Deep Bayesian inference for task-based Seismic Imaging

Ali Siahkoohi   Gabrio Rizzuti *   Mathias Louboutin   Felix J. Herrmann

Georgia Institute of Technology

*Now at Utrecht University

SLIM

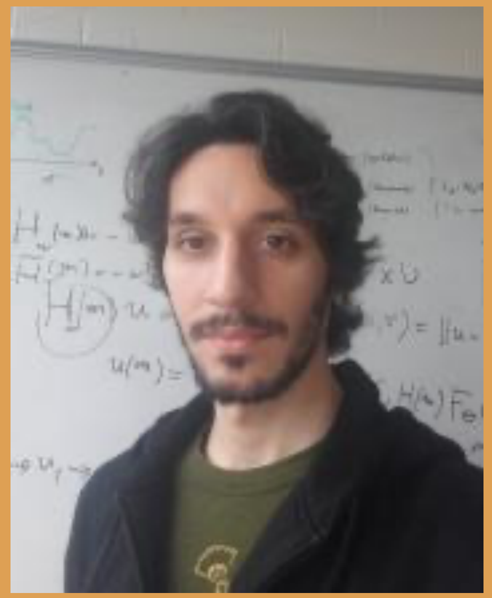
Georgia Institute of Technology

Deep Bayesian inference for task-based Seismic Imaging

Ali Siahkoohi Gabrio Rizzuti * Mathias Louboutin Felix J. Herrmann

Georgia Institute of Technology

*Now at Utrecht University


SLIM
Georgia Institute of Technology

Motivation

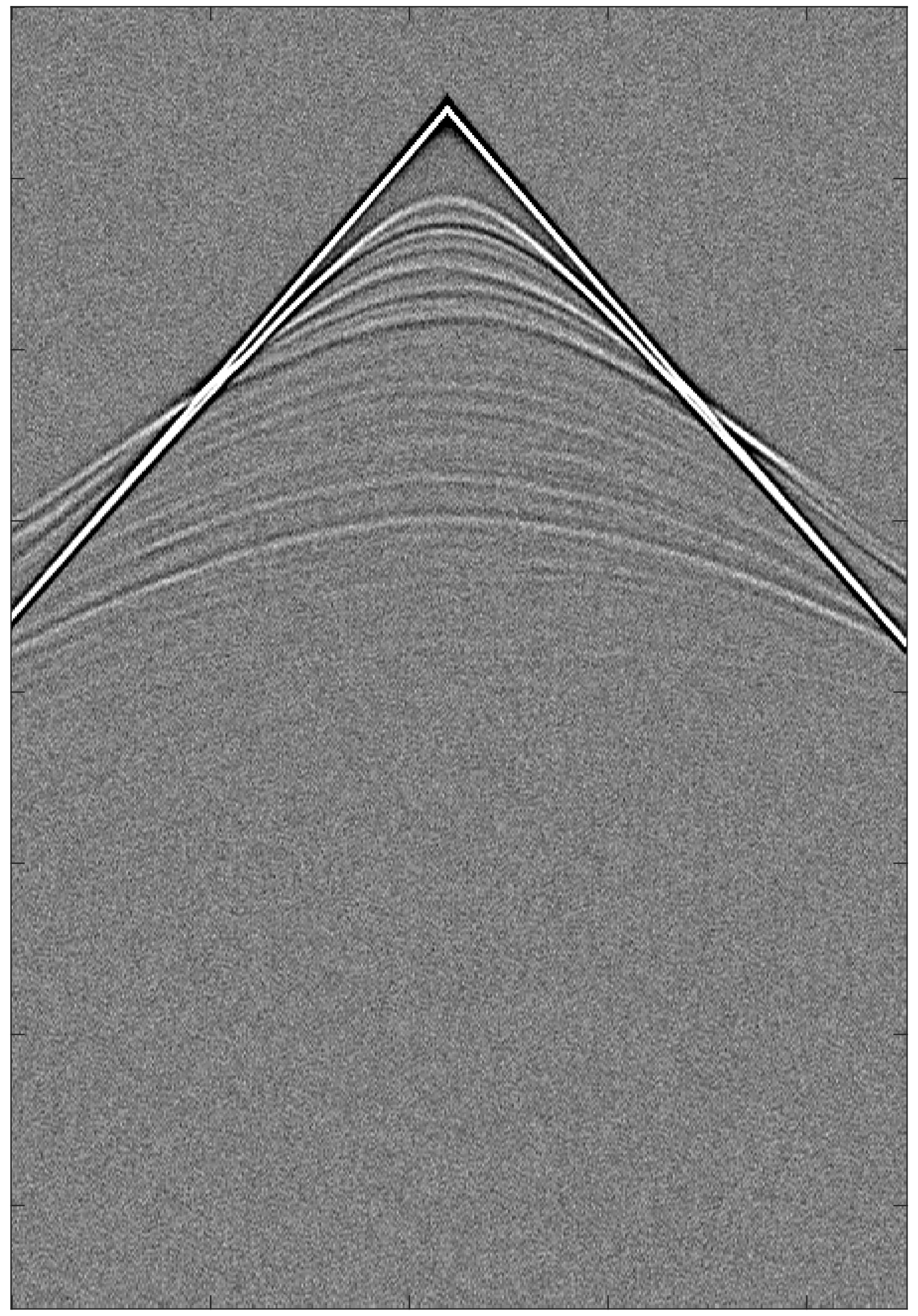
Increased demand for systematic Uncertainty Quantification (UQ) for Seismic

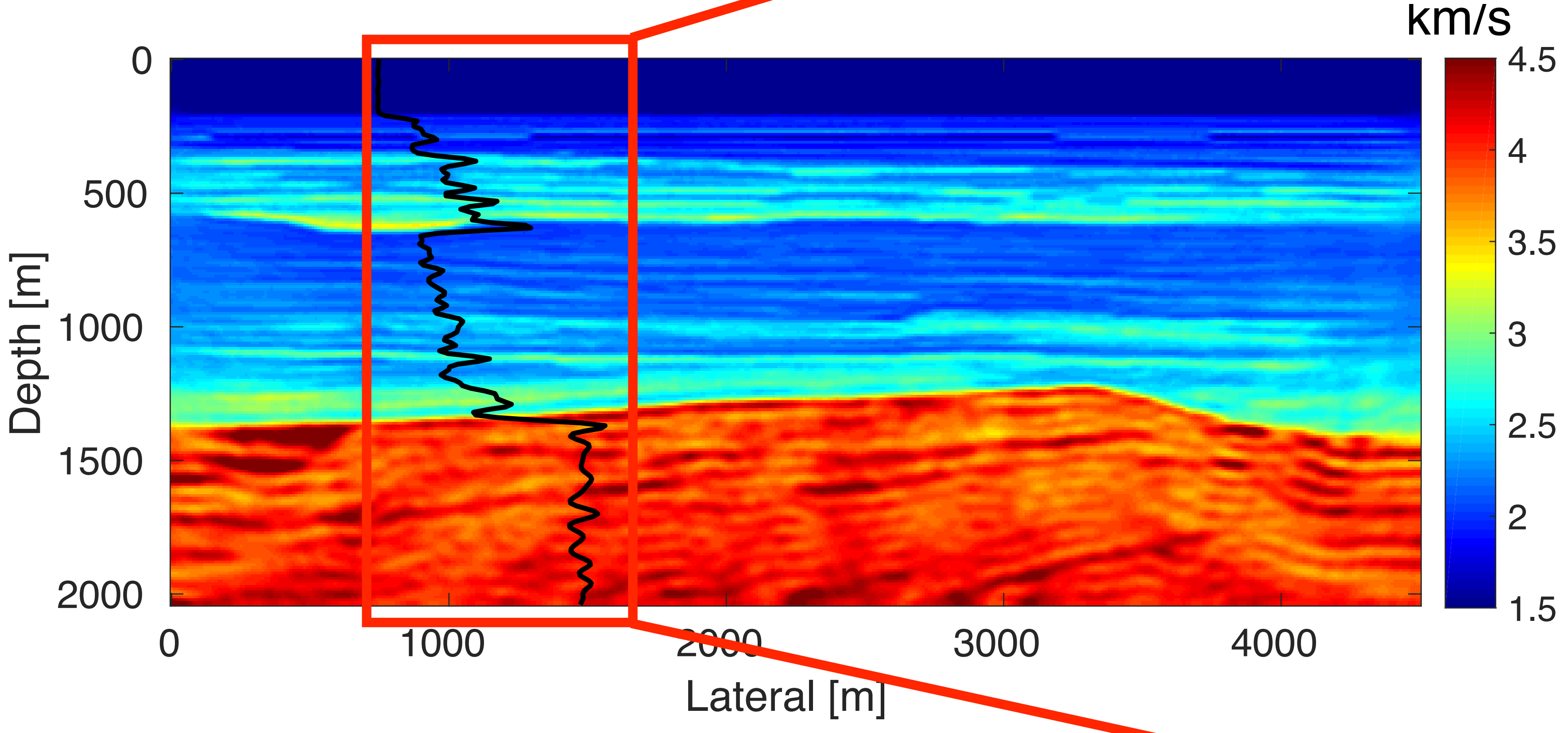
- understand how errors in the data propagate to seismic images
- how these errors affect subsequent tasks
- how to mitigate risks, e.g. of Carbon Capture Sequestration

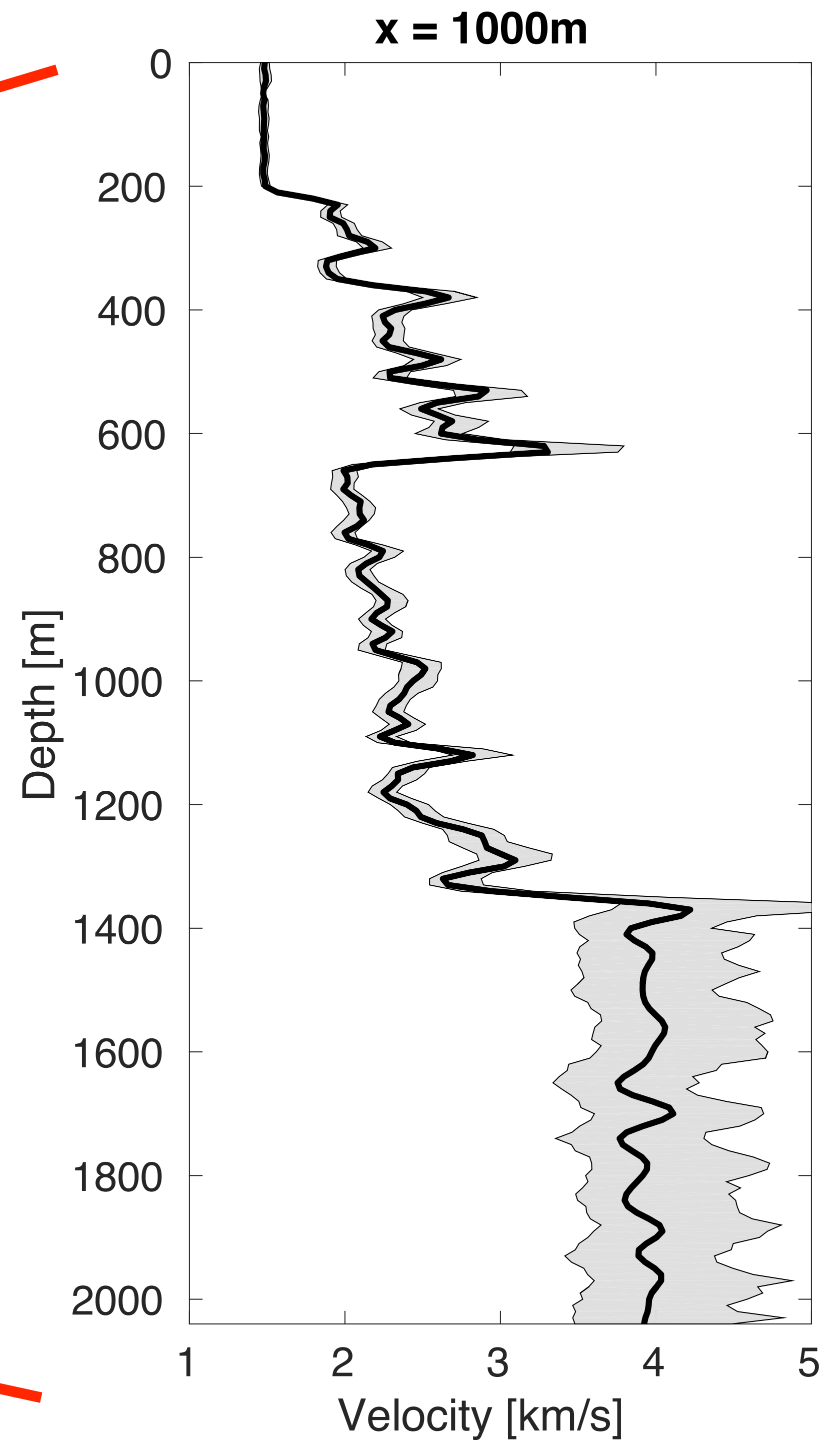
Many attempts have been dwarfed by dimensionality/complexity of seismic imaging.

- combine the task of seismic horizon tracking with UQ using deep priors
- introduce scalable formulation for Bayes inference with normalizing flows

Seismic exploration – uncertainty quantification







"Noise" in data / physics

Uncertainty in model

Confidence in model

Discretized acoustic wave equation

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

elementwise multiplication, $\odot$

Laplacian, $\boldsymbol{\nabla}^2$

squared slowness model, $\mathbf{m}$

Nonlinear forward model

$$\mathbf{d}_i = \mathbf{P}\mathbf{A}^{-1}[\mathbf{m}]\mathbf{q}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}), \;\; i = 1, \ldots, n_s$$

discretized wave equation, $\mathbf{A}[\mathbf{m}]$

restriction operator, $\mathbf{P}$

source signature, $\mathbf{q}_i$

noise variance, σ^2

Linearized Born modeling operator

$$\mathbf{J}(\mathbf{m}_0, \mathbf{q}_i) = \nabla_{\mathbf{m}} \mathbf{P}\mathbf{A}^{-1}[\mathbf{m}]\mathbf{q}_i \Big|_{\mathbf{m}=\mathbf{m}_0}$$

$$= -\mathbf{P}\,\underbrace{\mathbf{A}^{-1}[\mathbf{m}_0]\,\mathrm{diag}\left(\nabla\mathbf{A}[(\mathbf{m}_0]\,\overbrace{\mathbf{A}^{-1}[\mathbf{m}_0]\mathbf{q}_i}^{\text{PDE solve}}\right)}_{\text{PDE solve}}$$

Seismic imaging forward model

$$\delta\mathbf{d}_i = \mathbf{J}[\mathbf{m}_0, \mathbf{q}_i]\delta\mathbf{m} + \boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

linearized observed data, $\delta\mathbf{d}_i = \mathbf{d}_i - \mathbf{P}\mathbf{A}^{-1}[\mathbf{m}_0]\mathbf{q}_i$

reflectivity model, $\delta\mathbf{m} = \mathbf{m} - \mathbf{m}_0$

linearization error, $\boldsymbol{\eta}_i$

Challenges

high dimensionality

expensive forward operator

linearization error

inconsistent, mildly ill-conditioned

Example

Quasi real data derived from 2D image from the Parihaka field dataset

finite-difference simulations with `Devito`

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

M. Louboutin et al. "Devito (v3.1.0): an embedded domain-specific language for finite differences and geophysical exploration". In: *Geoscientific Model Development* 12.3 (2019), pp. 1165–1187. DOI: 10.5194/gmd-12-1165-2019.

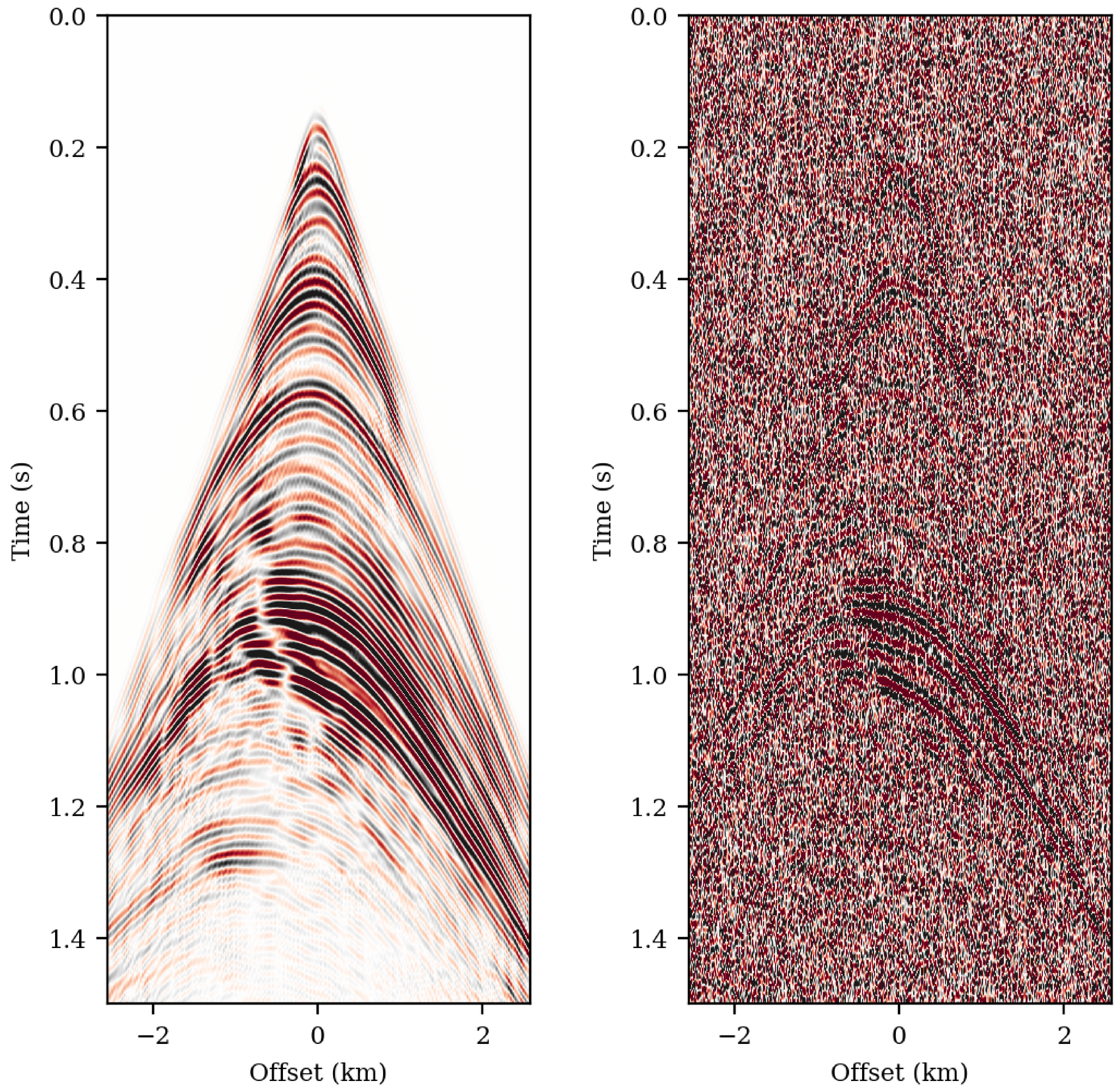
Acquisition details—Parihaka model

205 sources with $25\,\mathrm{m}$ sampling rate

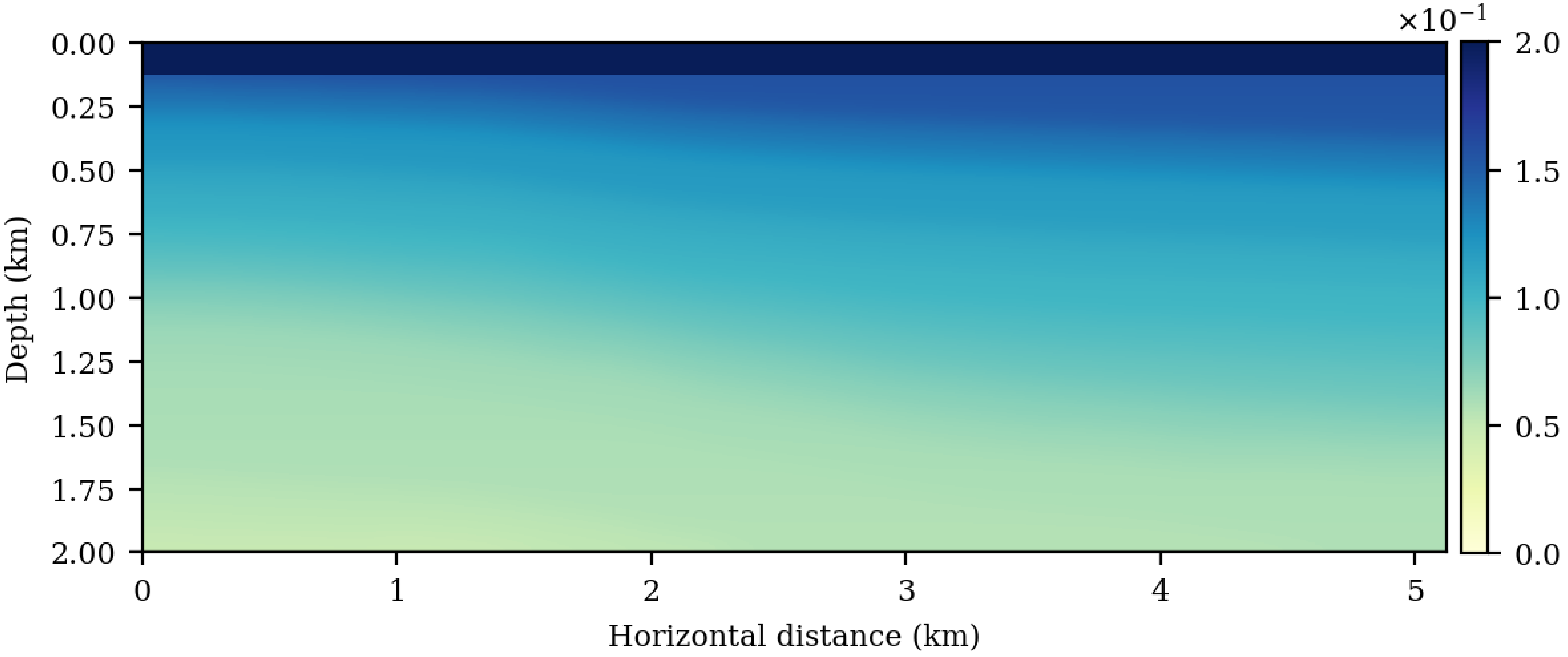
410 receivers with $12.5\,\mathrm{m}$ sampling rate

$1.5\,\mathrm{s}$ recording time

Ricker source wavelet with $30\,\mathrm{Hz}$ central frequency



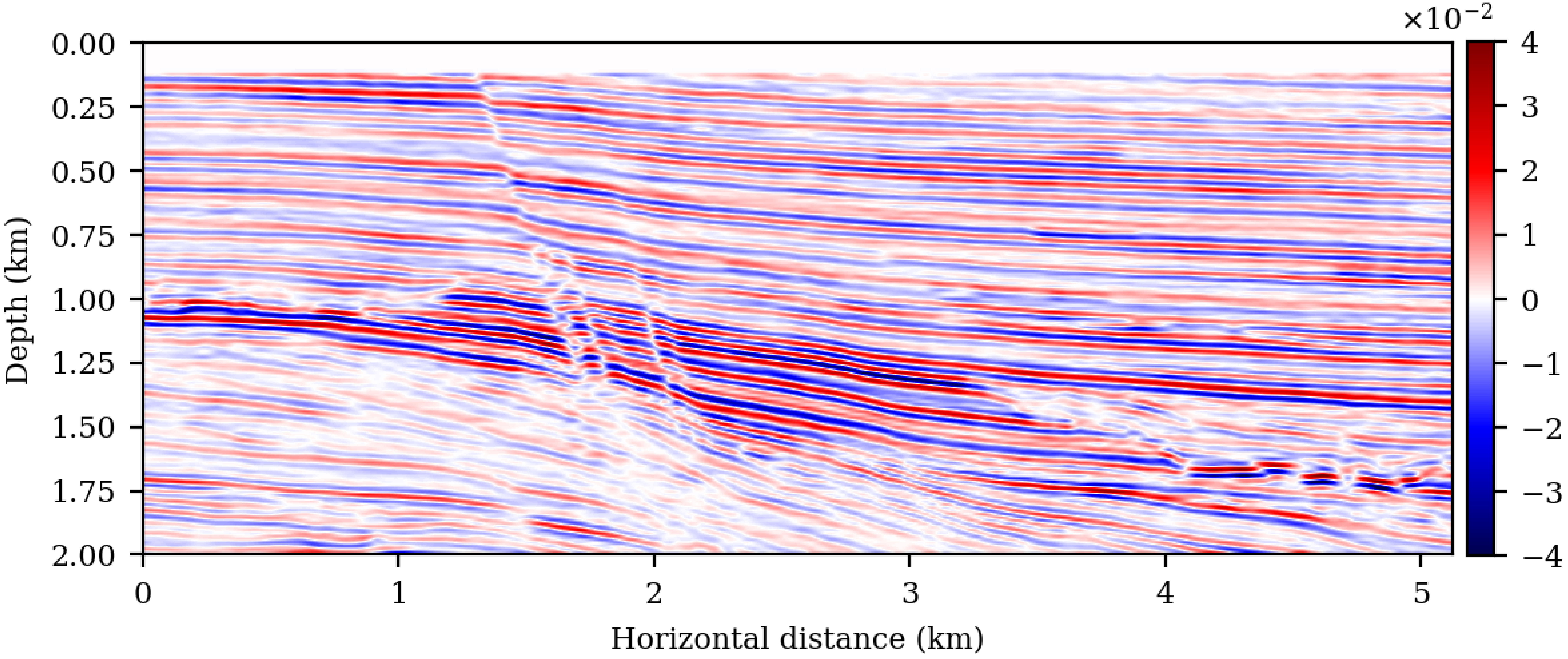
Noise-free (left) and noisy (right) linearized data—SNR: -8.75 dB



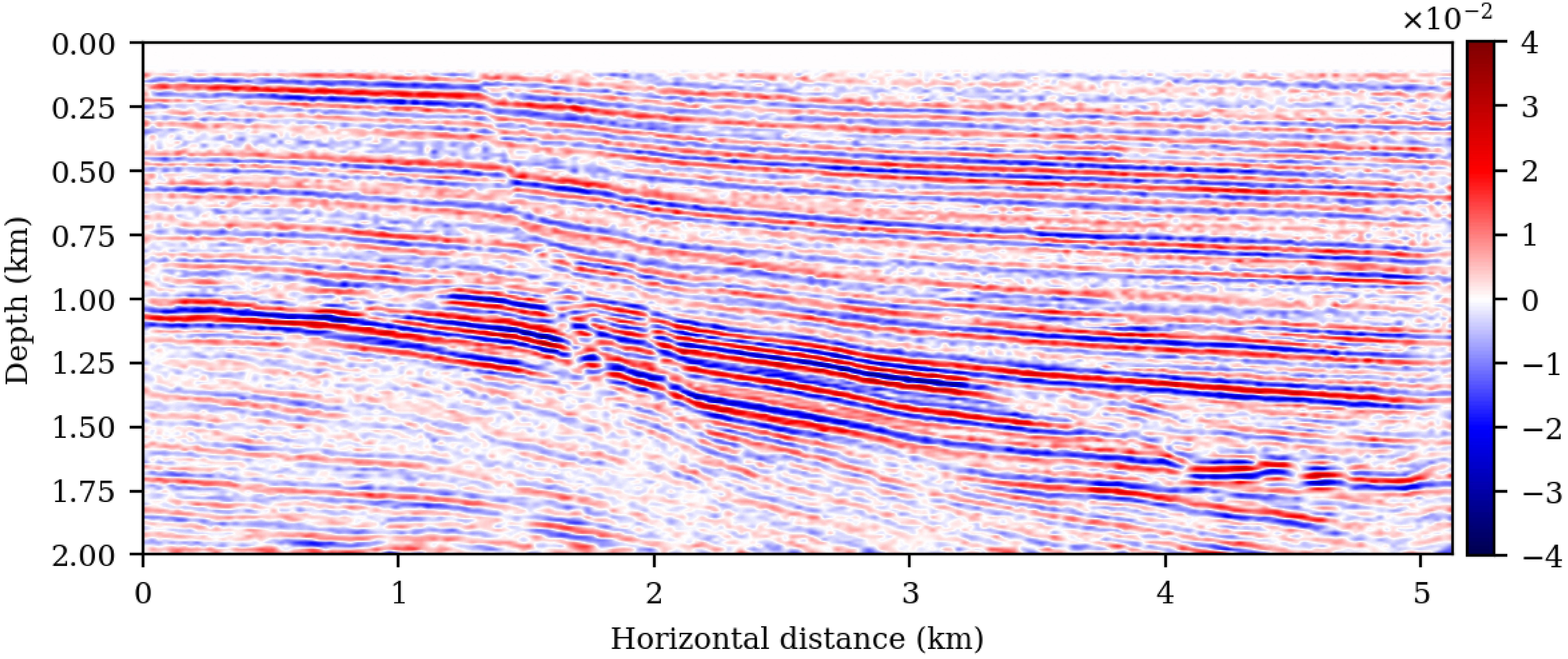
$\mathbf{m}_0$—made up background squared-slowness model $\left(\frac{\mathrm{s}^2}{\mathrm{km}^2}\right)$

Inversion without any regularization/prior

$$\min_{\delta \mathbf{m}} \sum_{i=1}^{n_s} \|\delta \mathbf{d}_i - \mathbf{J}[\mathbf{m}_0, \mathbf{q}_i]\delta \mathbf{m}\|_2^2$$



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



Estimation with no regularization/prior

Bayesian inverse problems

Represent the solution as a distribution over the model space

i.e., posterior distribution.

Bayes' rule

$$p_{\text{post}}\left(\delta\mathbf{m} \mid \delta\mathbf{d}\right) \propto p_{\text{like}}(\delta\mathbf{d} \mid \delta\mathbf{m})\, p_{\text{prior}}(\delta\mathbf{m})$$

posterior distribution, $p_{\text{post}}\left(\delta\mathbf{m} \mid \delta\mathbf{d}\right)$

likelihood function, $p_{\text{like}}(\delta\mathbf{d} \mid \delta\mathbf{m})$

prior distribution, $p_{\text{prior}}(\delta\mathbf{m})$

Likelihood function

$$-\log p_{\text{like}}\left(\delta\mathbf{d} \mid \delta\mathbf{m}\right) = -\sum_{i=1}^{n_s} \log p_{\text{like}}\left(\delta\mathbf{d}_i \mid \delta\mathbf{m}\right)$$

$$= \frac{1}{2\sigma^2}\sum_{i=1}^{n_s} \left\|\delta\mathbf{d}_i - \mathbf{J}[\mathbf{m}_0, \mathbf{q}_i]\delta\mathbf{m}\right\|_2^2 + \text{const},$$

source experiments, $\delta\mathbf{d} = \{\delta\mathbf{d}_i\}_{i=1}^{n_s}$

measures predicted data accuracy

based on Gaussian noise with variance σ^2

Prior distribution

encodes prior beliefs and information on $\delta\mathbf{m}$

probability of $\delta\mathbf{m}$ before observing data

handcrafted or data-driven

Posterior distribution

solution to the inverse problem

a distribution over the solution space, $p_{\text{post}}\left(\delta\mathbf{m} \mid \delta\mathbf{d}\right)$

assigns probability to different estimations

extracting information requires high-dimensional integration/sampling

Point estimators

maximimum likelihood, maximum a posteriori, conditional mean

pointwise variance, . . .

Maximum likelihood estimator

$$\underset{\delta\mathbf{m}}{\arg\max}\, p_{\text{like}}\left(\delta\mathbf{d} \mid \delta\mathbf{m}\right) = \underset{\delta\mathbf{m}}{\arg\min} - \sum_{i=1}^{n_s} \log p_{\text{like}}\left(\delta\mathbf{d}_i \mid \delta\mathbf{m}\right)$$

no high-dimensional integration/sampling

no prior/regularization

Maximum a posteriori estimate

$$\underset{\delta\mathbf{m}}{\arg\max}\;\; p_{\text{post}}\left(\delta\mathbf{m} \mid \delta\mathbf{d}\right) = \underset{\delta\mathbf{m}}{\arg\min}\left[-\sum_{i=1}^{n_s} \log p_{\text{like}}\left(\delta\mathbf{d}_i \mid \delta\mathbf{m}\right) - \log p_{\text{prior}}\left(\delta\mathbf{m}\right)\right]$$

no high-dimensional integration/sampling

uses prior information

requires choosing prior

Conditional mean estimate

$$\mathbb{E}\left[\delta\mathbf{m} \mid \delta\mathbf{d}\right] = \int \delta\mathbf{m} p_{\text{post}}\left(\delta\mathbf{m} \mid \delta\mathbf{d}\right) \mathrm{d}\delta\mathbf{m}$$

uses prior information

high-dimensional integration/sampling

requires choosing prior

Pointwise variance—uncertainty

SLIM

$$\mathbb{E}\left[\,\left(\delta\mathbf{m}-\mathbb{E}\left[\delta\mathbf{m}\mid\delta\mathbf{d}\right]\right)^{\circ 2}\mid\delta\mathbf{d}\right]=\int\left(\delta\mathbf{m}-\mathbb{E}\left[\delta\mathbf{m}\mid\delta\mathbf{d}\right]\right)^{\circ 2}p_{\text{post}}\left(\delta\mathbf{m}\mid\delta\mathbf{d}\right)\mathrm{d}\delta\mathbf{m}$$

uses prior information

a measure for uncertainty

high-dimensional integration/sampling

requires choosing prior

$^{\circ 2}$ stands for elementwise raising to the power of 2

Bayesian inversion with deep priors

regularization via an untrained CNN with Gaussian weights

V. Lempitsky, A. Vedaldi, and D. Ulyanov. "Deep Image Prior". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. June 2018, pp. 9446–9454. DOI: 10.1109/CVPR.2018.00984.

Deep priors in seismic

data reconstruction

denoising

Qun Liu, Lihua Fu, and Meng Zhang. "Deep-seismic-prior-based reconstruction of seismic data using convolutional neural networks". In: *arXiv preprint arXiv:1911.08784* (2019).

Francesco Picetti et al. "Anti-Aliasing Add-On for Deep Prior Seismic Data Interpolation". In: *arXiv preprint arXiv:2101.11361* (2021).

Yunzhi Shi, Xinming Wu, and Sergey Fomel. "Deep learning parameterization for geophysical inverse problems". In: *SEG 2019 Workshop: Mathematical Geophysics: Traditional vs Learning, Beijing, China, 5-7 November 2019*. Society of Exploration Geophysicists. 2020, pp. 36–40. DOI: 10.1190/iwmg2019_09.1.

Deep priors in seismic

imaging

full-waveform inversion

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.

Yulang Wu and George A McMechan. "Parametric convolutional neural network-domain full-waveform inversion". In: *GEOPHYSICS* 84.6 (2019), R881–R896. DOI: 10.1190/geo2018-0224.1.

Deep priors

$$\delta\mathbf{m} = g(\mathbf{z}, \mathbf{w}), \quad \mathbf{w} \sim \mathrm{N}(\mathbf{0}, \lambda^{-2}\mathbf{I})$$

untrained CNN, $g(\mathbf{z}, \mathbf{w})$

CNN weights, $\mathbf{w}$

fixed input, $\mathbf{z}$

variance of Gaussian prior on weight, λ^{-2}

V. Lempitsky, A. Vedaldi, and D. Ulyanov. “Deep Image Prior”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. June 2018, pp. 9446–9454. DOI: 10.1109/CVPR.2018.00984.

Deep-prior based likelihood function

$$-\log p_{\text{like}}\left(\delta\mathbf{d} \mid \mathbf{w}\right) = \frac{1}{2\sigma^2}\sum_{i=1}^{n_s}\|\delta\mathbf{d}_i - \mathbf{J}[\mathbf{m}_0, \mathbf{q}_i]g(\mathbf{z}, \mathbf{w})\|_2^2 + \underbrace{\text{const}}_{\text{Ind. of } \mathbf{w}}$$

source experiments, $\delta\mathbf{d} = \{\delta\mathbf{d}_i\}_{i=1}^{n_s}$

estimated noise variance, σ^2

Prior distribution

$$-\log p_{\text{prior}}\left(\mathbf{w}\right) = \frac{\lambda^2}{2}\|\mathbf{w}\|_2^2 + \underbrace{\text{const}}_{\text{Ind. of } \mathbf{w}}$$

Gaussian prior on $\mathbf{w}$

λ^2 to be tuned

Deep-prior based posterior distribution

$$-\log p_{\text{post}}(\mathbf{w} \mid \delta\mathbf{d}) = \underbrace{\frac{1}{2\sigma^2}\sum_{i=1}^{n_s}\|\delta\mathbf{d}_i - \mathbf{J}[\mathbf{m}_0, \mathbf{q}_i]g(\mathbf{z}, \mathbf{w})\|_2^2}_{\text{negative-log likelihood}} + \underbrace{\frac{\lambda^2}{2}\|\mathbf{w}\|_2^2}_{\substack{\text{negative-log} \\ \text{prior}}} + \underbrace{\text{const}}_{\text{Ind. of } \mathbf{w}}$$

Deep-prior based MAP

$$\mathbf{w}_{\mathsf{MAP}} := \underset{\mathbf{w}}{\arg\max} \;\; p_{\mathsf{post}} \left(\mathbf{w} \mid \delta \mathbf{d} \right)$$

MAP estimate, $\delta \mathbf{m}_{\mathsf{MAP}} = g(\mathbf{z}, \mathbf{w}_{\mathsf{MAP}})$

Deep-prior based conditional mean

$$\delta\mathbf{m}_{\mathsf{CM}} := \mathbb{E}\Big[g(\mathbf{z},\mathbf{w}) \mid \delta\mathbf{d}\Big] \approx \frac{1}{n_{\mathrm{w}}}\sum_{j=1}^{n_{\mathrm{w}}} g(\mathbf{z},\mathbf{w}_j),$$

requires samples from posterior, $\{\mathbf{w}_j\}_{j=1}^{n_{\mathrm{w}}} \sim p_{\mathsf{post}}(\mathbf{w} \mid \delta\mathbf{d})$

$n_{\mathrm{w}} = 5000$

Pointwise variance

SLIM

$$\sigma^2_{\text{post}} := \mathbb{E}\Big[\Big(g(\mathbf{z},\mathbf{w}) - \mathbb{E}\big[g(\mathbf{z},\mathbf{w}) \mid \delta\mathbf{d}\big]\Big)^{\circ 2} \mid \delta\mathbf{d}\Big]$$

$$\approx \frac{1}{n_{\text{w}}}\sum_{j=1}^{n_{\text{w}}}\big(g(\mathbf{z},\mathbf{w}_i) - \delta\mathbf{m}_{\text{CM}}\big)^{\circ 2}$$

requires samples from posterior, $\{\mathbf{w}_j\}_{j=1}^{n_{\text{w}}} \sim p_{\text{post}}(\mathbf{w} \mid \delta\mathbf{d})$

$^{\circ 2}$ stands for elementwise raising to the power of 2

Sampling the posterior, $p_{\text{post}}(\mathbf{w} \mid \delta\mathbf{d})$

stochastic gradient Langevin dynamics (SGLD)

SGLD

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

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

step size sequence, α_k

adaptive preconditioning matrix, $\mathbf{M}_k$

i chosen randomly without replacement from $\{1, \ldots, n_s\}$

n_s simultaneous sources, $\mathbf{q}_i$

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.

SGLD—pros vs cons

allows for stochastic likelihood approximation

exploits gradient information

requires decreasing step size sequence, $\alpha_k \to 0$

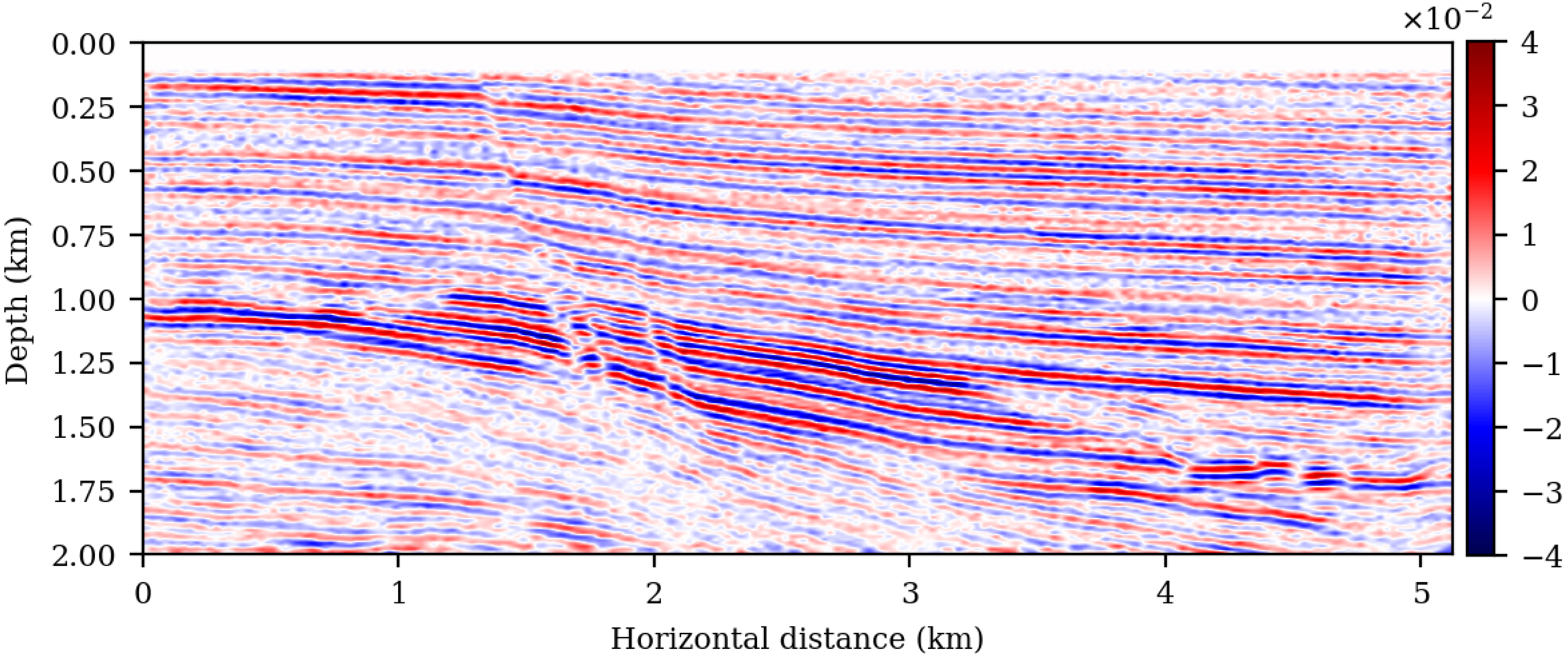
slow mixing rate, needs many iterations

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.

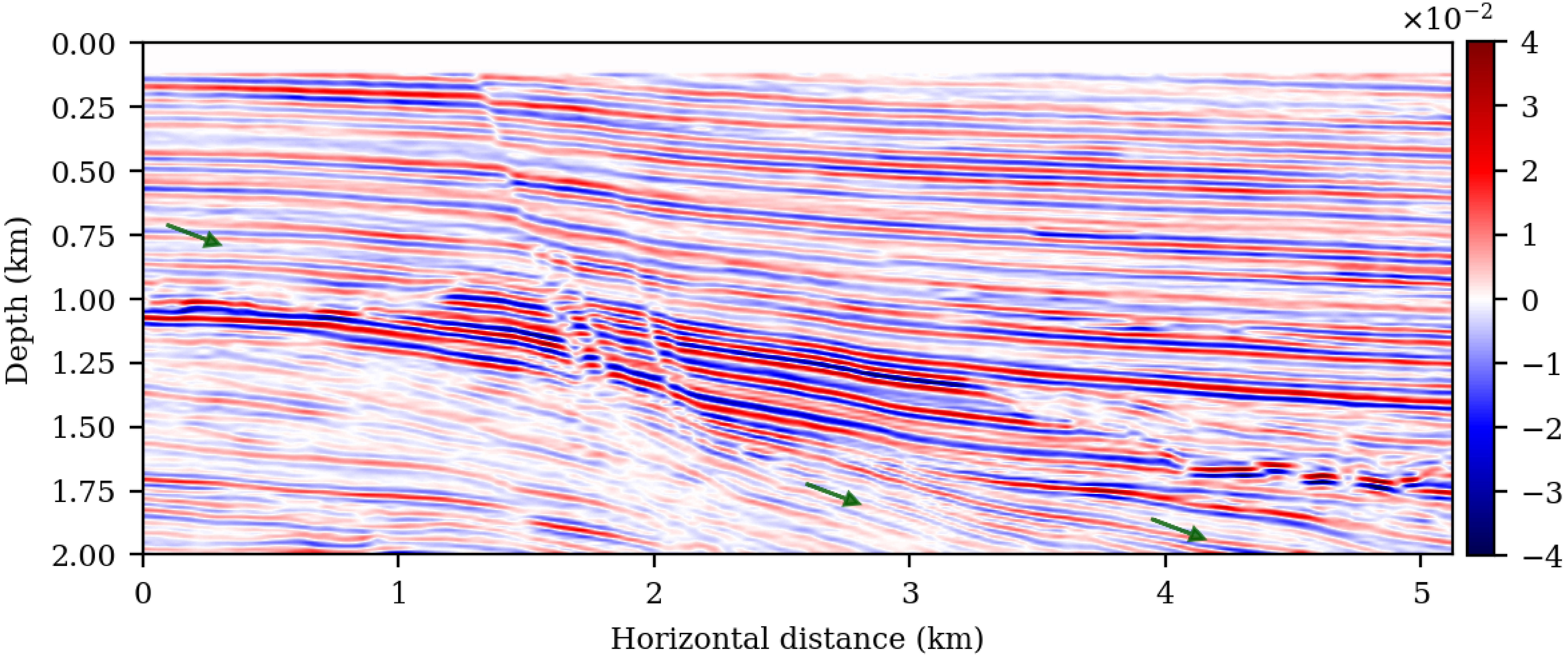
Nicolas Brosse, Alain Durmus, and Eric Moulines. "The promises and pitfalls of Stochastic Gradient Langevin Dynamics". In: *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 2018, pp. 8268–8278. URL: http://papers.nips.cc/paper/8048-the-promises-and-pitfalls-of-stochastic-gradient-langevin-dynamics.pdf.

Parihaka example

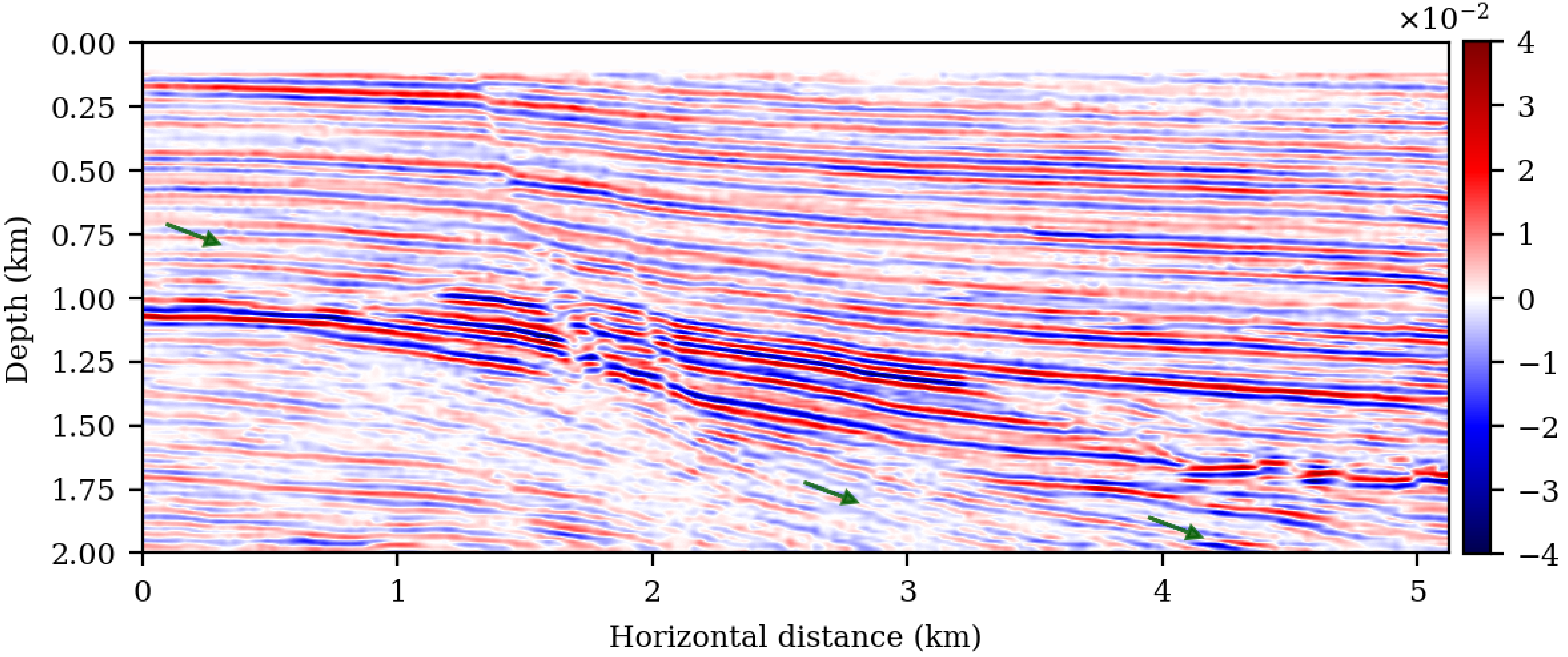
Comparing MLE (no regularization), MAP, and conditional mean



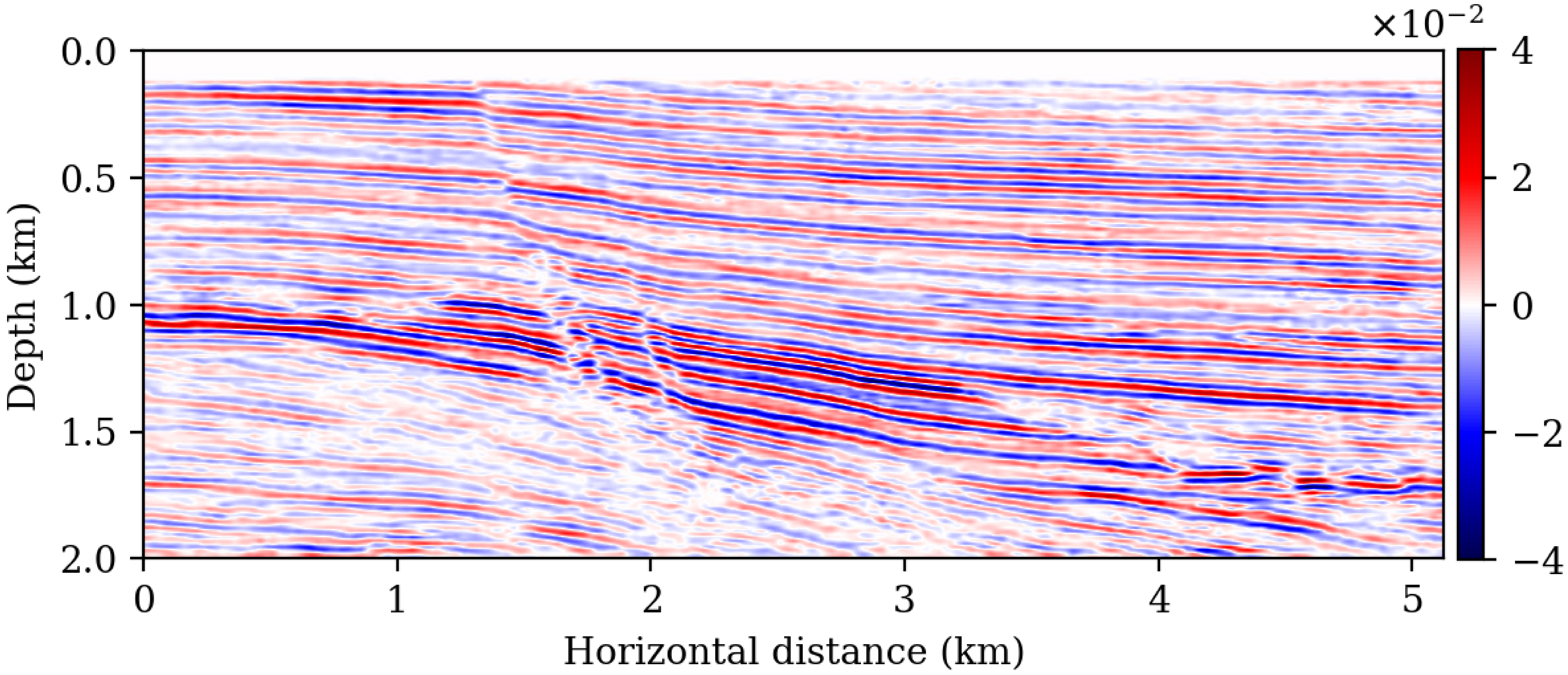
MLE—no regularization (prior)



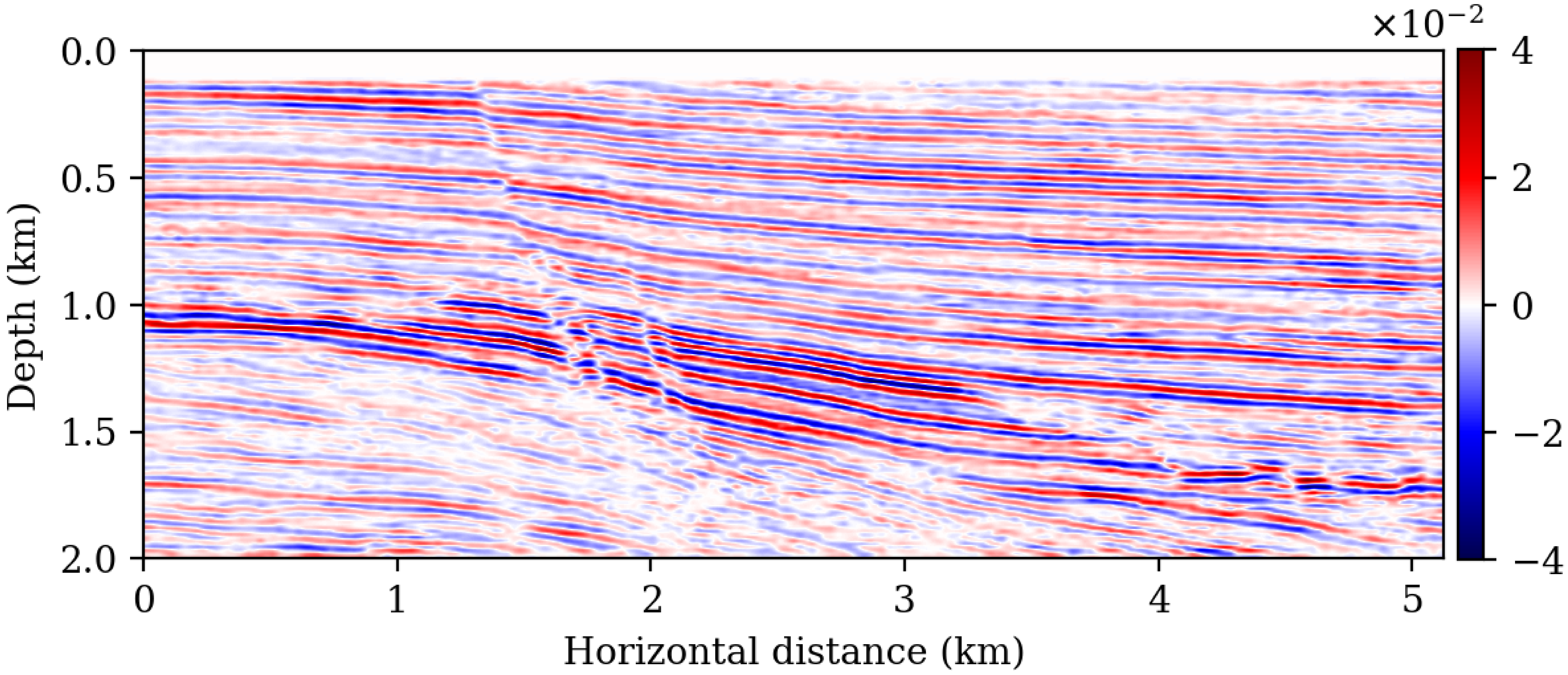
$\delta\mathbf{m}$—"true" reflectivity model obtain from Parihaka dataset



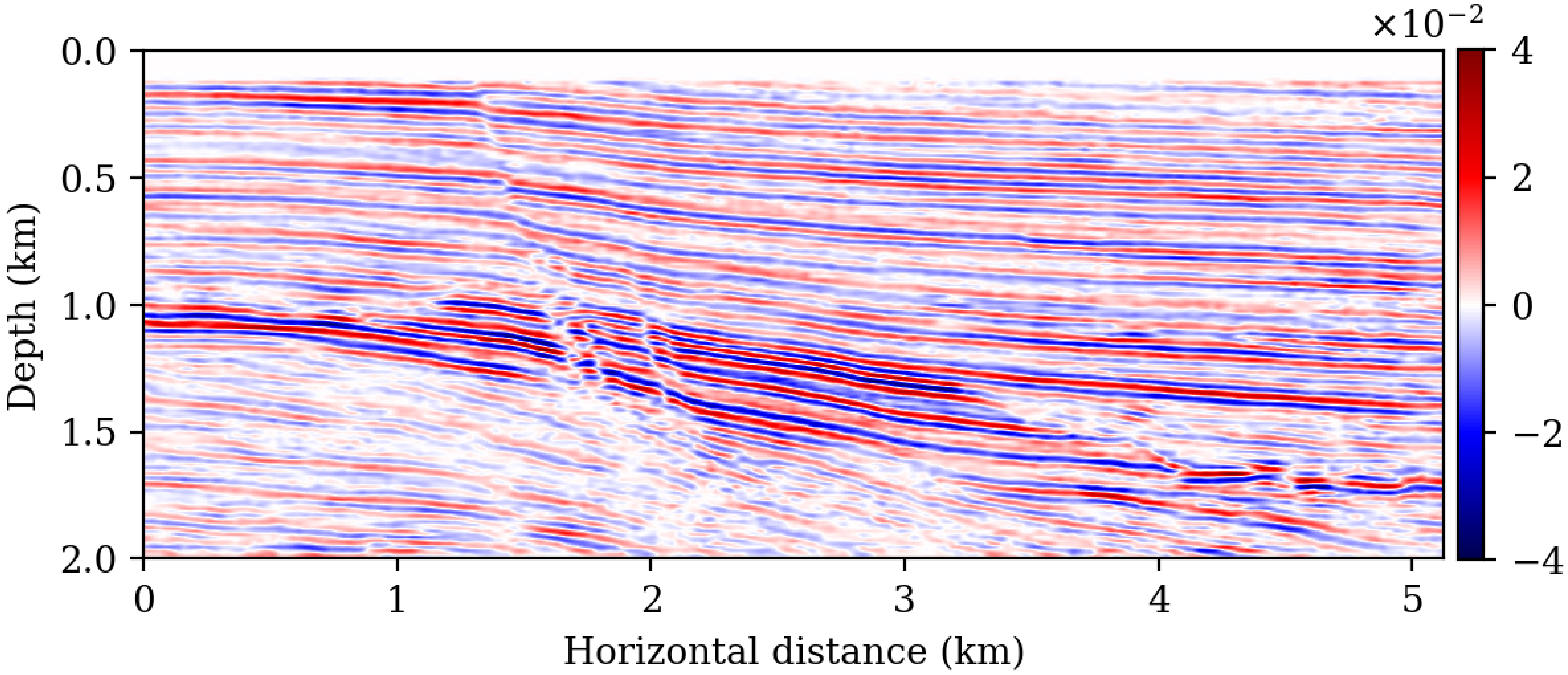
MAP estimate—SNR: 8.77 dB



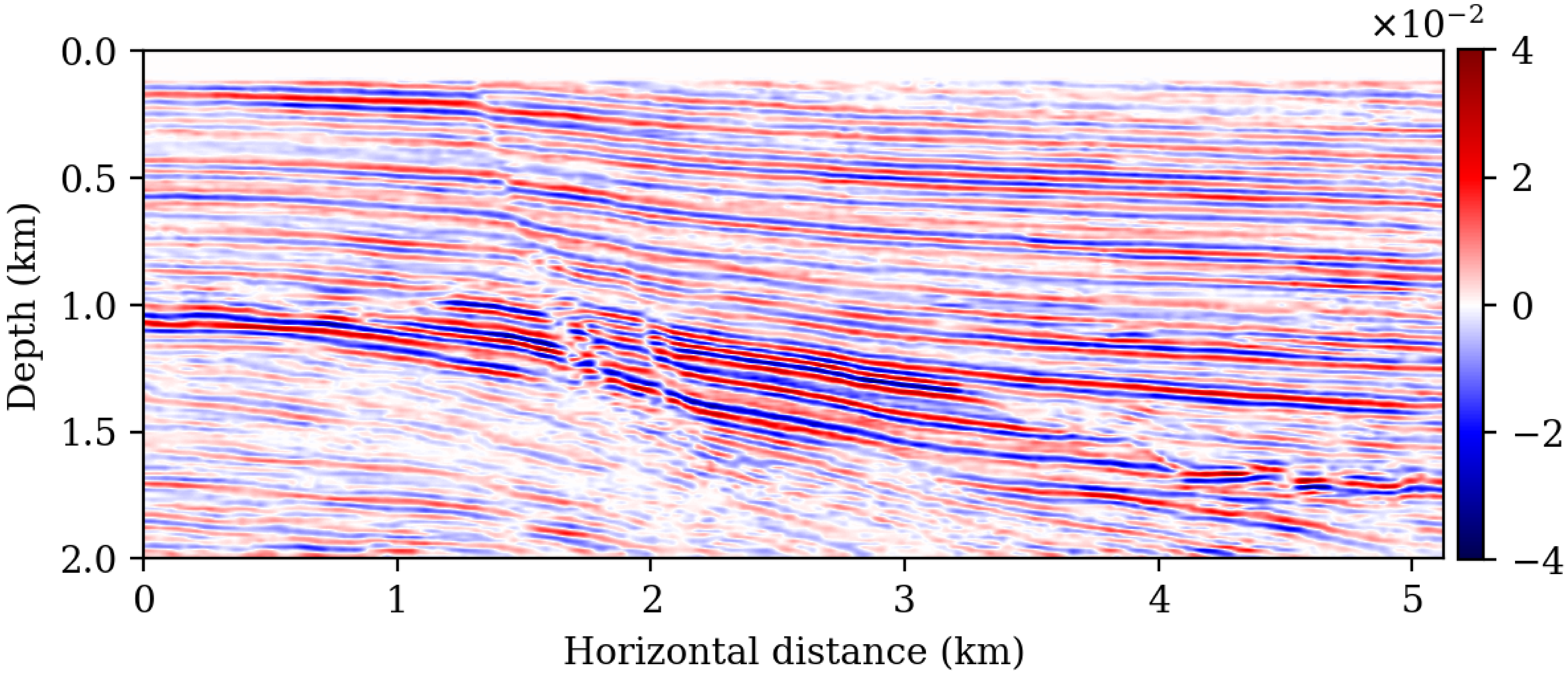
Posterior samples



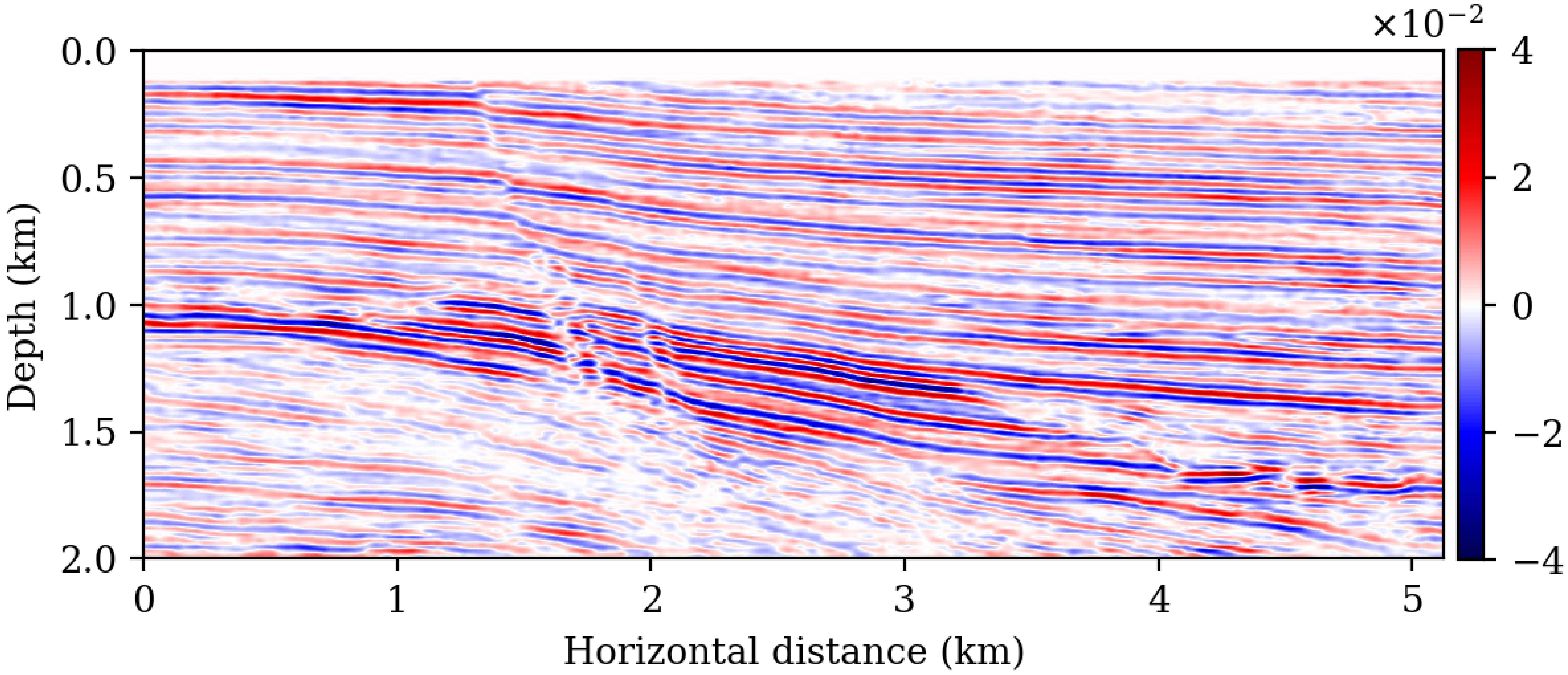
Posterior samples



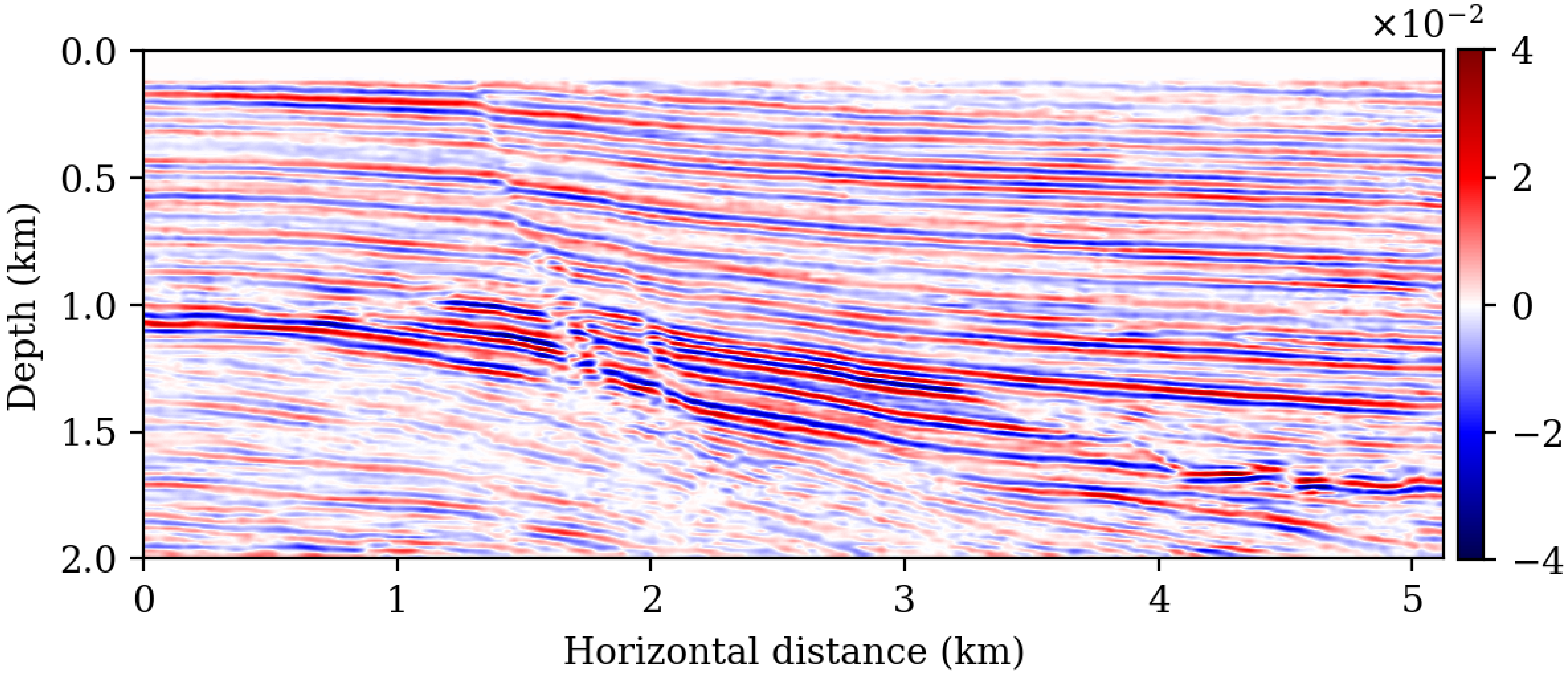
Posterior samples



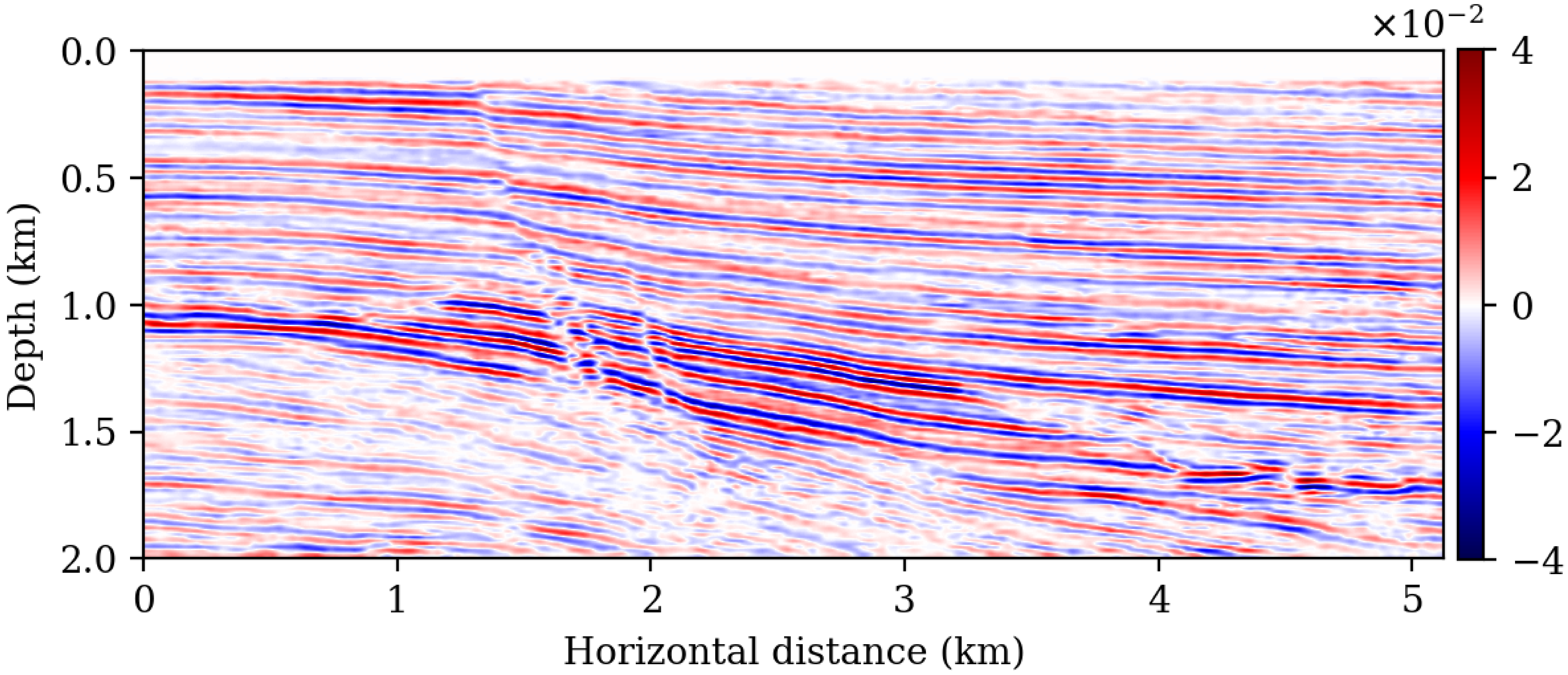
Posterior samples



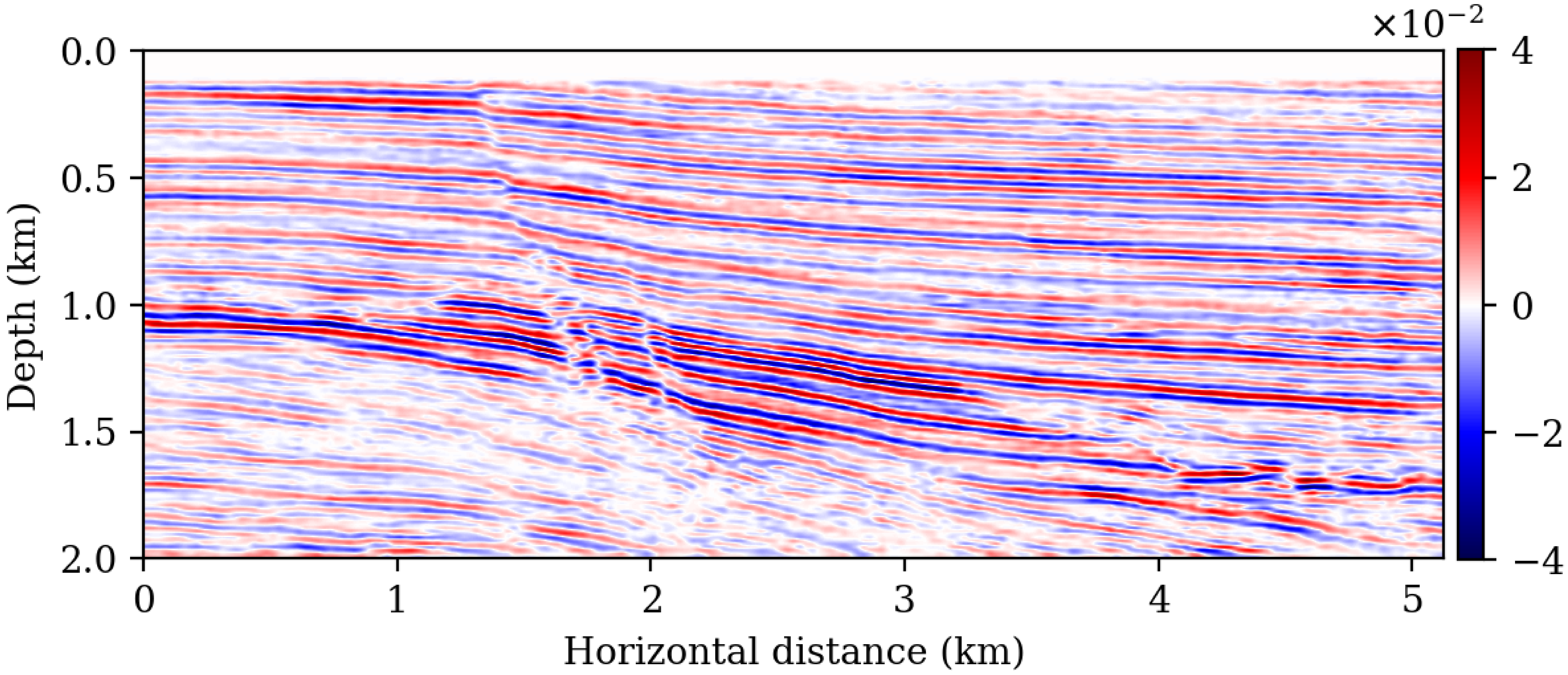
Posterior samples



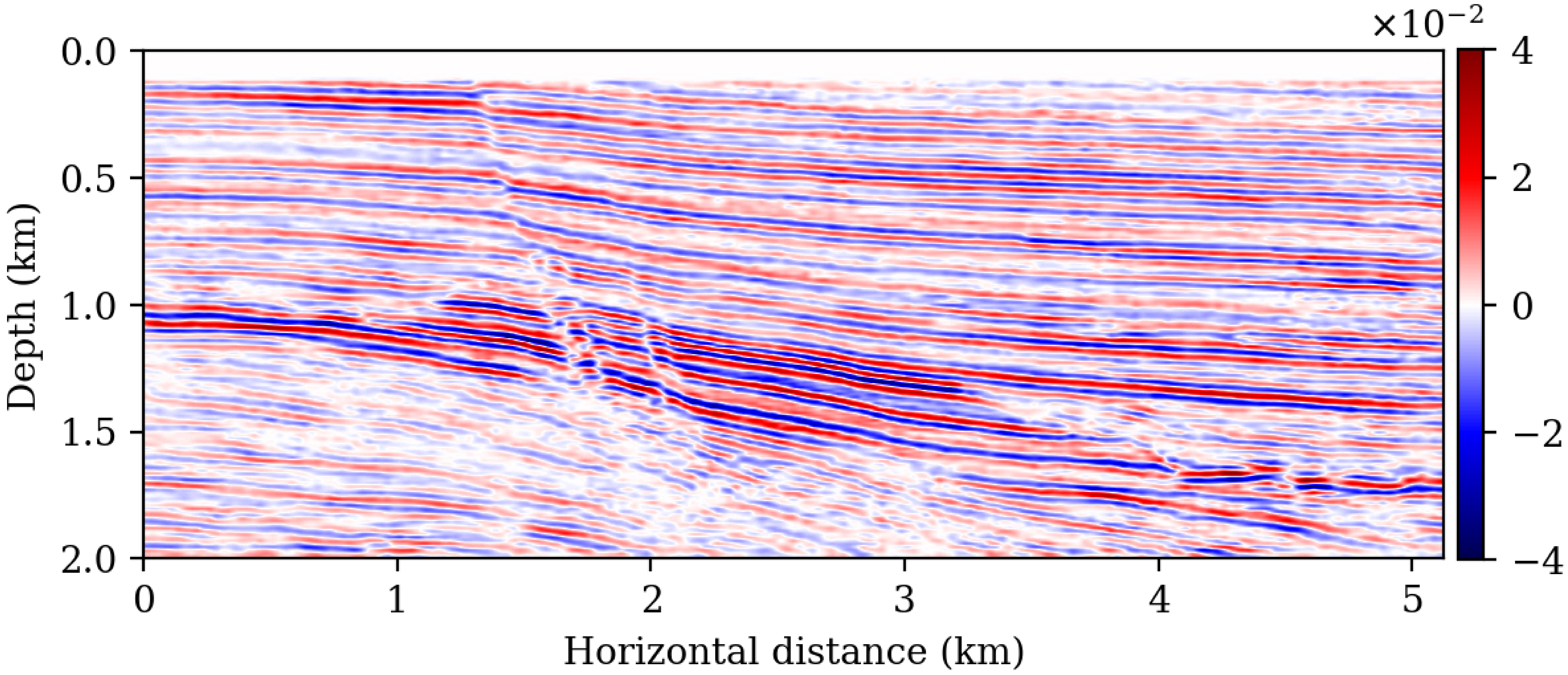
Posterior samples



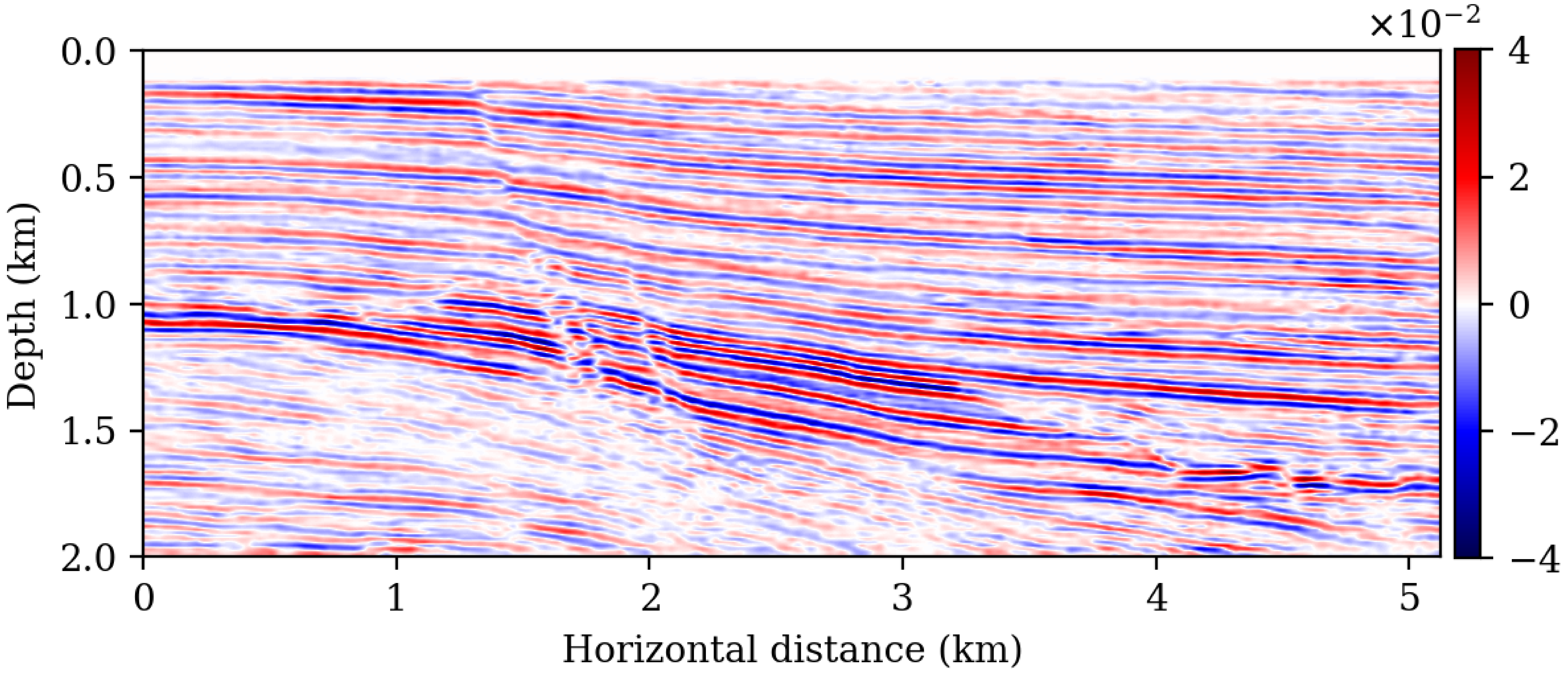
Posterior samples



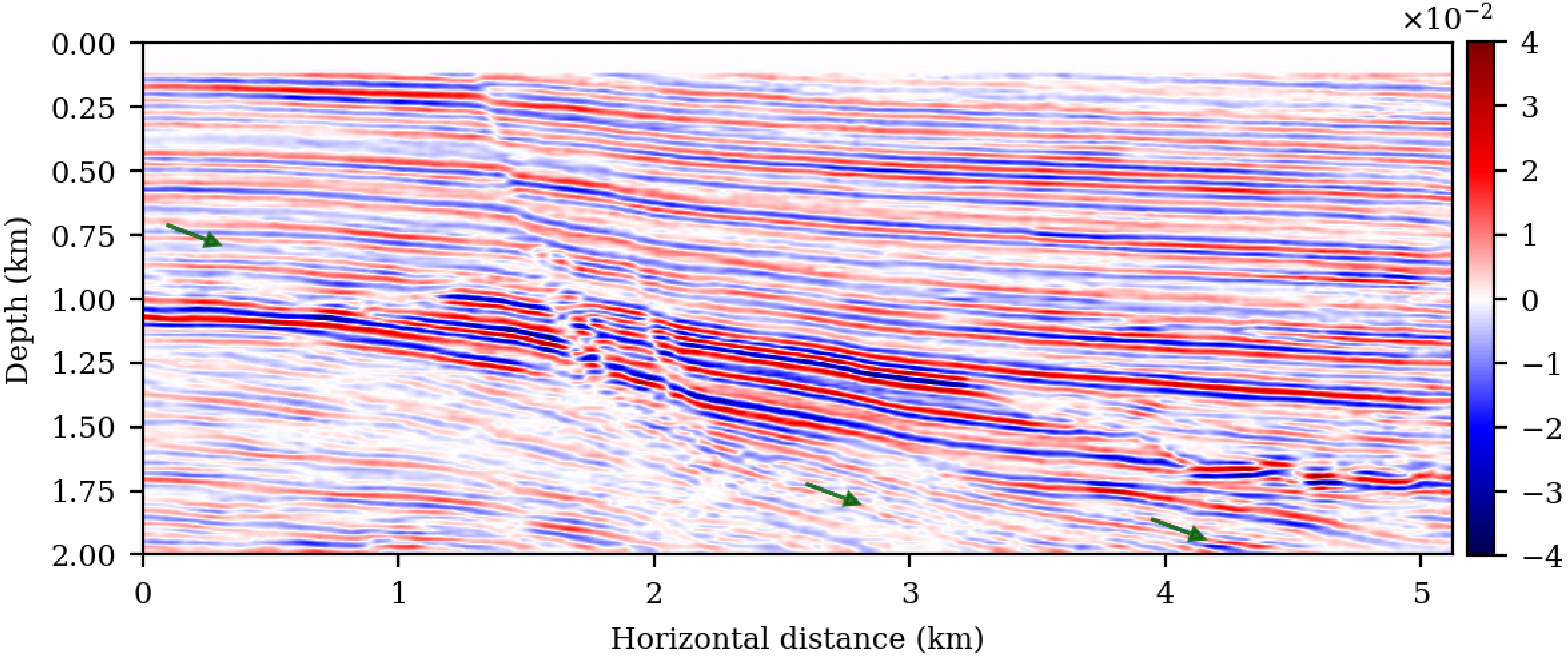
Posterior samples



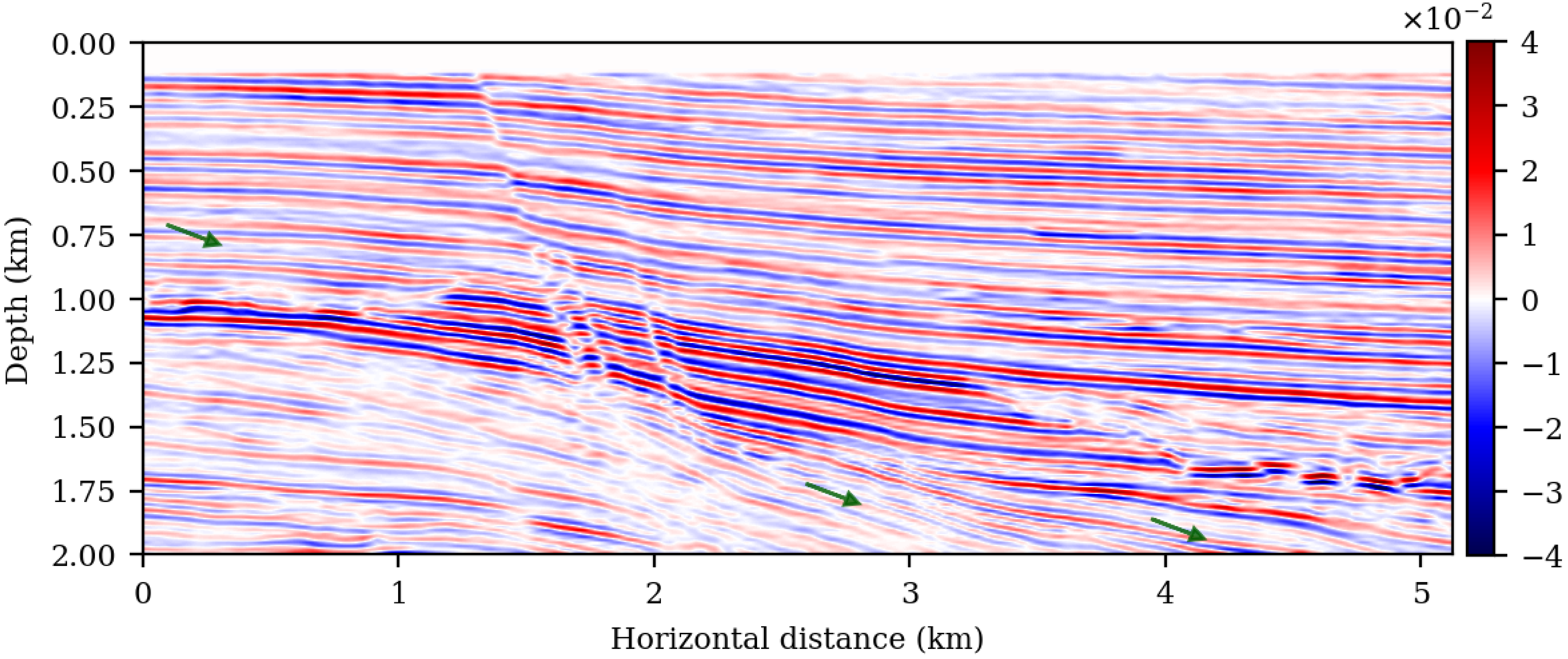
Posterior samples



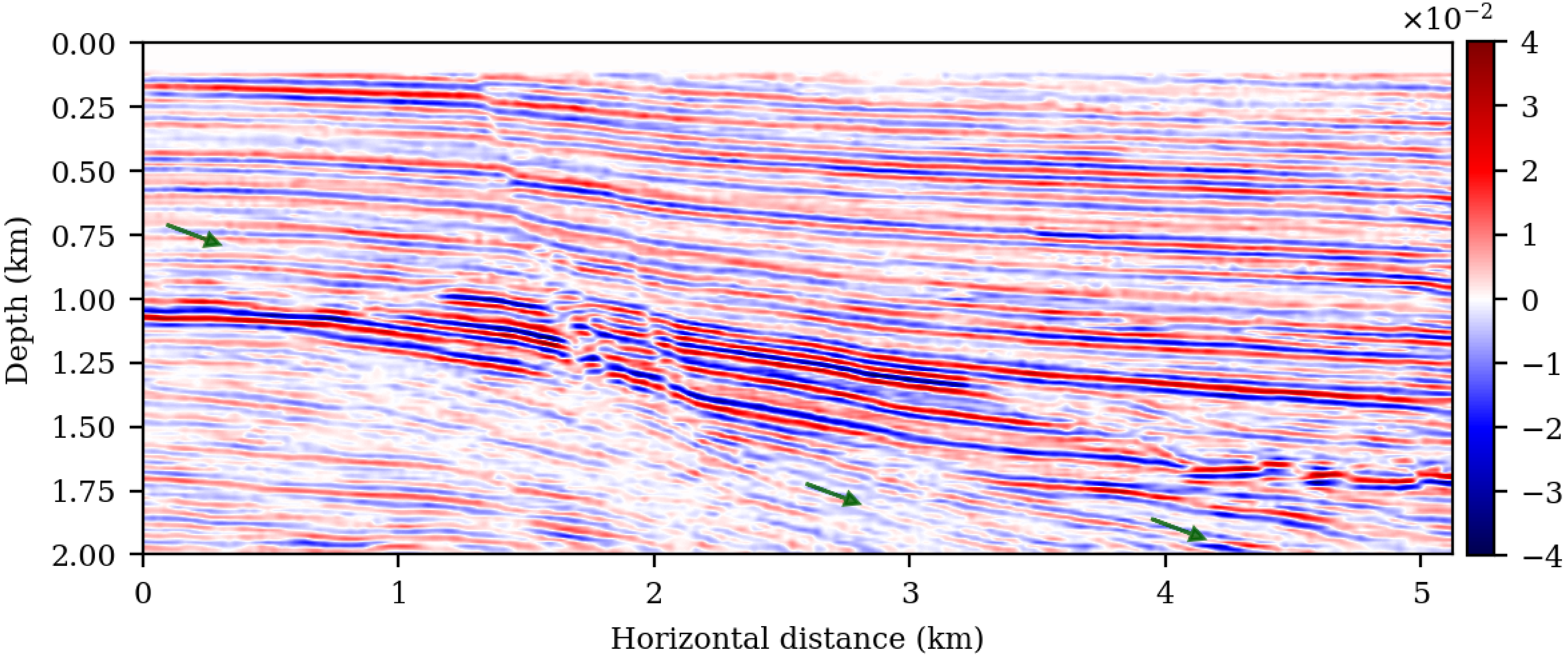
Posterior samples



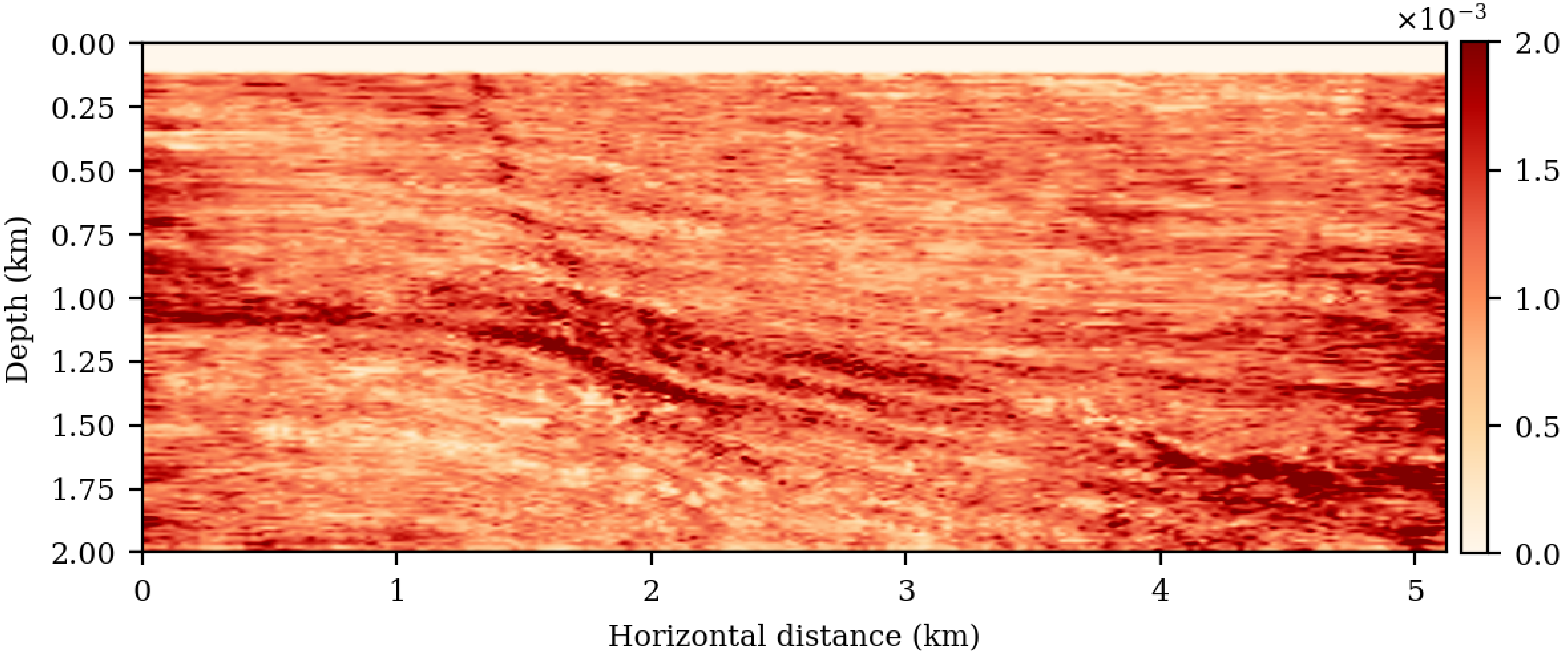
Conditional mean estimate—SNR: 9.66 dB



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

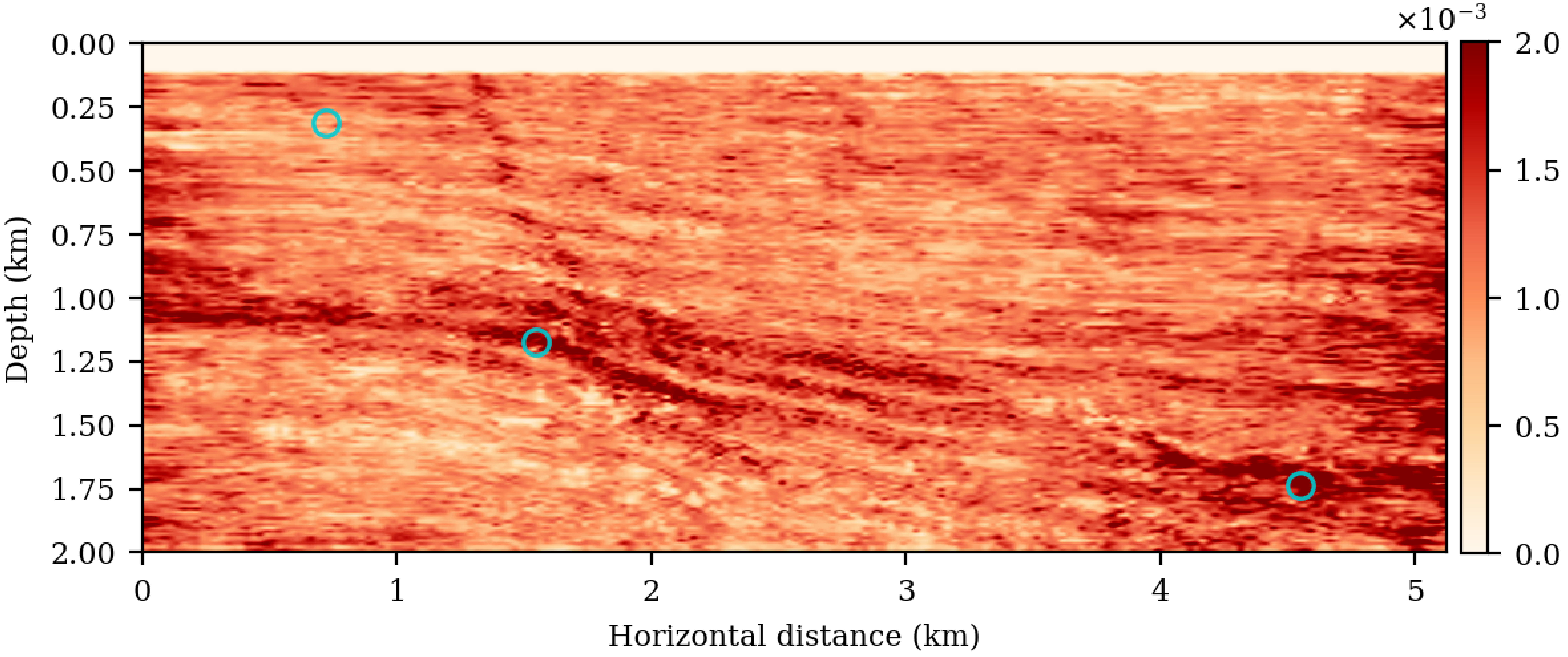


MAP estimate—SNR: 8.77 dB

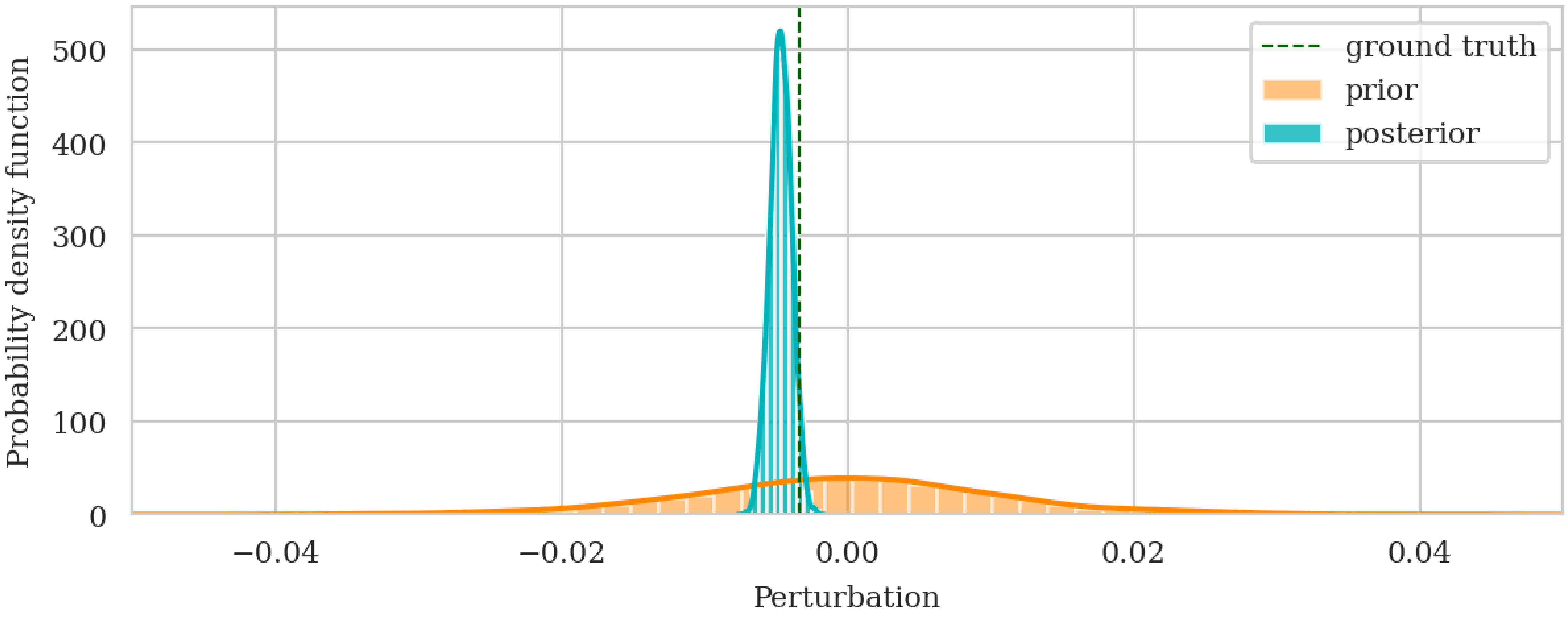


×10⁻³
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)
2.0
1.5
1.0
0.5
0.0

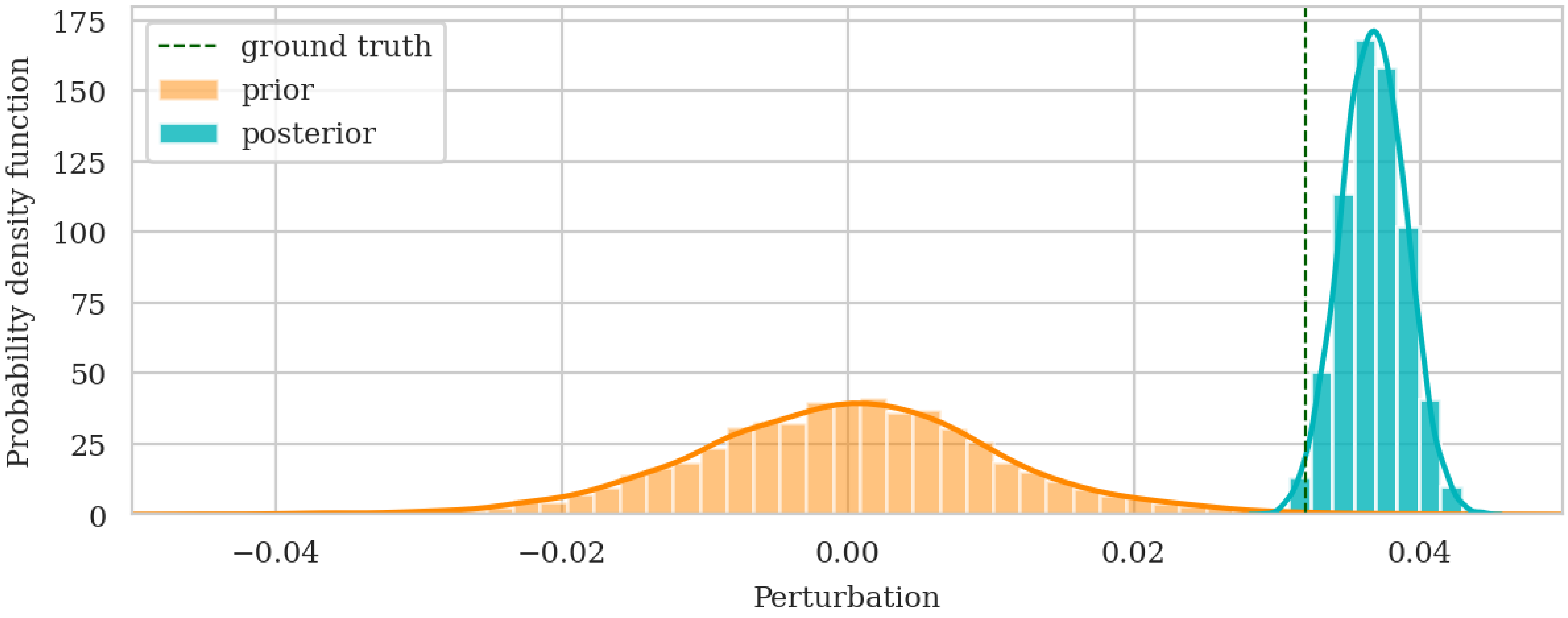
Imaging UQ


×10⁻³
2.0
1.5
1.0
0.5
0.0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)

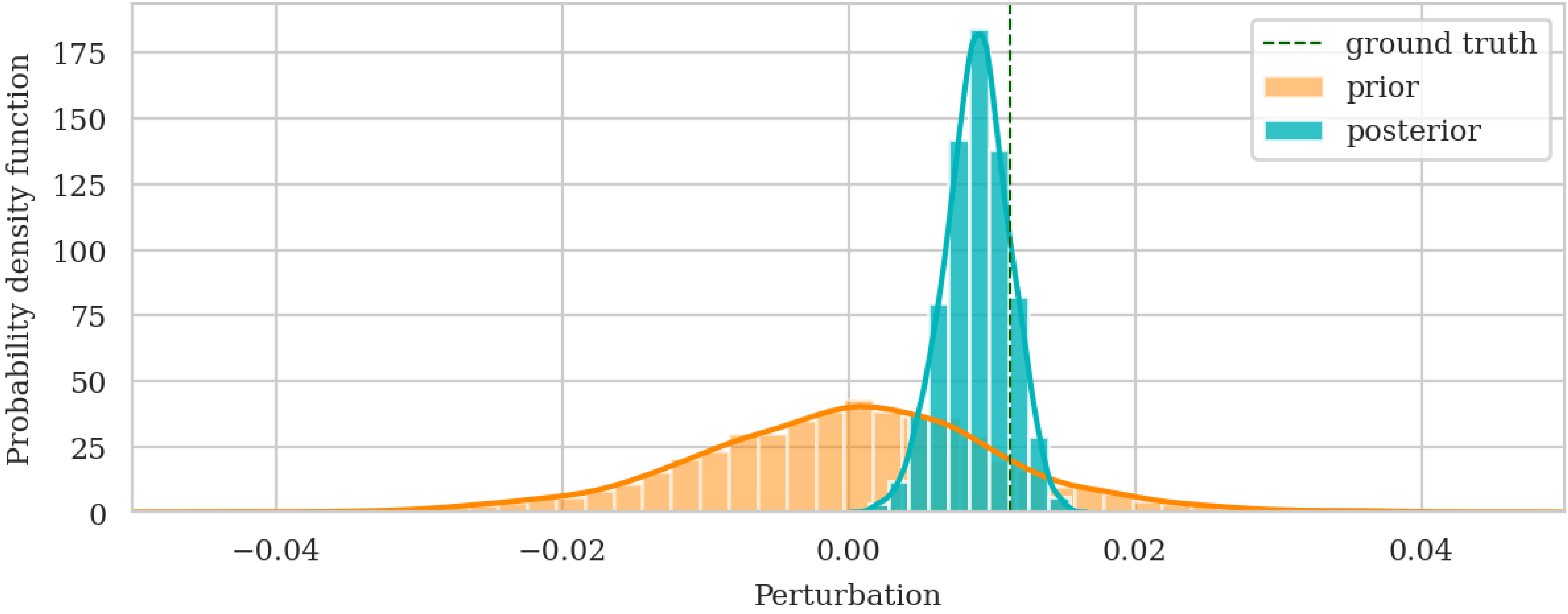
Imaging UQ



Pointwise marginals at $(0.73\,\mathrm{km},\ 0.32\,\mathrm{km})$



Pointwise marginals at $(1.55\,\mathrm{km},\ 1.18\,\mathrm{km})$



Pointwise marginals at $(4.55\,\mathrm{km},\ 1.74\,\mathrm{km})$

Downstream-task uncertainty quantification

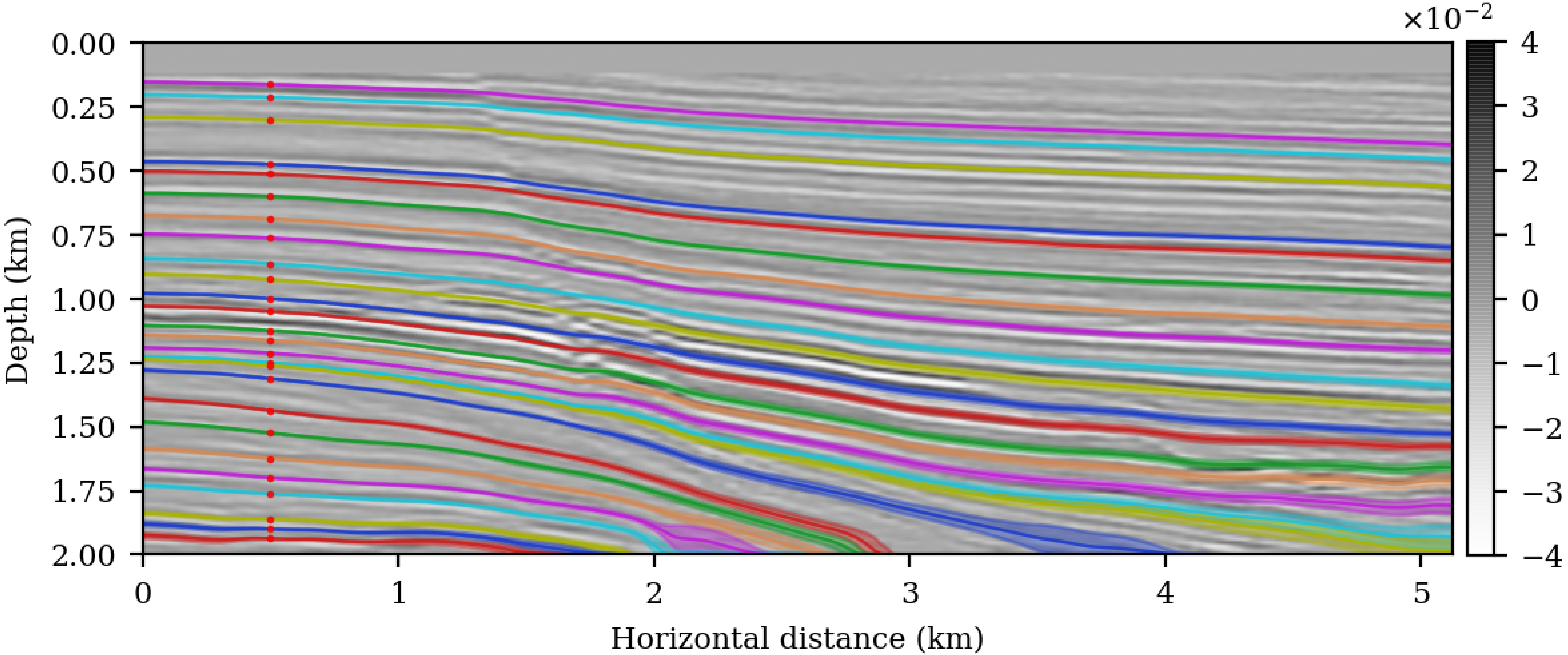
horizon tracking and uncertainty analysis

Deterministic horizon tracker

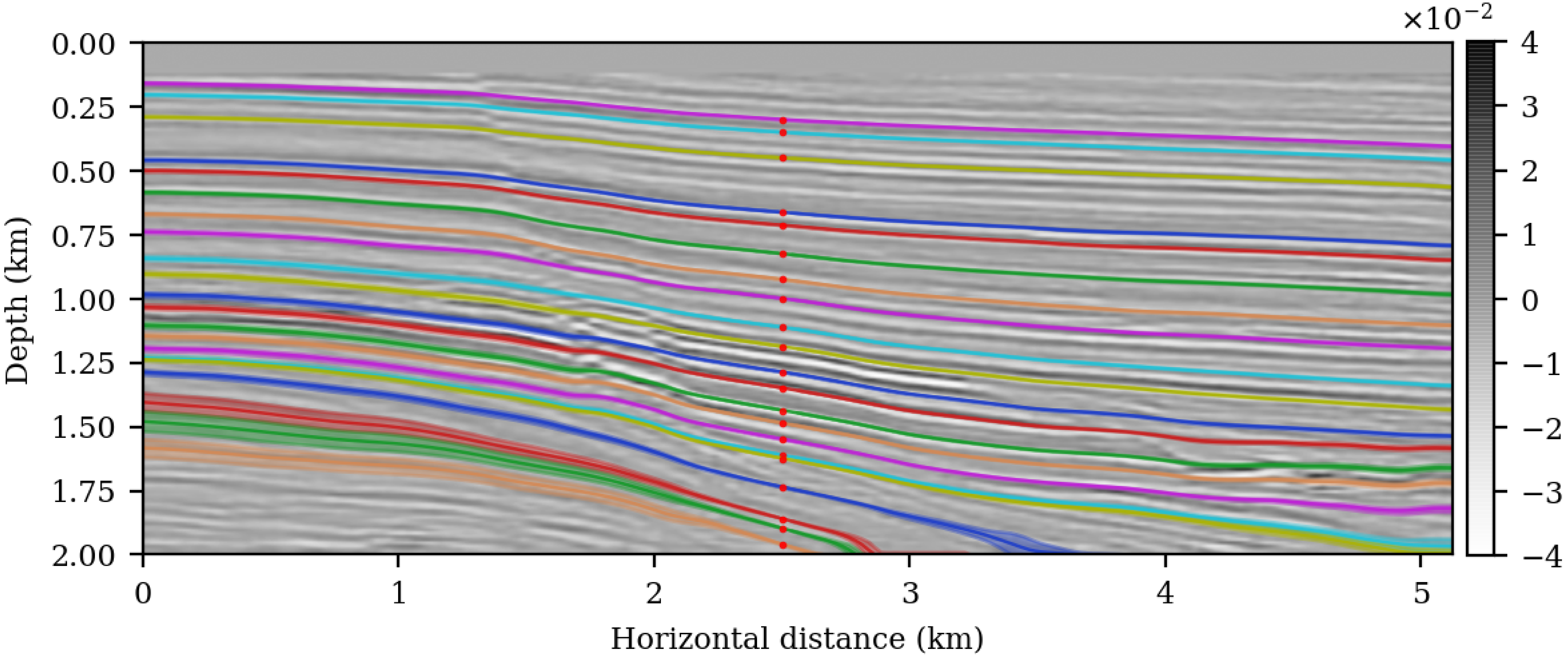
$$\mathbb{E}\left[f\left(\mathbf{h}\right) \mid \delta\mathbf{d}\right] = \iint f\left(\mathbf{h}\right) \delta_{[\mathbf{h}=\mathcal{H}(\delta\mathbf{m})]}\left(\mathbf{h}\right) p_{\text{post}}\left(\delta\mathbf{m} \mid \delta\mathbf{d}\right) \,\mathrm{d}\mathbf{h}\,\mathrm{d}\delta\mathbf{m}$$

$$\approx \frac{1}{n_{\mathrm{w}}} \sum_{j=1}^{n_{\mathrm{w}}} f\Big(\mathcal{H}(g\left(\mathbf{z}, \mathbf{w}_j\right))\Big)$$

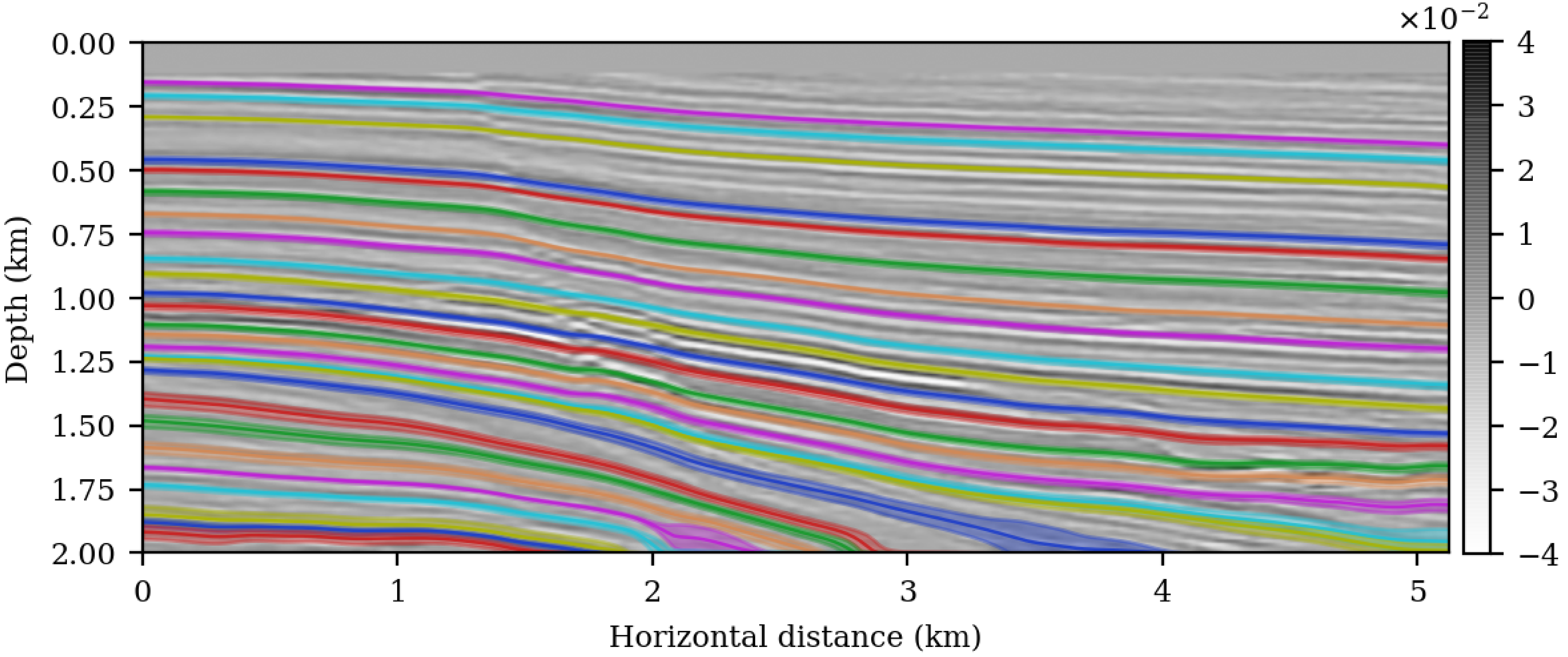
samples from posterior, $\{\mathbf{w}_j\}_{j=1}^{n_{\mathrm{w}}} \sim p_{\text{post}}(\mathbf{w} \mid \delta\mathbf{d})$



Uncertainty in horizon tracking due to uncertainties in imaging



Uncertainty in horizon tracking due to uncertainties in imaging



Uncertainty in horizon tracking due uncertainties in imaging and control points

Conclusions

Imaging with deep prior—albeit expensive circumvents artifacts and does not bia

Optimization with SGLD—yields reasonable first and second moments of the posterior while working with 205 sim. shots

Conditional mean—contains less artifacts and is more robust to noise

Uncertainties are reasonable in deep, close to boundary, and geologically complex areas

Computational costs (10k iterations) are unfeasible in practice

Contributions

Introduced regularization with deep priors, which only relies on network architecture

Conducted imaging UQ via SGLD

`Devito4PyTorch`—integration of Devito's PDE solvers into PyTorch

`github.com/slimgroup/Software.SEG2020`

Ali Siahkoohi, Mathias Louboutin, and Felix Herrmann. *slimgroup/Devito4PyTorch*. 2020. URL: `https://github.com/slimgroup/Devito4PyTorch`.

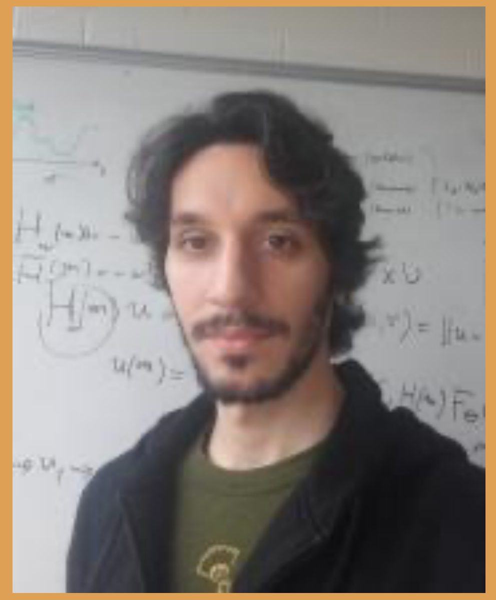
Preconditioned training of normalizing flows for variational inference in inverse problems

Ali Siahkoohi [1] Gabrio Rizzuti [1*] Mathias Louboutin [1] Philipp A. Witte [2]
Felix J. Herrmann [1]

[1] Georgia Institute of Technology

[2] Microsoft

[*] Now at Utrecht University



SLIM

Georgia Institute of Technology

Normalizing flows (NFs)

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

change of variable formula

latent distribution, $p_z(\mathbf{z}) = \mathrm{N}\left(\mathbf{z} \mid \mathbf{0}, \mathbf{I}\right)$

bijective transformation, $G_\theta : \mathcal{X} \rightarrow \mathcal{Z}$

explicitly models the unknown density

Danilo Rezende and Shakir Mohamed. "Variational Inference with Normalizing Flows". In: vol. 37. Proceedings of Machine Learning Research. PMLR, 2015, pp. 1530–1538. URL: http://proceedings.mlr.press/v37/rezende15.html.

Training NFs

$$\arg\max_{\theta} \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})} \left[\log p_\theta(\mathbf{x}) \right]$$

$$= \arg\min_{\theta} \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})} \left[\frac{1}{2} \left\| G_\theta(\mathbf{x}) \right\|^2 - \log \left| \det \nabla_x G_\theta(\mathbf{x}) \right| \right]$$

maximum likelihood estimate for $\boldsymbol{\theta}$

$\det \nabla_x G_\theta(\mathbf{x})$ comes for free with *affine coupling layers*

Danilo Rezende and Shakir Mohamed. "Variational Inference with Normalizing Flows". In: vol. 37. Proceedings of Machine Learning Research. PMLR, 2015, pp. 1530–1538. URL: http://proceedings.mlr.press/v37/rezende15.html.

NFs—pros vs cons

fast joint and conditional sampling

exact and tractable density (likelihood) estimation

allows for memory efficient gradient computation

often have less representation power

unable to alter dimensionality

Variational inference with NFs

sampling the posterior distribution directly via NFs

Problem setup

Find $\mathbf{x}$ such that:

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

observed data, $\mathbf{y}$

measurement noise, $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$

expensive (nonlinear) forward operator, F

limited access to training data (marginal or joint)

Training a NF for posterior inference

$$
\begin{aligned}
&\arg\min_{\theta} \mathbb{D}_{\mathsf{KL}} \left(p_\theta \;||\; p_{\mathsf{post}}(\,\cdot \mid \boldsymbol{y})\right) \\
&= \arg\min_{\theta} \mathbb{E}_{\boldsymbol{x}\sim p_\theta(\boldsymbol{x}|\boldsymbol{y})} \Big[-\log p_{\mathsf{post}}(\boldsymbol{x} \mid \boldsymbol{y}) + \log p_\theta(\boldsymbol{x}) \Big].
\end{aligned}
$$

KL divergence, $\mathbb{D}_{\mathsf{KL}}$, a distance between two distributions

distribution encoded by the NF, $p_\theta(\boldsymbol{x}) = p_z(\boldsymbol{z}) \left| \det \nabla_x T_\theta^{-1}(\boldsymbol{x}) \right|, \quad \mathbf{z} = T_\theta^{-1}(\mathbf{x})$

Posterior inference with NFs

purely physics-based approach

- tied to the physics
- no training data needed
- requires handcrafted priors

purely data-driven approach

- needs supervised pairs of model and data
- heavily relies on the training data to generalize
- ignores the physics

Data-driven variational inference

$$\min_{\phi} \mathbb{E}_{\mathbf{y},\mathbf{x}\sim p_{y,x}(\mathbf{y},\mathbf{x})} \left[\frac{1}{2}\left\|G_{\phi}(\mathbf{y},\mathbf{x})\right\|^2 - \log\left|\det \nabla_{y,x} G_{\phi}(\mathbf{y},\mathbf{x})\right|\right]$$

cheap posterior sampling for all $\mathbf{y}$: $G^{-1}_{\phi_x}(G_{\phi_y}(\mathbf{y}),\mathbf{z}) \sim p_{\text{post}}(\boldsymbol{x} \mid \boldsymbol{y})$ for $\mathbf{z} \sim \mathrm{N}(\mathbf{0},\boldsymbol{I})$

does not involve the expensive forward operator, F

does not require a prior density, $p_{\text{prior}}(\mathbf{x})$

heavily relies on access to training pairs, $\widehat{p}_{y,x} \neq p_{y,x}$

Ricardo Baptista, Olivier Zahm, and Youssef Marzouk. *An adaptive transport framework for joint and conditional density estimation*. 2020. arXiv: 2009.10303 [stat.ML].

Block-triangular map

SLIM

$$G_\phi(\mathbf{y}, \mathbf{x}) = \begin{bmatrix} G_{\phi_y}(\mathbf{y}) \\ G_{\phi_x}(\mathbf{y}, \mathbf{x}) \end{bmatrix}, \quad \phi = \begin{bmatrix} \phi_y \\ \phi_x \end{bmatrix}$$

conditional sampling: $G_{\phi_x}^{-1}(G_{\phi_y}(\mathbf{y}), \mathbf{z}) \sim p_{\text{post}}\left(\boldsymbol{x} \mid \boldsymbol{y}\right)$ for $\mathbf{z} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{I})$

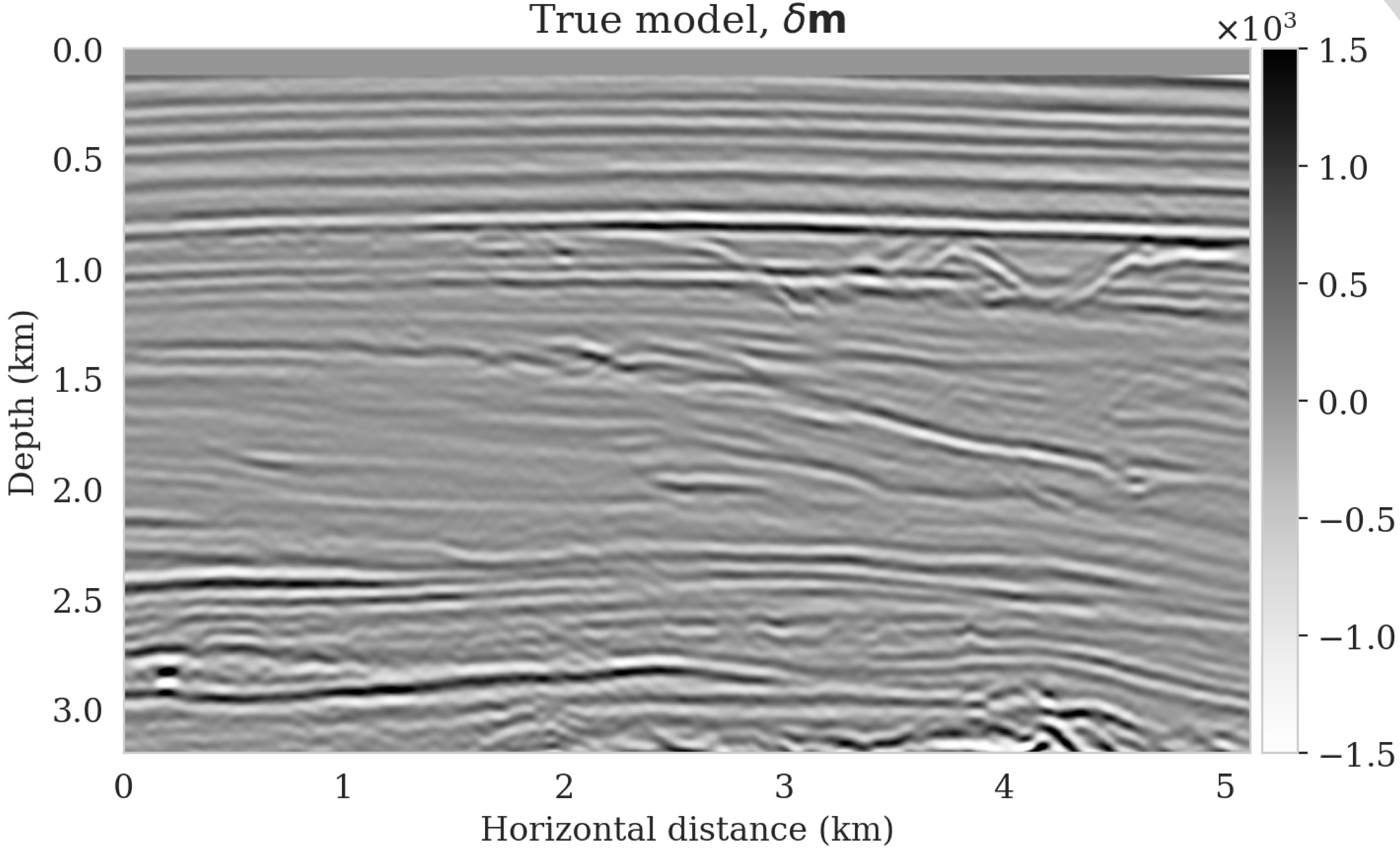
posterior density estimation: $p_G(\boldsymbol{x} \mid \boldsymbol{y}) = p_z\big(G_{\phi_x}(\boldsymbol{y}, \boldsymbol{x})\big) \left| \det \nabla_x G_{\phi_x}(\boldsymbol{y}, \boldsymbol{x}) \right|.$

Ricardo Baptista, Olivier Zahm, and Youssef Marzouk. *An adaptive transport framework for joint and conditional density estimation*. 2020. arXiv: 2009.10303 [stat.ML].

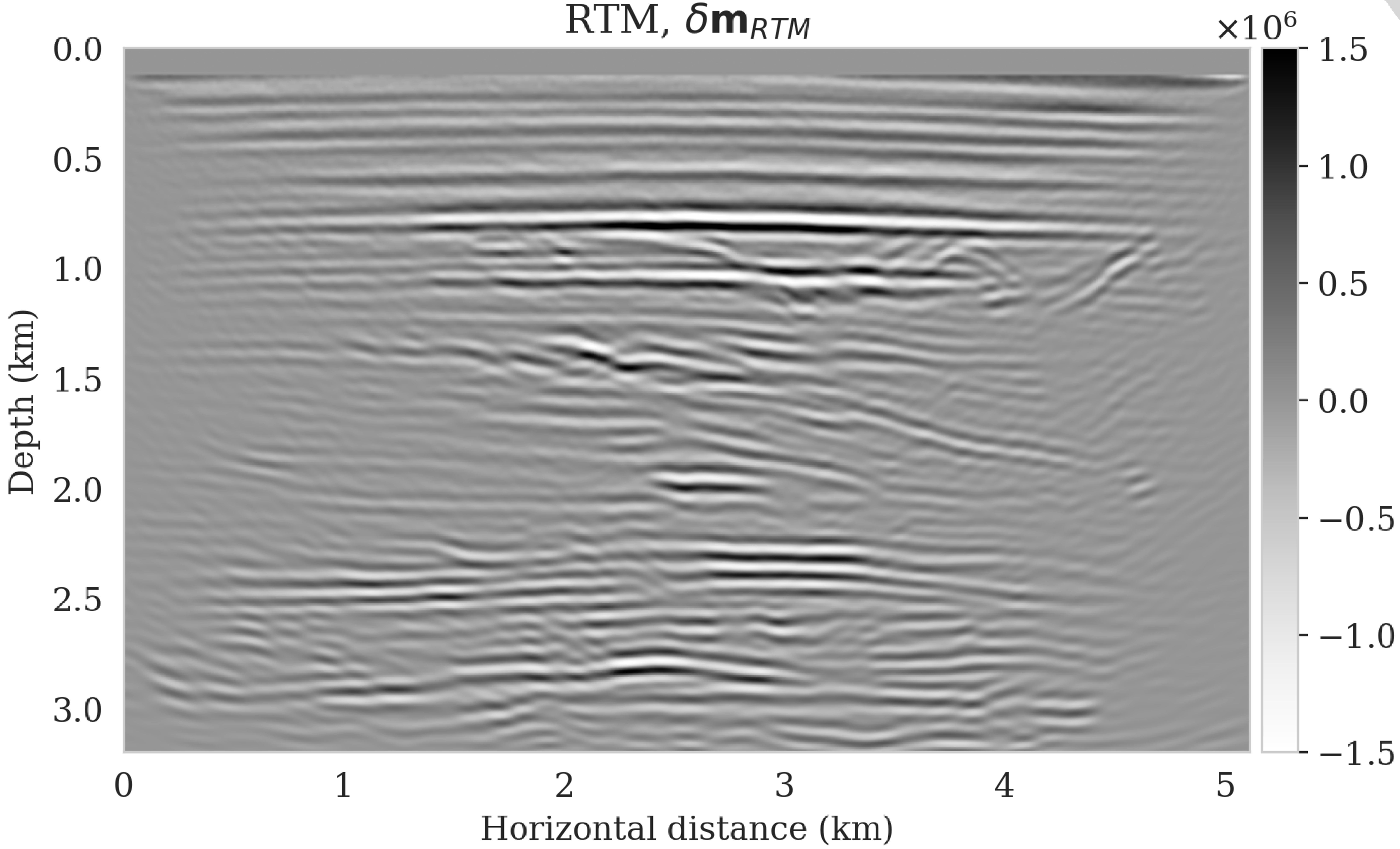
Jakob Kruse et al. "HINT: Hierarchical Invertible Neural Transport for Density Estimation and Bayesian Inference". In: *arXiv preprint arXiv:1905.10687* (2019).

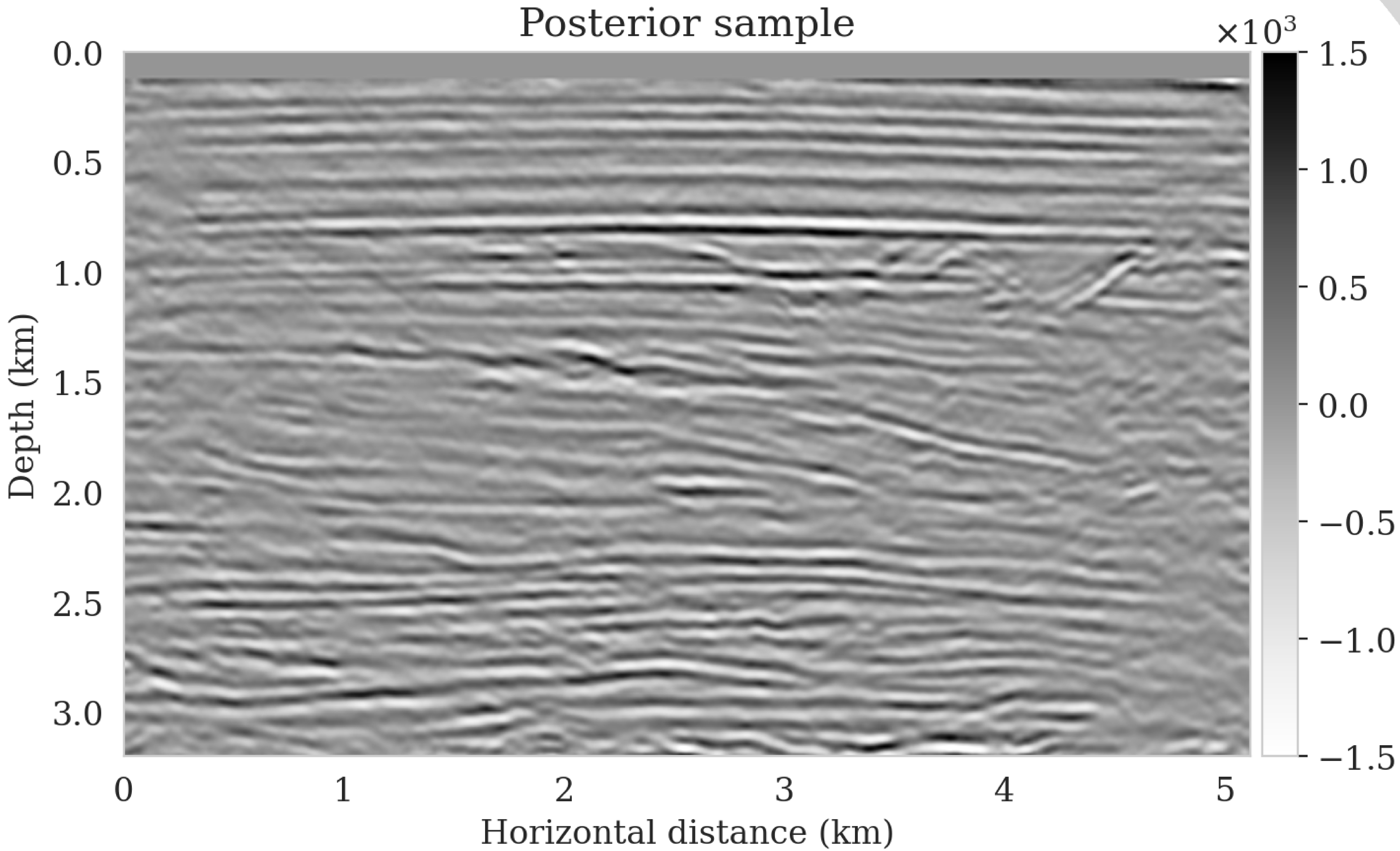
Seismic imaging example 1

True model, $\delta\mathbf{m}$

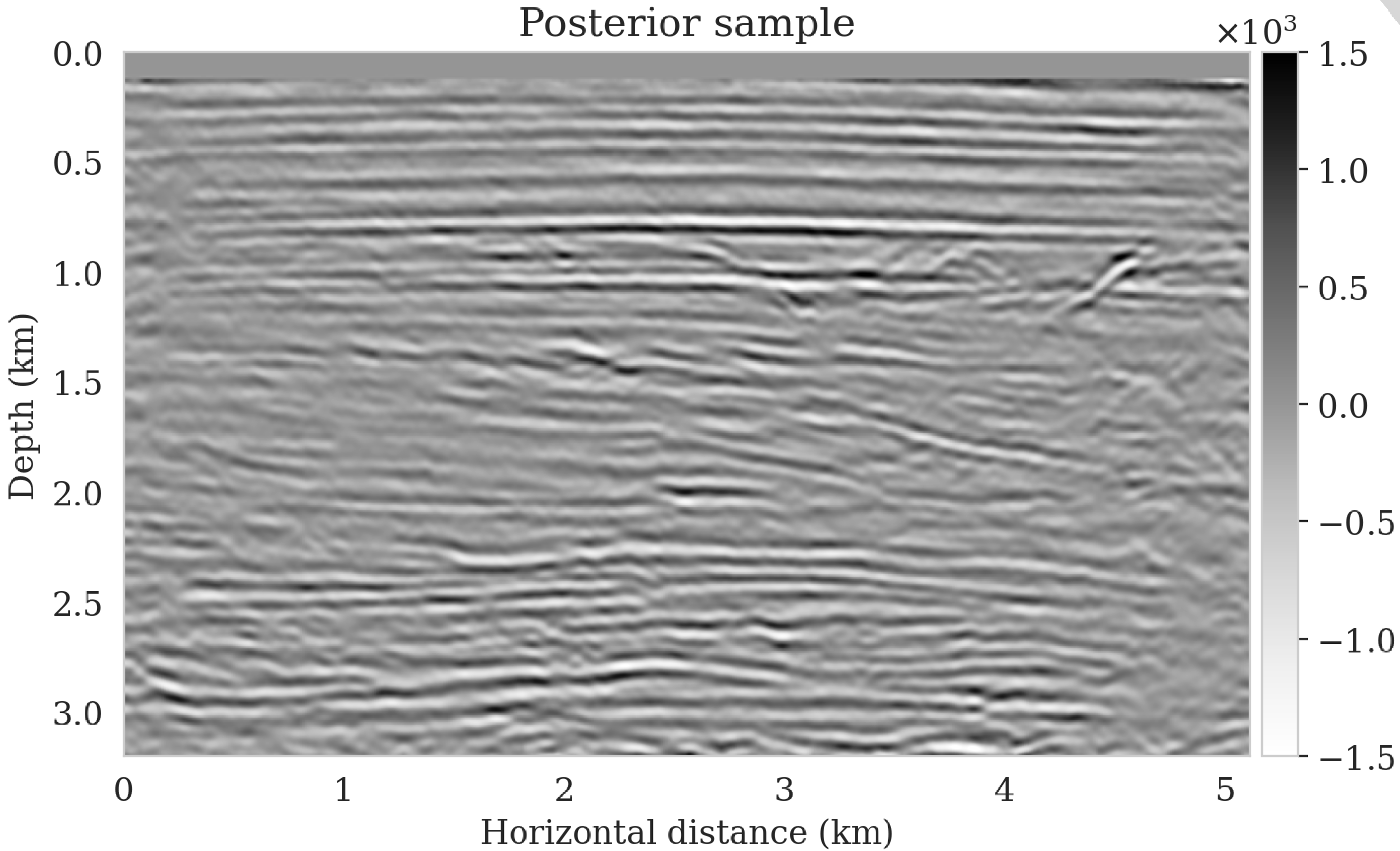


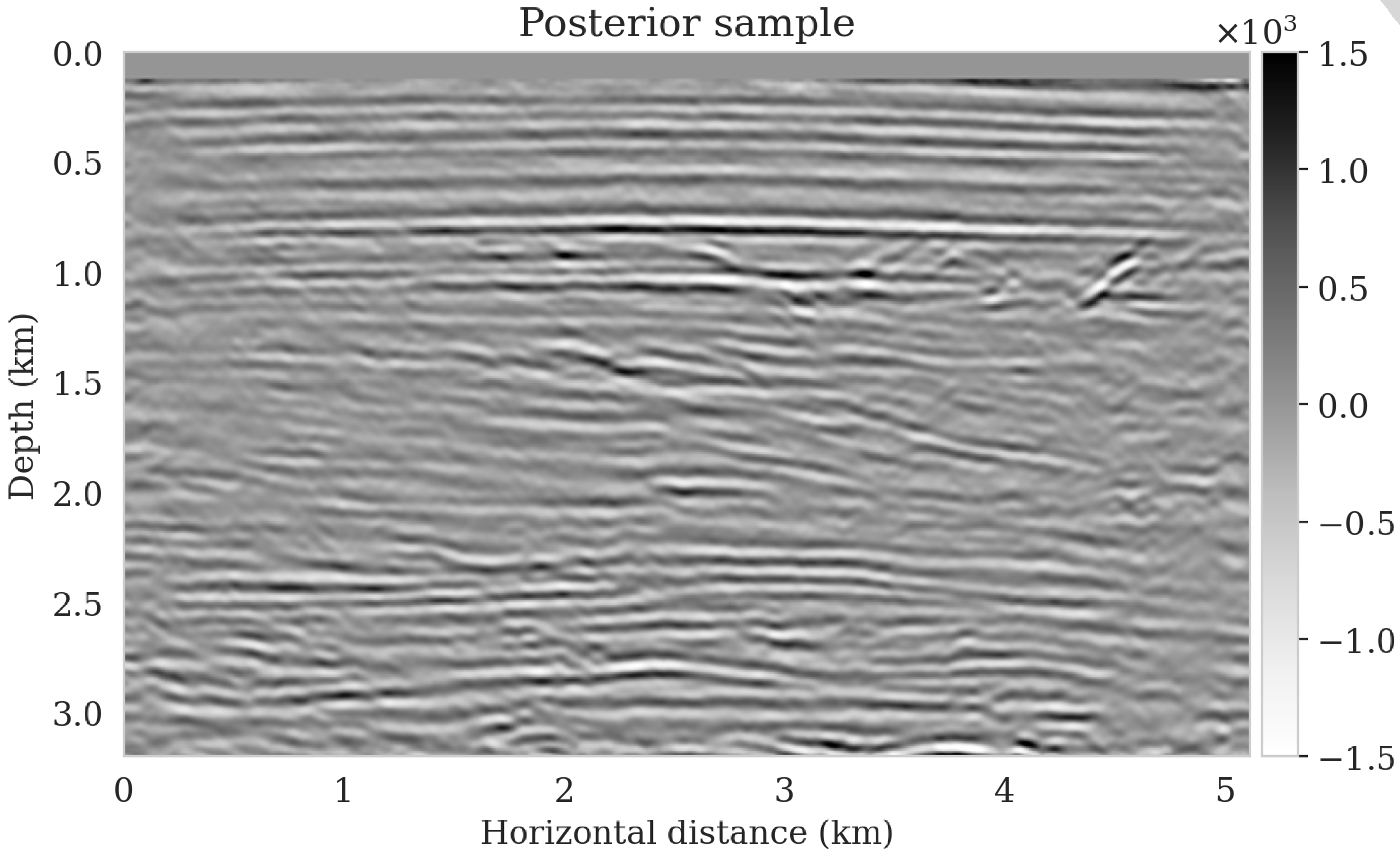
RTM, $\delta \mathbf{m}_{RTM}$





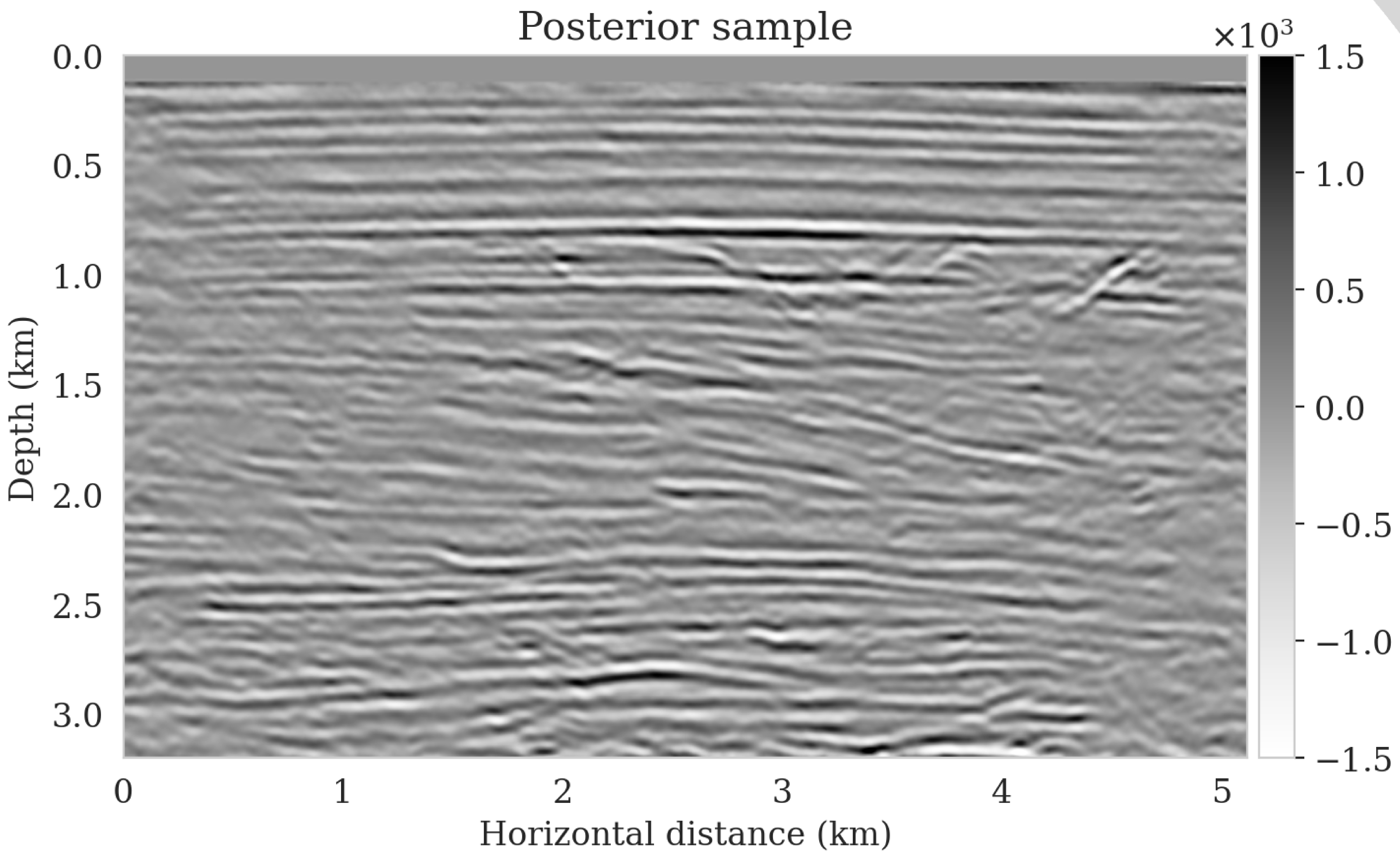
Posterior sample
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Horizontal distance (km)
0
1
2
3
4
5
×10^3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


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

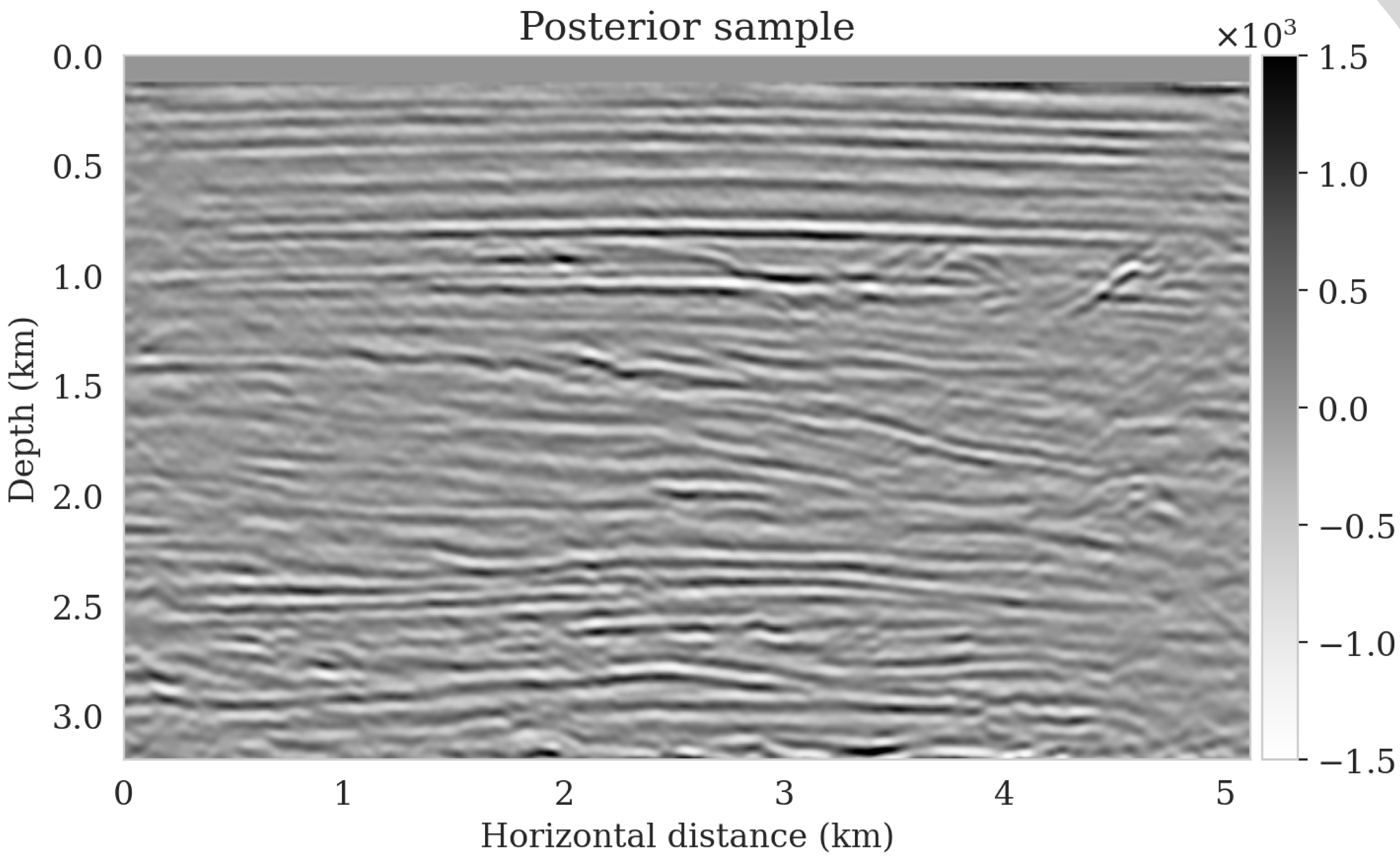


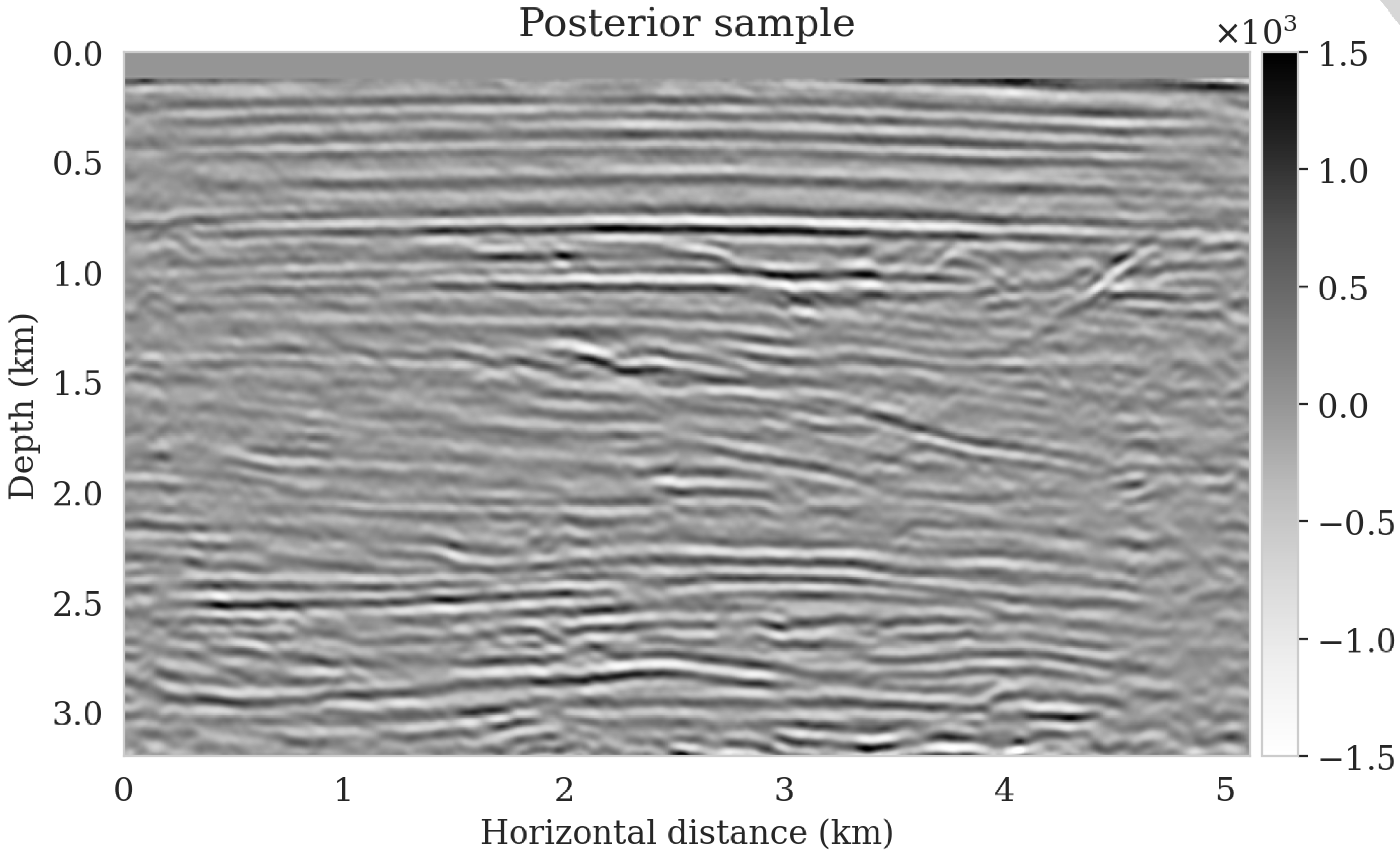
Posterior sample
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

SLIM

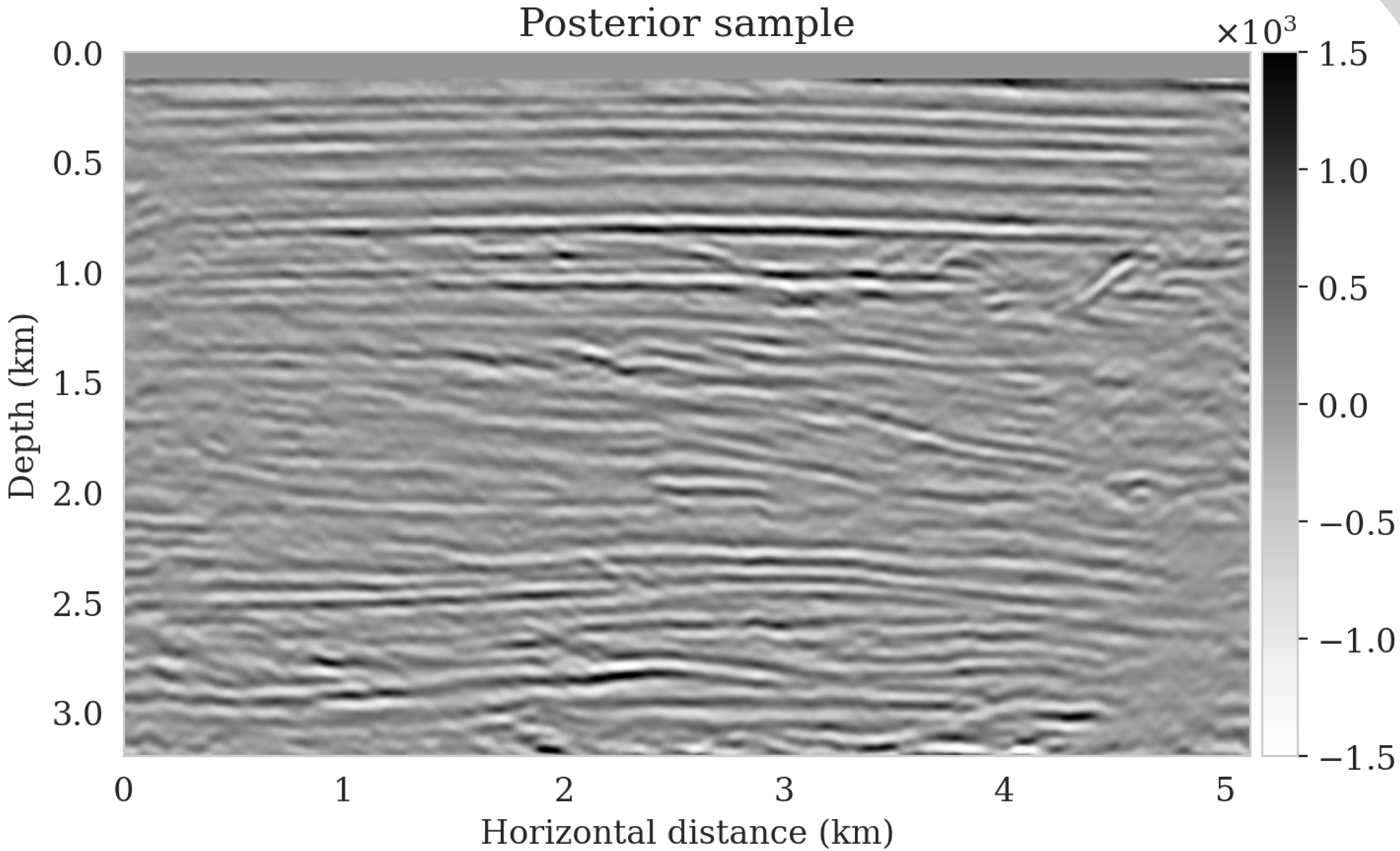


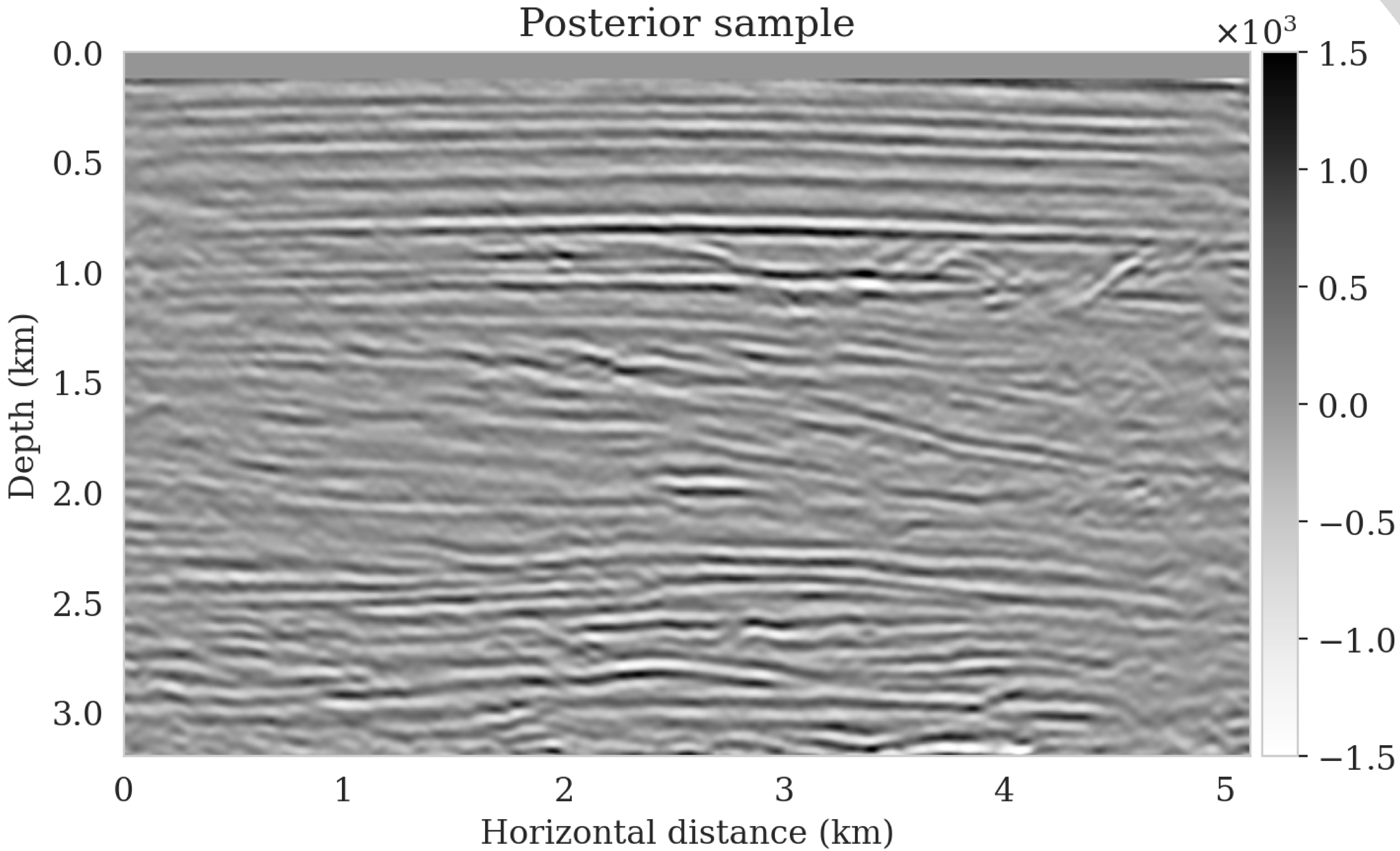
Posterior sample
$\times 10^3$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)


Posterior sample
×10^3
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

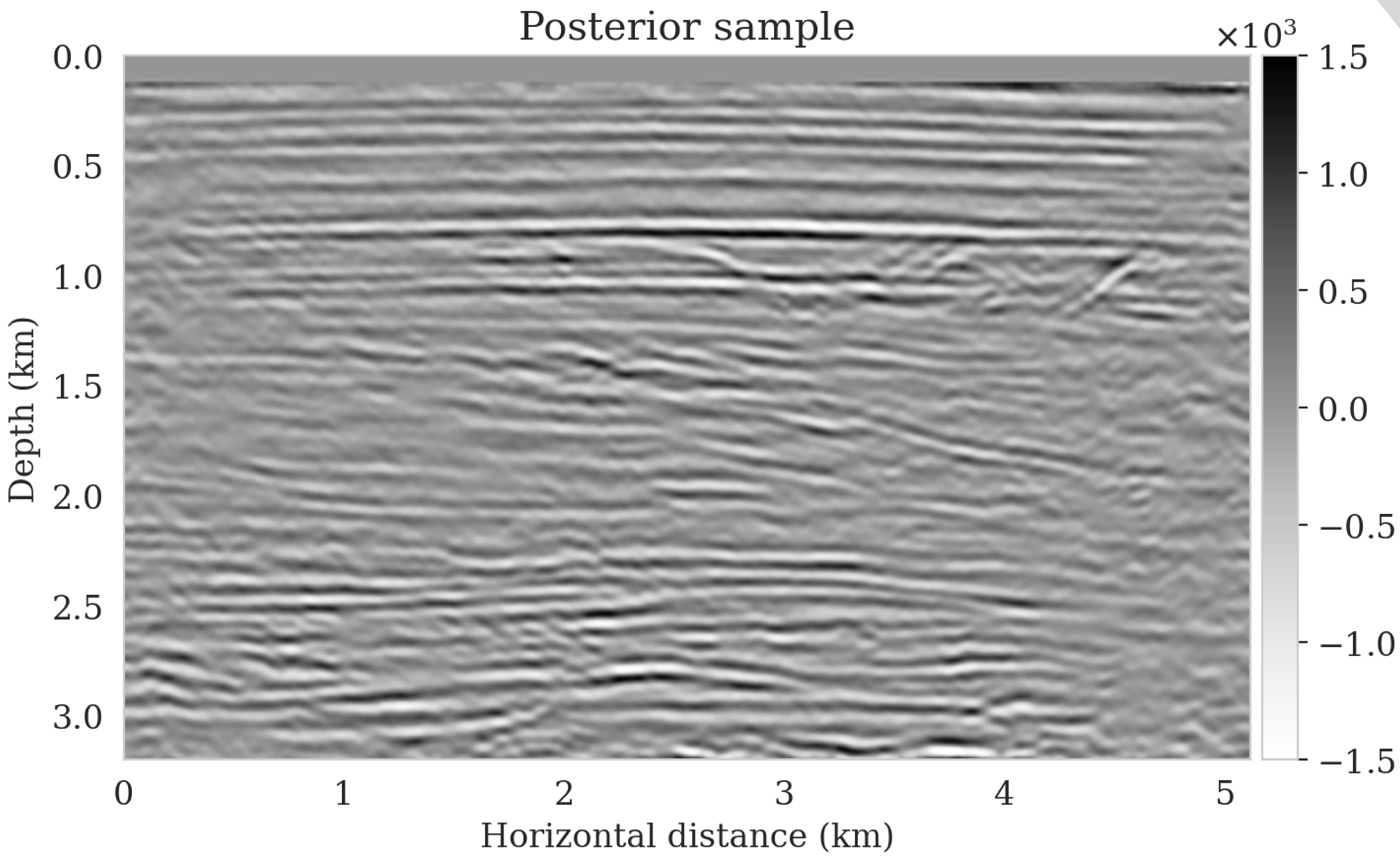

Posterior sample
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Horizontal distance (km)
0
1
2
3
4
5

SLIM

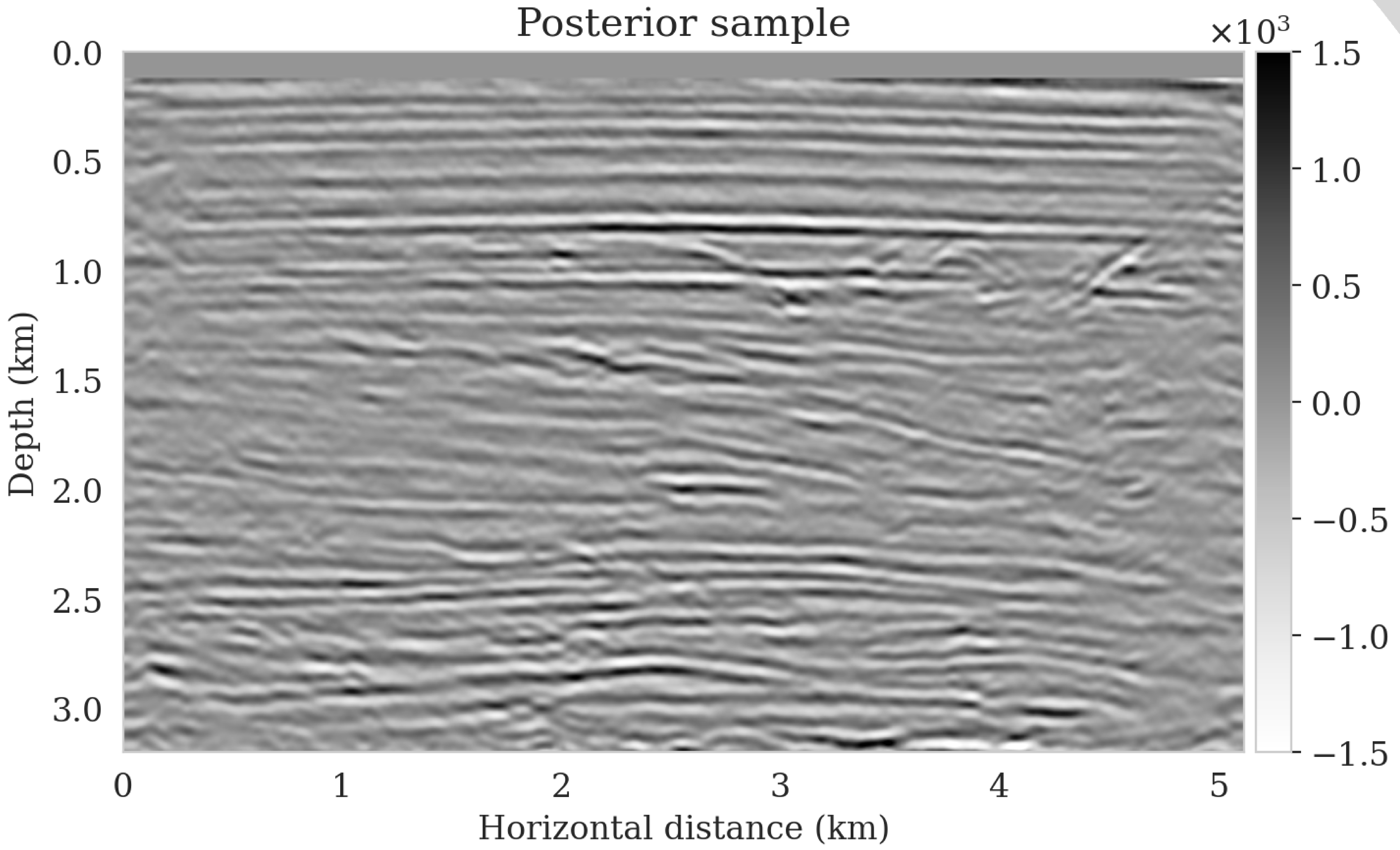

Posterior sample
Depth (km)
Horizontal distance (km)
×10^3
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


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)
×10^3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

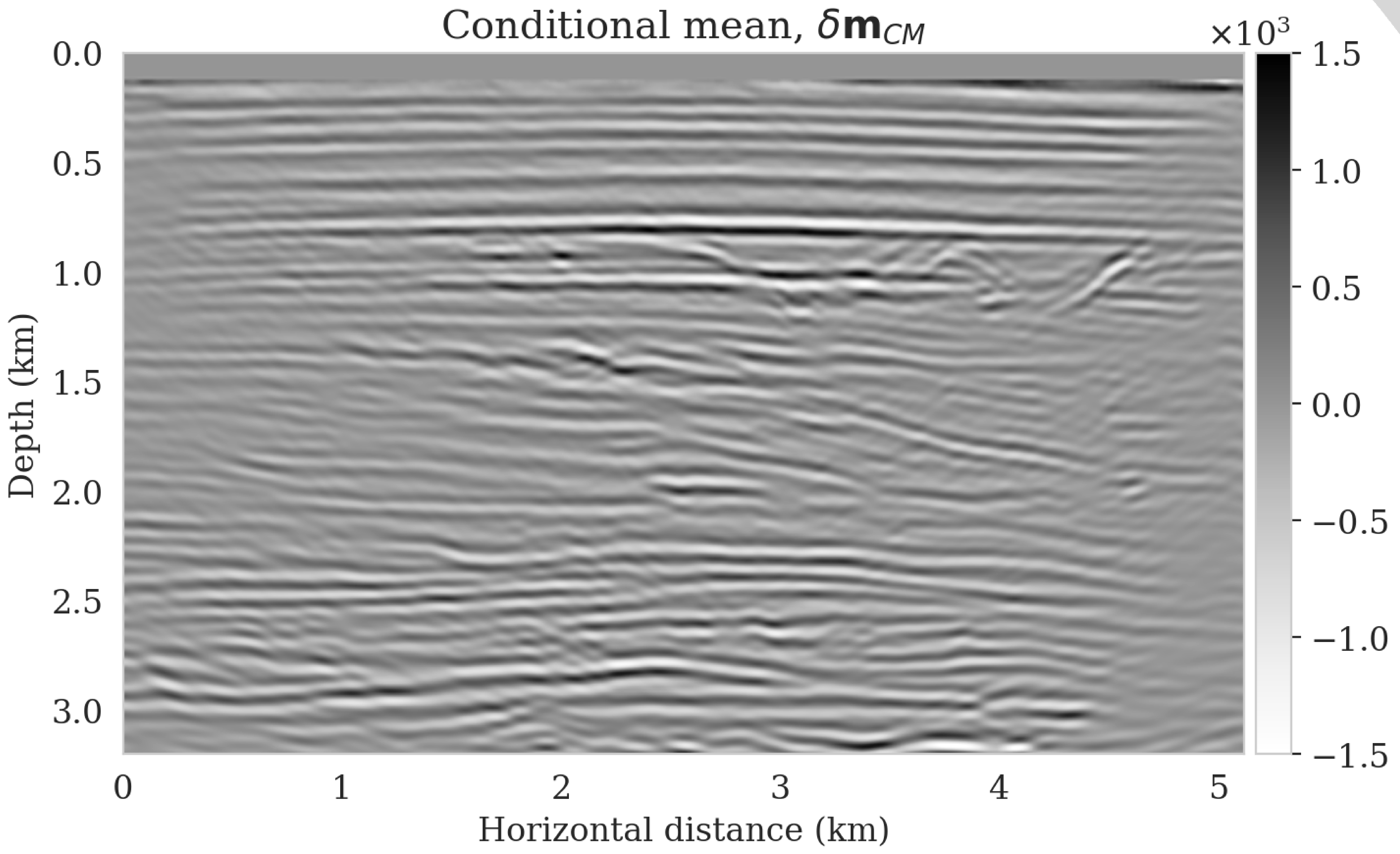
SLIM



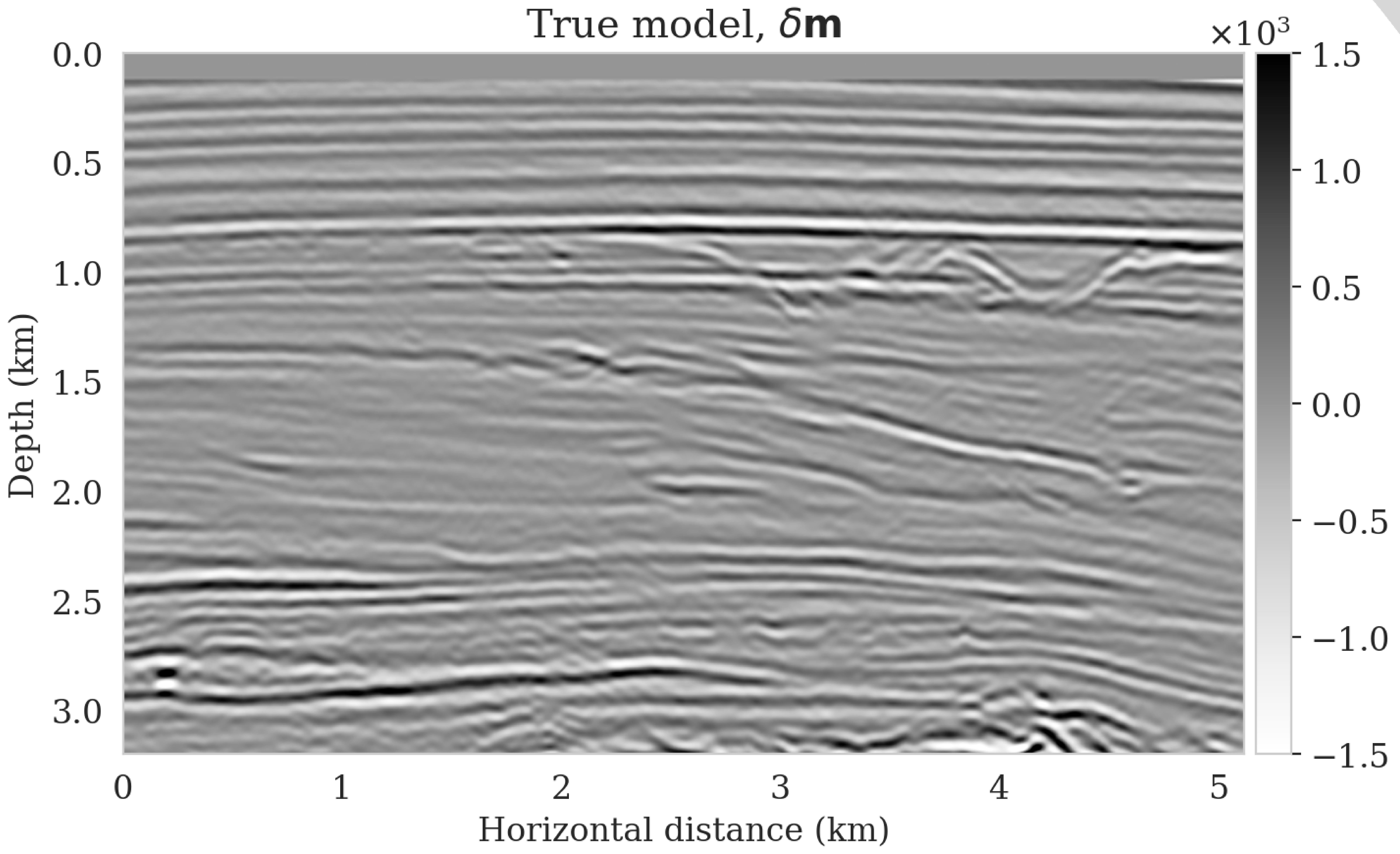
Posterior sample
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)


Posterior sample
×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

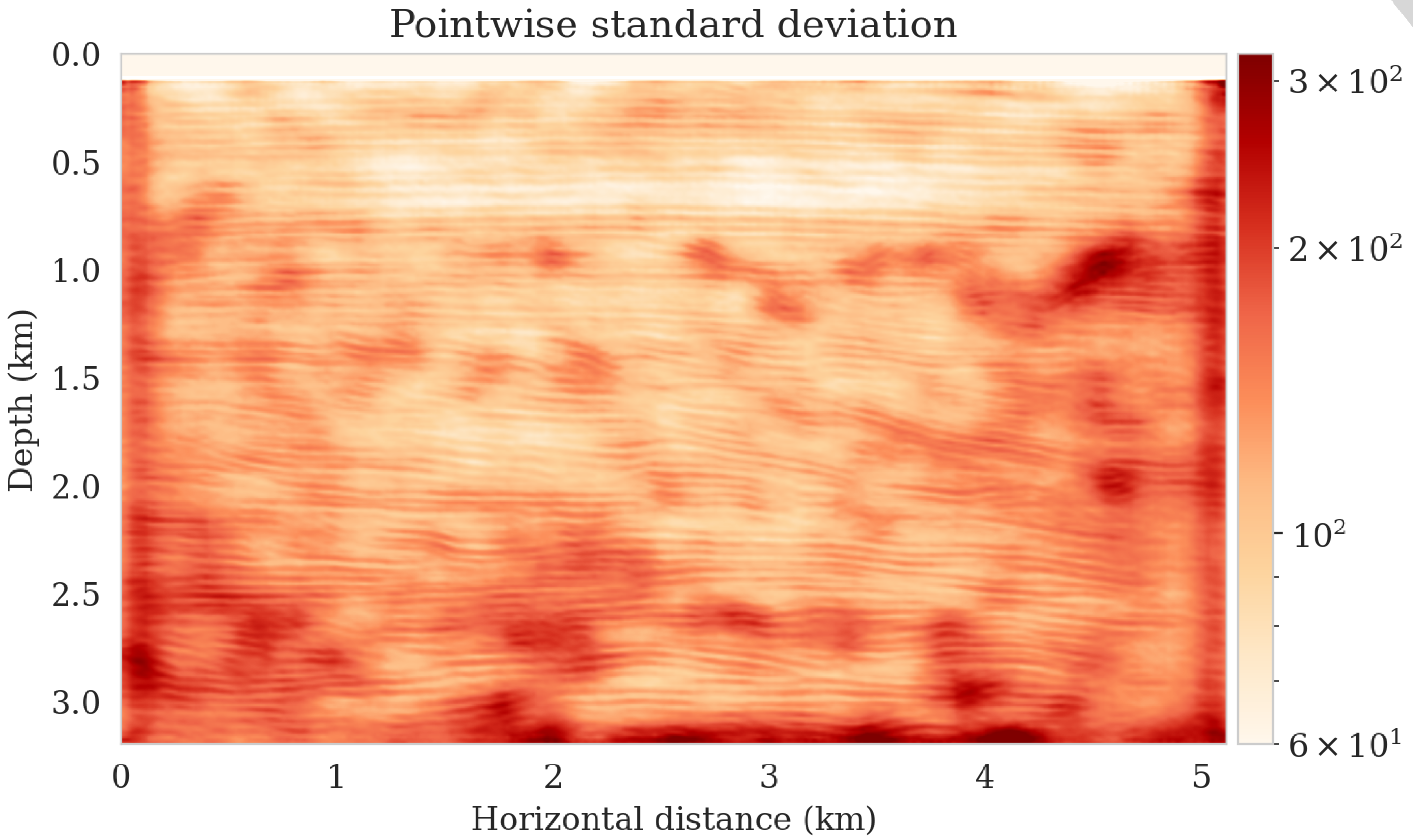

Conditional mean, $\delta\mathbf{m}_{CM}$
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
$\times 10^3$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

SLIM



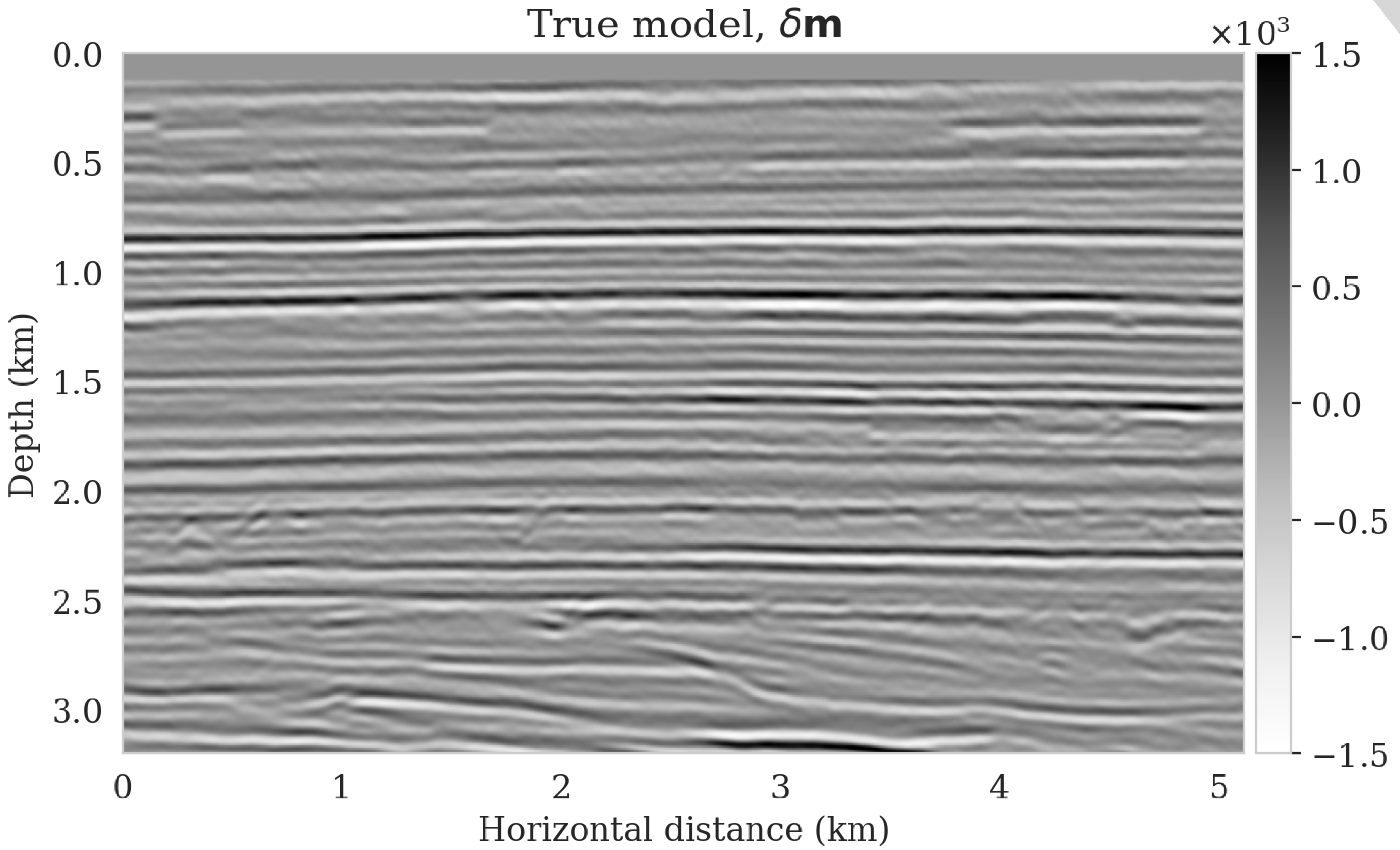
True model, $\delta \mathbf{m}$
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
$\times 10^3$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

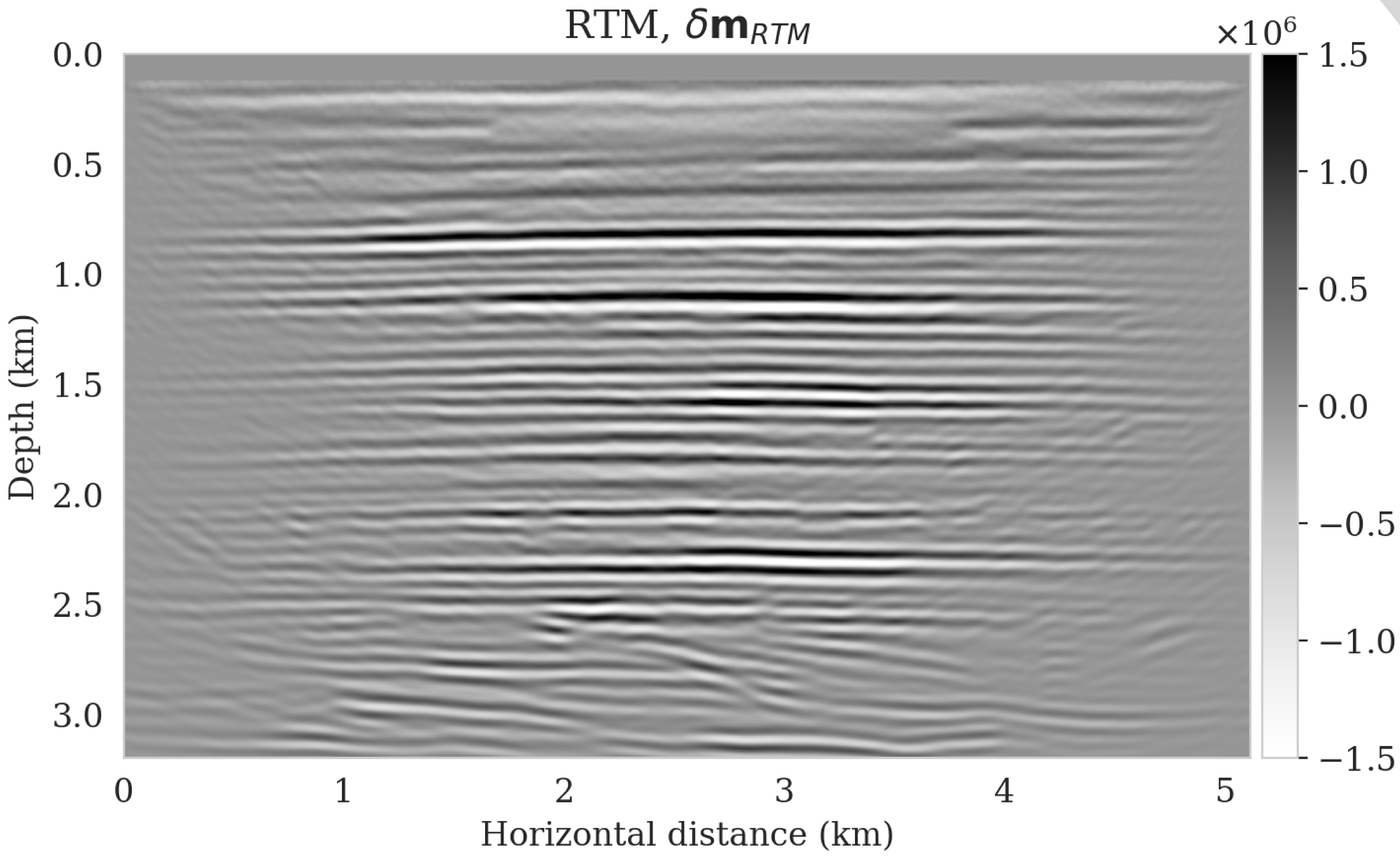
SLIM

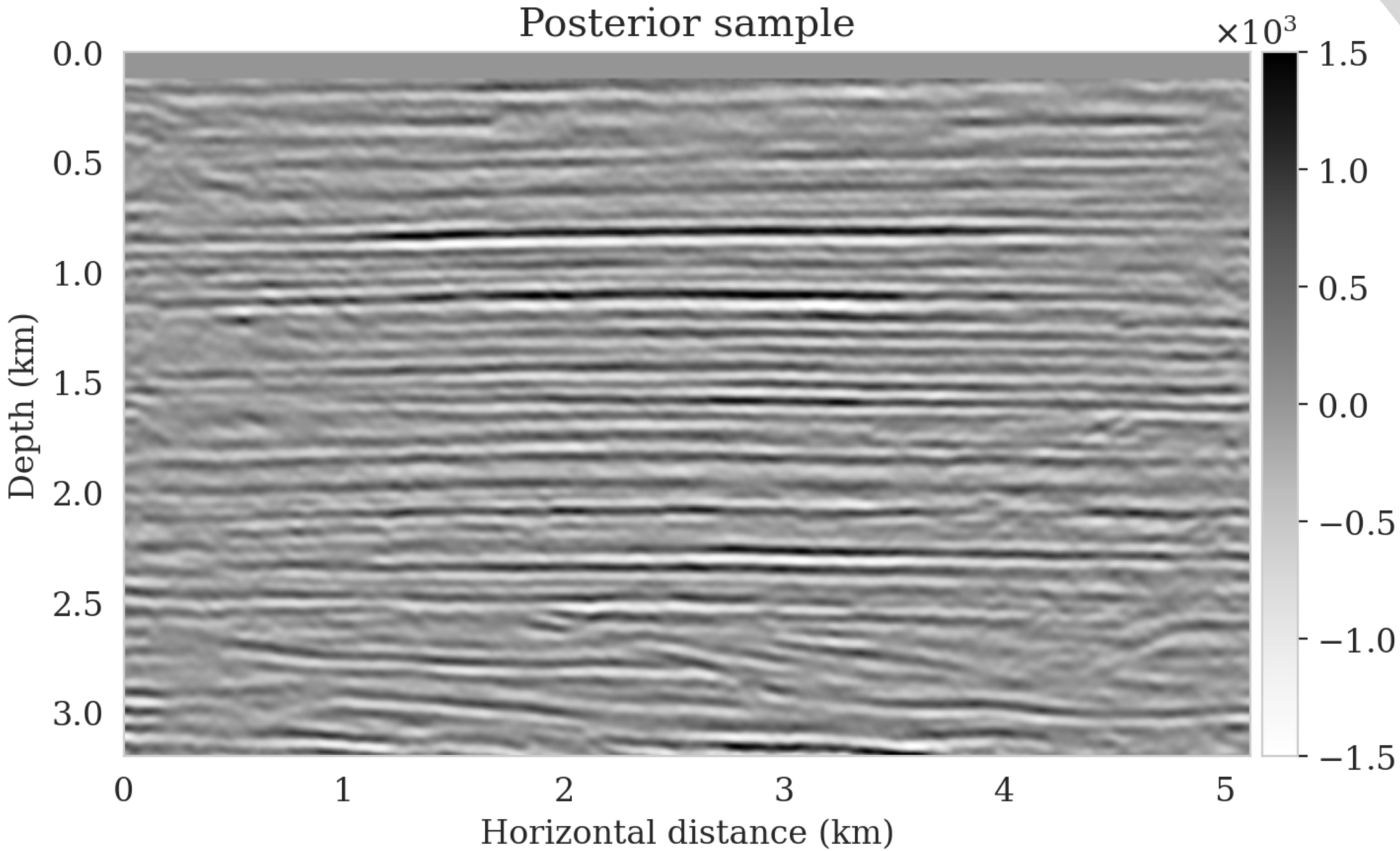


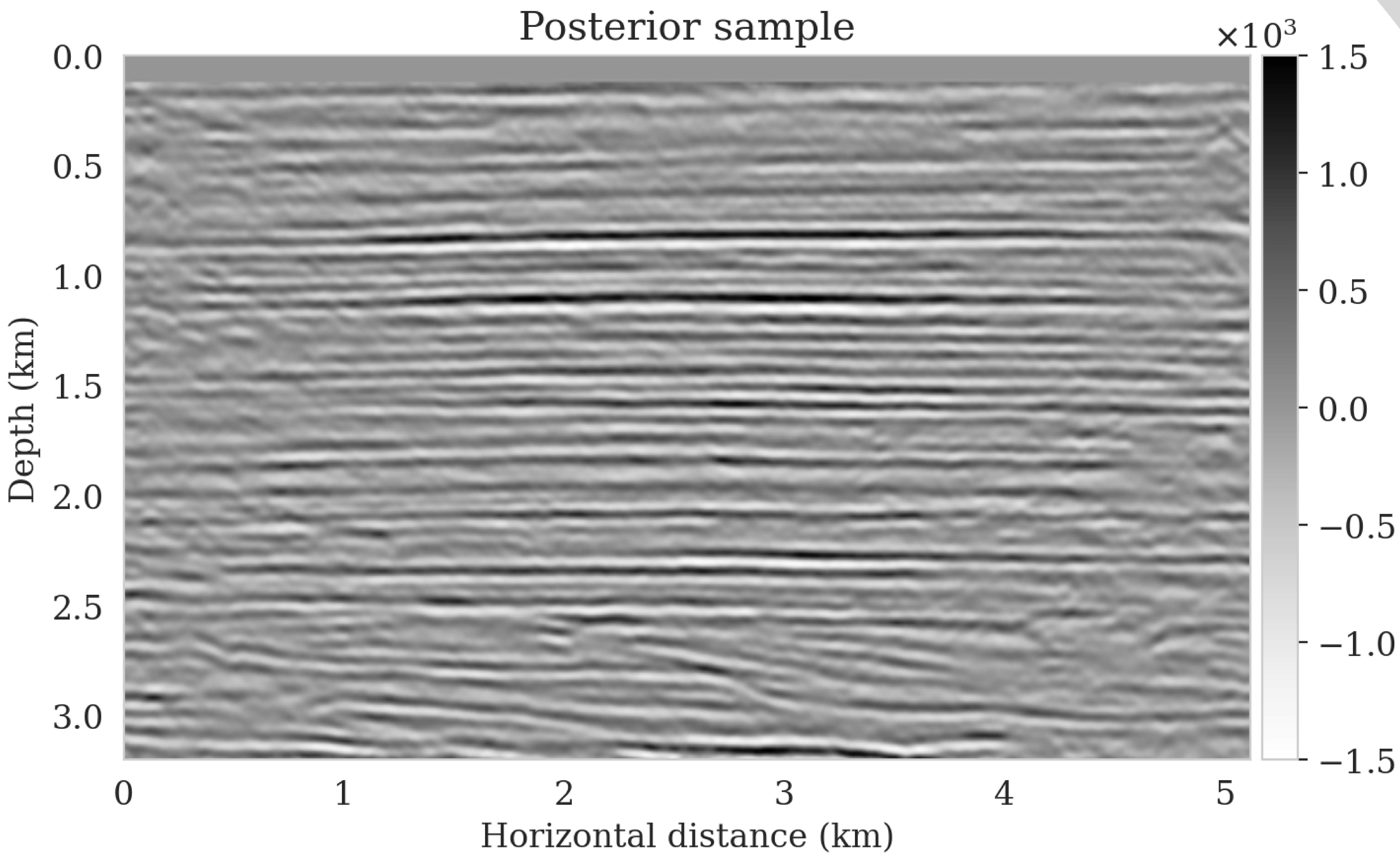
Pointwise standard deviation
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
3×10^2
2×10^2
10^2
6×10^1

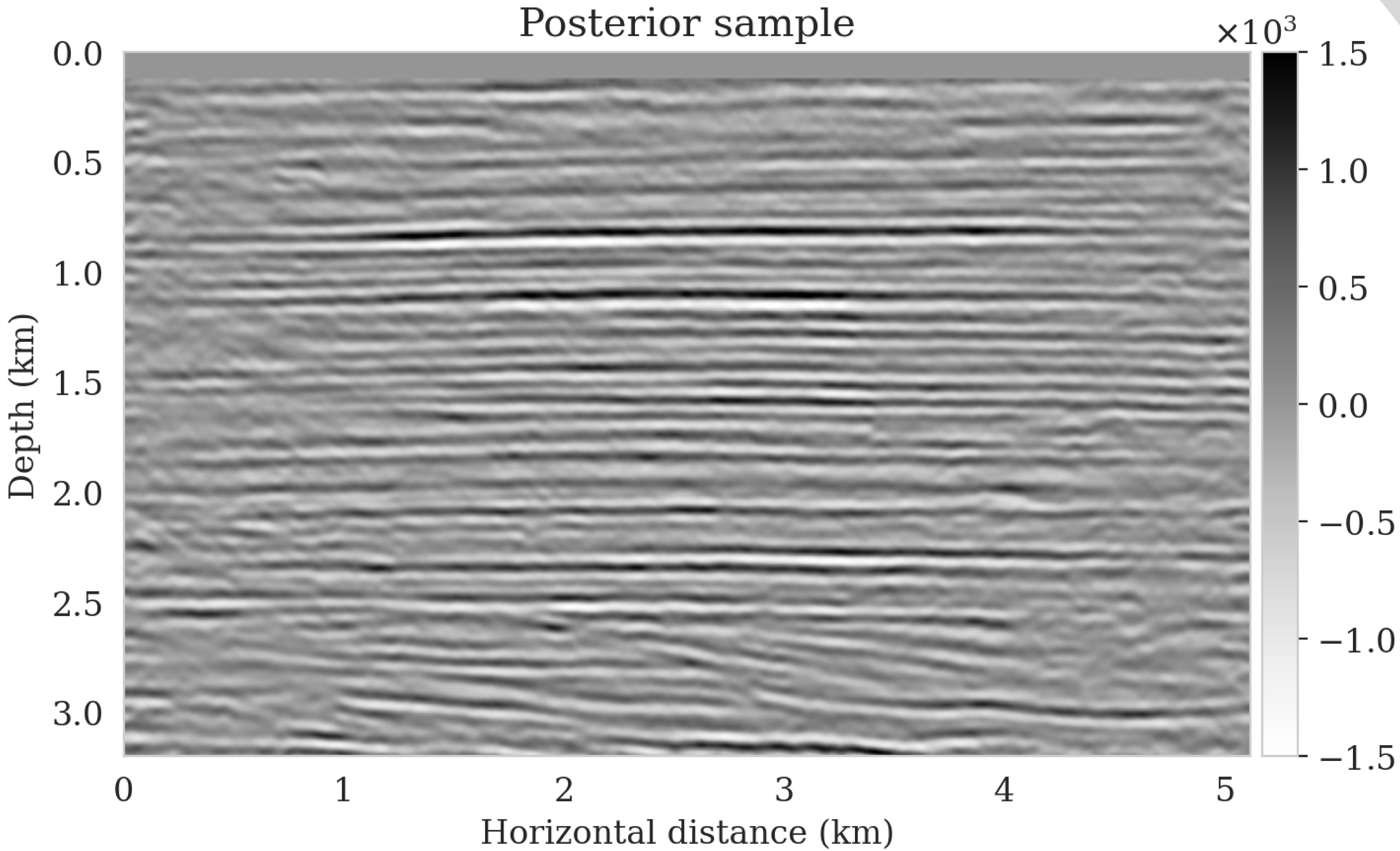
Seismic imaging example 2

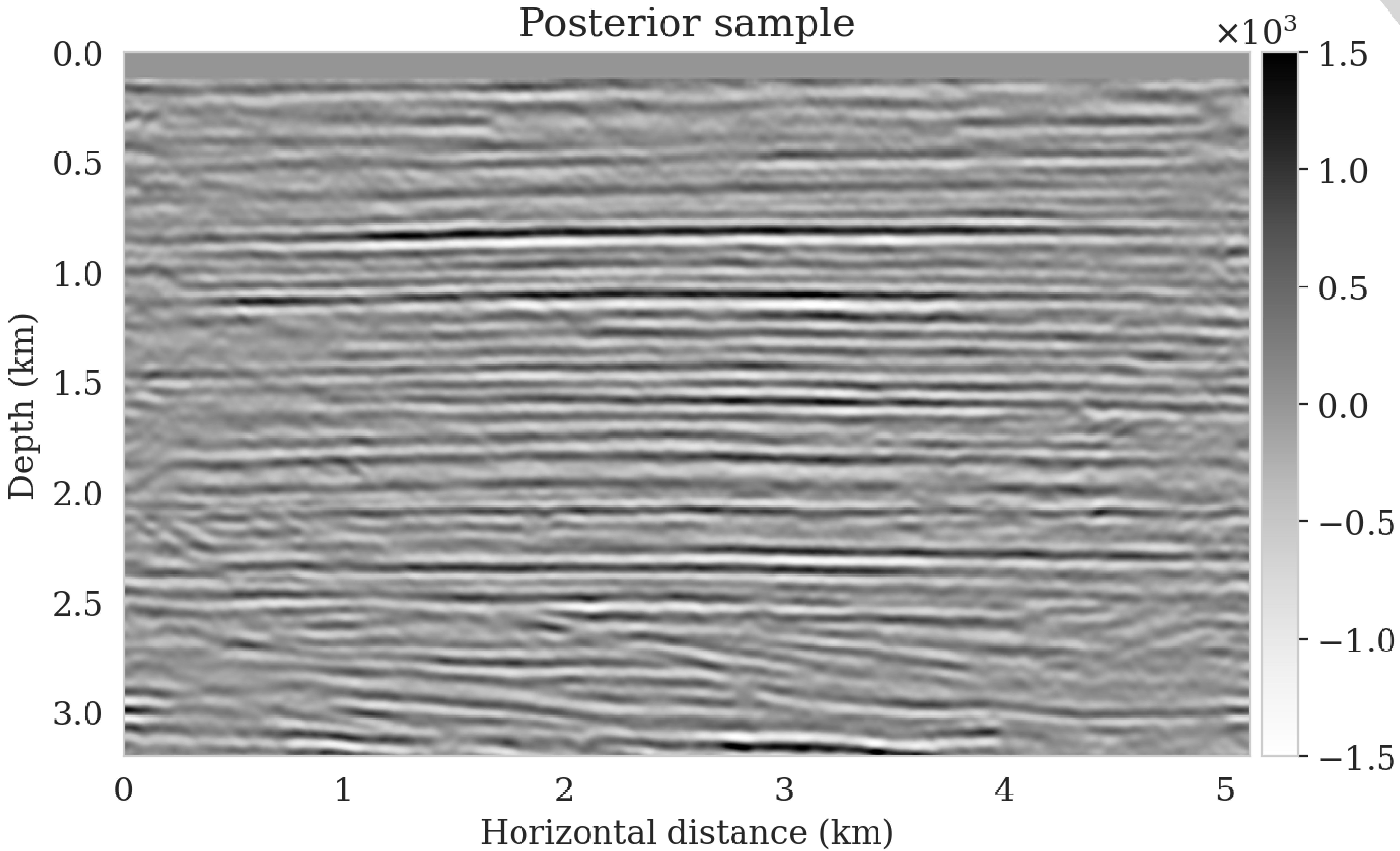

True model, $\delta\mathbf{m}$
$\times 10^3$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)

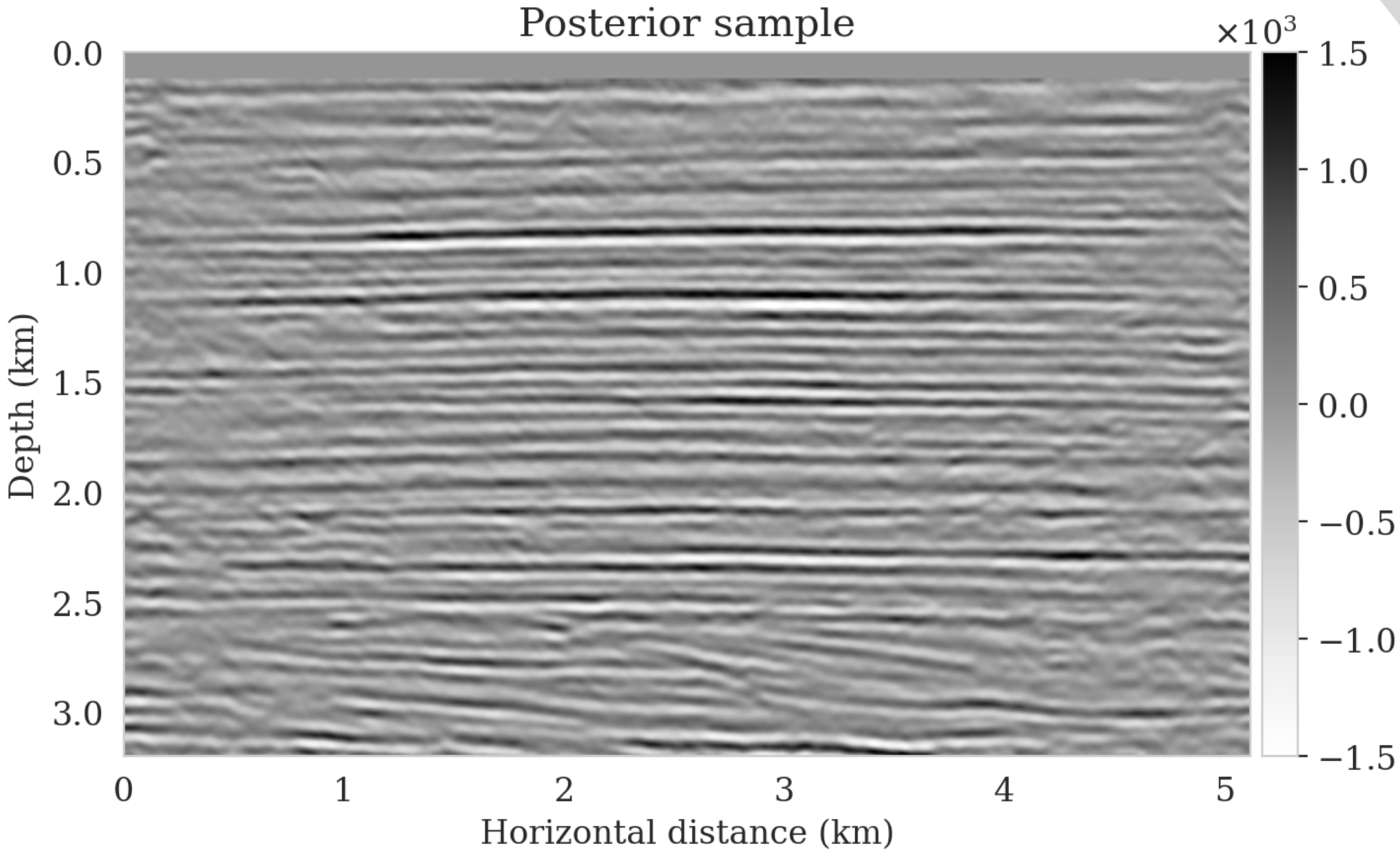

RTM, $\delta\mathbf{m}_{RTM}$
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
$\times 10^6$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

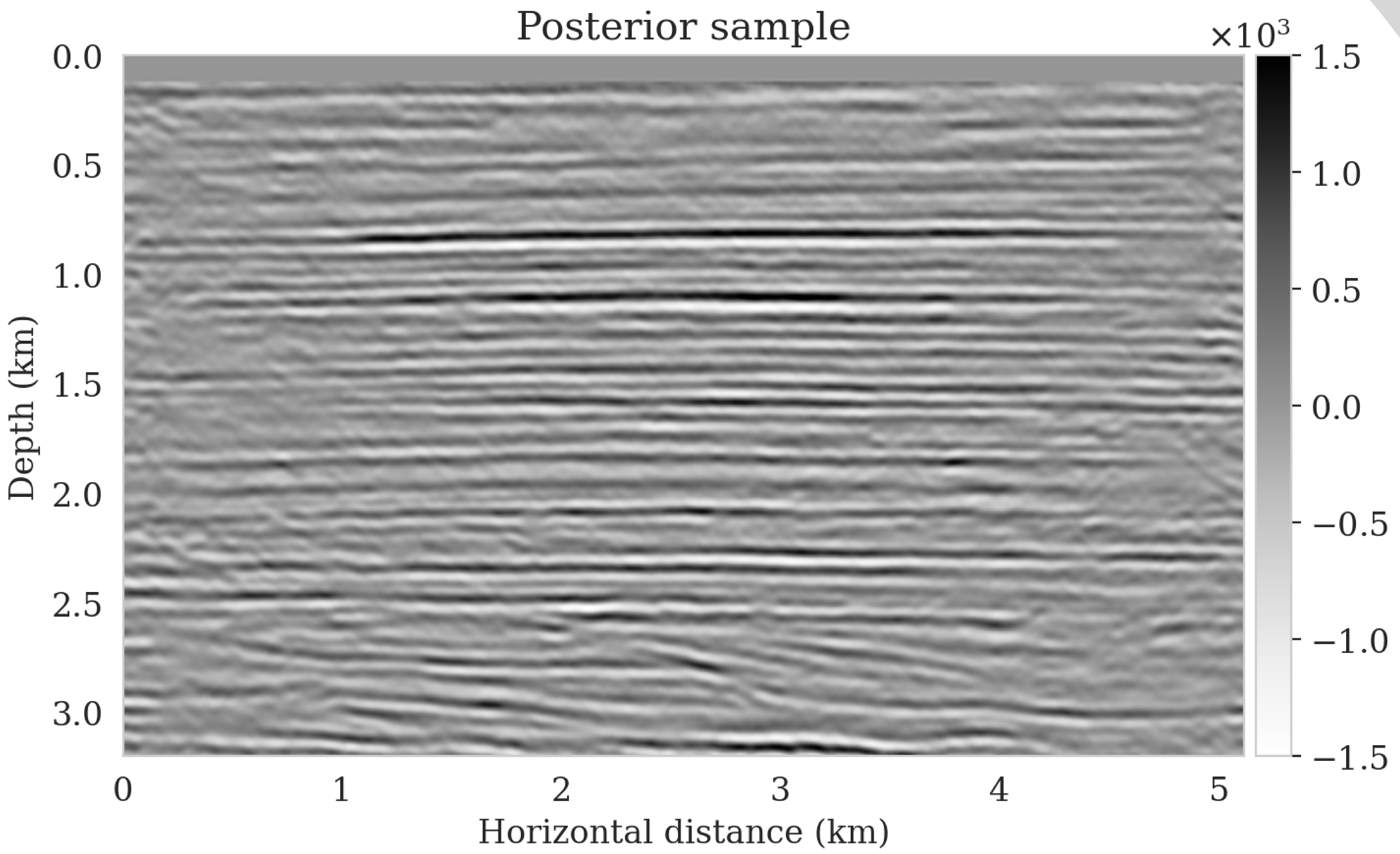

Posterior sample
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)

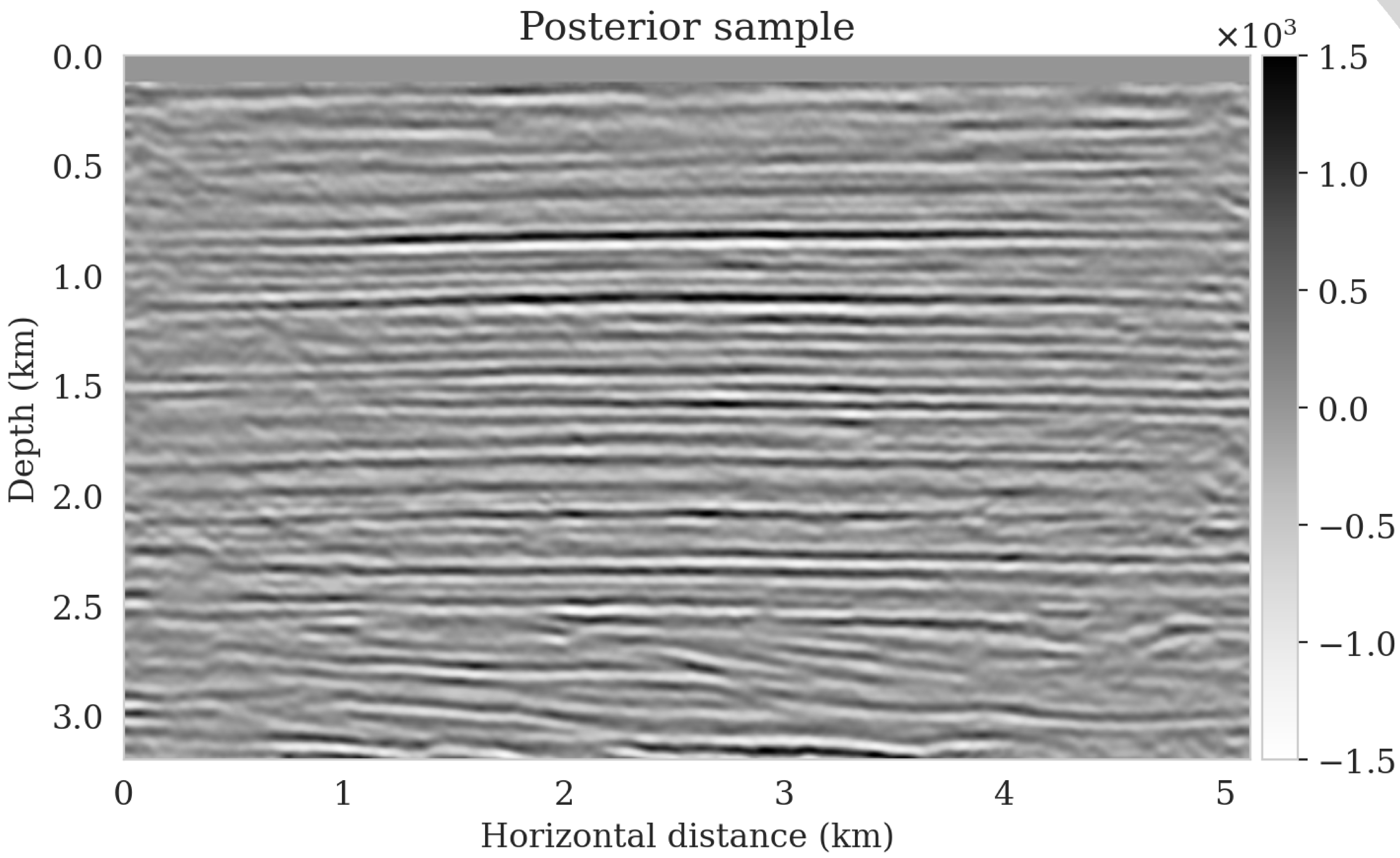

Posterior sample
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
×10^3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


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

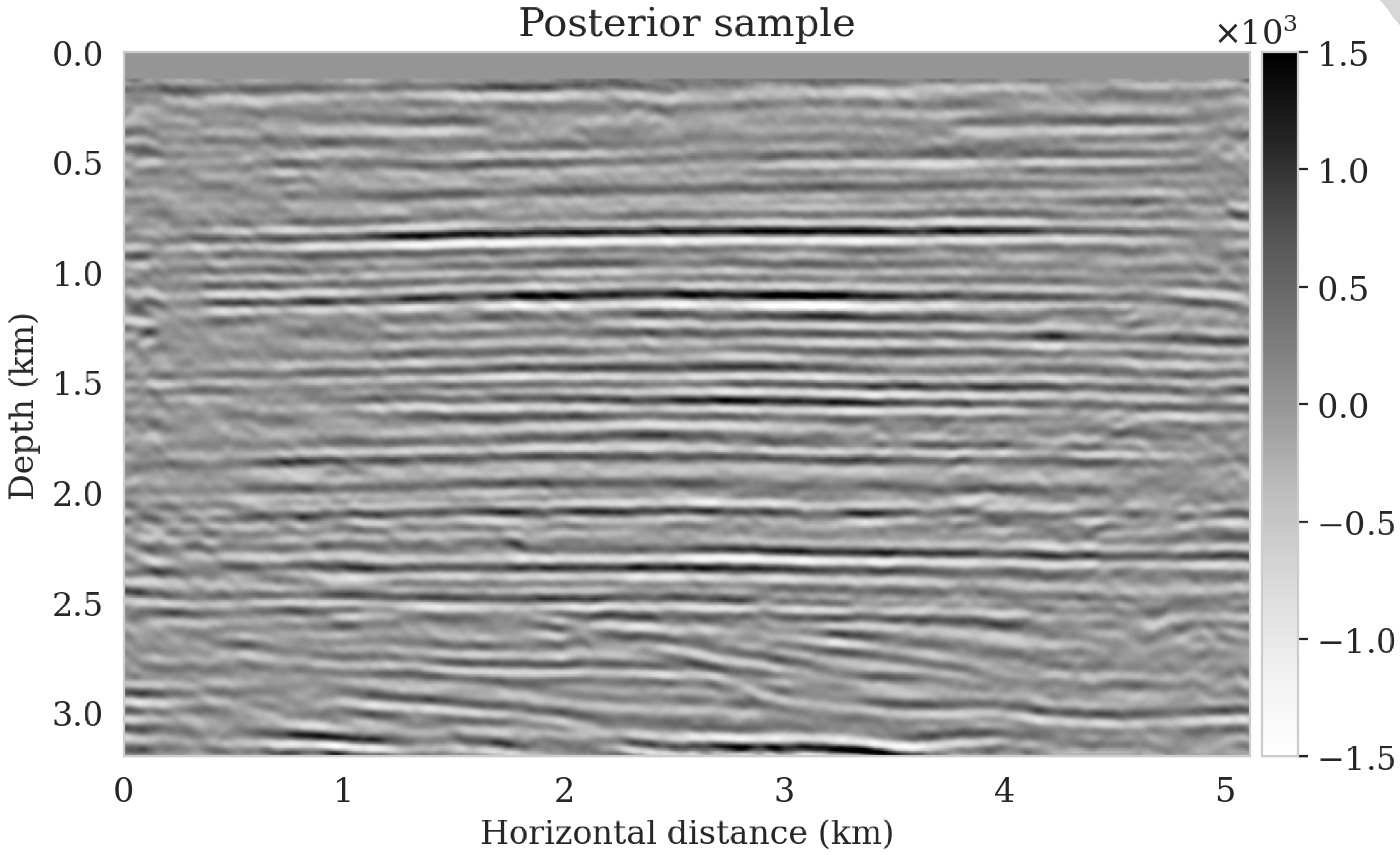

Posterior sample
Depth (km)
Horizontal distance (km)
×10³
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

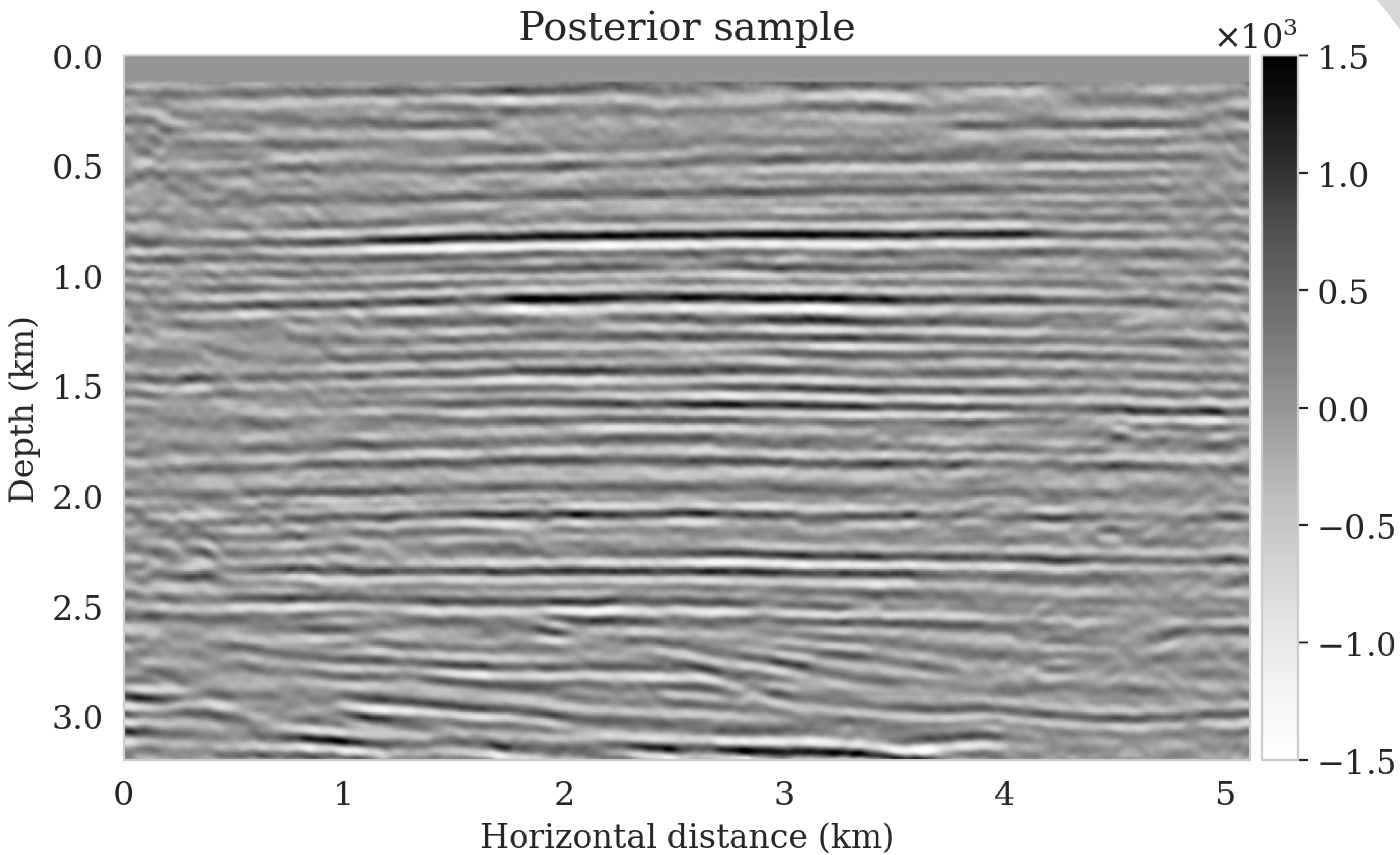

Posterior sample
Depth (km)
Horizontal distance (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
×10^3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


Posterior sample
×10³
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)

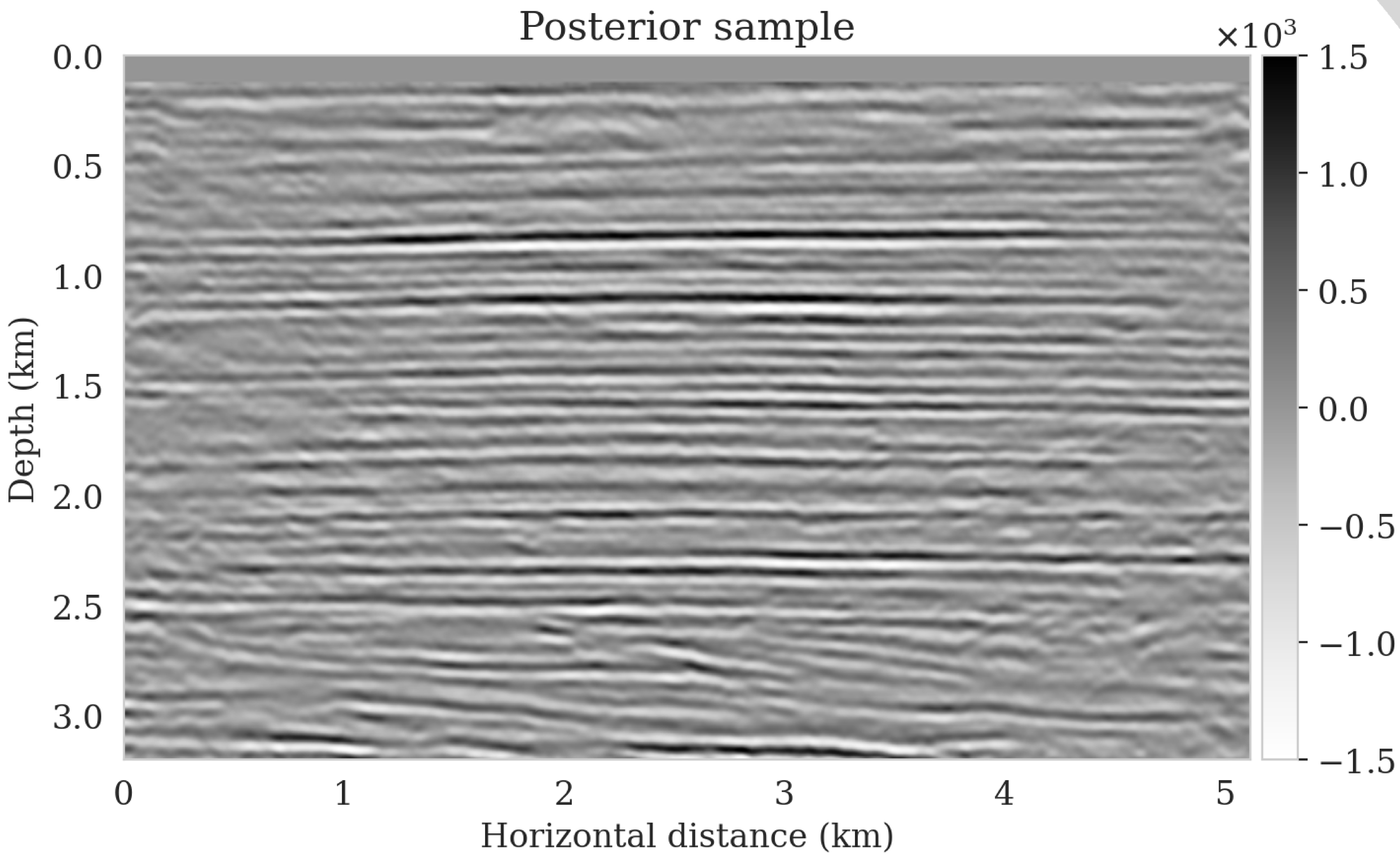


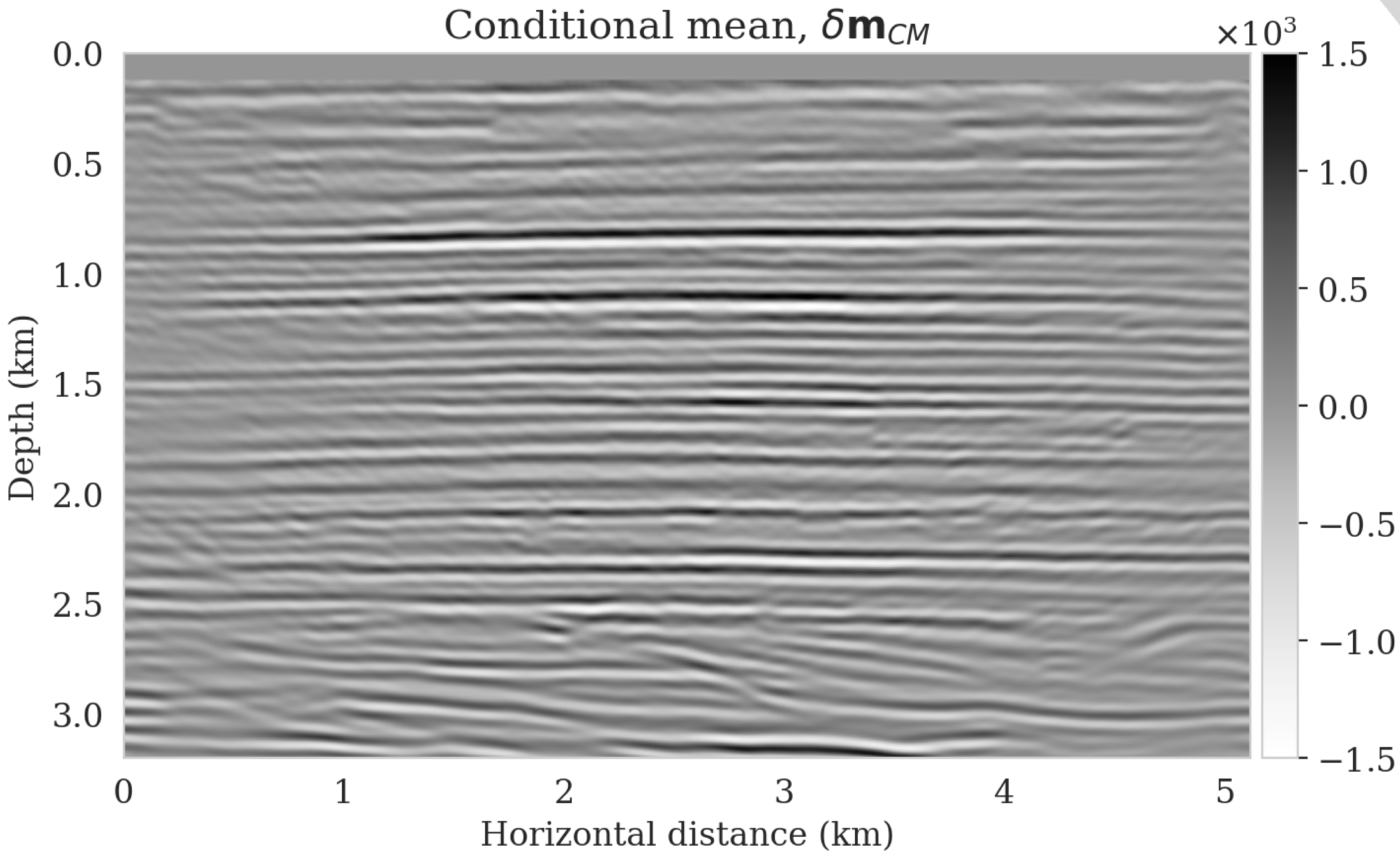
Posterior sample
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)
×10^3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

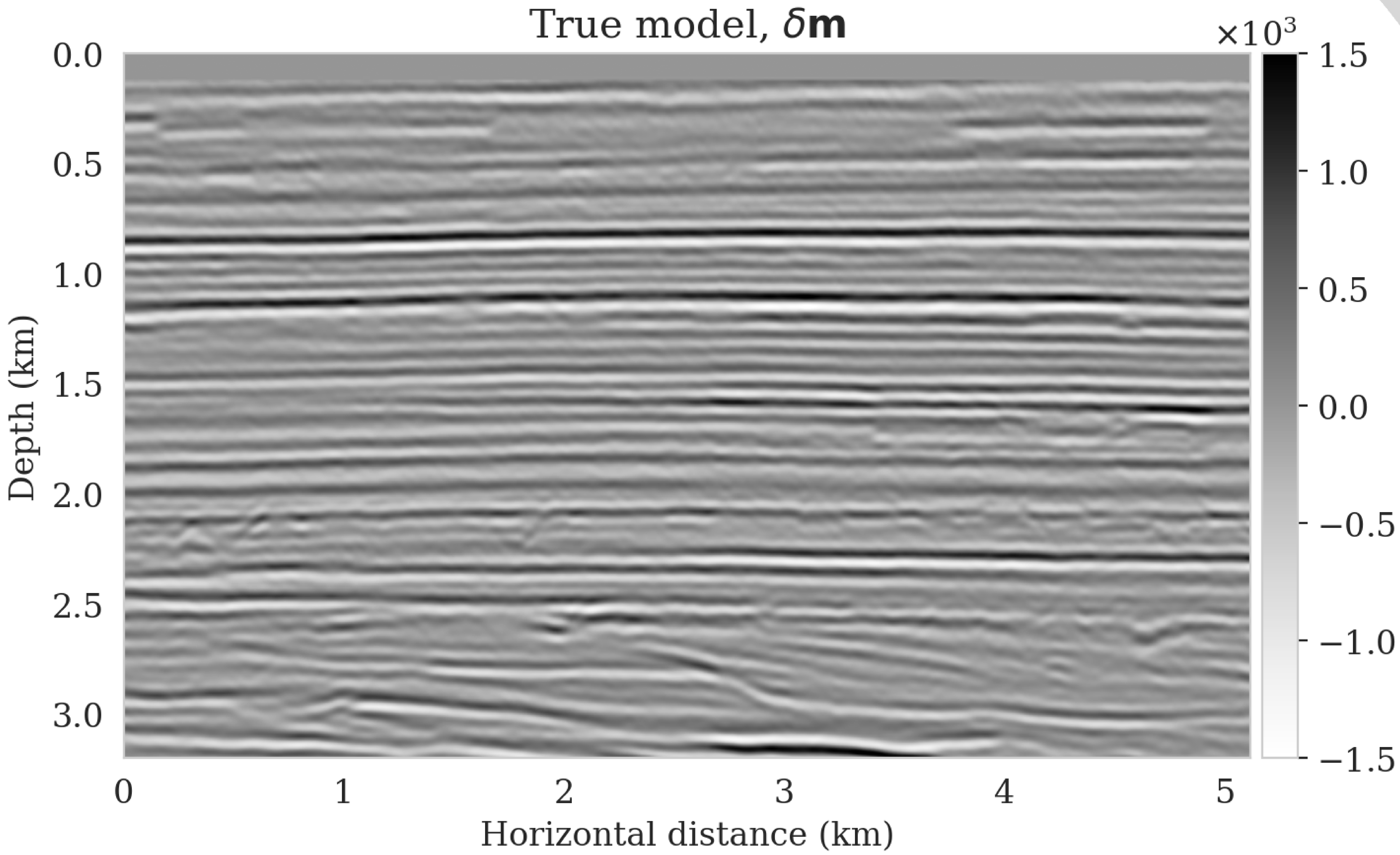

Posterior sample
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)

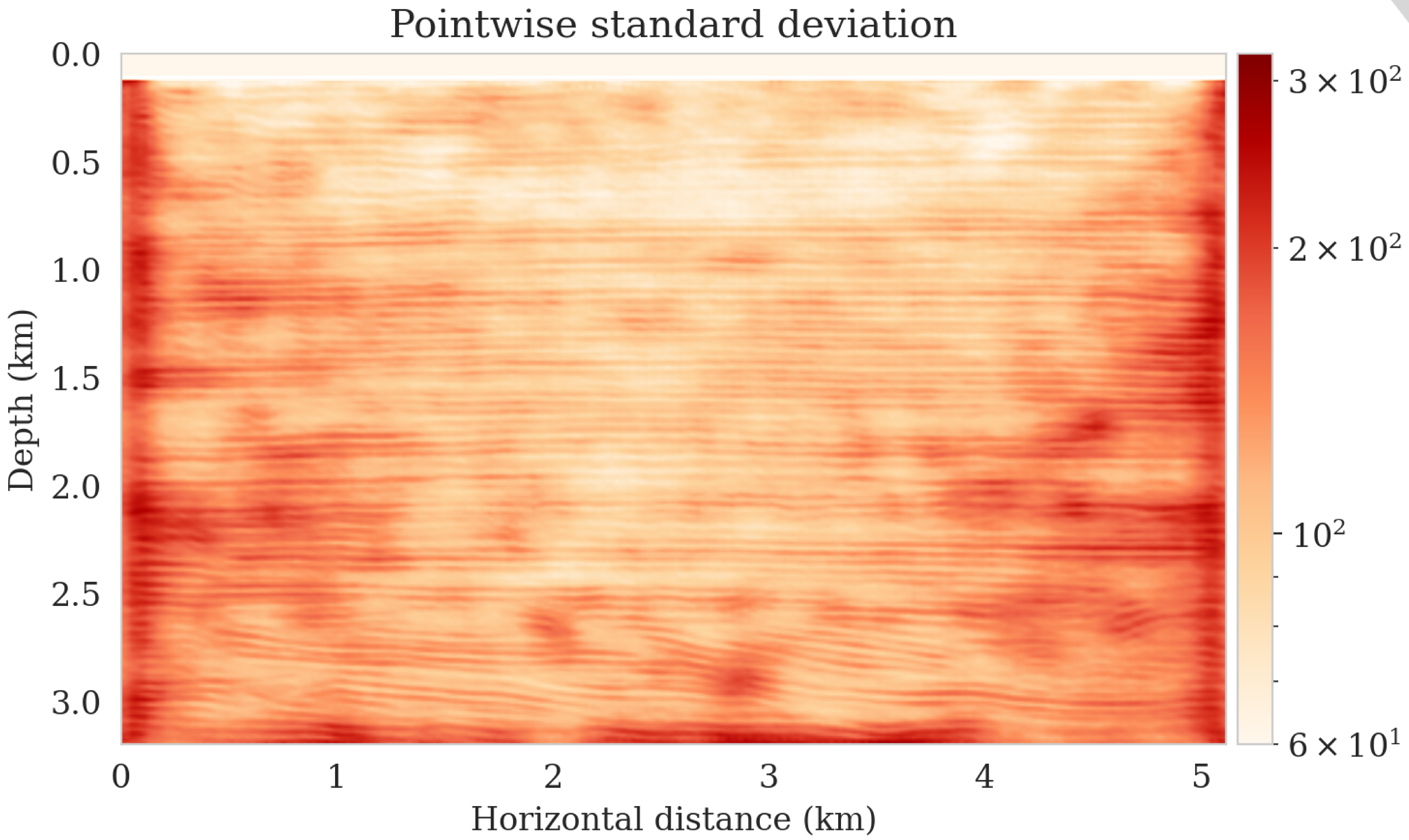


Posterior sample
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

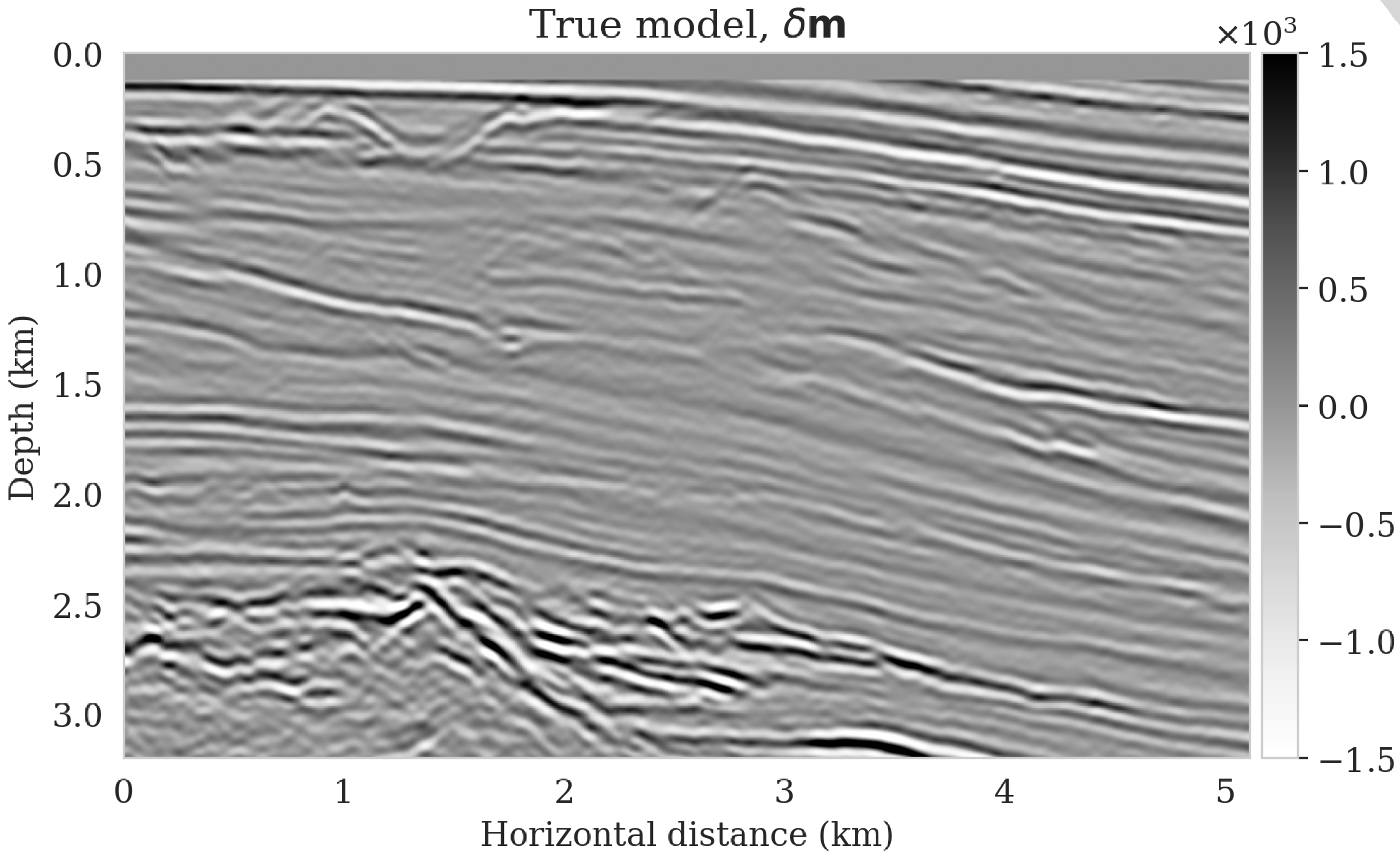

Posterior sample
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

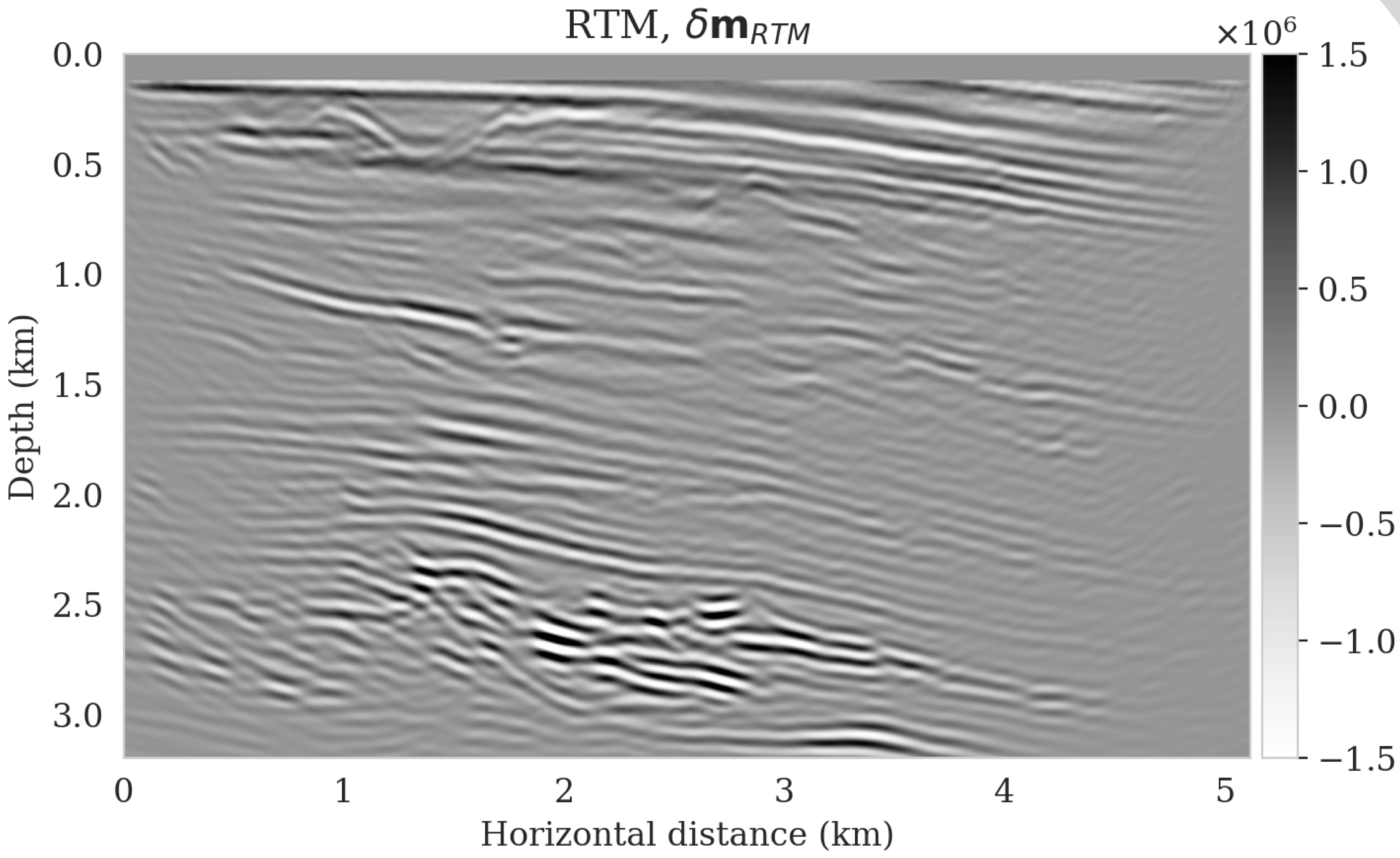

Conditional mean, $\delta \mathbf{m}_{CM}$
$\times 10^3$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)

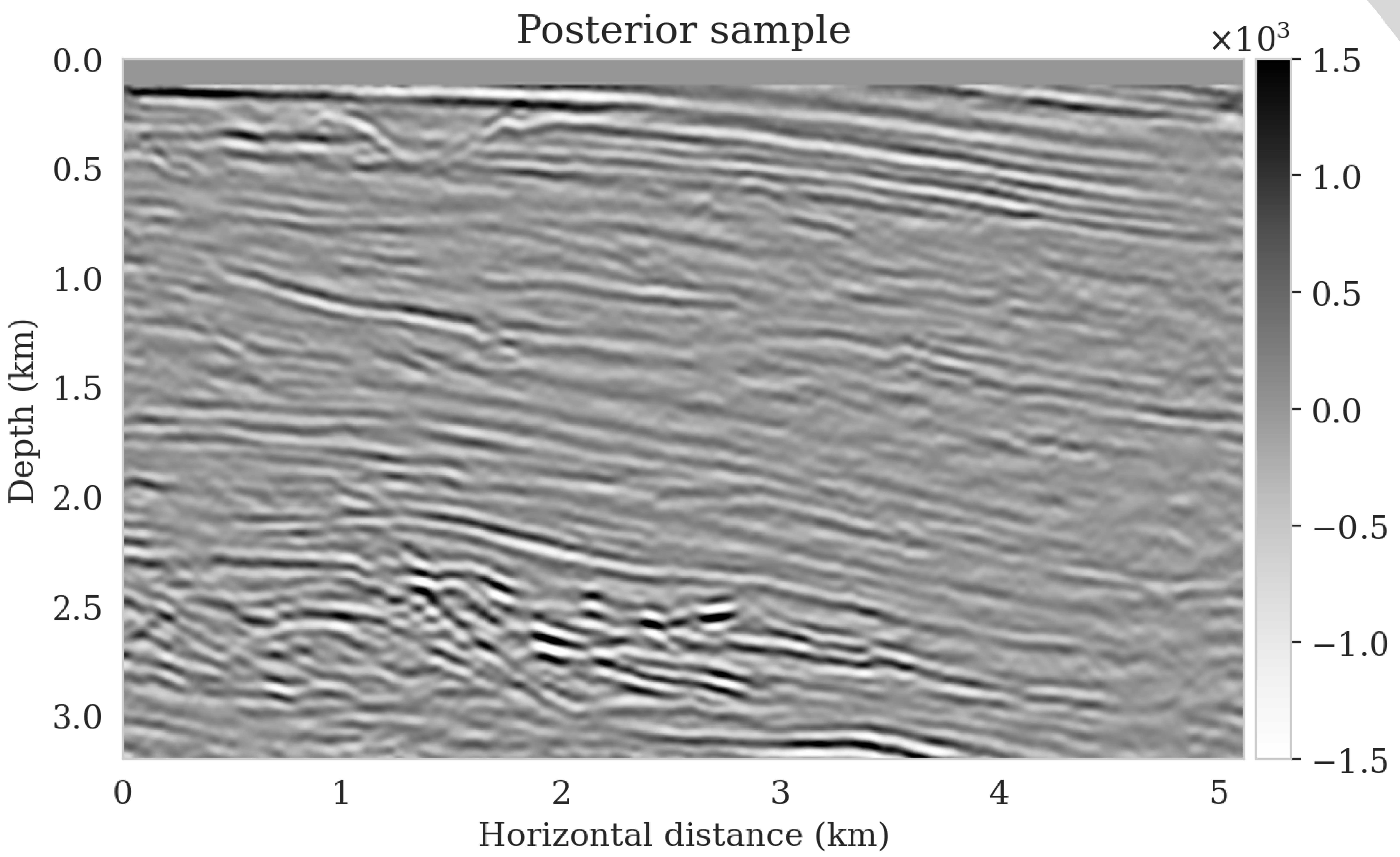

True model, $\delta \mathbf{m}$
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
$\times 10^3$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


Pointwise standard deviation
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)
3×10^2
2×10^2
10^2
6×10^1

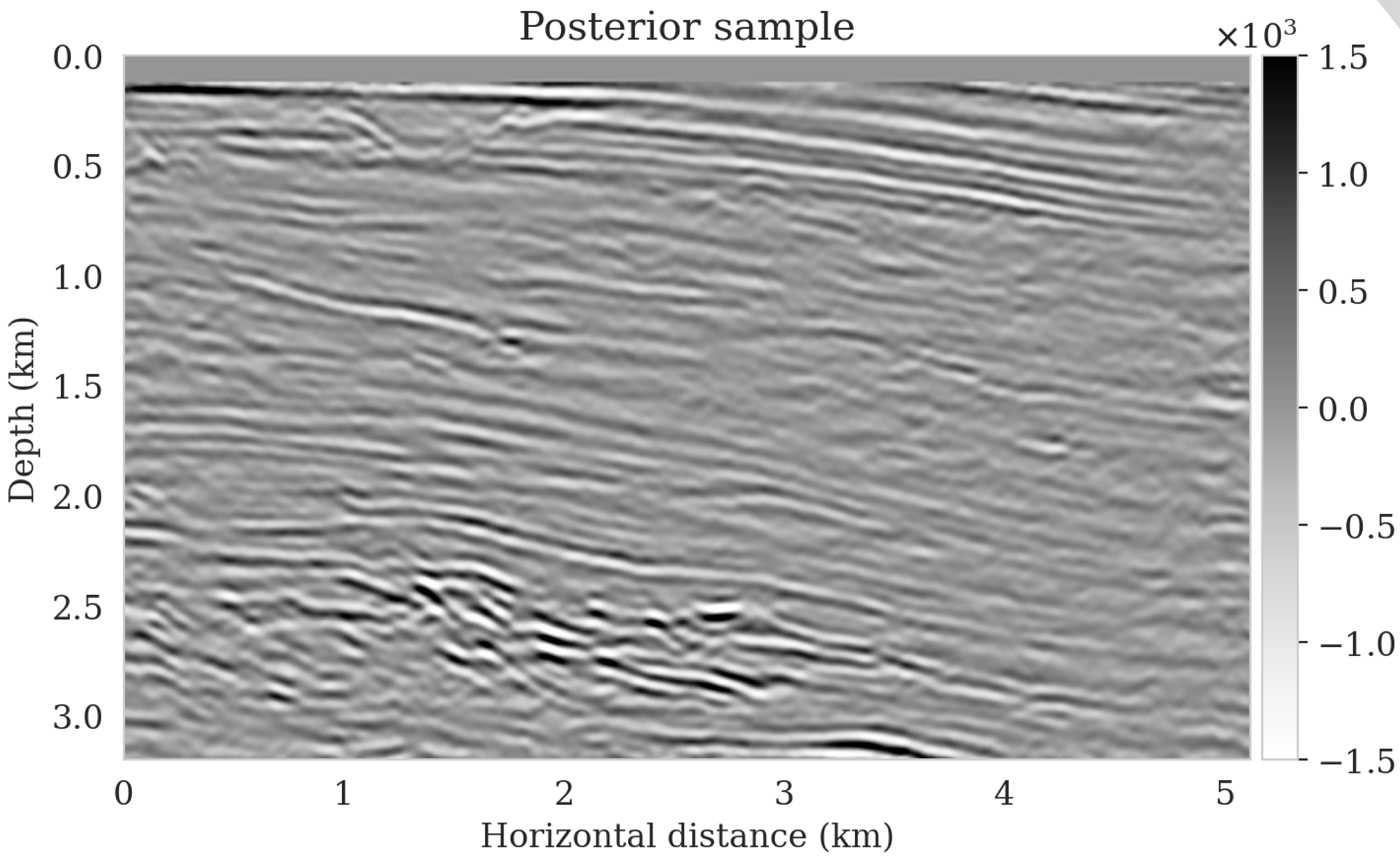
Seismic imaging example 3

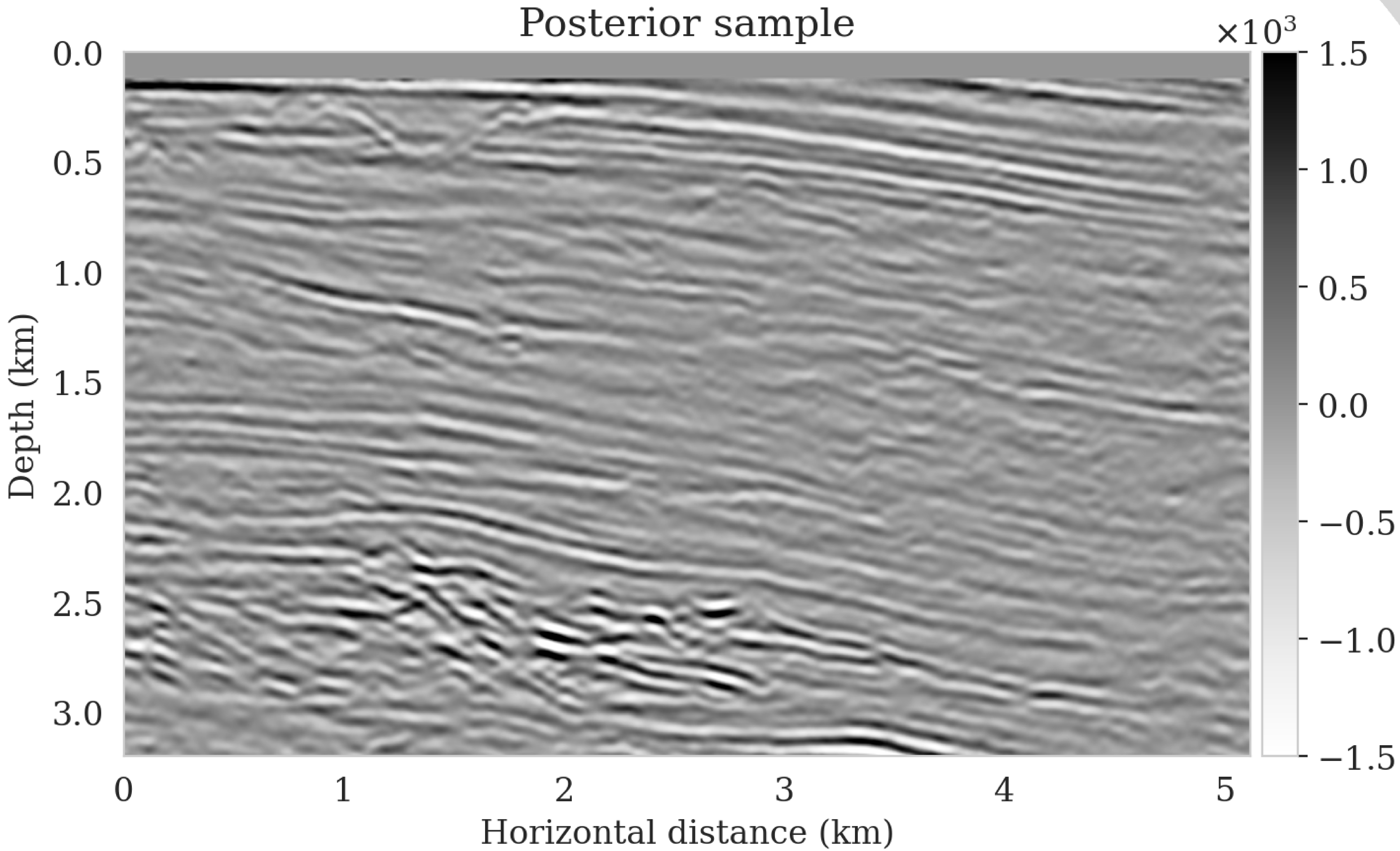

True model, $\delta \mathbf{m}$
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Horizontal distance (km)
0
1
2
3
4
5
$\times 10^3$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

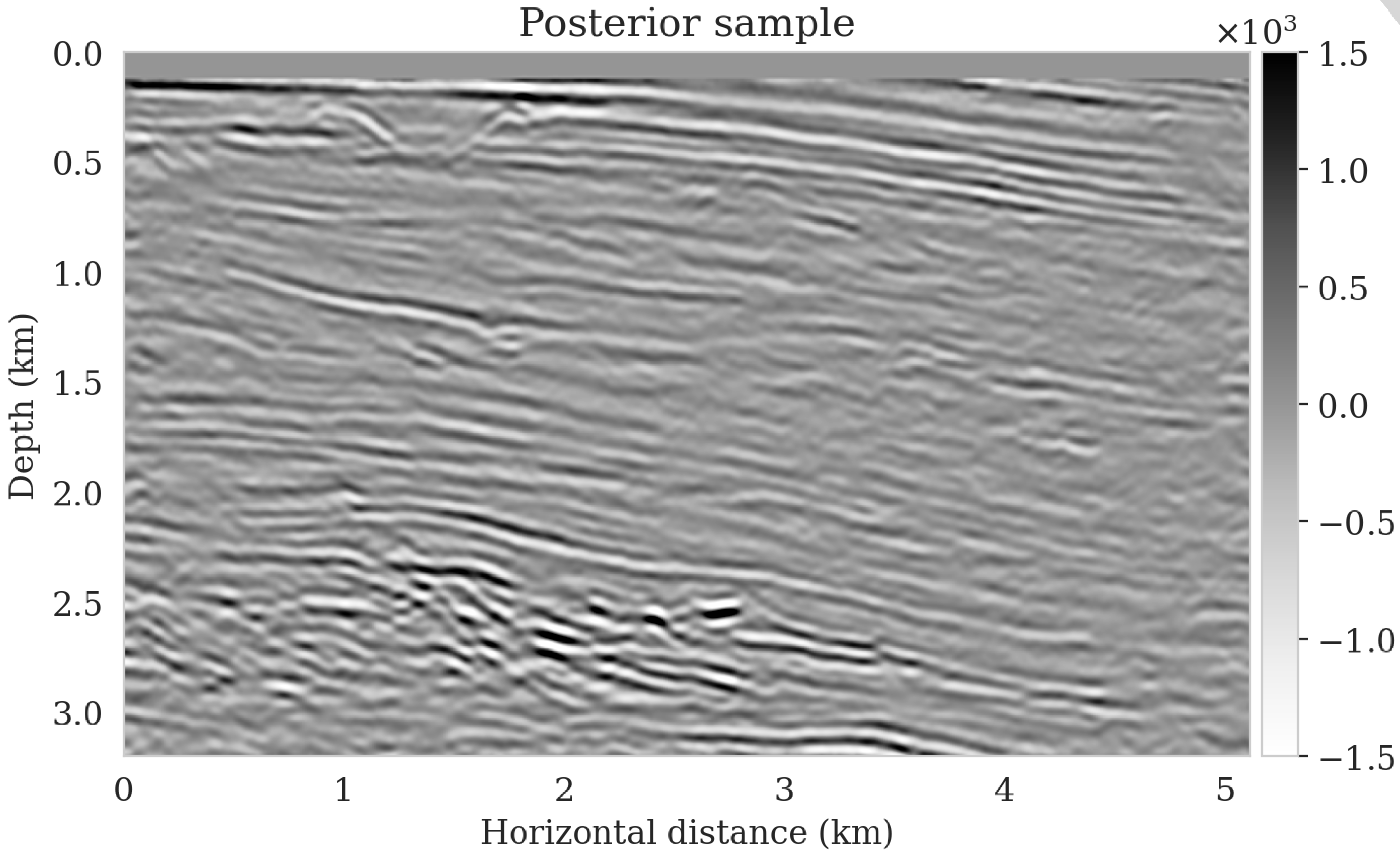

RTM, $\delta \mathbf{m}_{RTM}$
$\times 10^6$
Depth (km)
Horizontal distance (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5



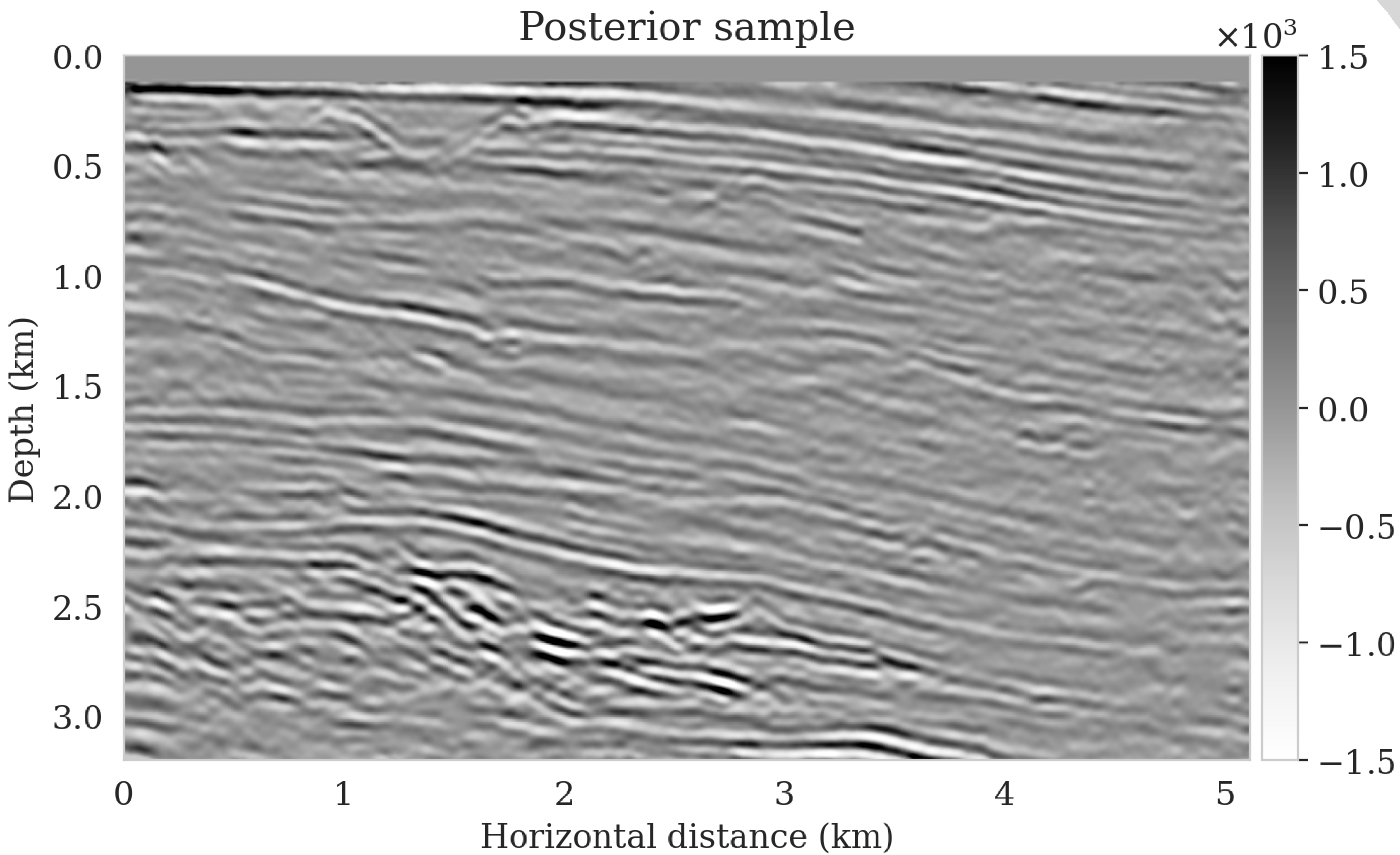
Posterior sample
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)
×10^3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

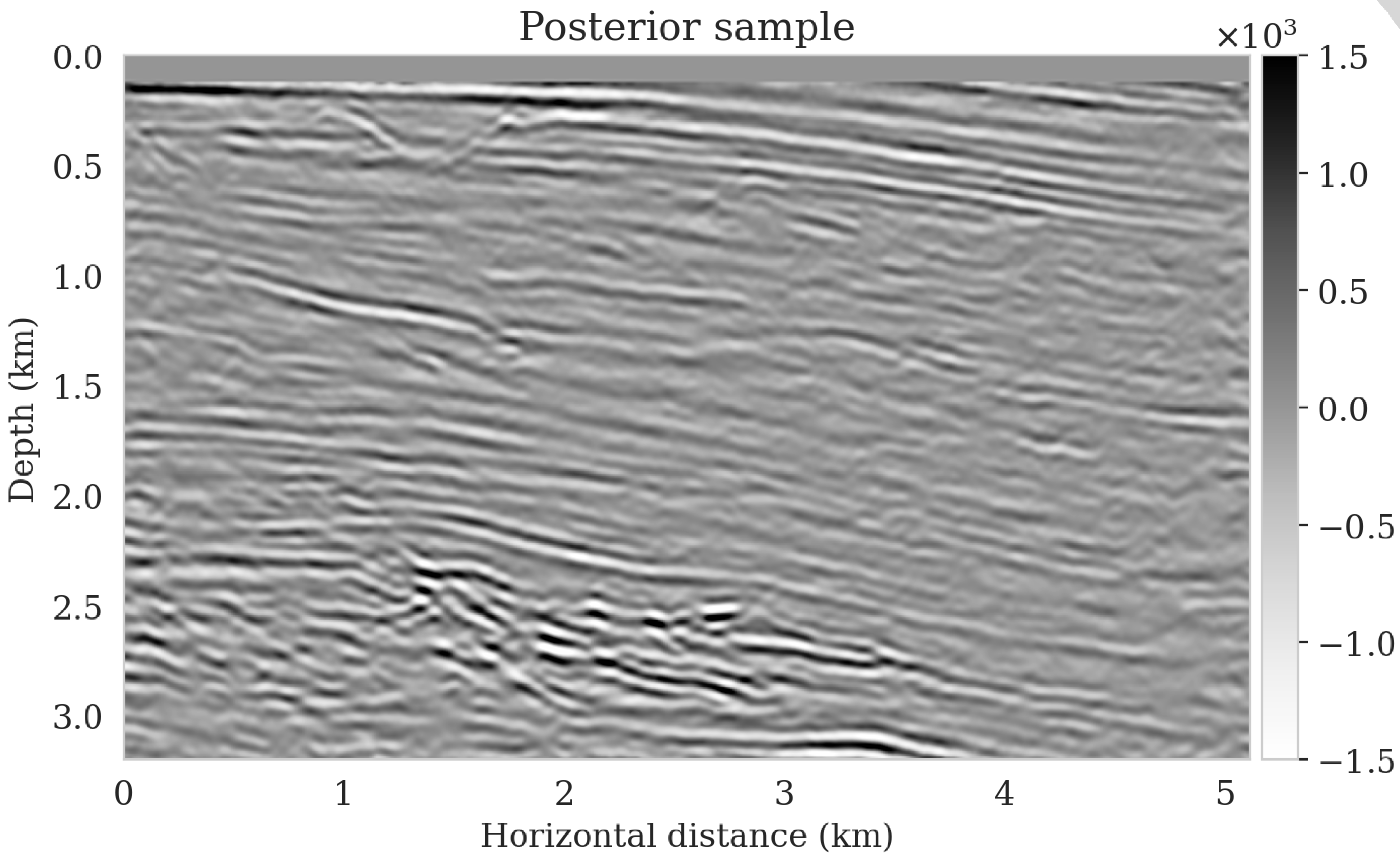

Posterior sample
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)


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
×10^3
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
Horizontal distance (km)
0
1
2
3
4
5
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

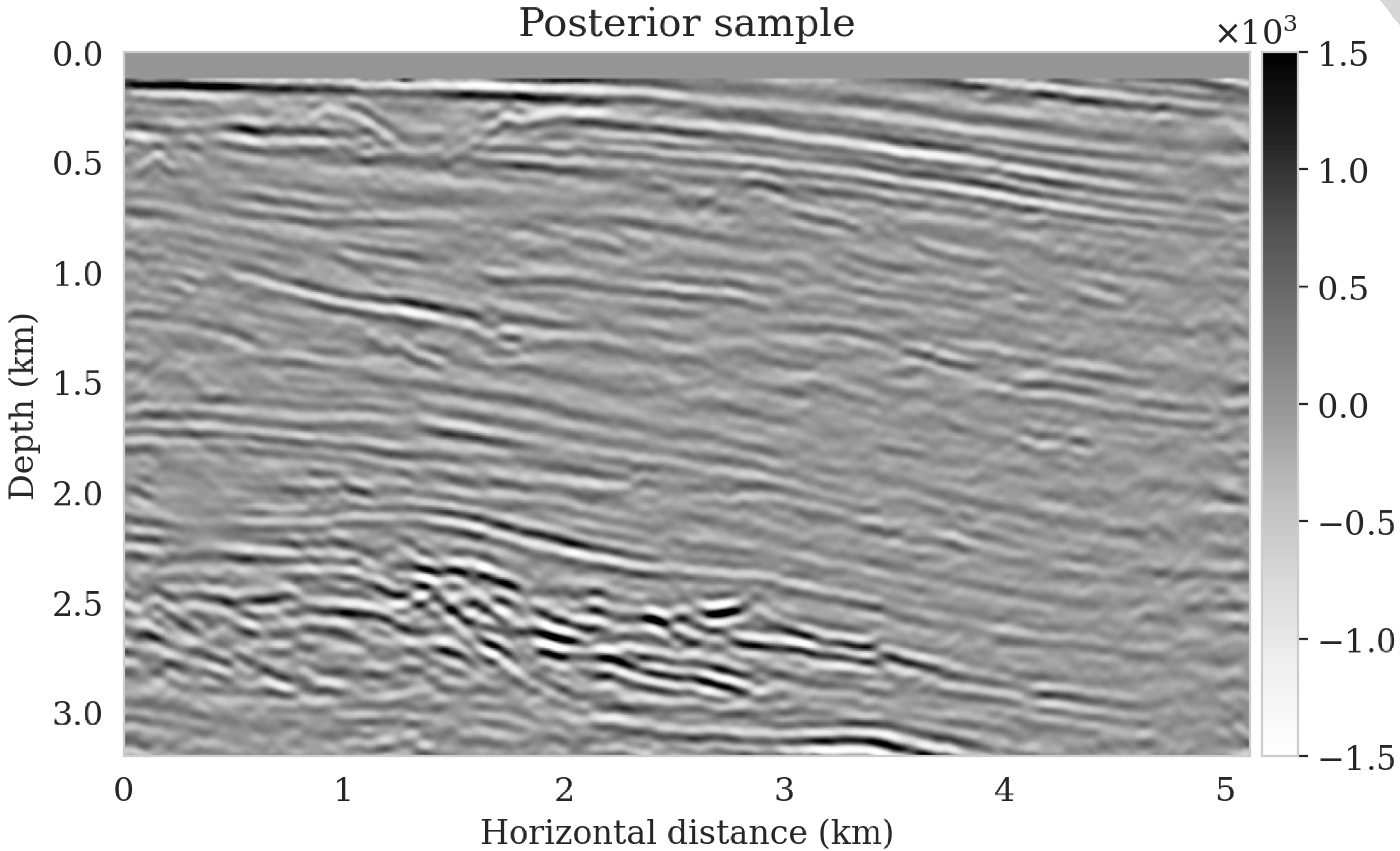
SLIM

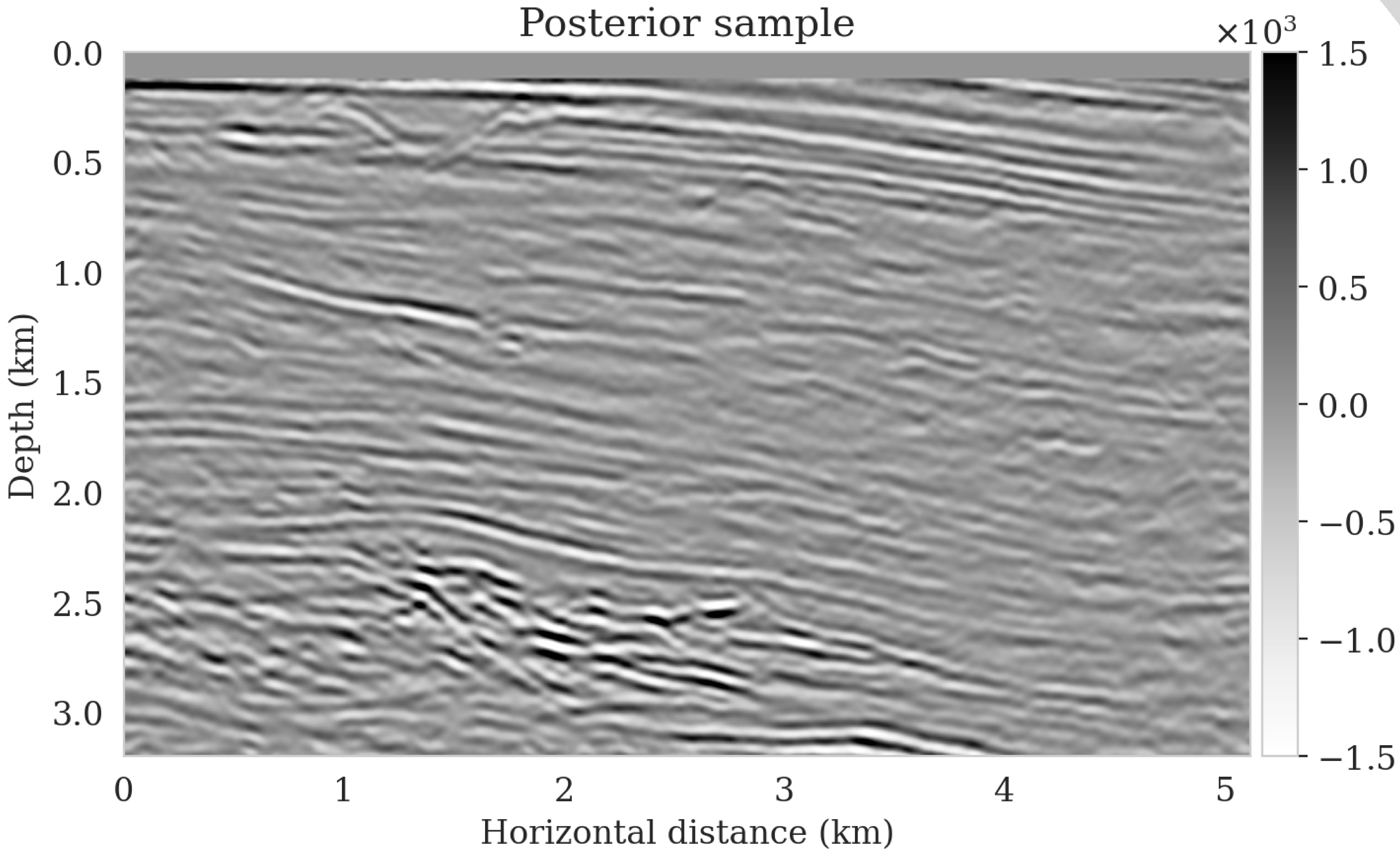

Posterior sample
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

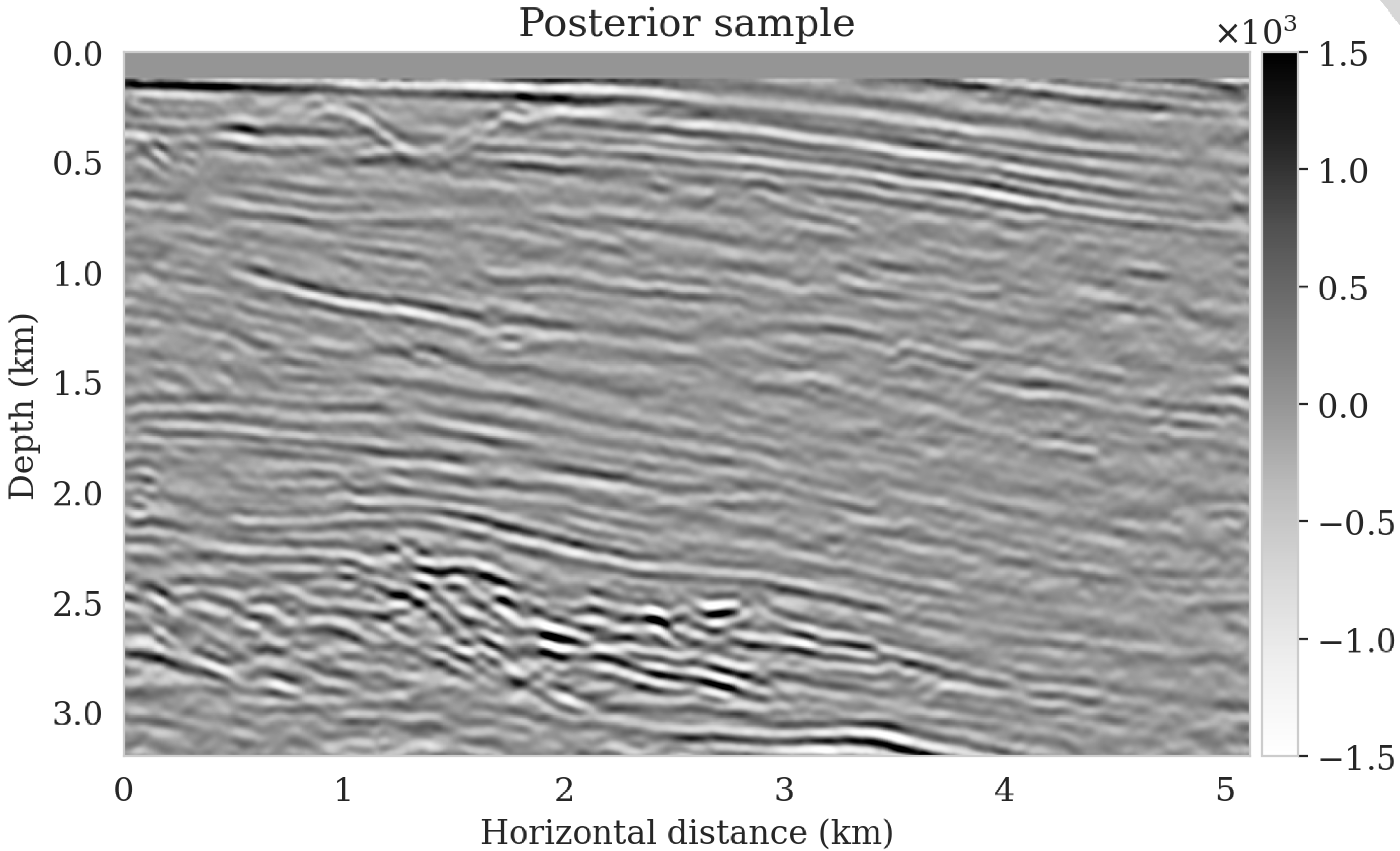


Posterior sample
×10³
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

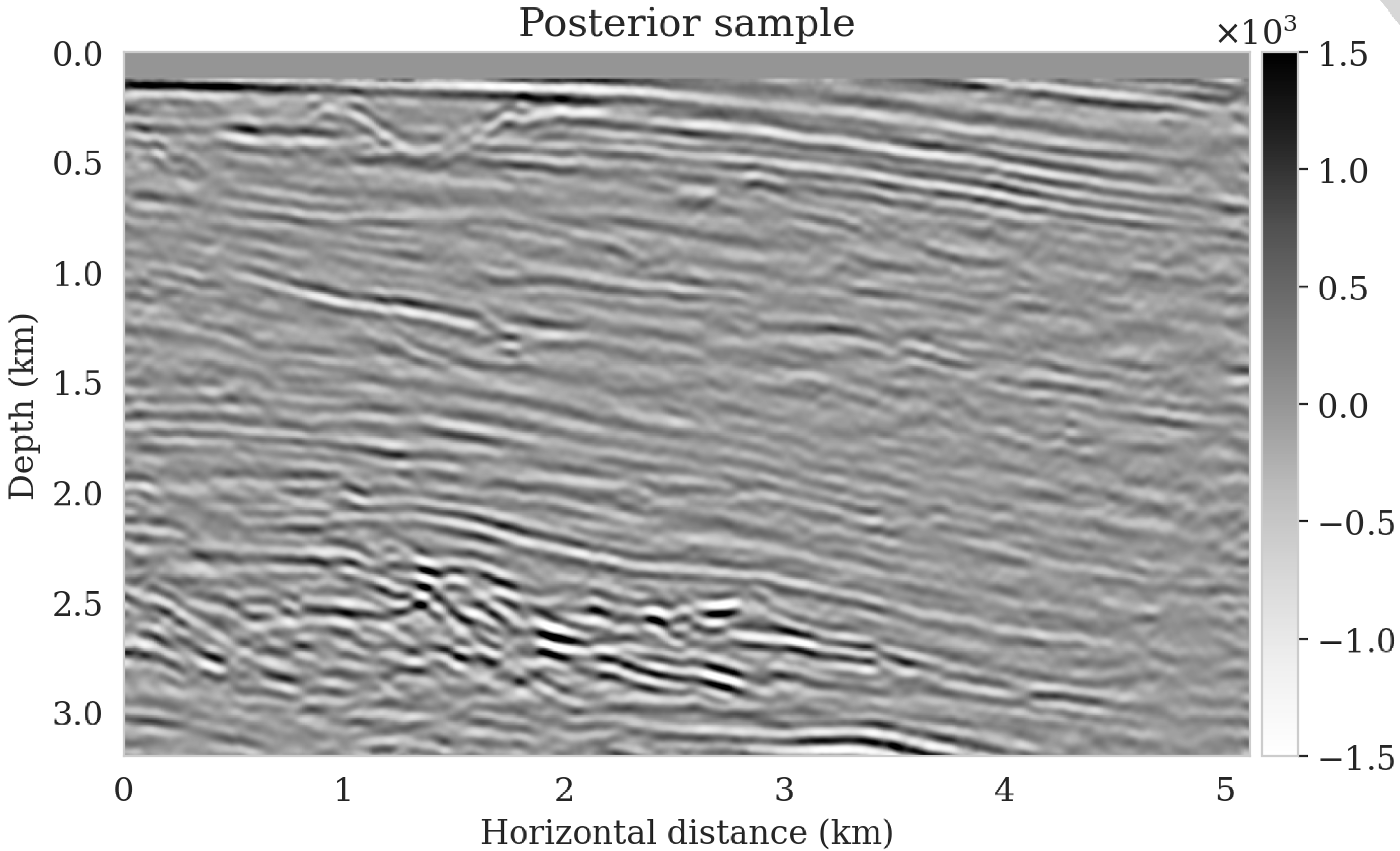
SLIM

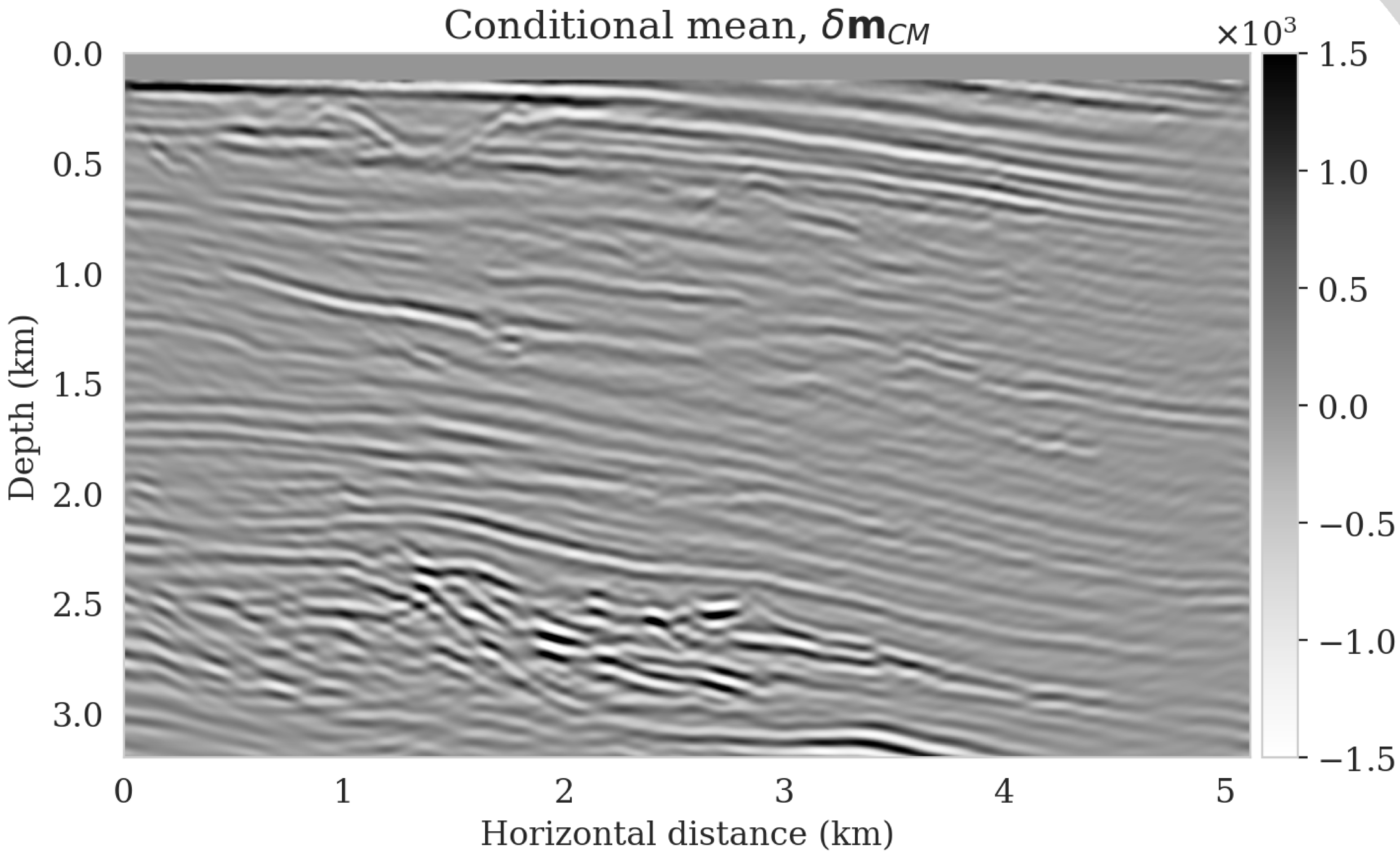

Posterior sample
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


Posterior sample
×10³
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)



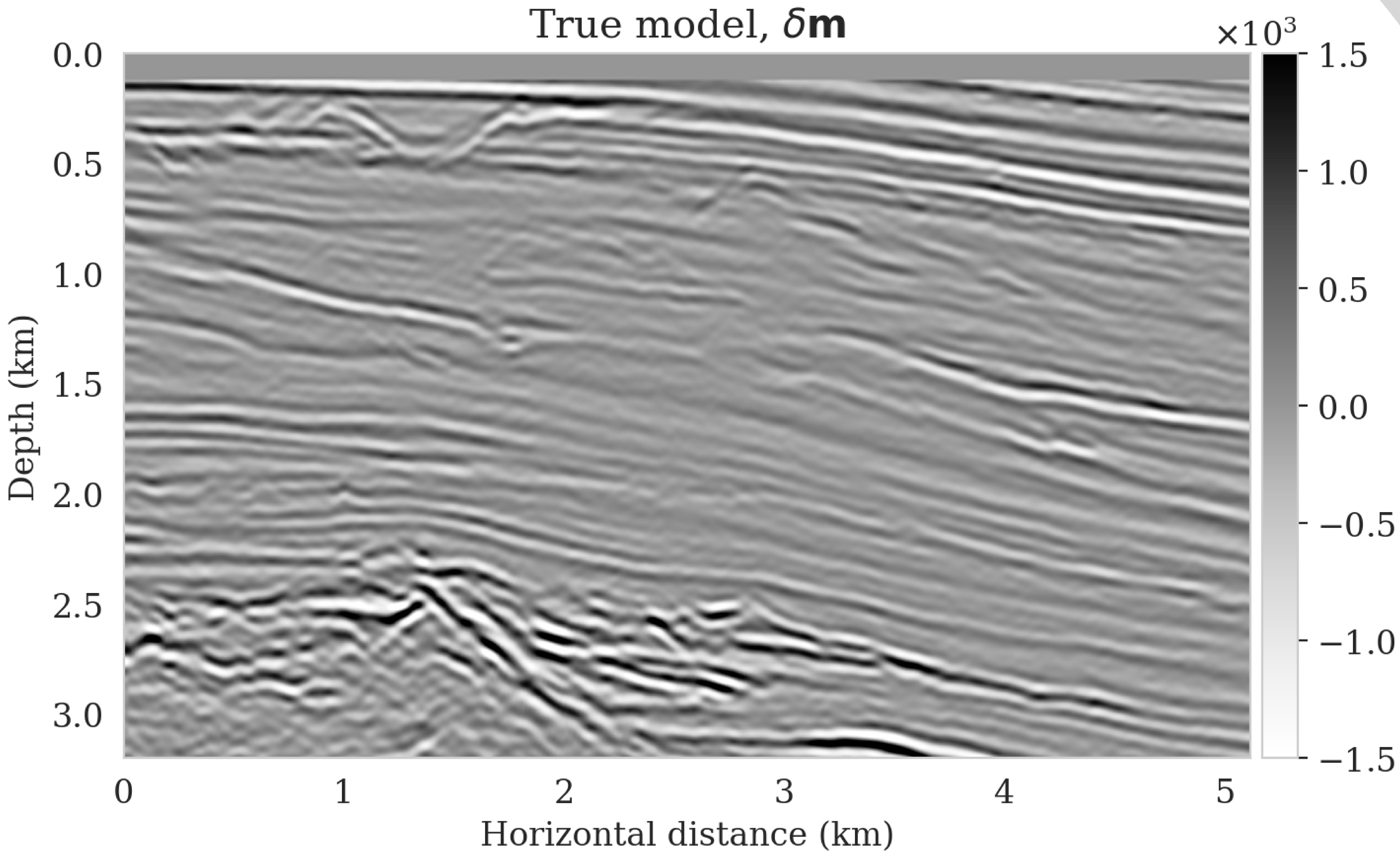
Posterior sample
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


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

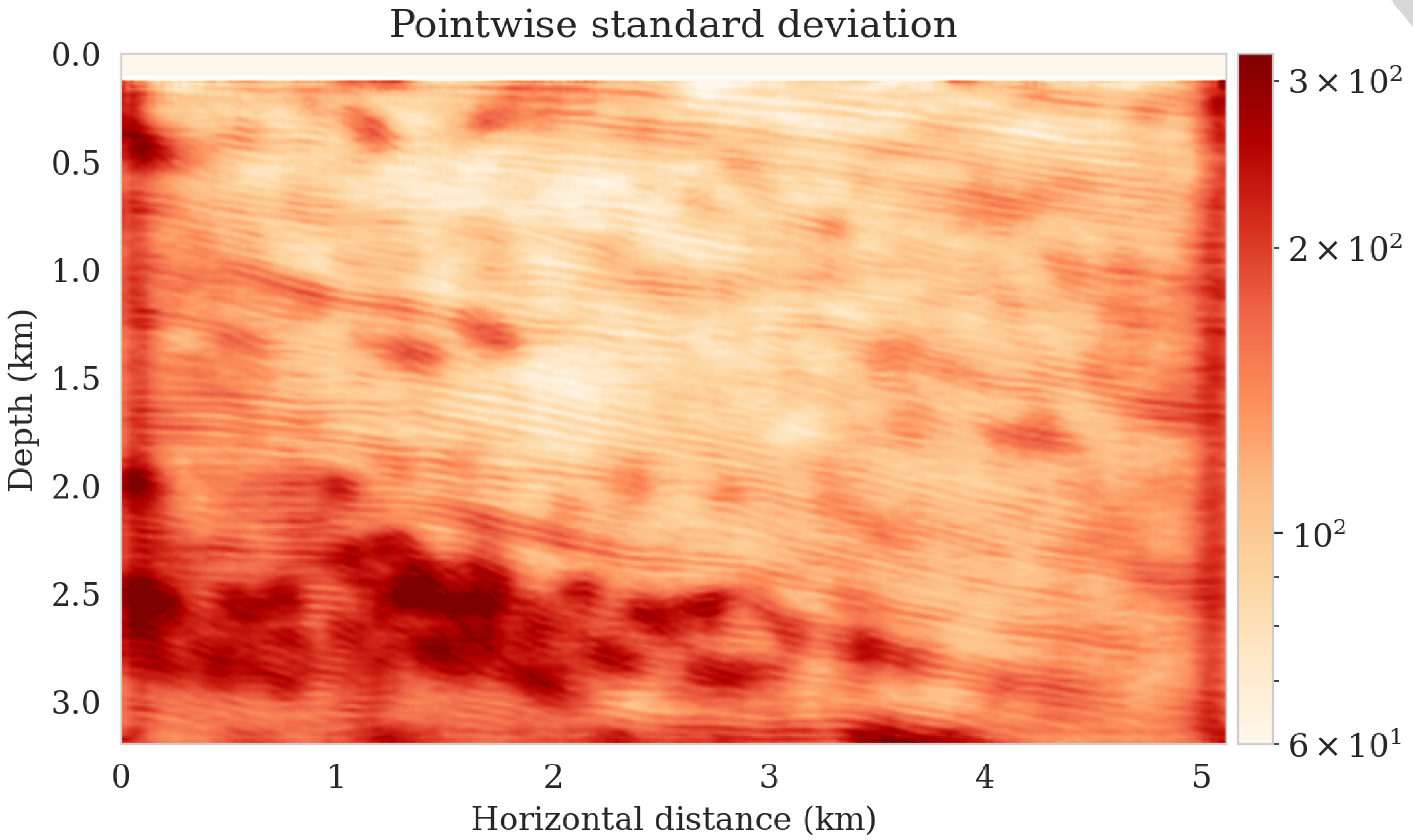


Conditional mean, $\delta \mathbf{m}_{CM}$
$\times 10^3$
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

SLIM



True model, $\delta\mathbf{m}$
$\times 10^3$
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
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)
3×10^2
2×10^2
10^2
6×10^1

Physics-based variational inference

$$\min_{\theta} \mathbb{E}_{\mathbf{z}\sim p_z(\boldsymbol{z})} \left[\frac{1}{2\sigma^2} \left\| F(T_\theta(\mathbf{z})) - \mathbf{y} \right\|_2^2 - \log p_{\text{prior}}\left(T_\theta(\mathbf{z})\right) - \log \left| \det \nabla_z T_\theta(\mathbf{z}) \right| \right]$$

no training data required

cheap posterior sampling: $T_\theta(\mathbf{z}) \sim p_\theta(\mathbf{x} \mid \mathbf{y}) \approx p_{\text{post}}(\mathbf{x} \mid \mathbf{y})$

requires a prior density, $p_{\text{prior}}(\mathbf{x})$

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

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

Qiang Liu and Dilin Wang. "Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm". In: *Advances in Neural Information Processing Systems*. Vol. 29. 2016, pp. 2378–2386. URL: https://proceedings.neurips.cc/paper/2016/file/b3ba8f1bee1238a2f37603d90b58898d-Paper.pdf.

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

Multi-fidelity preconditioned scheme

initialize $T_{\phi_x}(\mathbf{z}) := G_{\phi_x}^{-1}(G_{\phi_y}(\mathbf{y}), \mathbf{z})$, followed by transfer learning

use $p_G(\mathbf{x})$—density encoded by $G_{\phi_x}^{-1}$—as a (conditional) prior

Variational inference with the preconditioned scheme

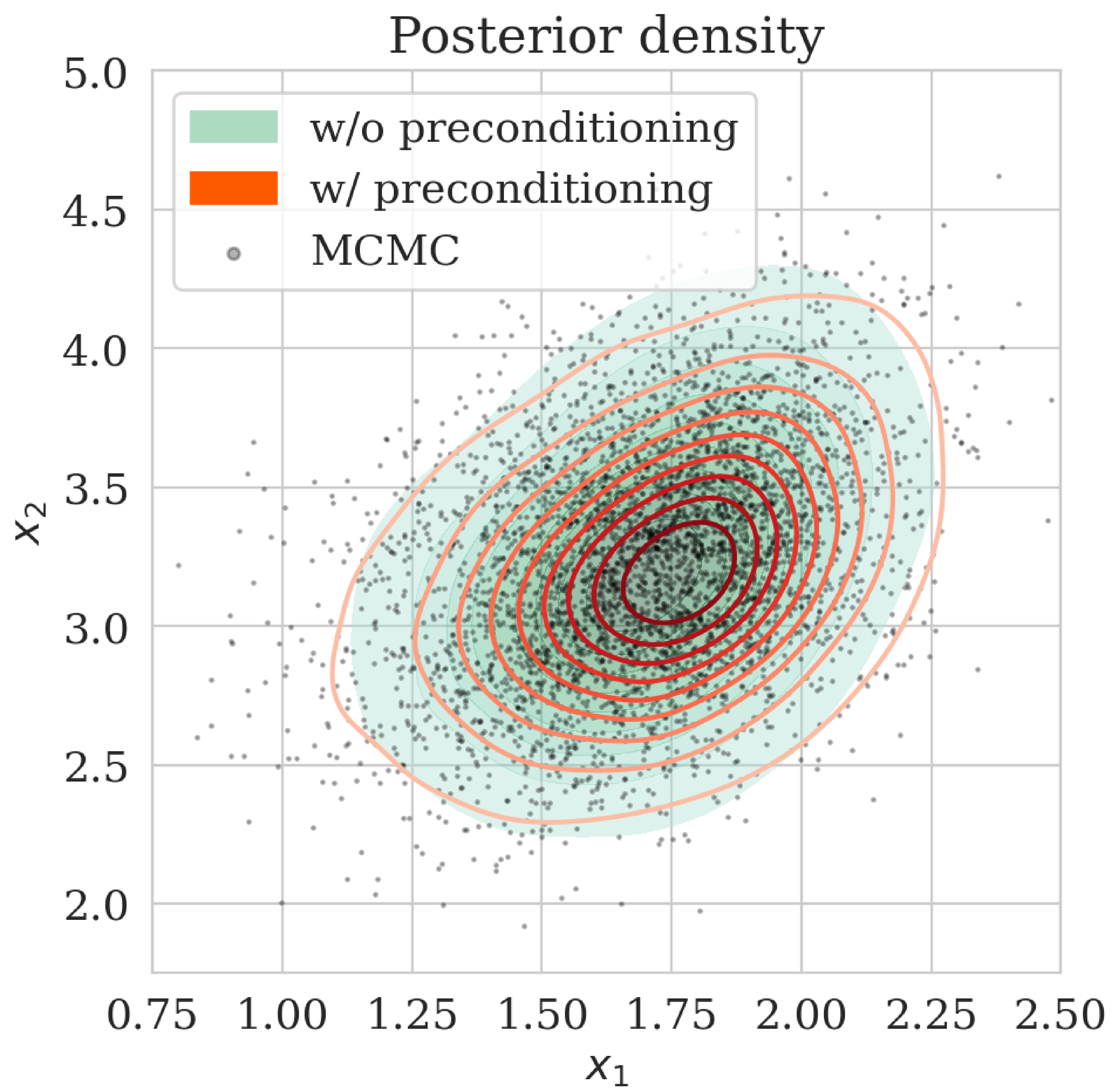
$$\min_{\phi_x} \mathbb{E}_{\mathbf{z}\sim \mathrm{N}(\mathbf{0},\mathbf{I})} \left[\frac{1}{2\sigma^2} \left\| F\big(T_{\phi_x}(\mathbf{z})\big) - \mathbf{y} \right\|_2^2 - \log p_G\big(T_{\phi_x}(\mathbf{z})\big) - \log \left| \det \nabla_z T_{\phi_x}(\mathbf{z}) \right| \right]$$

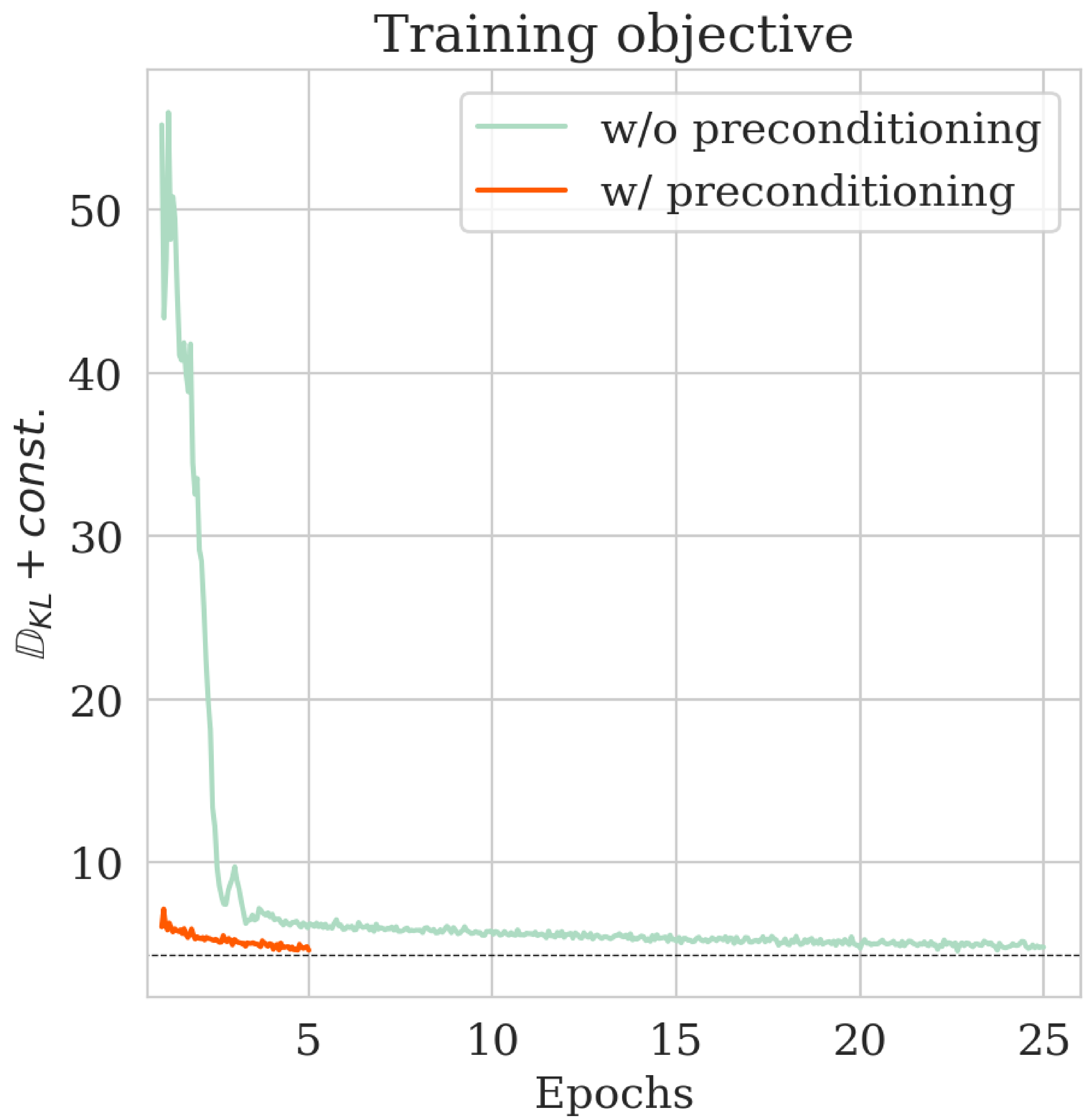
transfer learning:

- corrects for low-fidelity training data, $p_{y,x} \neq p_{y,x}$
- less evaluations of F and $\nabla F^\top$

learned prior density, $p_G(\mathbf{x}) := p_{\mathbf{z}}\big(G_{\phi_x}(\mathbf{y},\mathbf{x})\big) \left| \det \nabla_x G_{\phi_x}(\mathbf{y},\mathbf{x}) \right|$

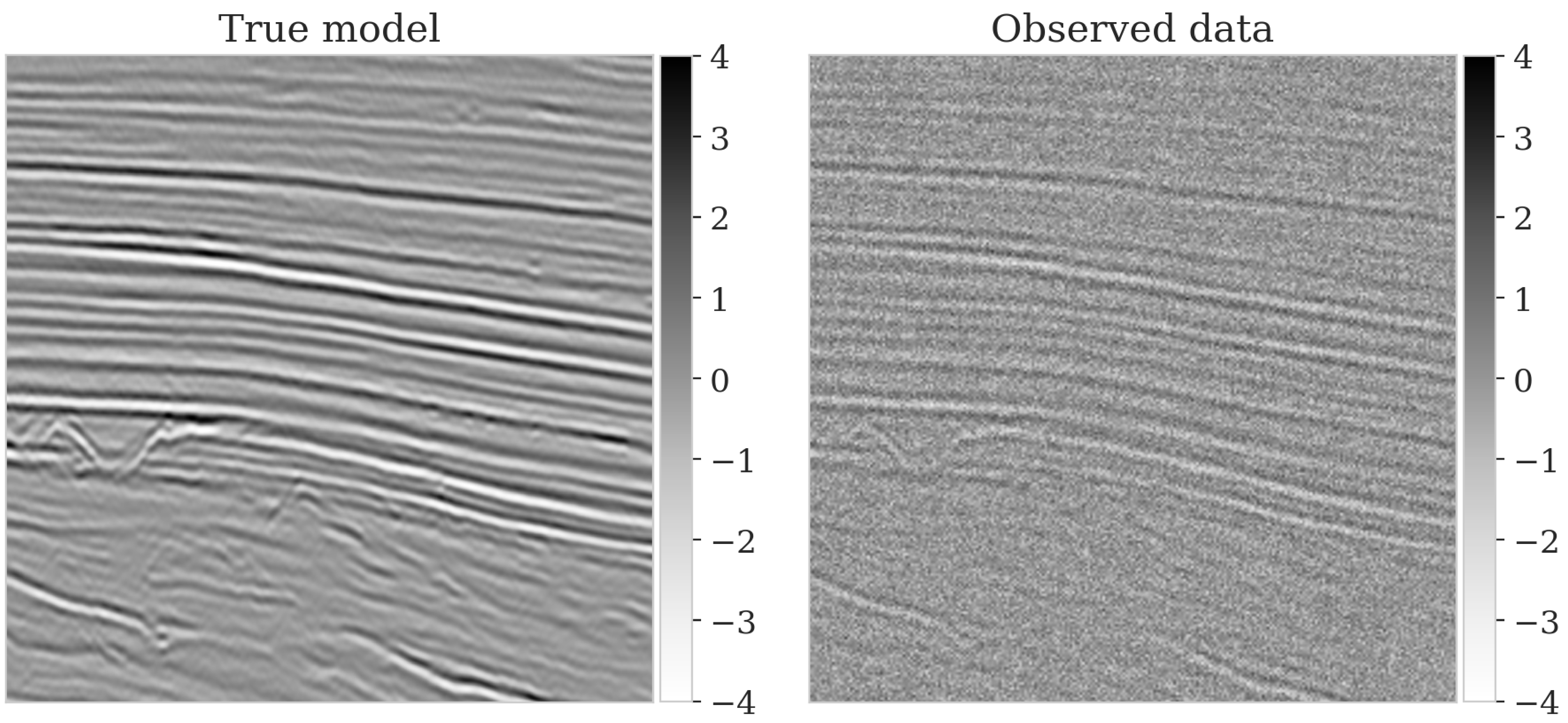
fast to adapt to new observation



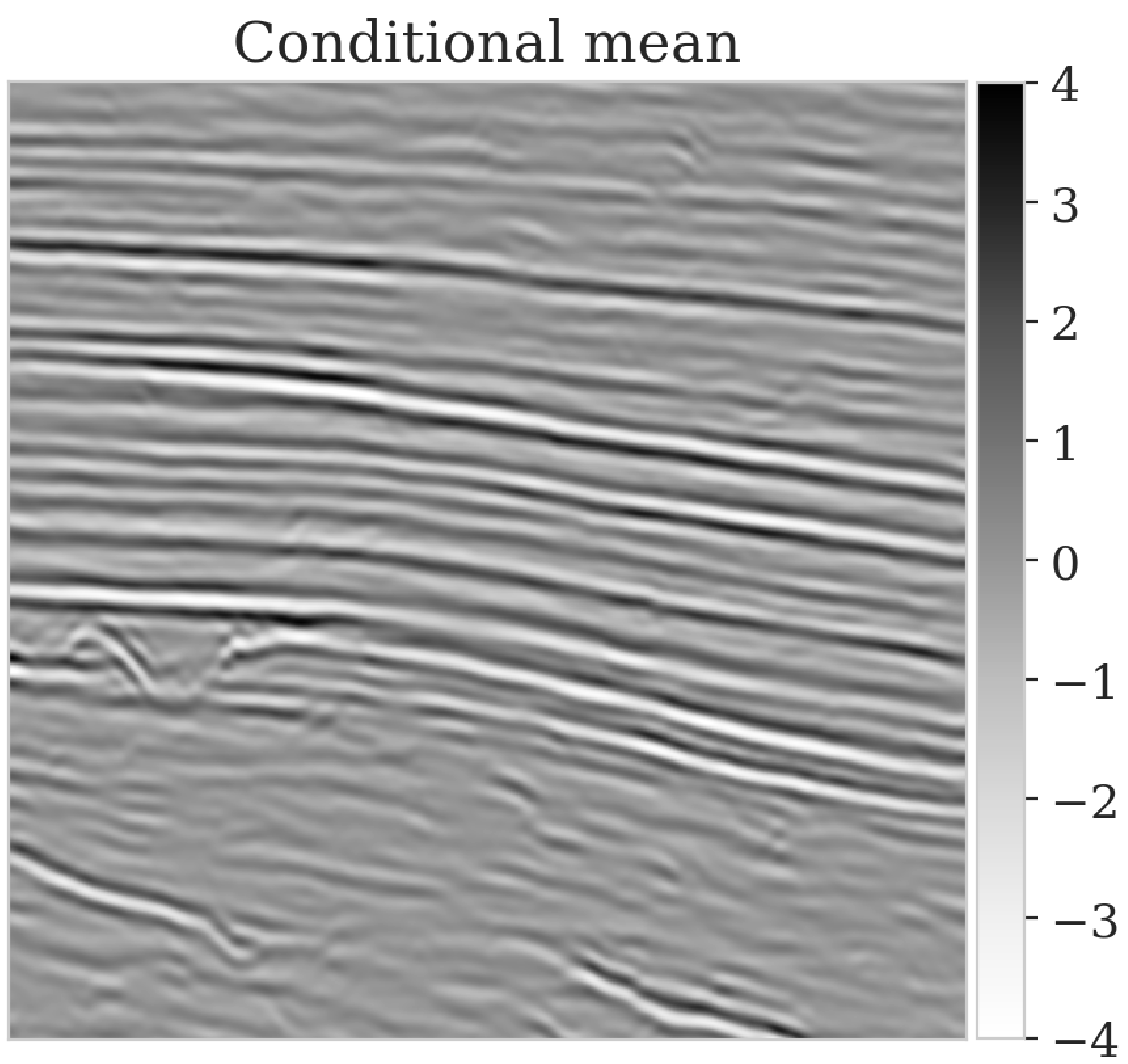


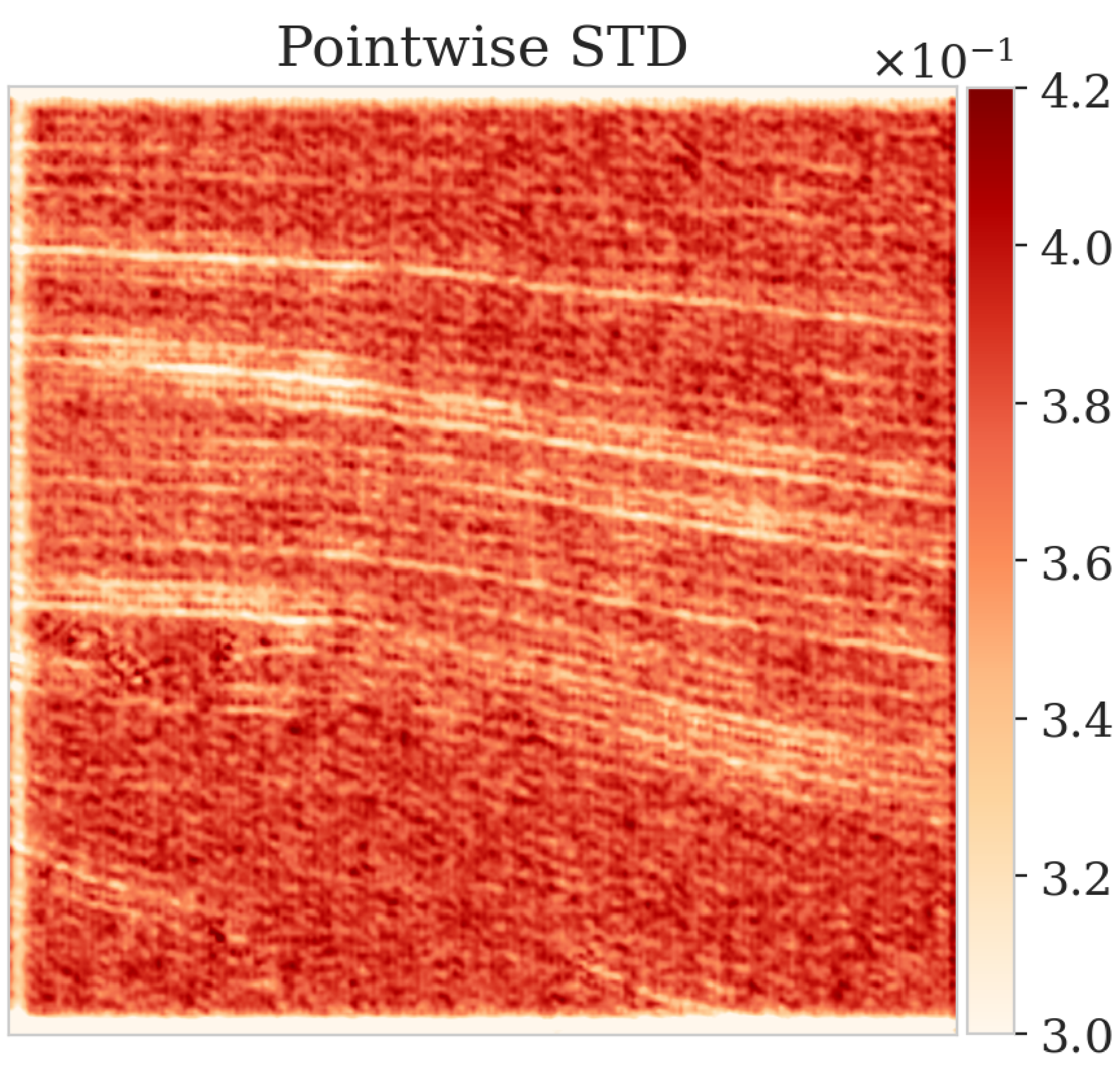
(left) without vs with preconditioning vs MCMC, (right) loss with and without preconditioning

Seismic compressed sensing example

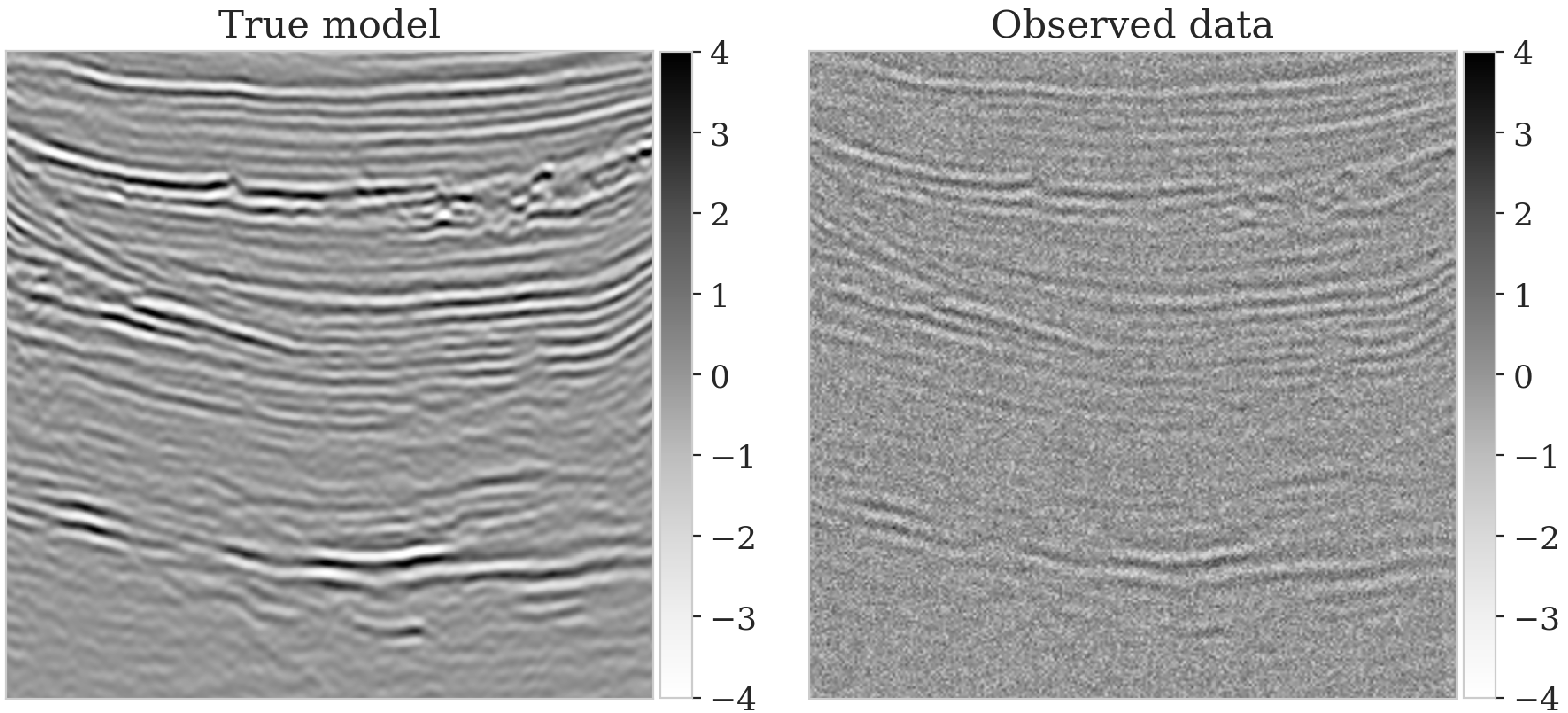


(left) true model, (right) “observed” data

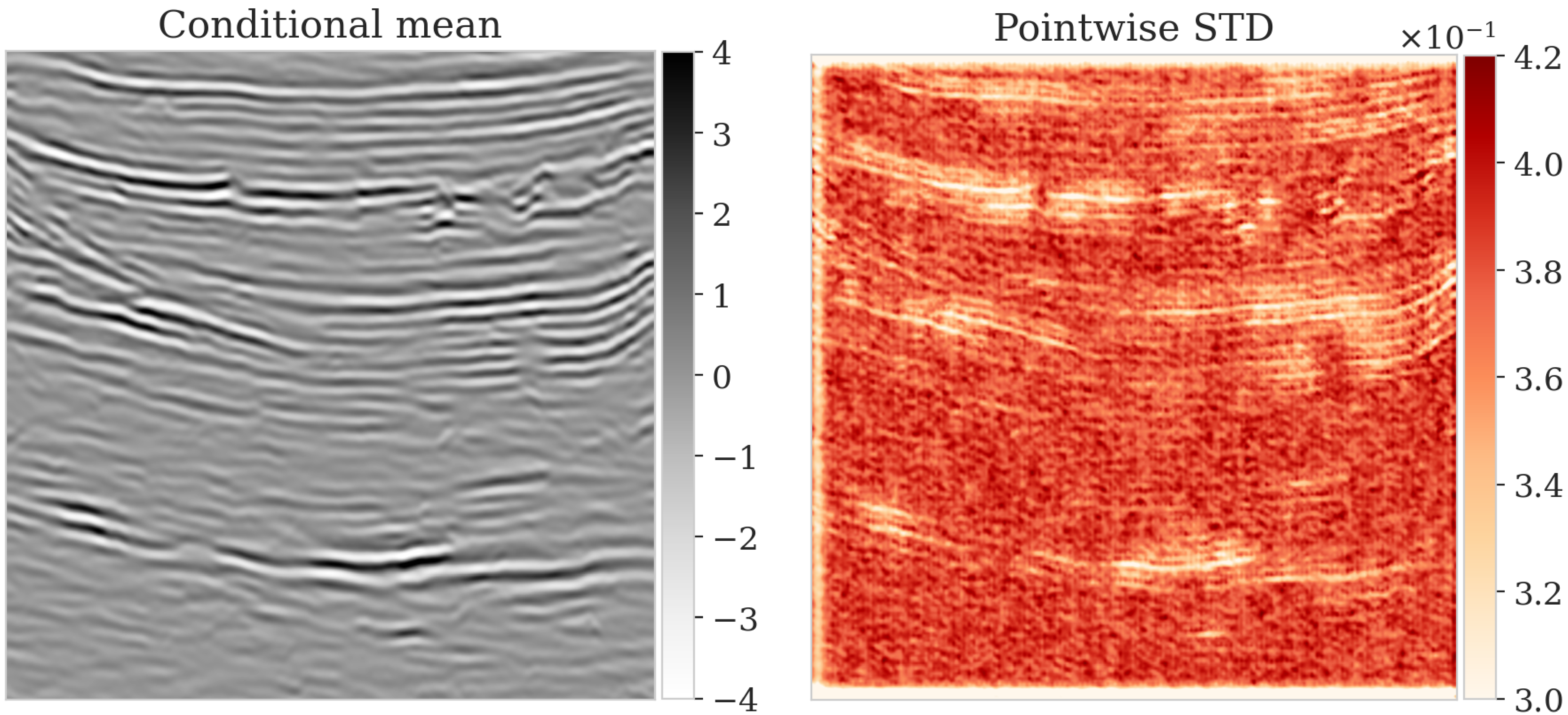




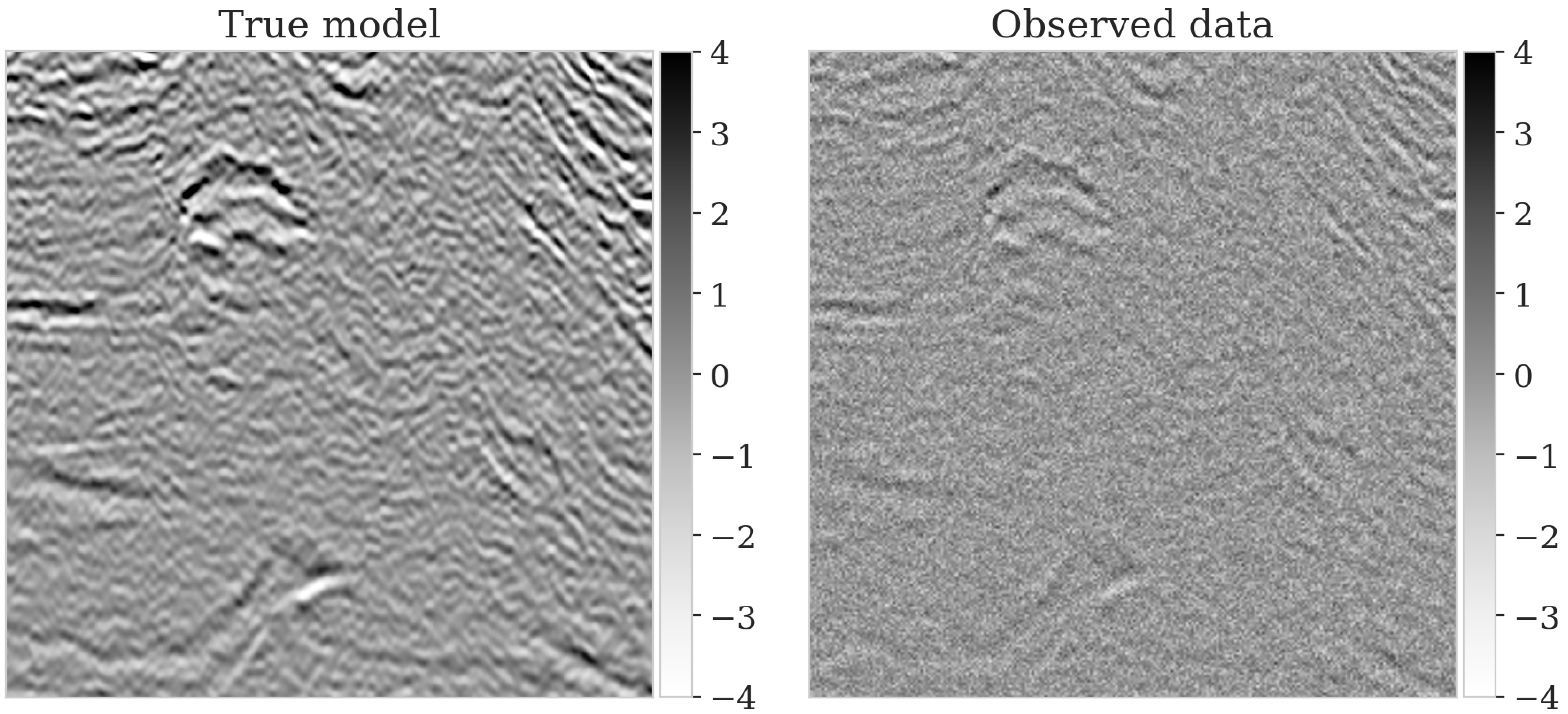
conditional mean (left) and pointwise standard deviation (right)



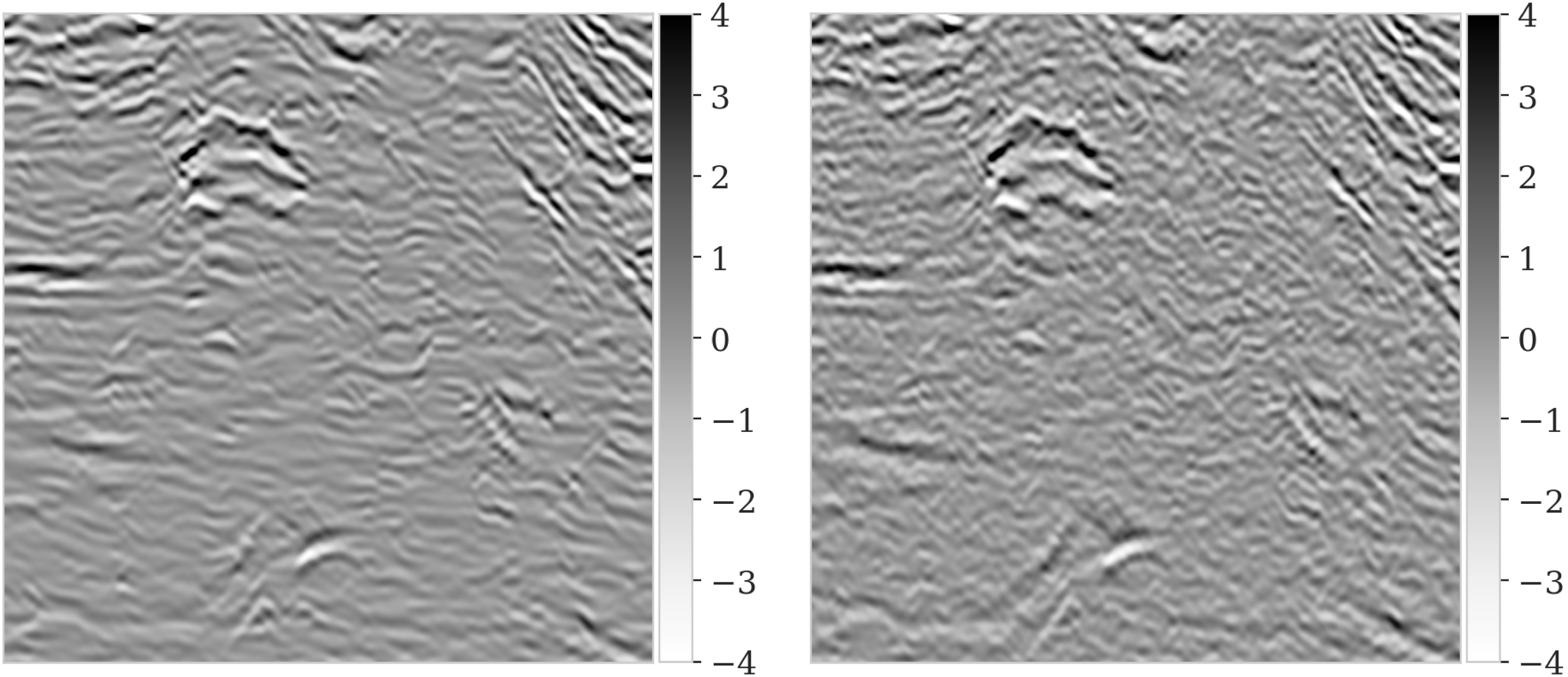
(left) true model, (right) "observed" data



conditional mean (left) and pointwise standard deviation (right)



(left) true model, (right) observed data



conditional mean before (left) and after (right) transfer learning

Contributions

Take full advantage of existing training data to:

- speed-up variational inference
- train a learned (conditional) prior density
- pretrained conditional neural net can serve as a preconditioner

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

Conclusions

UQ can be propagated to tasks and integrated with errors in the interpretation.

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 improvements
- preconditioning with a pretrained conditional normalizing flow can lead to another order of magnitude speed up