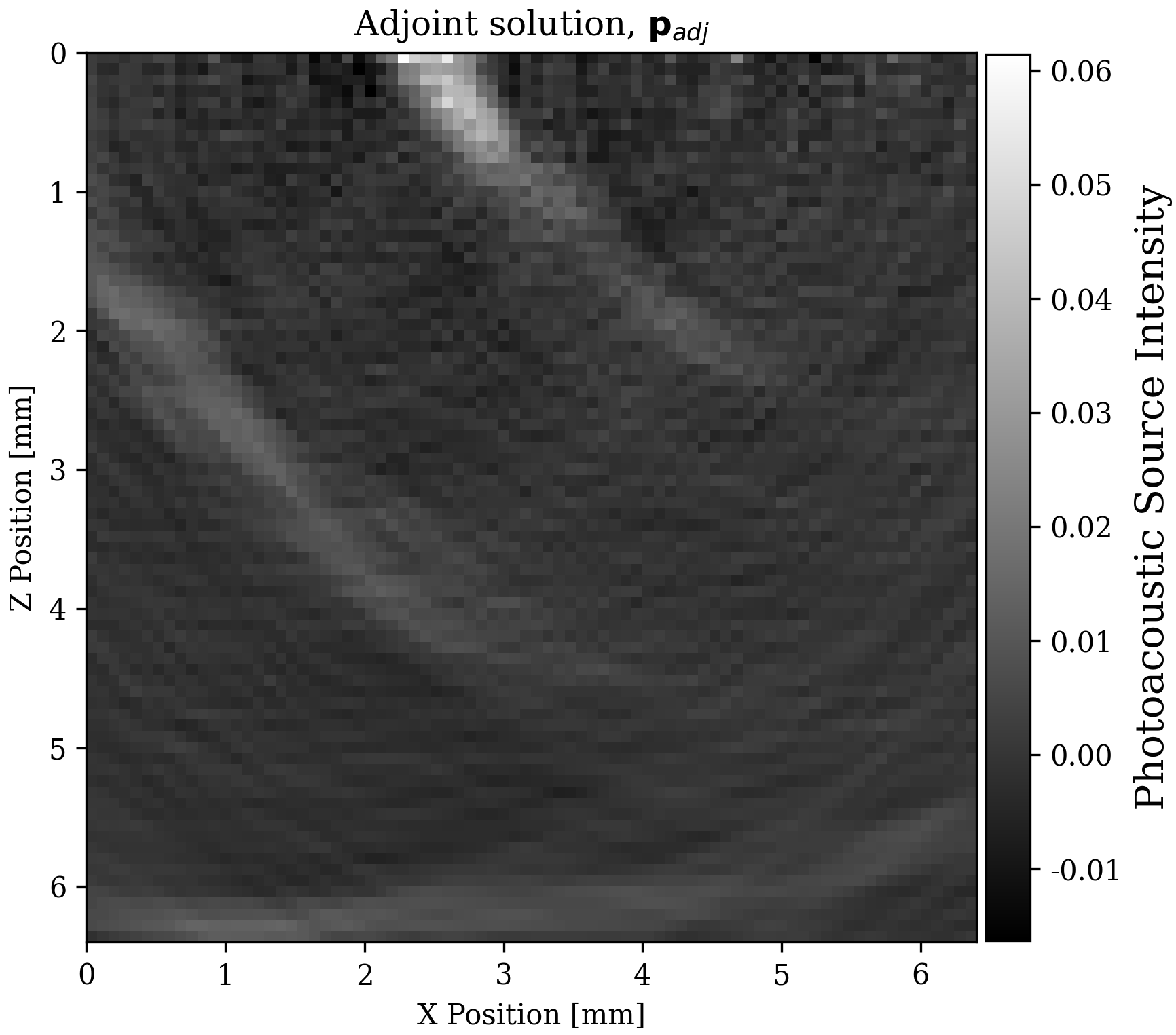
$$(\mathbf{y}_i, \mathbf{x}_i) = \left(\mathbf{A}^\top(\mathbf{A}\mathbf{p}_i + \epsilon_i), \mathbf{p}_i\right)$$

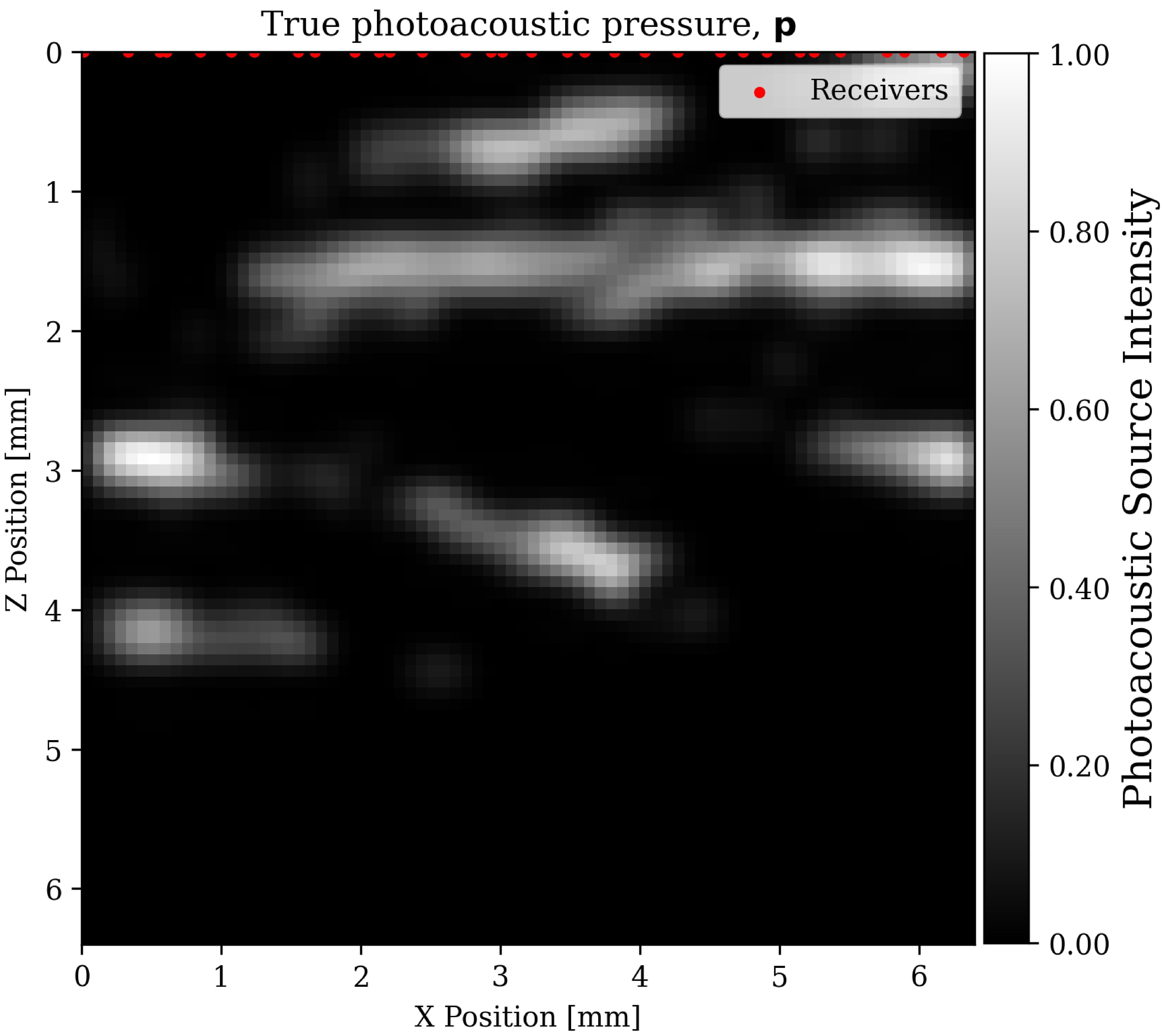
Simon R Arridge, Marta M Betcke, Ben T Cox, Felix Lucka, and Brad E Treeby. "On the adjoint operator in photoacoustic tomography". In: *Inverse Problems* 32.11 (2016), p. 115012.

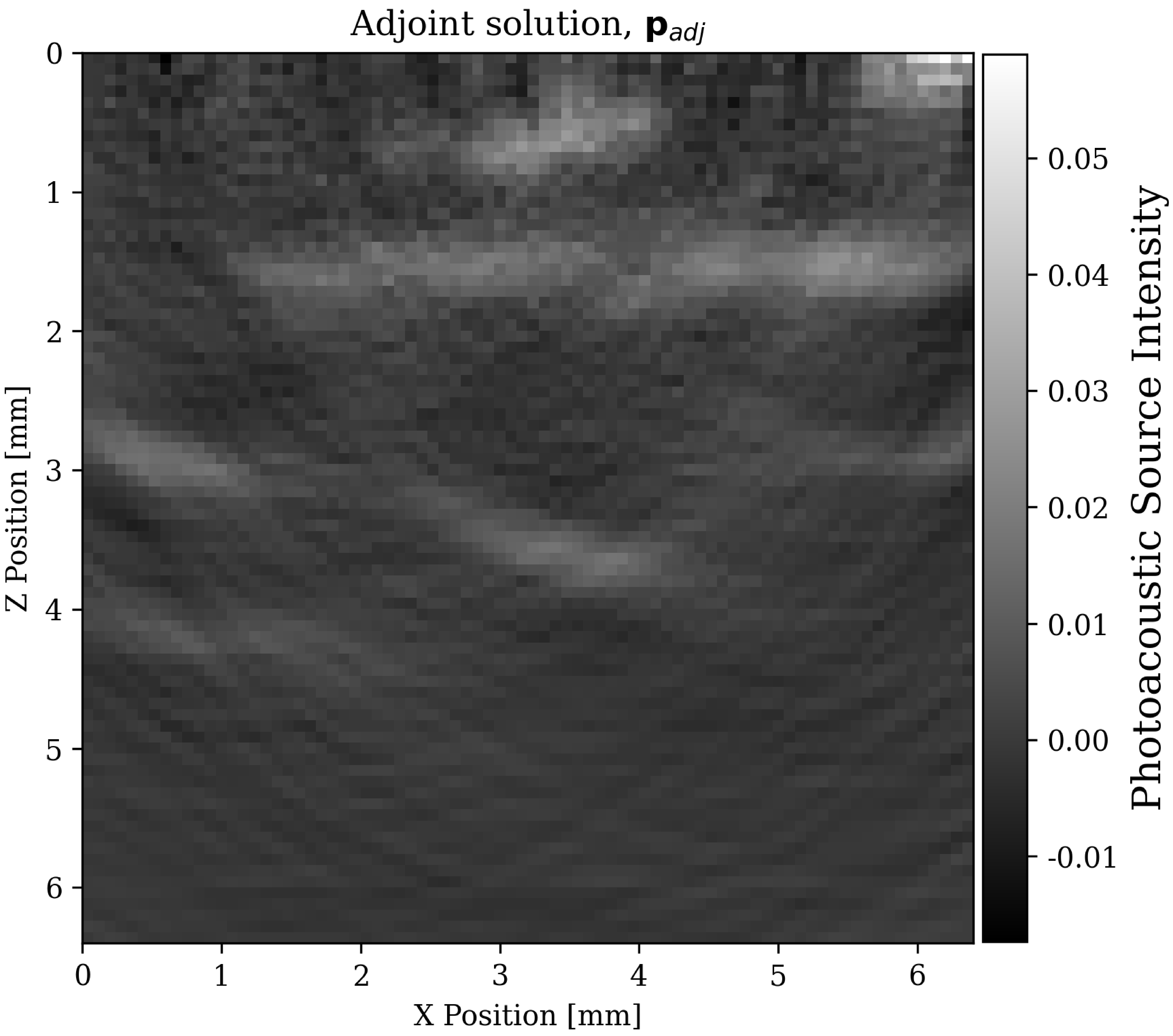
Some data pairs from the training set



True photoacoustic pressure, **p**
Receivers
Z Position [mm]
X Position [mm]
0
1
2
3
4
5
6
Photoacoustic Source Intensity
1.00
0.80
0.60
0.40
0.20
0.00

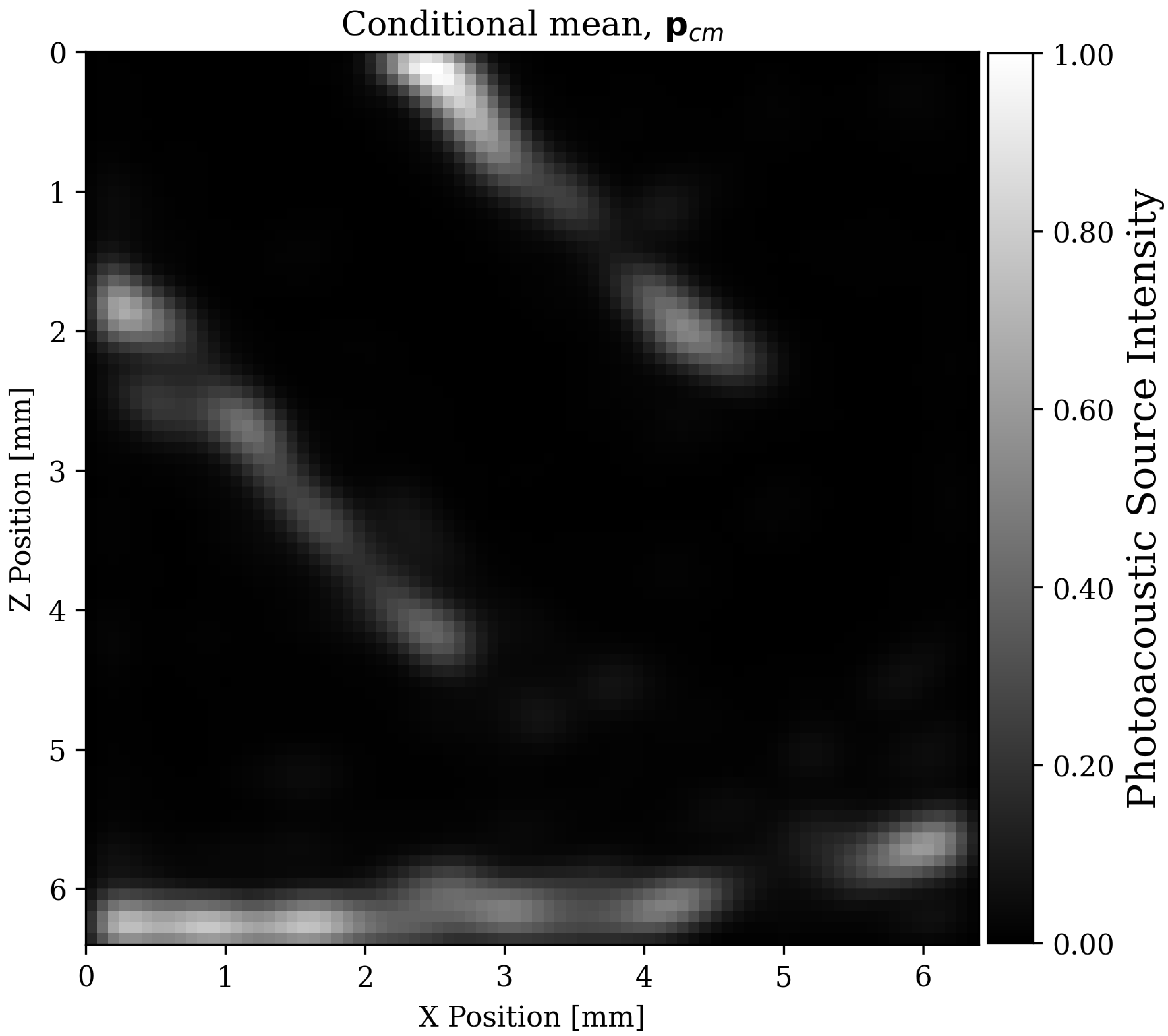

Adjoint solution, $\mathbf{p}_{adj}$
Z Position [mm]
X Position [mm]
Photoacoustic Source Intensity
0.06
0.05
0.04
0.03
0.02
0.01
0.00
-0.01


True photoacoustic pressure, **p**
Receivers
Z Position [mm]
X Position [mm]
Photoacoustic Source Intensity
0
1
2
3
4
5
6
0.00
0.20
0.40
0.60
0.80
1.00

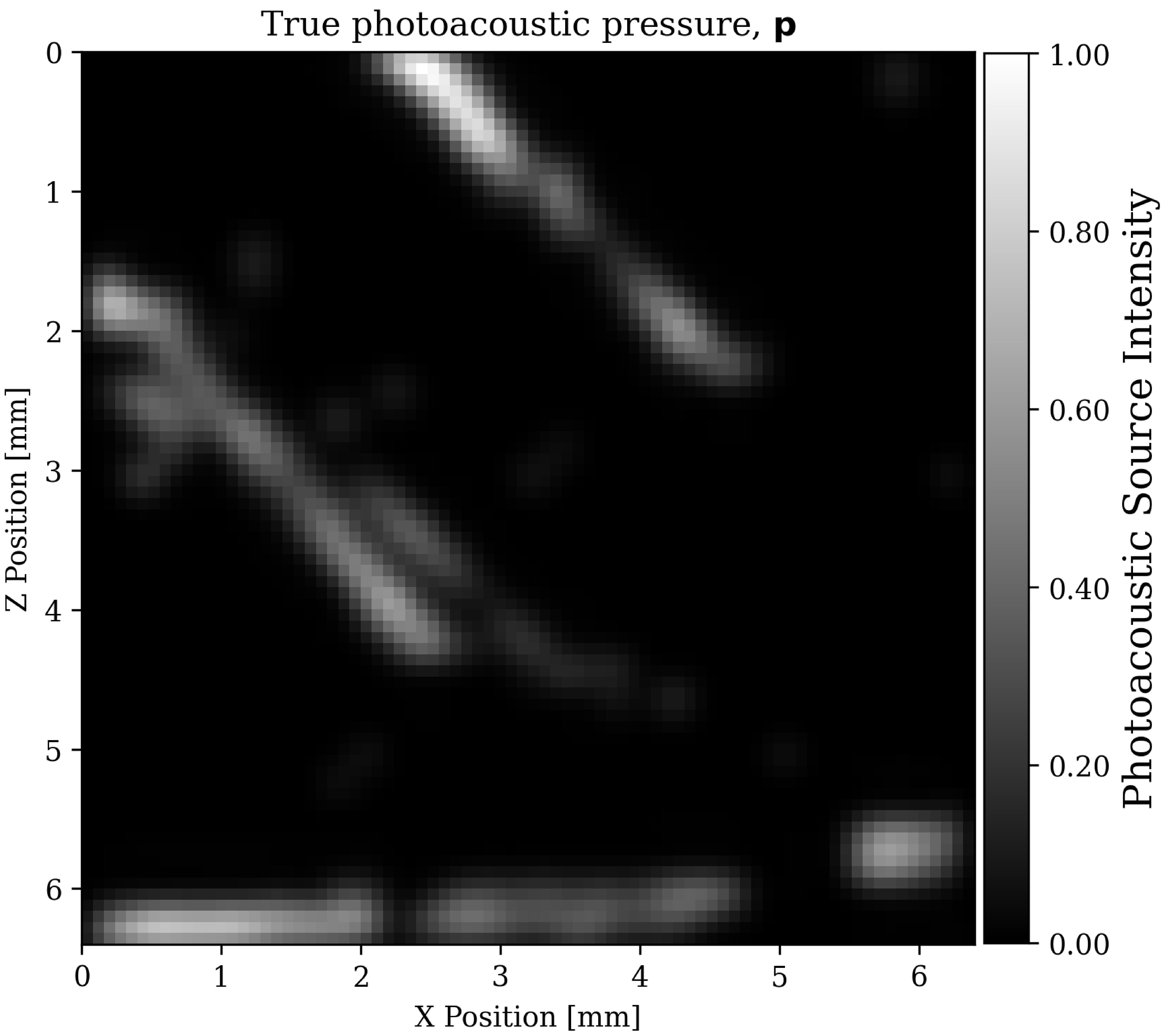


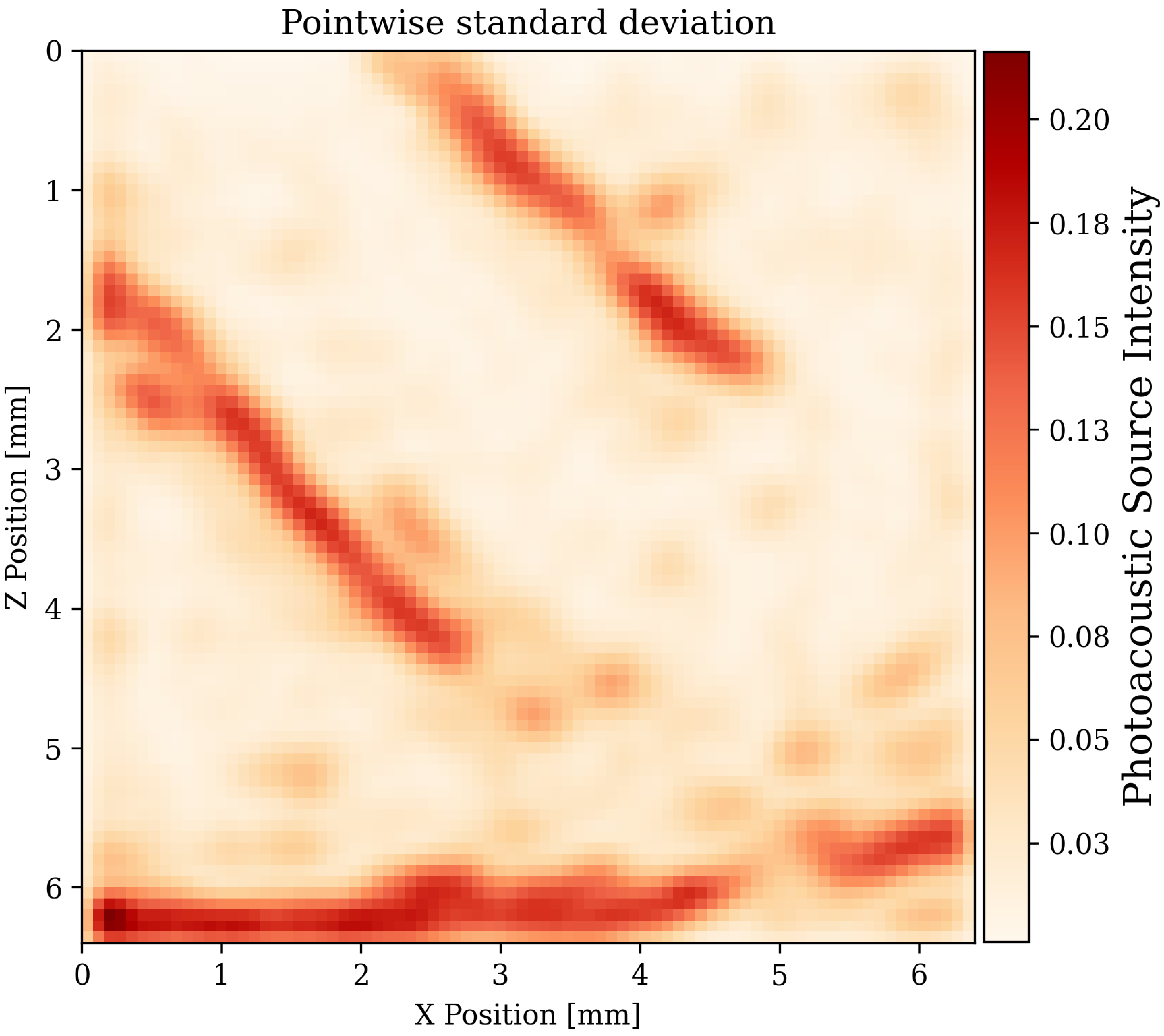
Adjoint solution, $\mathbf{p}_{adj}$
Z Position [mm]
X Position [mm]
Photoacoustic Source Intensity
0
1
2
3
4
5
6
0.05
0.04
0.03
0.02
0.01
0.00
-0.01

Bayesian inference w/ the data-driven approach



Conditional mean, $\mathbf{p}_{cm}$
0
1
2
3
4
5
6
Z Position [mm]
X Position [mm]
1.00
0.80
0.60
0.40
0.20
0.00
Photoacoustic Source Intensity
SLIM


True photoacoustic pressure, **p**
Z Position [mm]
X Position [mm]
Photoacoustic Source Intensity
0
1
2
3
4
5
6
0.00
0.20
0.40
0.60
0.80
1.00



Pointwise standard deviation
Z Position [mm]
X Position [mm]
Photoacoustic Source Intensity
0
1
2
3
4
5
6
0.20
0.18
0.15
0.13
0.10
0.08
0.05
0.03

Purely data-driven approach

SLIM

learns the prior distribution from training data

samples from the posterior virtually for free in test time

not specific to one observation $\mathbf{y}$

needs supervised pairs of model and data

heavily relies on the training data to generalize

not directly tied to physics/data

Purely physics-based approach

$$\arg\min_{\phi} \mathbb{D}_{\mathsf{KL}}\left(p_\phi \,||\, p(\,\cdot \mid \mathbf{y})\right) = \arg\min_{\phi} \mathbb{E}_{\mathbf{x}\sim p_\phi(\mathbf{x})}\Big[-\log p(\mathbf{x} \mid \mathbf{y}) + \log p_\phi(\mathbf{x})\Big]$$

$$= \arg\min_{\phi} \mathbb{E}_{\mathbf{z}\sim \mathrm{N}(\mathbf{0},\mathbf{I})}\left[\frac{1}{2\sigma^2}\left\|F\big(T_\phi(\mathbf{z})\big) - \mathbf{y}\right\|_2^2 - \log p\big(T_\phi(\mathbf{z})\big) - \log\left|\det \nabla_z T_\phi(\mathbf{z})\right|\right]$$

NF $T_\phi : \mathcal{Z} \to \mathcal{X}$

prior distribution $p(\mathbf{x})$

specific to one observation, $\mathbf{y}$

Gabrio Rizzuti, Ali Siahkoohi, Philipp A. Witte, and Felix J. Herrmann. "Parameterizing uncertainty by deep invertible networks, an application to reservoir characterization". In: *SEG Technical Program Expanded Abstracts*. Sept. 2020, pp. 1541–1545. DOI: 10.1190/segam2020-3428150.1.

Ali Siahkoohi, Gabrio Rizzuti, Mathias Louboutin, Philipp Witte, and Felix J. Herrmann. "Preconditioned training of normalizing flows for variational inference in inverse problems". In: *3rd Symposium on Advances in Approximate Bayesian Inference*. Jan. 2021. URL: https://openreview.net/pdf?id=P9m1sMaNQ8T.

Xuebin Zhao, Andrew Curtis, and Xin Zhang. "Bayesian Seismic Tomography using Normalizing Flows". In: (2020). DOI: 10.31223/X53K6G. URL: https://eartharxiv.org/repository/view/1940/.

Purely physics-based approach

tied to the physics and data

no training data needed

requires choosing a prior distribution $p(\mathbf{x})$

repeated evaluations of F and $\nabla F^{\top}$

specific to one observation, $\mathbf{y}$

Multi-fidelity preconditioned scheme

exploit the information in the pretrained conditional NF (data-driven approach)

tie the results to physics and data

Multi-fidelity preconditioned scheme

Use the density encoded by the pretrained conditional NF as a (conditional) prior

$$p_{\text{prior}}(\mathbf{x}) := p_{\mathbf{z}}\big(G_{\theta_x}(\mathbf{y}, \mathbf{x})\big) \left| \det \nabla_x G_{\theta_x}(\mathbf{y}, \mathbf{x}) \right|$$

allows for using a data-driven prior

removes the bias introduced by hand-crafted priors

can be used as a prior density in inverse problems

Ali Siahkoohi, Gabrio Rizzuti, Mathias Louboutin, Philipp Witte, and Felix J. Herrmann. "Preconditioned training of normalizing flows for variational inference in inverse problems". In: *3rd Symposium on Advances in Approximate Bayesian Inference*. Jan. 2021. URL: https://openreview.net/pdf?id=P9m1sMaNQ8T.

Preconditioned MAP estimation

SLIM

$$\min_{\mathbf{z}} \frac{1}{2\sigma^2} \left\| F\big(T_{\theta_x}(\mathbf{z};\mathbf{y})\big) - \mathbf{y} \right\|_2^2 + \frac{1}{2} \|\mathbf{z}\|_2^2$$

initializing $\mathbf{z} = \mathbf{0}$ acts as an implicit prior

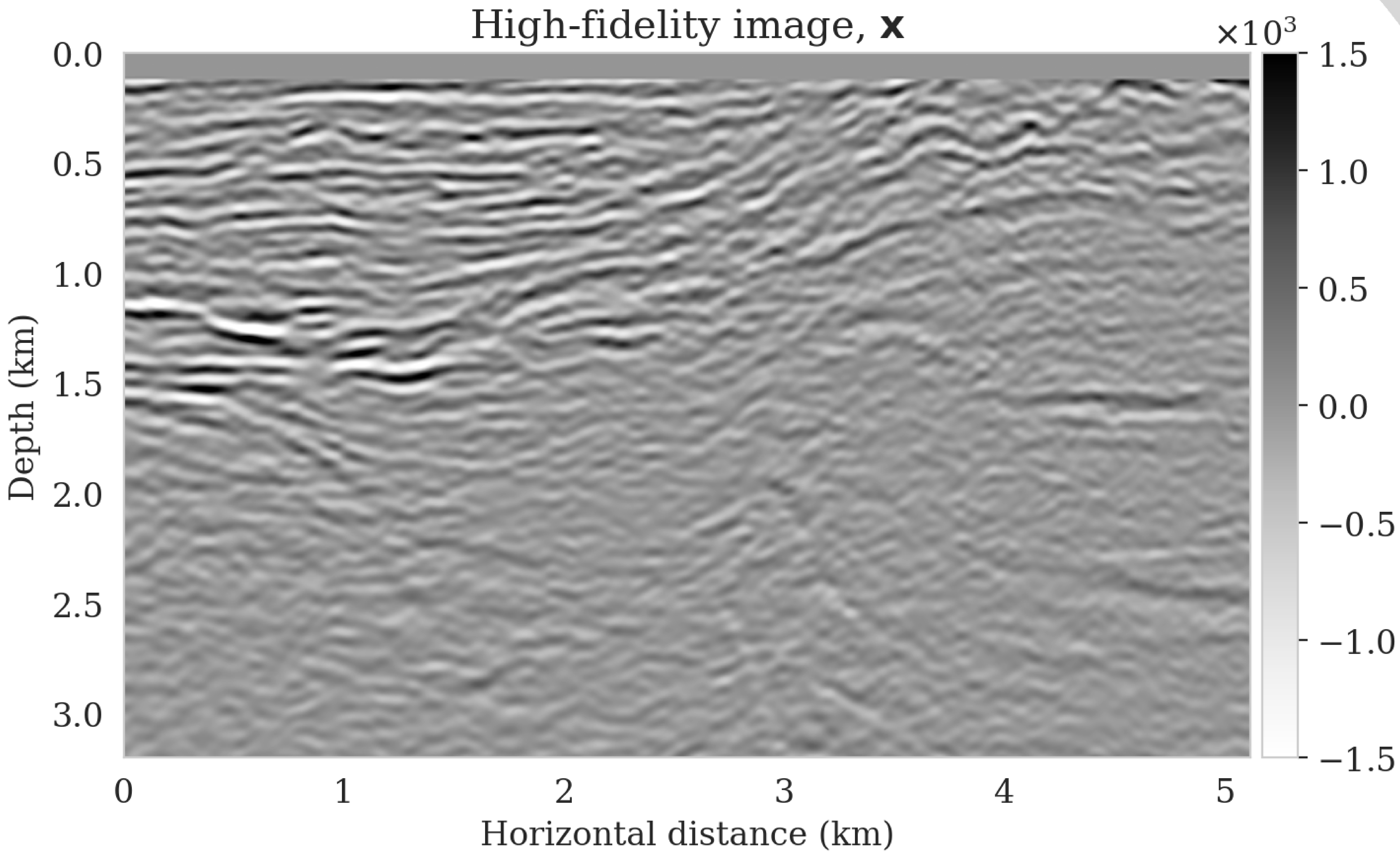
since T is invertible, in principle it can represent any unknown $\mathbf{x} \in \mathcal{X}$

can be used if $\mathbf{y}$ is out-of-distribution

final estimate is tied to the physics/data

Muhammad Asim, Max Daniels, Oscar Leong, Ali Ahmed, and Paul Hand. "Invertible generative models for inverse problems: mitigating representation error and dataset bias". In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 399–409. URL: http://proceedings.mlr.press/v119/asim20a.html.

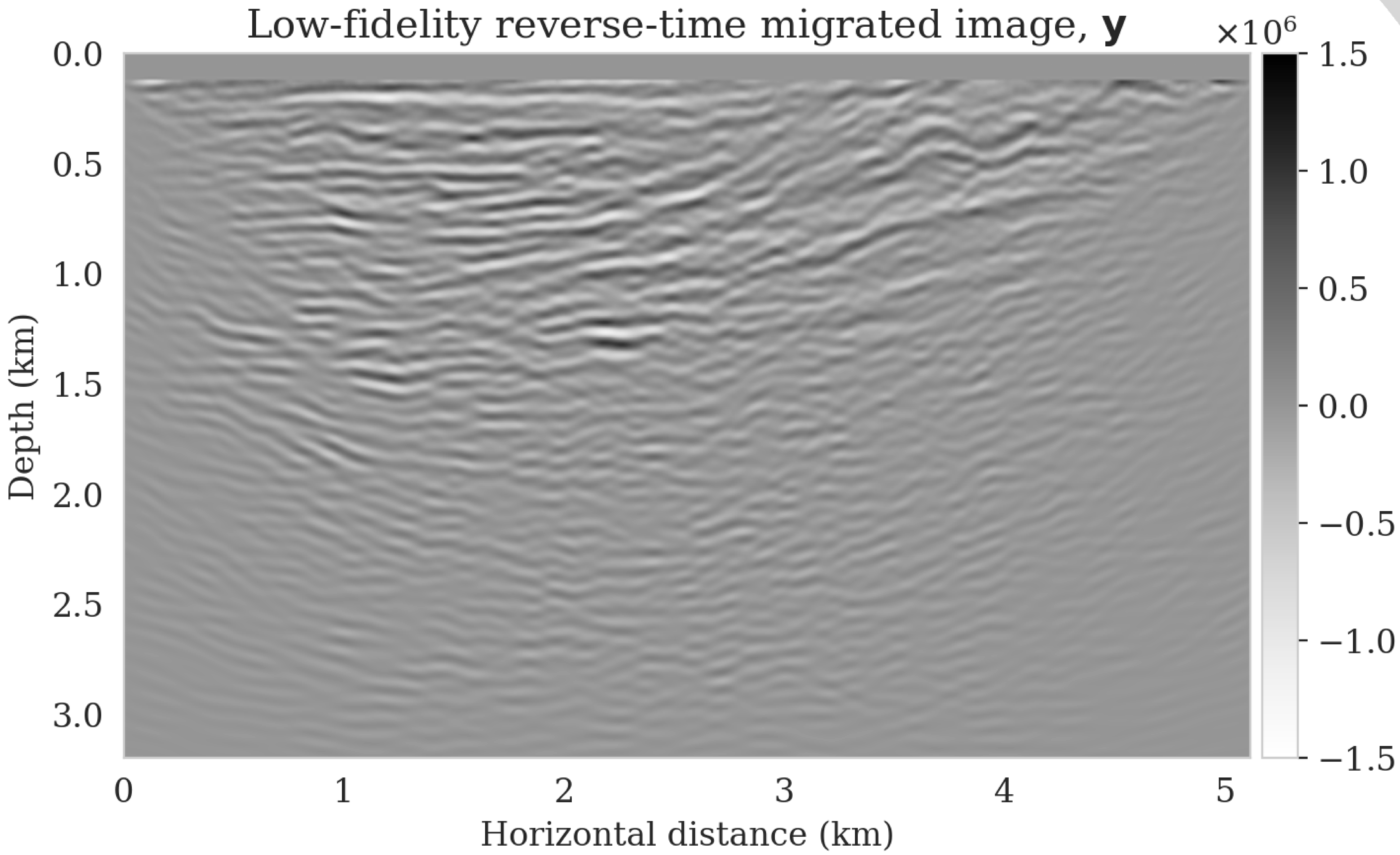
Seismic imaging example



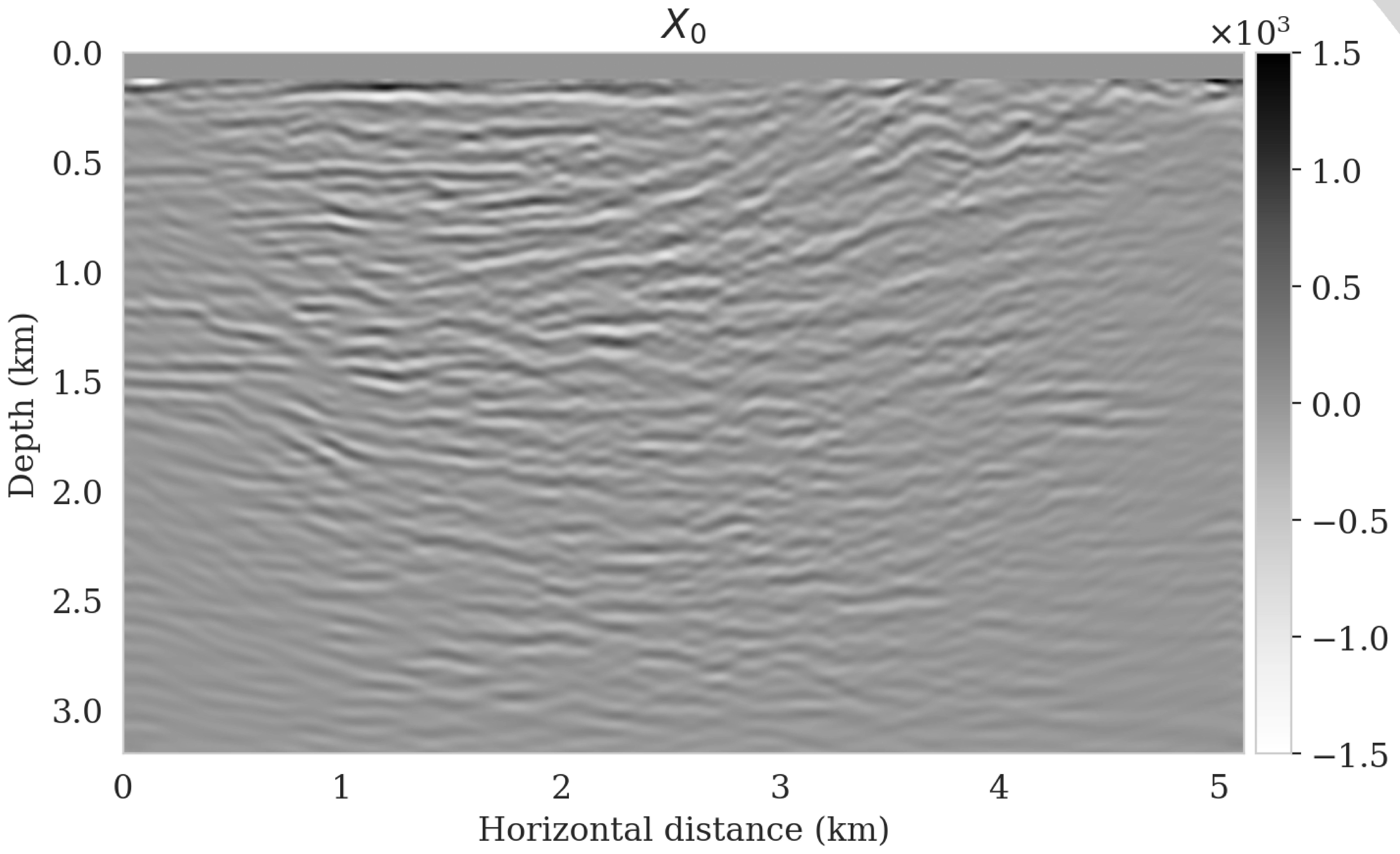
High-fidelity image, **x**
×10³
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

SLIM

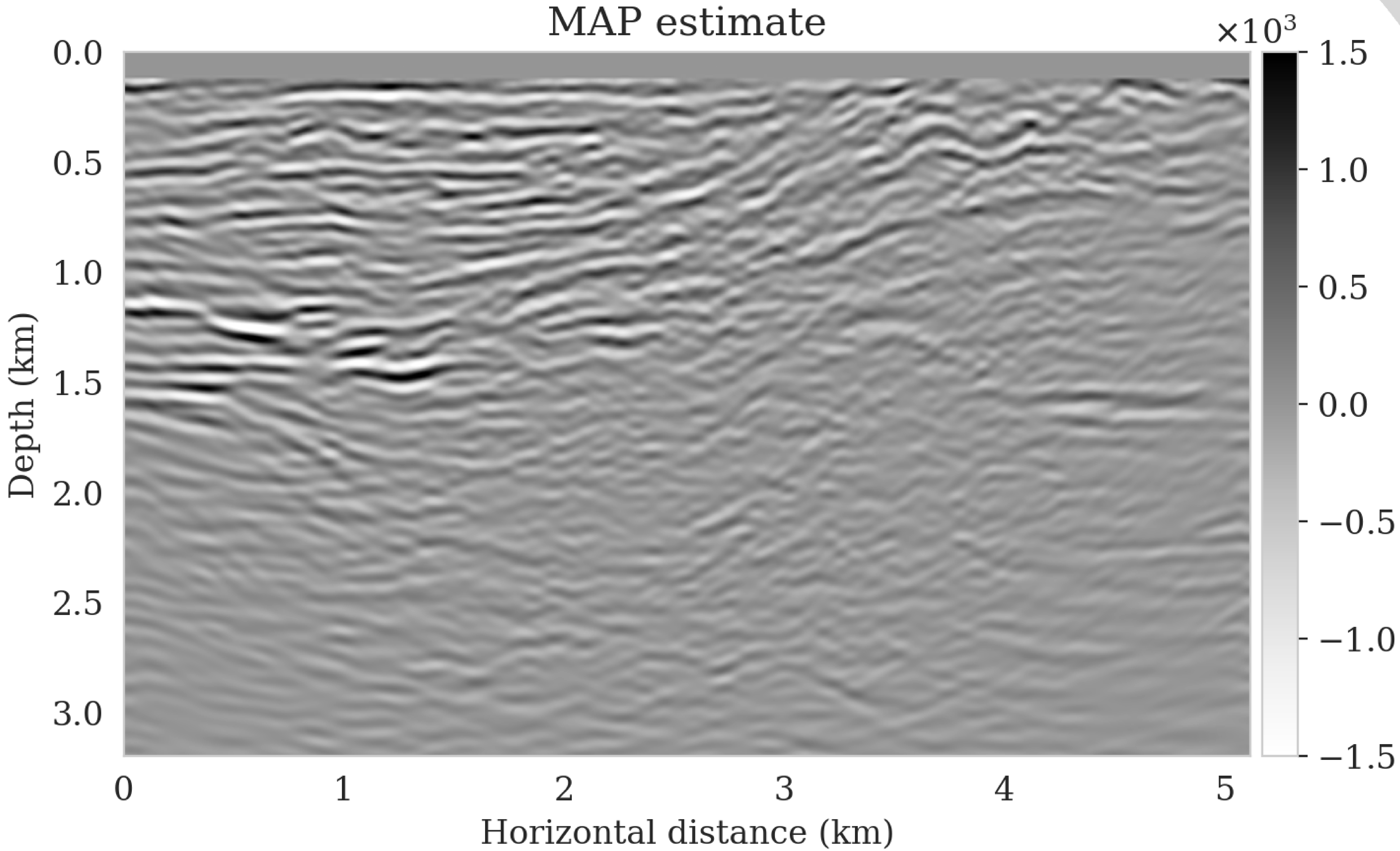
Out of distribution



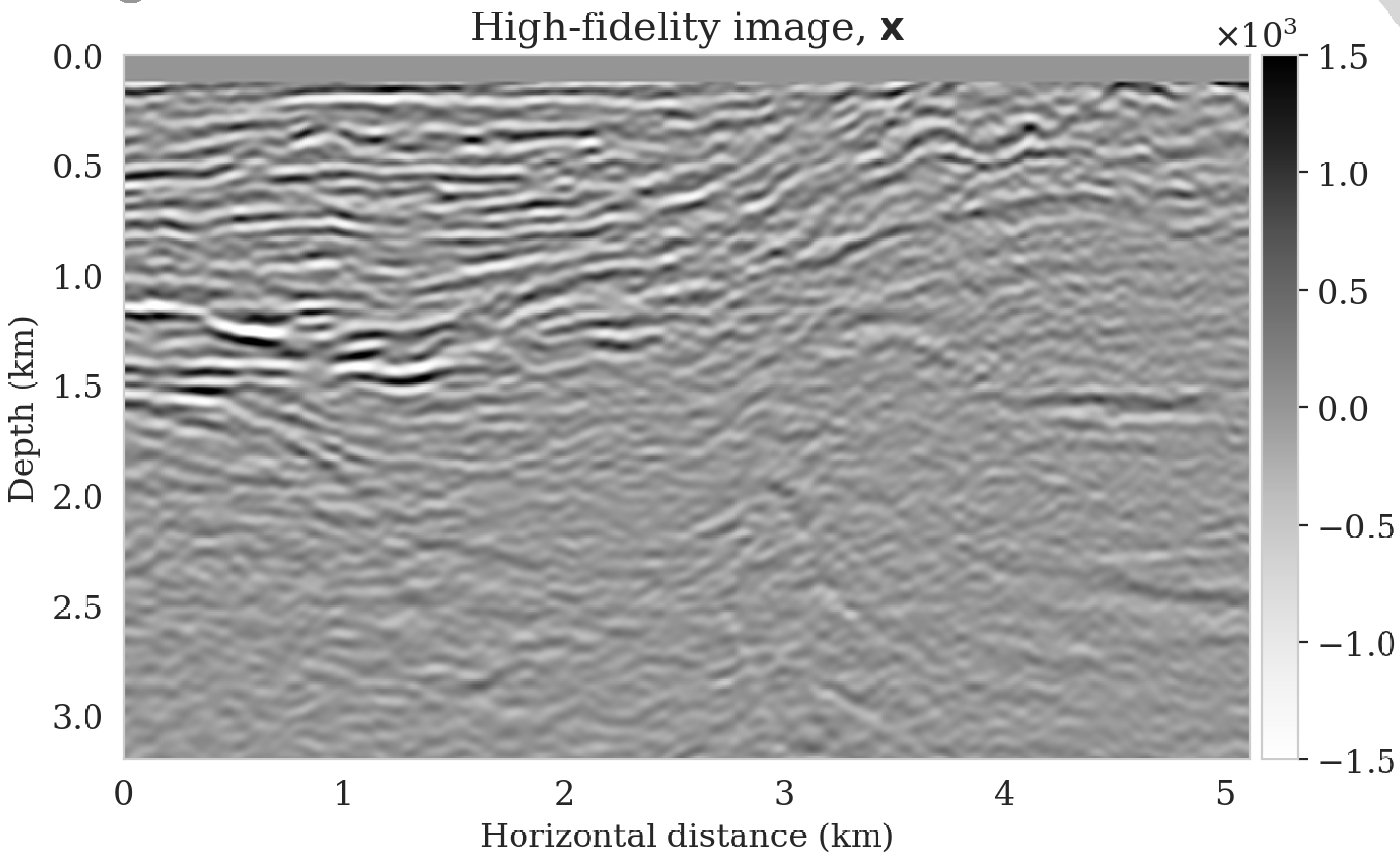
Maximum *a posteriori* estimate $\lambda = 0$



Maximum *a posteriori* estimate $\lambda = 1$

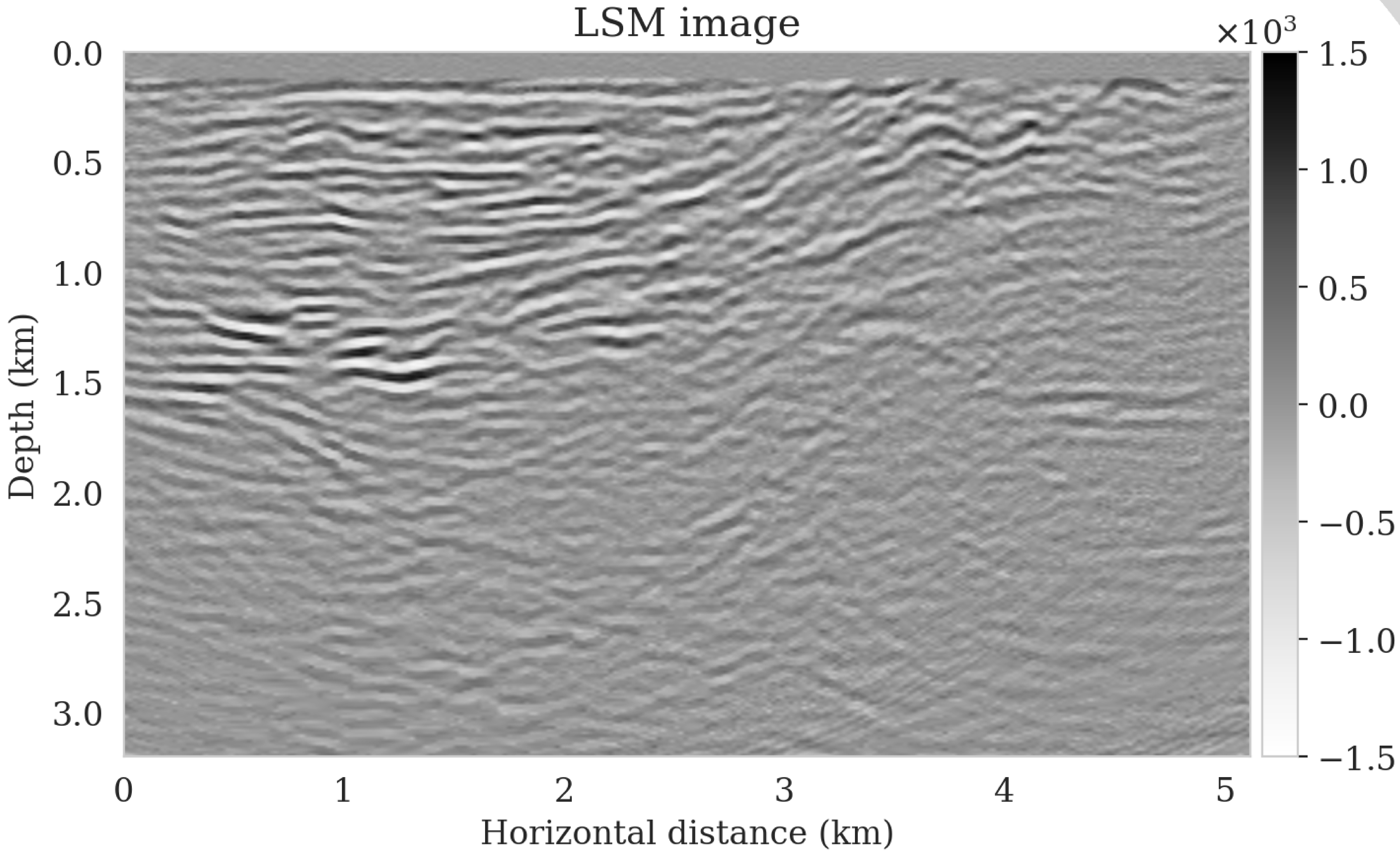


True image



High-fidelity image, **x**
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

Maximum Likelihood Estimate



Signal-to-noise ratio comparison

	SNR (dB)
RTM	-54.83
MLE	7.40
Preconditioned MAP	11.29
Data-driven CM	4.34

SNR comparison

Preconditioned physics-based approach

$$\min_{\theta_x} \mathbb{E}_{\mathbf{z}\sim \mathrm{N}(\mathbf{0},\mathbf{I})} \left[\frac{1}{2\sigma^2} \left\| F\left(T_{\theta_x}(\mathbf{z};\mathbf{y}) \right) - \mathbf{y} \right\|_2^2 - \log p_{\text{prior}}\left(T_{\theta_x}(\mathbf{z};\mathbf{y}) \right) - \log \left| \det \nabla_z T_{\theta_x}(\mathbf{z};\mathbf{y}) \right| \right]$$

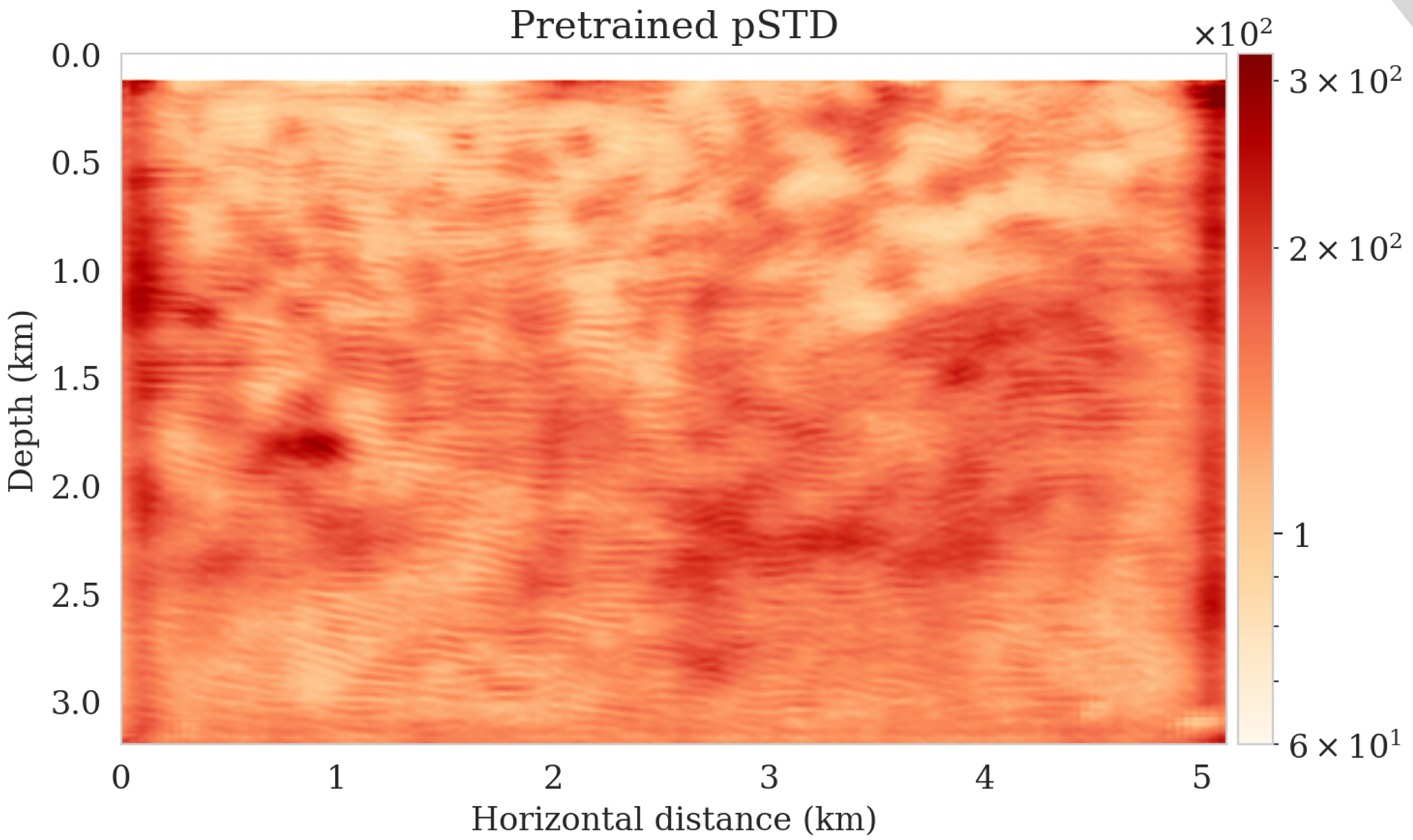
transfer learning $T_{\theta_x}(\mathbf{z};\mathbf{y}) := G_{\theta_x}^{-1}(G_{\theta_y}(\mathbf{y}), \mathbf{z})$

- corrects for errors due to out-of-distribution data
- less evaluations of F and $\nabla F^\top$

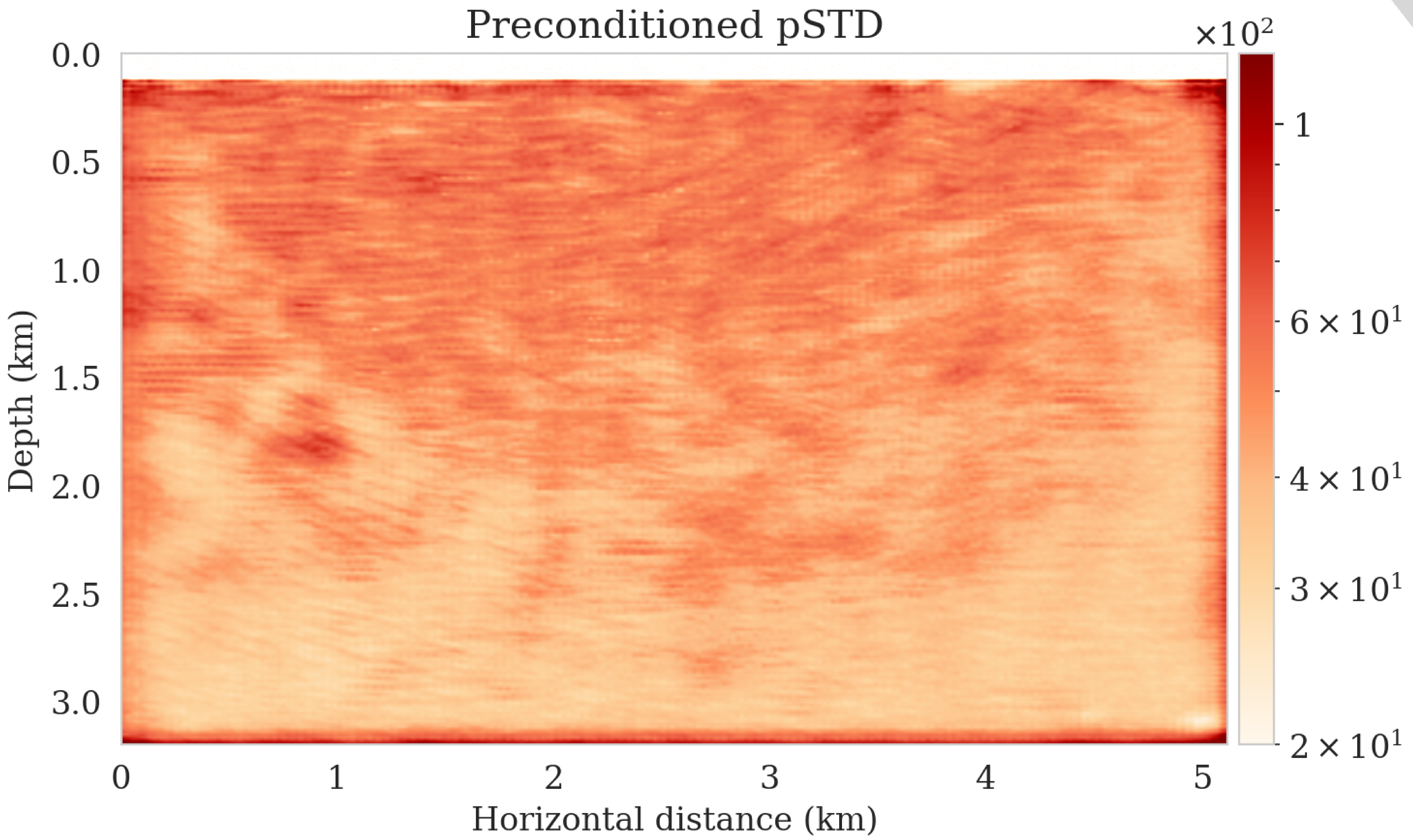
learned (conditional) prior density, $p_{\text{prior}}(\mathbf{x}) := p_{\mathbf{z}}\left(G_{\theta_x}(\mathbf{y},\mathbf{x})\right) \left| \det \nabla_x G_{\theta_x}(\mathbf{y},\mathbf{x}) \right|$

fast to adapt to new observation

Ali Siahkoohi, Gabrio Rizzuti, Mathias Louboutin, Philipp Witte, and Felix J. Herrmann. "Preconditioned training of normalizing flows for variational inference in inverse problems". In: *3rd Symposium on Advances in Approximate Bayesian Inference*. Jan. 2021. URL: https://openreview.net/pdf?id=P9m1sMaNQ8T.



Pretrained pSTD
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
$\times 10^2$
3×10^2
2×10^2
1
6×10^1



Preconditioned pSTD
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
×10^2
1
6 × 10^1
4 × 10^1
3 × 10^1
2 × 10^1

Number of PDE solves

	Number of PDE solves	
	F	$\nabla F^{\top}$
RTM	102	102
MLE	204	204
Preconditioned MAP	510	510
Preconditioned VI approach	$2,000$	$2,000$
MCMC w/ deep prior	$10,000$	$10,000$

Comparison of number of PDE solves in the seismic imaging example

Related work

Purely physics-based approach NF-based Bayesian inference

show orders of magnitude speed up compared to traditional MCMC methods

need to train the NF from scratch for a new observation $\mathbf{y}$

use handcrated priors, often negatively bias the inversion

Gabrio Rizzuti, Ali Siahkoohi, Philipp A. Witte, and Felix J. Herrmann. "Parameterizing uncertainty by deep invertible networks, an application to reservoir characterization". In: *SEG Technical Program Expanded Abstracts*. Sept. 2020, pp. 1541–1545. DOI: 10.1190/segam2020-3428150.1.

Xin Zhang and Andrew Curtis. "Seismic tomography using variational inference methods". In: *Journal of Geophysical Research: Solid Earth* 125.4 (2020), e2019JB018589.

Xuebin Zhao, Andrew Curtis, and Xin Zhang. "Bayesian Seismic Tomography using Normalizing Flows". In: (2020). DOI: 10.31223/X53K6G. URL: https://eartharxiv.org/repository/view/1940/.

Related work

Data-driven approaches for directly sampling from the posterior distribution

fast Bayesian inference given new observation $\mathbf{y}$

sample the posterior virtually for free

not directly tied to the physics/data

not reliable when applied to out-of-distribution data

not trivial to scale GAN-based approaches to large-scale problems

---

Jakob Kruse, Gianluca Detommaso, Robert Scheichl, and Ullrich Köthe. "HINT: Hierarchical Invertible Neural Transport for Density Estimation and Bayesian Inference". In: *Proceedings of AAAI-2021* (2021). URL: https://arxiv.org/pdf/1905.10687.pdf.

Jonas Adler and Ozan Öktem. "Deep Bayesian Inversion". In: *arXiv preprint arXiv:1811.05910* (2018).

Nikola Kovachki, Ricardo Baptista, Bamdad Hosseini, and Youssef Marzouk. *Conditional Sampling With Monotone GANs*. 2021. arXiv: 2006.06755 [stat.ML].

Jan-Hinrich Nölke et al. "Invertible neural networks for uncertainty quantification in photoacoustic imaging". In: *arXiv preprint arXiv:2011.05110* (2020).

Related work

Injective network for inverse problems and uncertainty quantification

use an injective map to map data to a low-dimensional space

can be used as a prior (projection operator) in inverse problems

heavily relies on availability of training data to generalize

unclear how the learned projection operator behaves when applied to out-of-distribution data

for new observation $\mathbf{y}$, Bayesian inference requires training an additional NF, involving the forward operator

Konik Kothari, AmirEhsan Khorashadizadeh, Maarten de Hoop, and Ivan Dokmanić. *Trumpets: Injective Flows for Inference and Inverse Problems*. 2021. arXiv: 2102.10461 [cs.LG].

Contributions

Take full advantage of existing training data to provide

low-fidelity but fast conditional mean estimate

a first assessment of the image's reliability

preconditioned physics-based high-fidelity MAP estimate via the learned prior

preconditioned, scalable, and physics-based Bayesian inference

Conclusions

Obtaining UQ information is rendered impractical when

- the forward operators are expensive to evaluate
- the problem is high dimensional

There are strong indications that

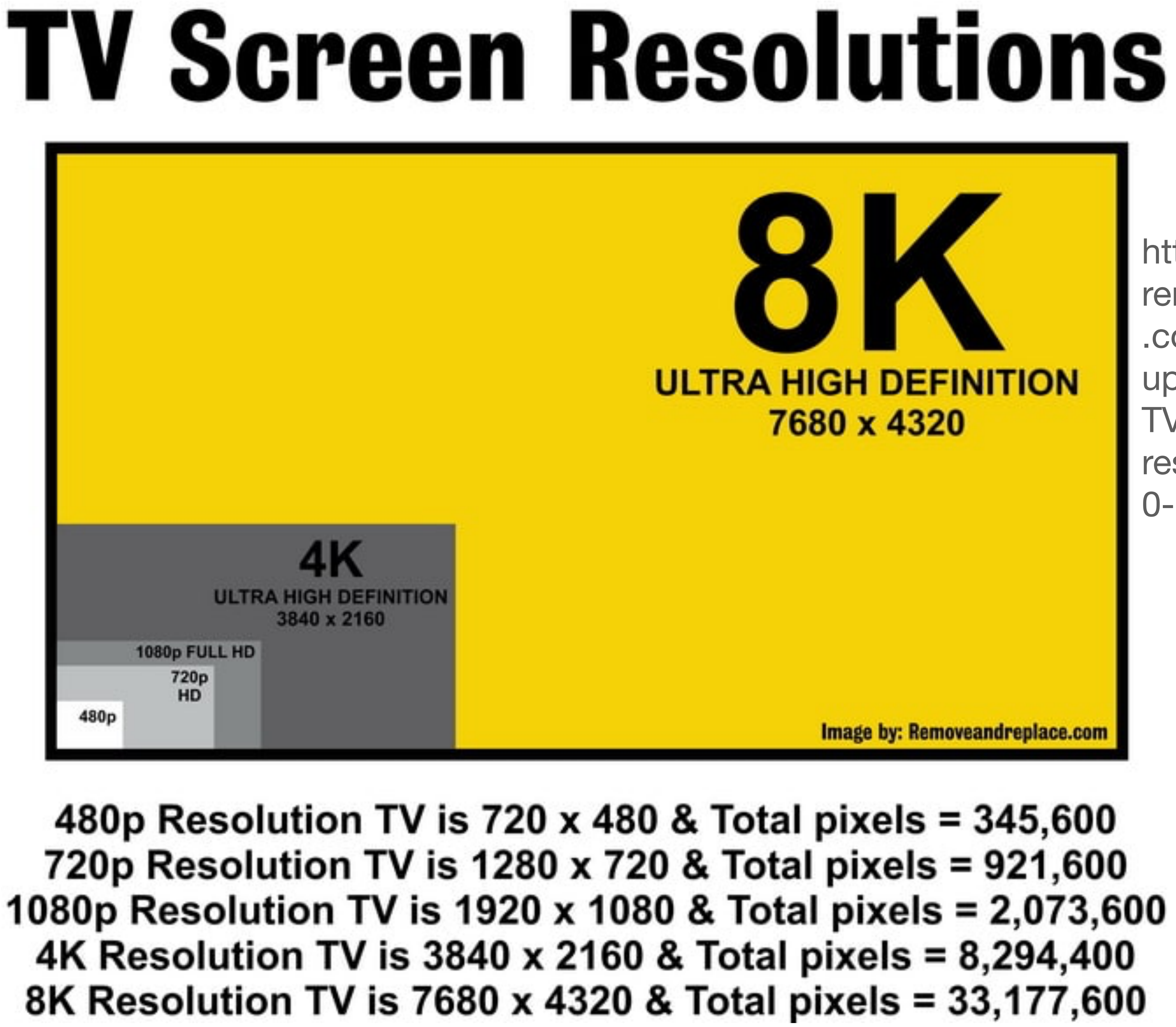
- Bayesian inference with normalizing flows can lead to orders of magnitude computational improvements compared to MCMC methods
- preconditioning with a pretrained conditional normalizing flow can lead to another order of magnitude speed up

Future work

- characterize uncertainties due to modeling errors

Scaling – on CPU

https://www.pcmag.com/news/what-is-8k-should-you-buy-a-new-tv-or-wait



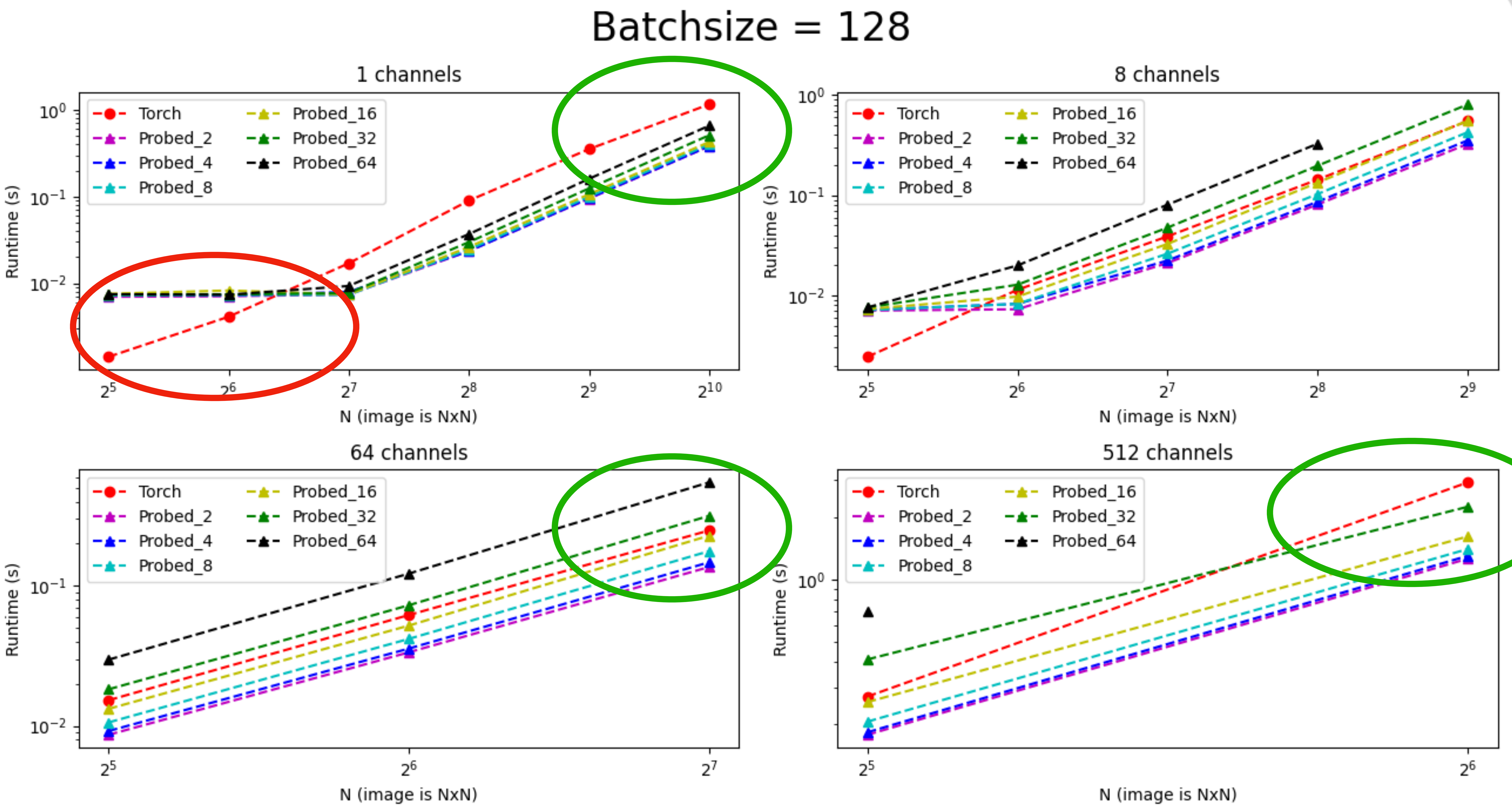
https://removeandreplace.com/wp-content/uploads/2016/04/TV-screen-resolutions-480-720-1080-4K-8K.jpg

- Validated scaling to size $N \times N \times 4 \times 1$ for $N = 2^{16}$ on Azure's HBv2 instance (AMD EPYC)
- 2^{32} pixels equals $64 \times$ 8K images (4×1024^3)
- Runtime 54 seconds 120 threads
- 480 Gb memory not enough for PyTorch...

https://github.com/slimgroup/XConv/blob/master/benchmark/perf_pyxconv.py

GPU

- sizes
  $N \times N \times (1-512) \times 128$
  for $N = 2^5 - 2^{10}$
- **linear scaling in PyTorch**
- **up to $3\times$ speedup on GPUs**
- not good for small sizes & small # of channels



Code

https://github.com/slimgroup/InvertibleNetworks.jl

https://github.com/slimgroup/FastApproximateInference.jl

https://github.com/slimgroup/Software.SEG2021

Acknowledgment

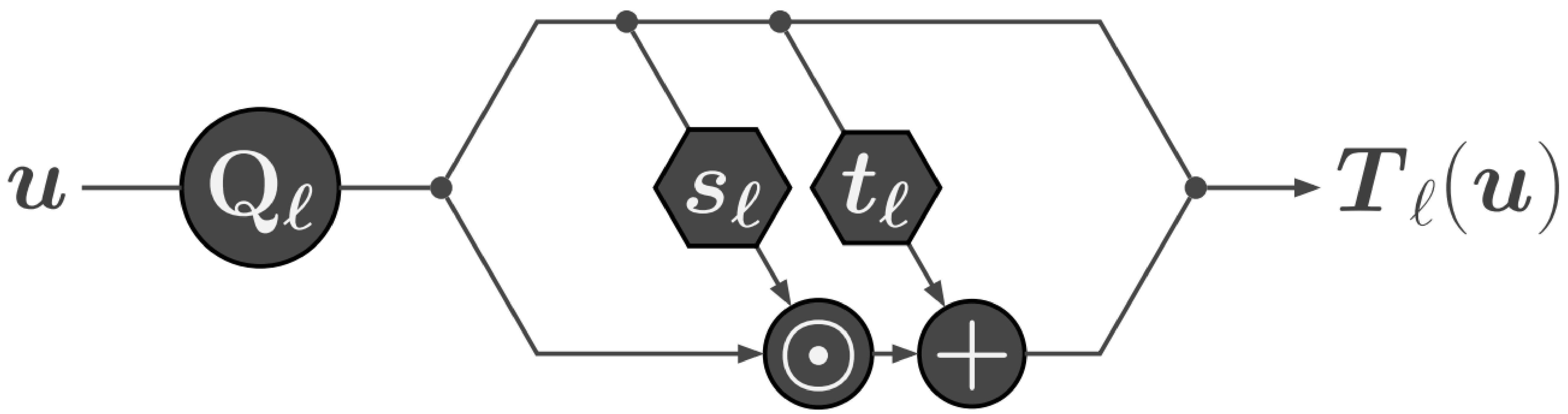
This research was carried out with the support of

Georgia Research Alliance and partners of the ML4Seismic Center

Affine coupling layer

a basic "layer" with an analytic inverse

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using Real NVP". In: *International Conference on Learning Representations, ICLR*. 2016. URL: http://arxiv.org/abs/1605.08803.



- an invertible layer with input $\boldsymbol{u}$
- orthogonal Householder reflections $\boldsymbol{Q}_\ell$
- arbitrary neural networks $\boldsymbol{s}_\ell$, $\boldsymbol{t}_\ell$
- output of ℓ^{th} layer $\boldsymbol{T}_\ell(\boldsymbol{u})$

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using Real NVP". In: *International Conference on Learning Representations, ICLR*. 2016. URL: `http://arxiv.org/abs/1605.08803`.

Jakob Kruse, Gianluca Detommaso, Robert Scheichl, and Ullrich Köthe. "HINT: Hierarchical Invertible Neural Transport for Density Estimation and Bayesian Inference". In: *Proceedings of AAAI-2021* (2021). URL: `https://arxiv.org/pdf/1905.10687.pdf`.

Affine coupling layer—forward

$$\boldsymbol{T}_\ell(\boldsymbol{u}) = \begin{bmatrix} \boldsymbol{T}_\ell^1\left(\tilde{\boldsymbol{u}}_1\right) \\ \boldsymbol{T}_\ell^2\left(\tilde{\boldsymbol{u}}_1, \tilde{\boldsymbol{u}}_2\right) \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{u}}_1 \\ \tilde{\boldsymbol{u}}_2 \odot \sigma\left(\boldsymbol{s}_\ell\left(\tilde{\boldsymbol{u}}_1\right)\right) + \boldsymbol{t}_\ell\left(\tilde{\boldsymbol{u}}_1\right) \end{bmatrix},$$

$$\text{with } \tilde{\boldsymbol{u}} = \begin{bmatrix} \tilde{\boldsymbol{u}}_1 \\ \tilde{\boldsymbol{u}}_2 \end{bmatrix} = \boldsymbol{Q}_\ell \boldsymbol{u}$$

sigmoid function $\sigma\left(\cdot\right)$

elementwise multiplication $\odot$

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using Real NVP". In: *International Conference on Learning Representations, ICLR*. 2016. URL: `http://arxiv.org/abs/1605.08803`.

Affine coupling layer—inverse

$$
\tilde{\boldsymbol{u}} = \begin{bmatrix} \tilde{\boldsymbol{u}}_1 \\ \tilde{\boldsymbol{u}}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{T}_\ell^1\left(\tilde{\boldsymbol{u}}_1\right) \\ \left(\boldsymbol{T}_\ell^2\left(\tilde{\boldsymbol{u}}_1, \tilde{\boldsymbol{u}}_2\right) - \boldsymbol{t}_\ell\left(\boldsymbol{T}_\ell^1\left(\tilde{\boldsymbol{u}}_1\right)\right)\right) \oslash \sigma\left(\boldsymbol{s}_\ell\left(\boldsymbol{T}_\ell^1\left(\tilde{\boldsymbol{u}}_1\right)\right)\right) \end{bmatrix},
$$

$$
\boldsymbol{u} = \boldsymbol{Q}_\ell^{-1} \begin{bmatrix} \tilde{\boldsymbol{u}}_1 \\ \tilde{\boldsymbol{u}}_2 \end{bmatrix}
$$

orthogonal learnable matrix $\boldsymbol{Q}_\ell^{-1} = \boldsymbol{Q}_\ell^\top$

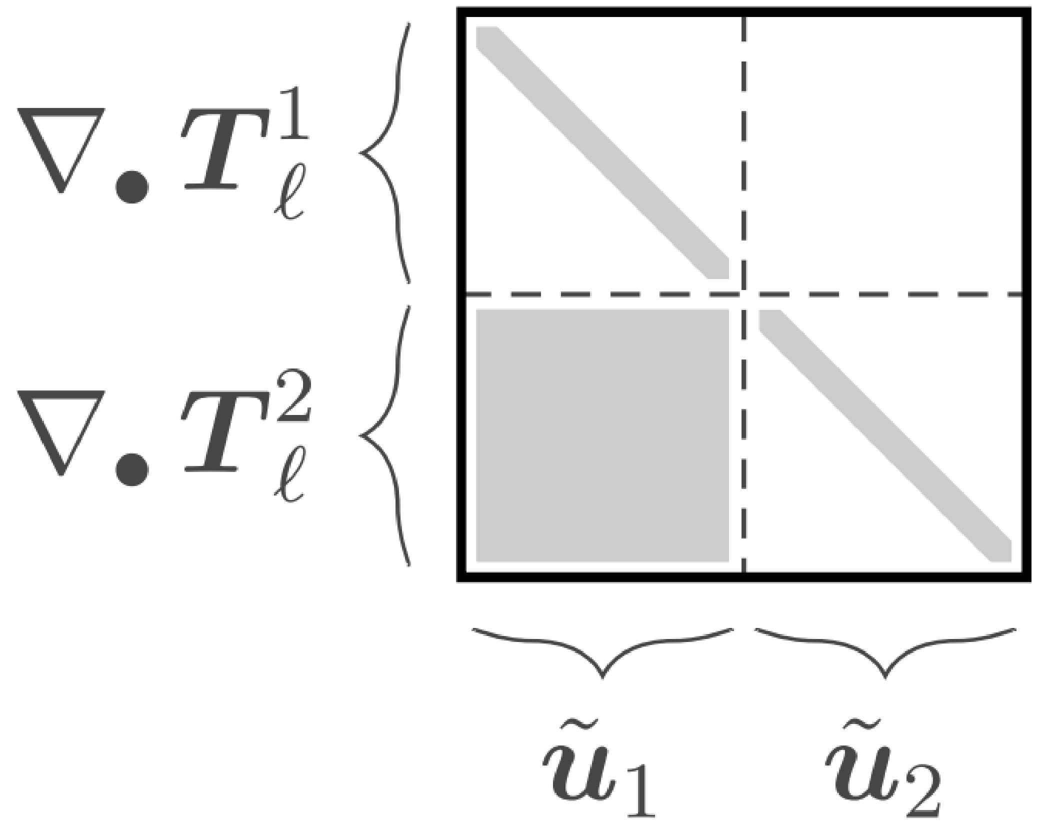
elementwise division $\oslash$

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. “Density estimation using Real NVP”. In: *International Conference on Learning Representations, ICLR*. 2016. URL: http://arxiv.org/abs/1605.08803.

Determinant of the Jacobian

$$\nabla_{\boldsymbol{u}} \boldsymbol{T}_\ell(\boldsymbol{u}) = \nabla_{\boldsymbol{u}} \begin{bmatrix} \boldsymbol{T}_\ell^1\left(\tilde{\boldsymbol{u}}_1\right) \\ \boldsymbol{T}_\ell^2\left(\tilde{\boldsymbol{u}}_1, \tilde{\boldsymbol{u}}_2\right) \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \nabla_{\boldsymbol{u}_1} \boldsymbol{T}_\ell^2\left(\tilde{\boldsymbol{u}}_1, \tilde{\boldsymbol{u}}_2\right) & \mathrm{diag}\Big(\sigma\left(\boldsymbol{s}_\ell\left(\tilde{\boldsymbol{u}}_1\right)\right)\Big) \end{bmatrix}$$

$$\Rightarrow \log \Big| \det \nabla_{\boldsymbol{u}} \boldsymbol{T}_\ell(\boldsymbol{u}) \Big| = \mathrm{sum}\left(\log \sigma\left(\boldsymbol{s}_\ell\left(\tilde{\boldsymbol{u}}_1\right)\right)\right)$$



sparsity pattern of Jacobian

Jakob Kruse, Gianluca Detommaso, Robert Scheichl, and Ullrich Köthe. "HINT: Hierarchical Invertible Neural Transport for Density Estimation and Bayesian Inference". In: *Proceedings of AAAI-2021* (2021). URL: https://arxiv.org/pdf/1905.10687.pdf.

Challenges of affine coupling layers

lower representation power due

- sparse triangular Jacobian

requires many layer compositions and orthogonal transforms

- to capture all pixel-wise dependencies

may become numerically non-invertible

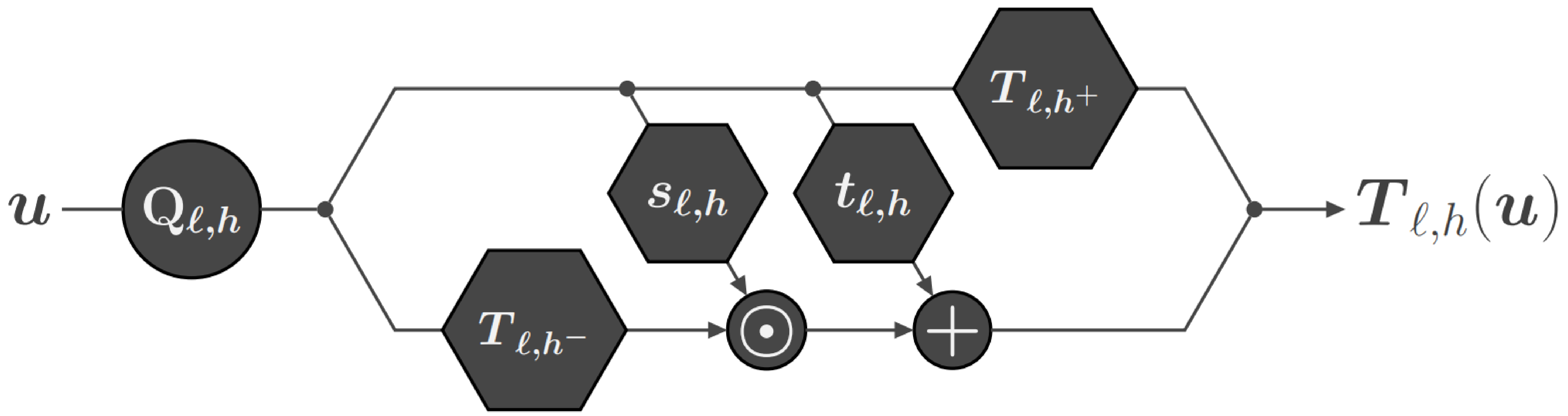
- e.g. due to very small Jacobian singular values

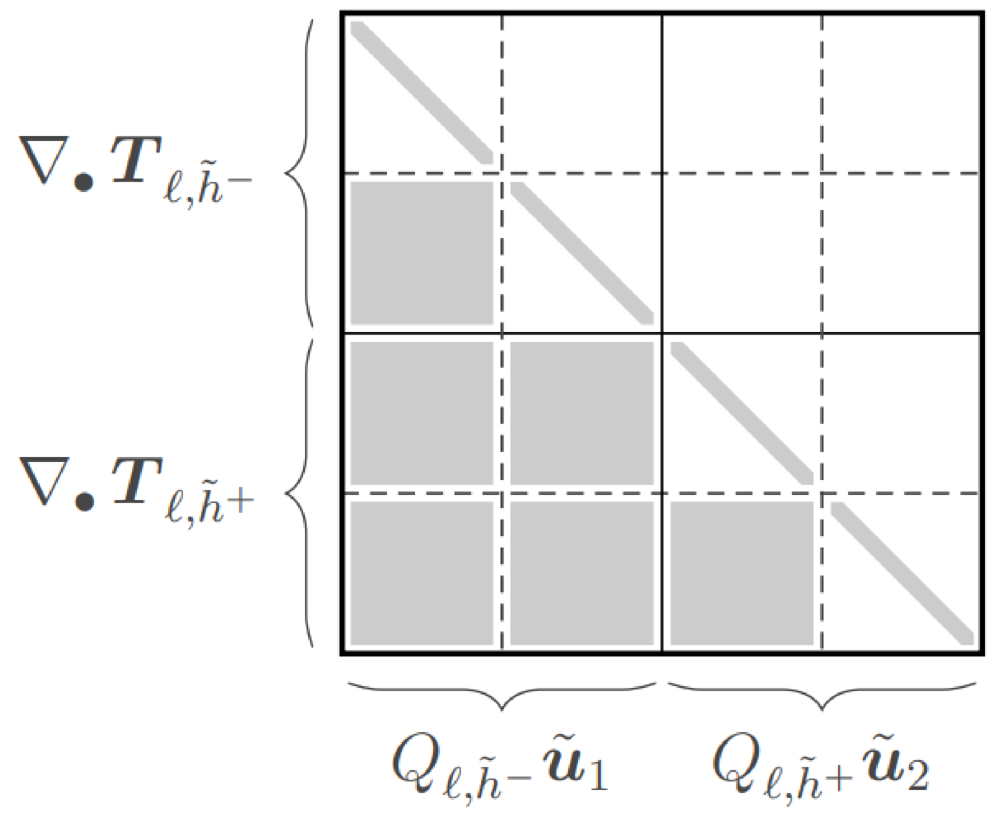
Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. "Why Normalizing Flows Fail to Detect Out-of-Distribution Data". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 20578–20589. URL: https://proceedings.neurips.cc/paper/2020/file/ecb9fe2fbb99c31f567e9823e884dbec-Paper.pdf.

Jens Behrmann, Paul Vicol, Kuan-Chieh Wang, Roger Grosse, and Jörn-Henrik Jacobsen. "Understanding and mitigating exploding inverses in invertible neural networks". In: *arXiv preprint arXiv:2006.09347* (2020).

Didrik Nielsen, Priyank Jaini, Emiel Hoogeboom, Ole Winther, and Max Welling. "SurVAE Flows: Surjections to Bridge the Gap between VAEs and Flows". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 12685–12696. URL: https://proceedings.neurips.cc/paper/2020/file/9578a63fbe545bd82cc5bbe749636af1-Paper.pdf.

Hierarchical coupling layers





dense triangular Jacobian

hierarchical architecture

- recursive affine coupling layers
- until input can not be split any further

Jakob Kruse, Gianluca Detommaso, Robert Scheichl, and Ullrich Köthe. "HINT: Hierarchical Invertible Neural Transport for Density Estimation and Bayesian Inference". In: *Proceedings of AAAI-2021* (2021). URL: https://arxiv.org/pdf/1905.10687.pdf.