An efficient penalty method for PDE-constrained optimization problem with source estimation and stochastic optimization

Zhilong Fang*, Curt Da Silva* and Felix J. Herrmann*
*Seismic Laboratory for Imaging and Modeling (SLIM), University of British Columbia

SLIM

University of British Columbia

Motivation

Consider the following problem:

$$\min_{\mathbf{m}} \frac{1}{N} \sum_{i=1}^{N} f_i(\mathbf{m})$$

where $\mathbf{m} \in \mathcal{C}^{n_{\mathrm{m}}}$

Motivation

Consider the following problem:

$$\min_{\mathbf{m}} \frac{1}{N} \sum_{i=1}^{N} f_i(\mathbf{m})$$

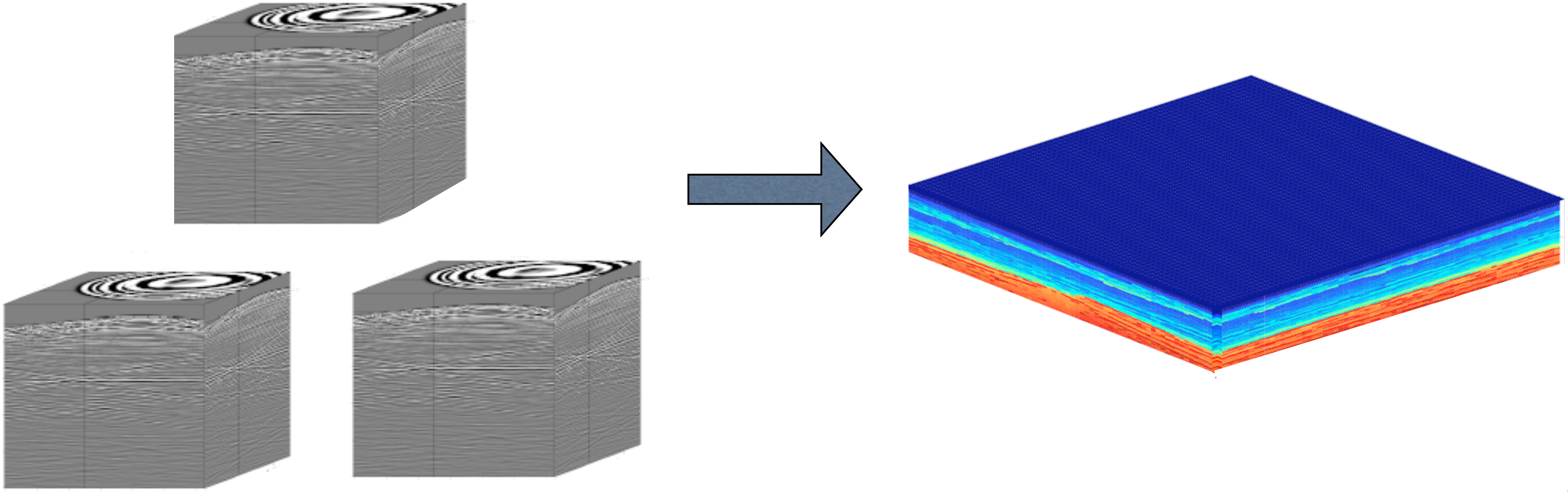
where $\mathbf{m} \in \mathcal{C}^{n_\mathrm{m}}$

Challenge: n_m and N are large

Motivation

Inverse problem in seismic exploration

- $\mathcal{O}(10^{12})$ data
- $\mathcal{O}(10^{9})$ parameters

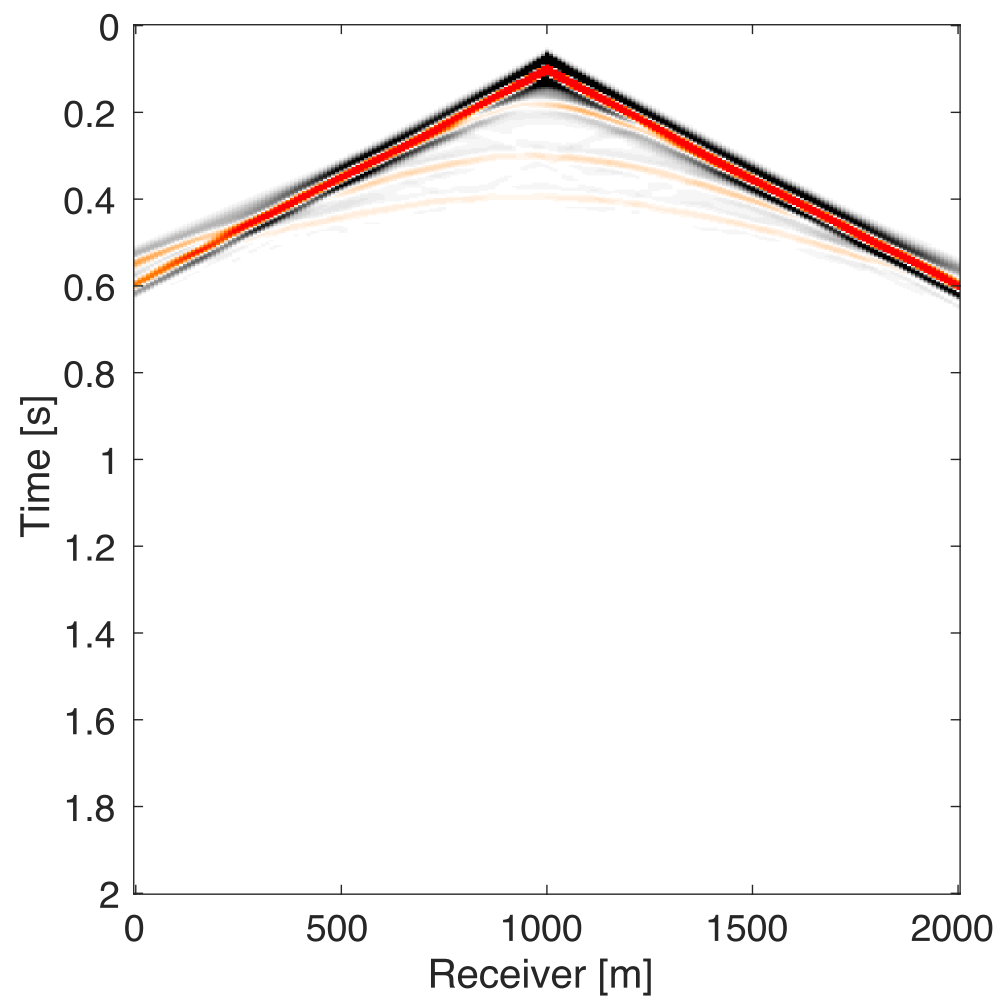


Motivation

Questions:

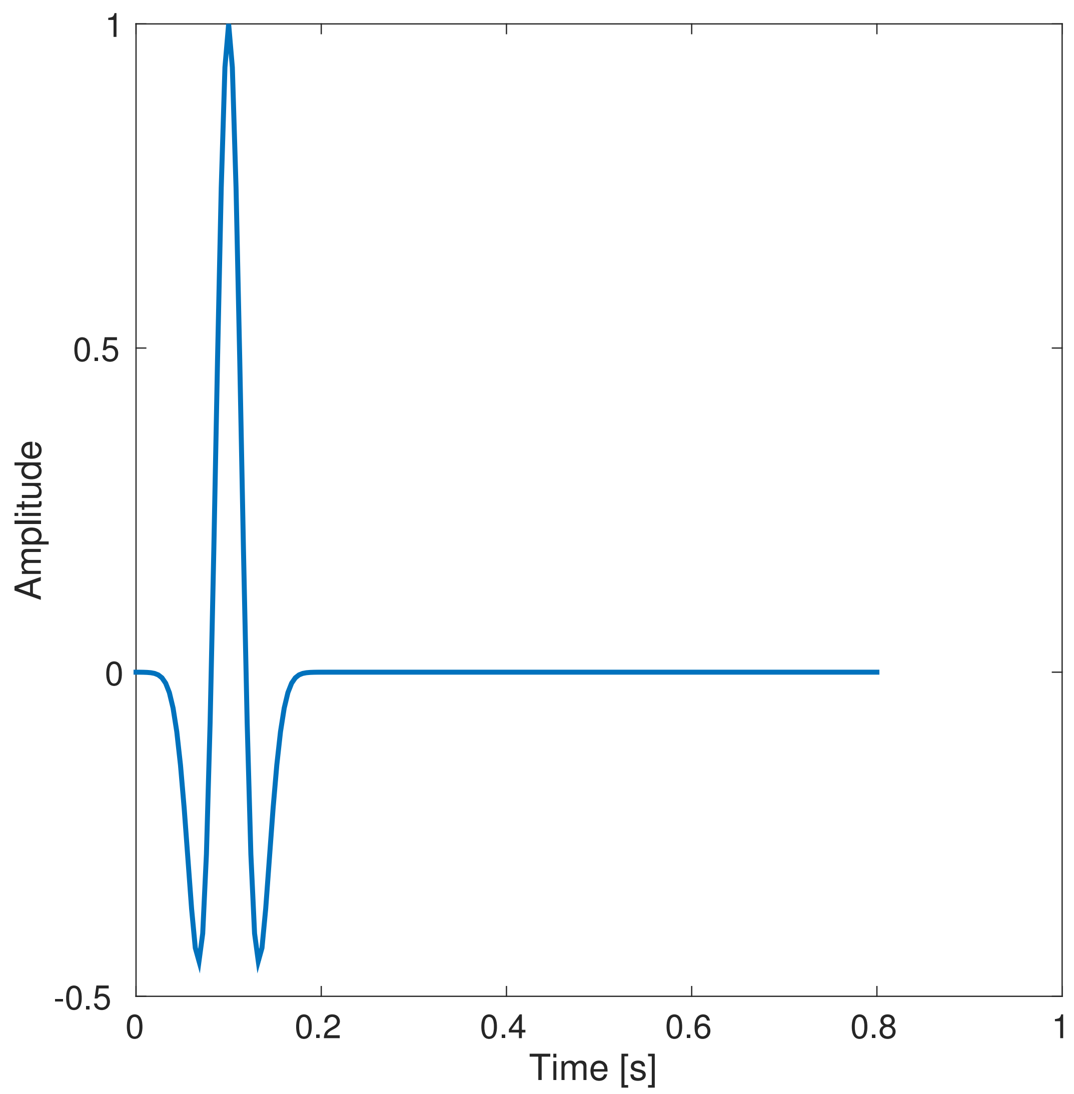
1. How can we reduce the computational cost?

Motivation



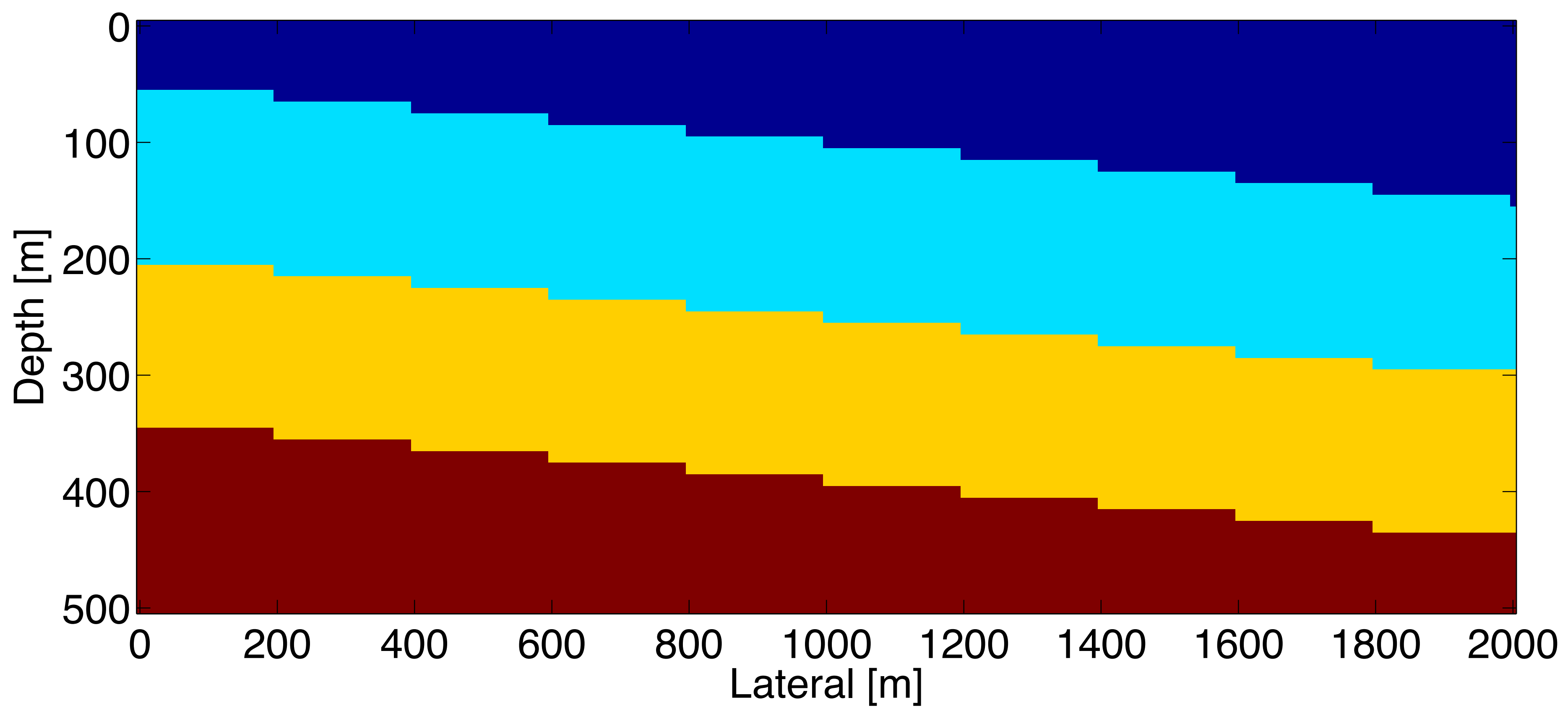
Data

$F(\mathbf{m})$
=



Source wavelet

*



Velocity model

Motivation

Forward map $F(\mathbf{m})$:

$$\mathbf{A}(\mathbf{m})\mathbf{u} = \mathbf{q}, \;\; \mathbf{d} = \mathbf{P}\mathbf{u} \longrightarrow \text{wavefields}$$

$\mathbf{A}(\mathbf{m})$ → $\Delta + \omega^2\mathbf{m}$

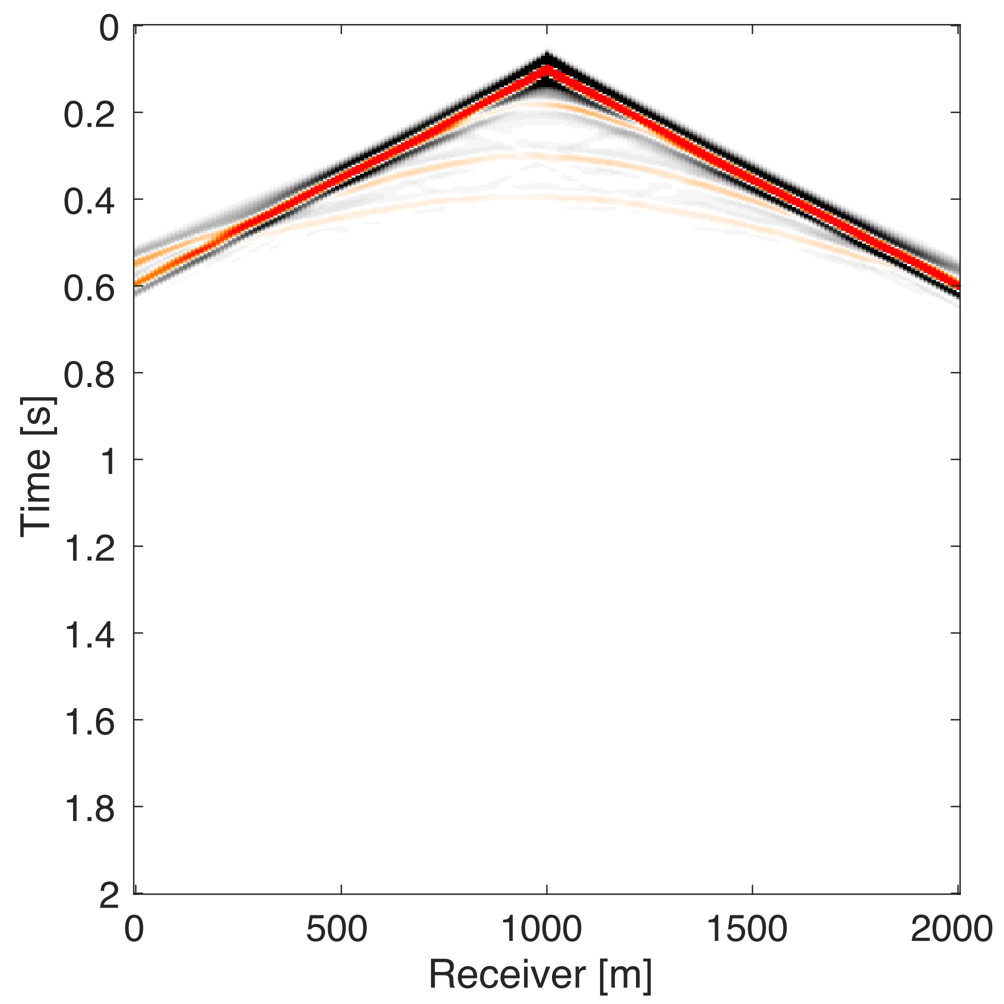
$\mathbf{q}$ → source

$\mathbf{P}$ → projection operator

ω → frequency

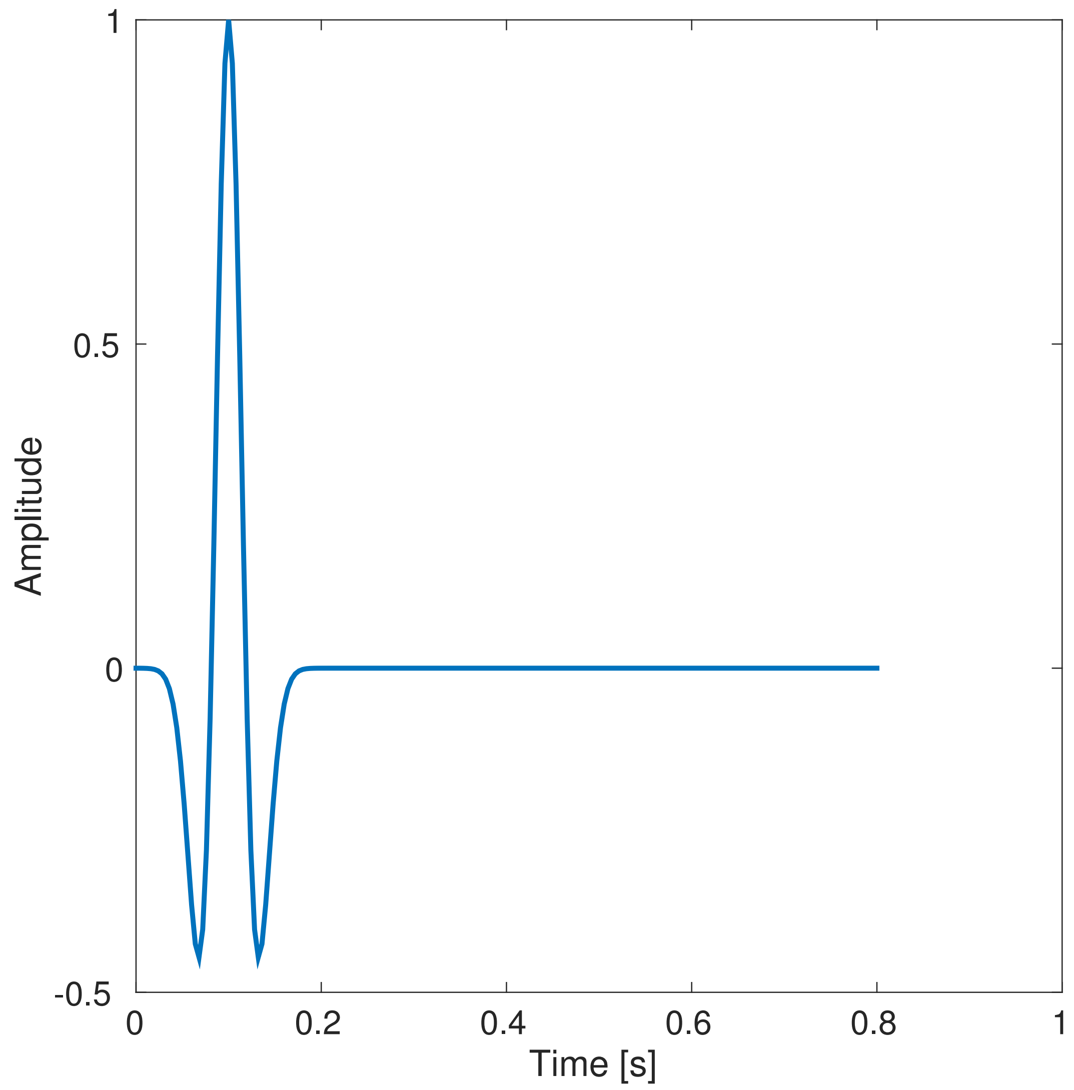
$\mathbf{m}$ → squared-slowness

Motivation



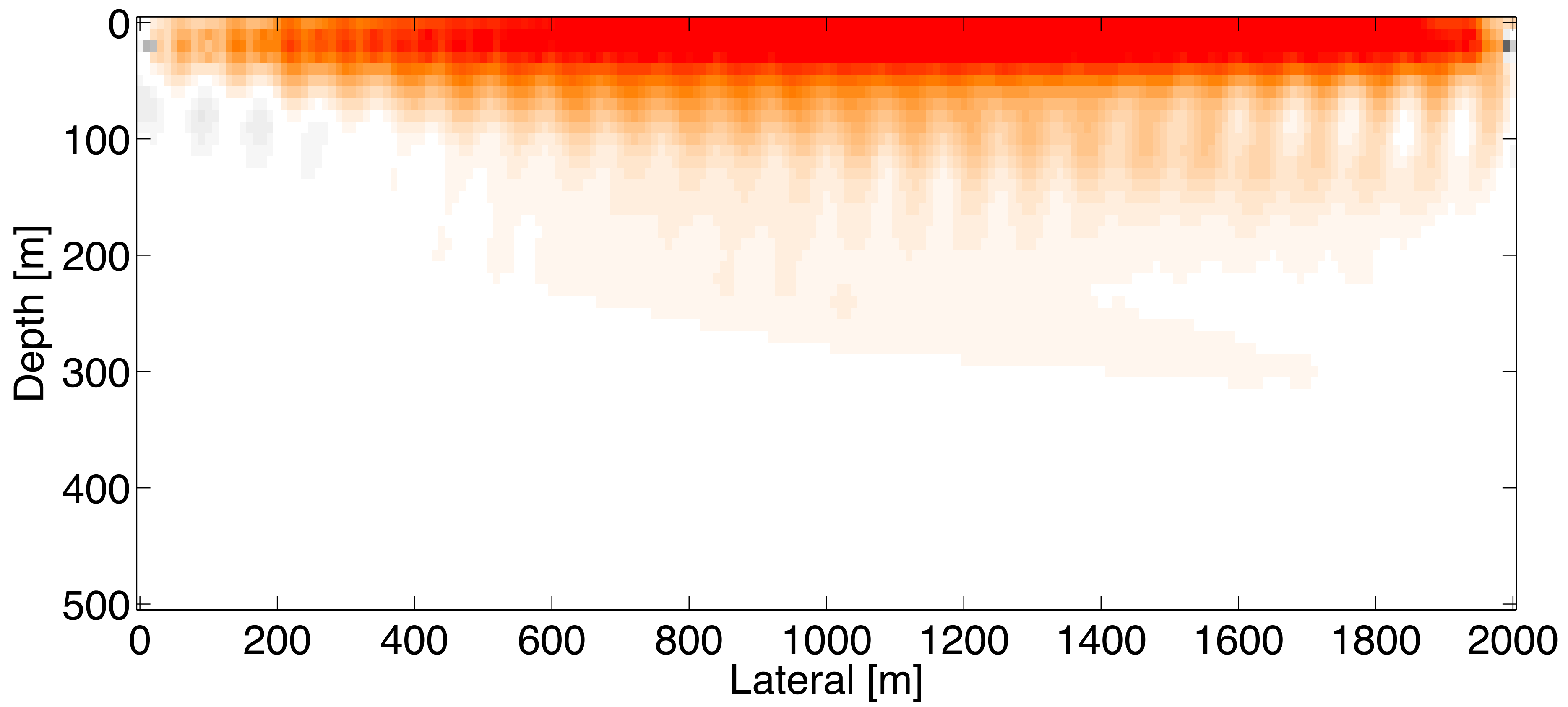
Data

+



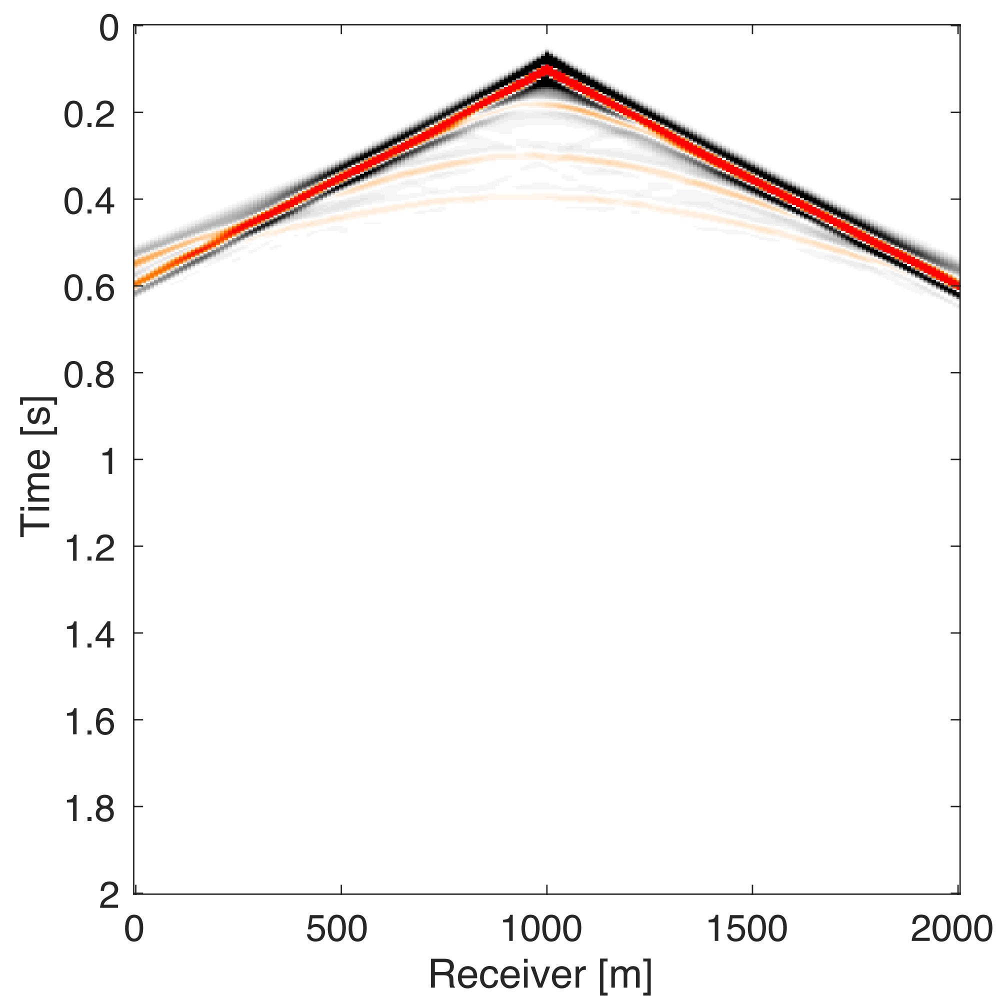
Correct source wavelet

→



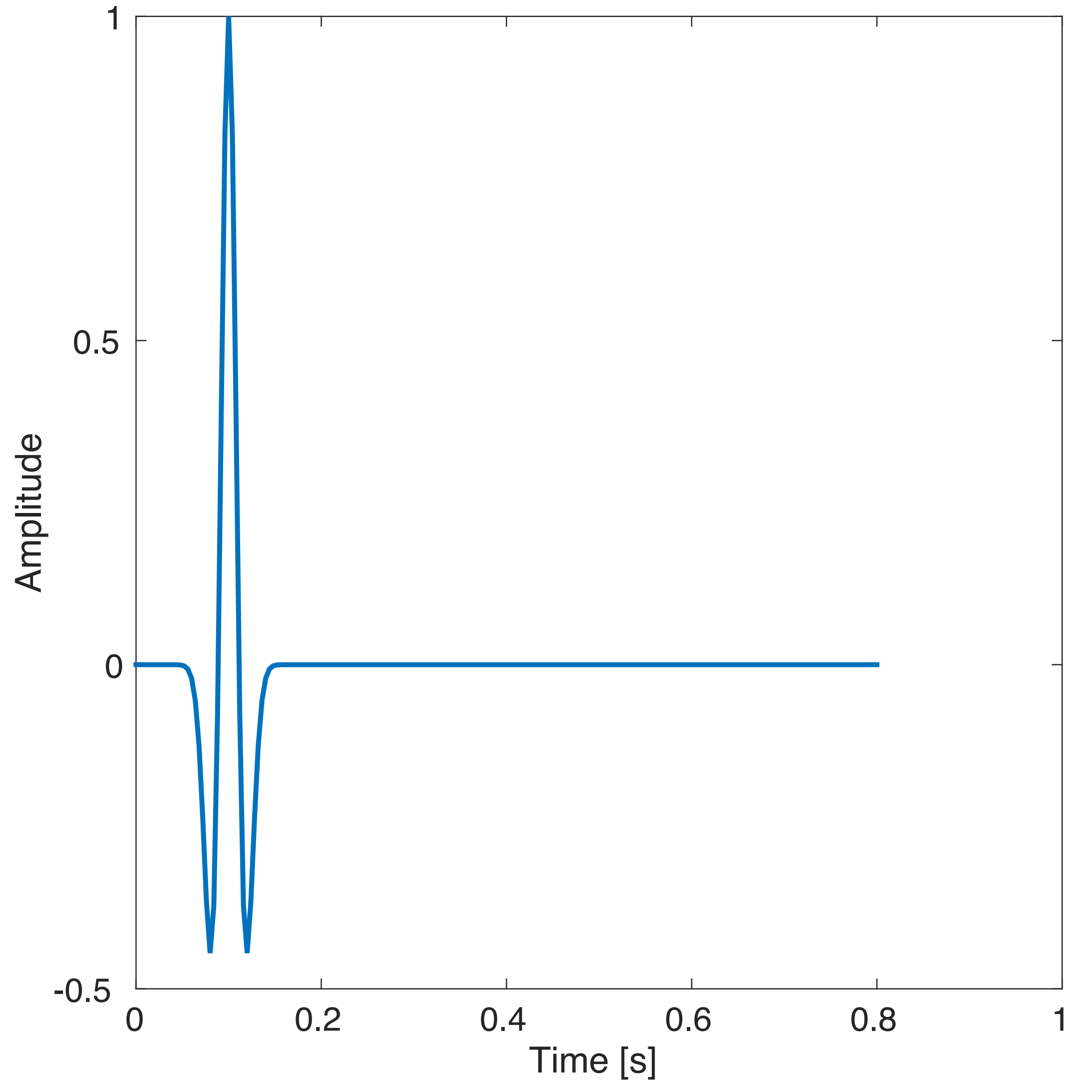
Correct gradient

Motivation

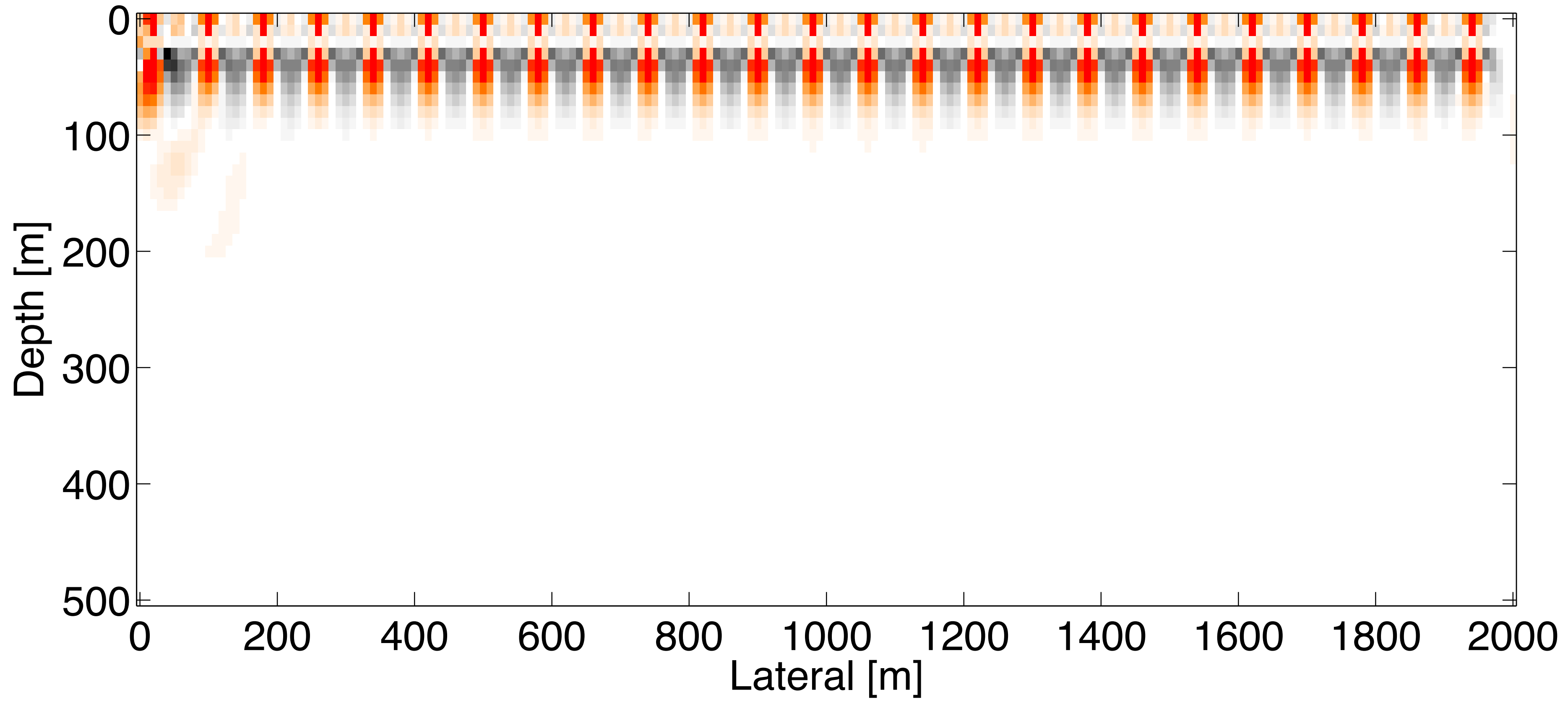


Data

+



Wrong source wavelet



Wrong gradient

Motivation

Questions:

1. How can we reduce the computational cost?

2. How can we obtain the correct source wavelets?

[Tarantola, 1984]

[J. Virieux and S. Operto, 2009]

SLIM

Full-waveform inversion

PDE-constrained optimization problem:

$$\min_{\mathbf{u},\mathbf{m}} \frac{1}{2N} \sum_{i=1}^{N_{\text{src}}} \sum_{l=1}^{N_{\text{freq}}} \|\mathbf{P}_i \mathbf{u}_{i,l} - \mathbf{d}_{i,l}\|_2^2$$

$$\text{subject to } \mathbf{A}_{i,l}(\mathbf{m})\mathbf{u}_{i,l} = \mathbf{q}_{i,l}$$

where $N = N_{\text{src}} N_{\text{freq}}$

[Tarantola, 1984]

[J. Virieux and S. Operto, 2009]

SLIM

Full-waveform inversion

Reduced/adjoint-state method:

$$\min_{\mathbf{m}} \frac{1}{2N} \sum_{i=1}^{N_{\text{src}}} \sum_{l=1}^{N_{\text{freq}}} \|\mathbf{P}_i \mathbf{A}_{i,l}(\mathbf{m})^{-1} \mathbf{q}_{i,l} - \mathbf{d}_{i,l}\|_2^2$$

with the gradient given by

$$\mathbf{g} = \frac{1}{N} \sum_{i=1}^{N_{\text{src}}} \sum_{l=1}^{N_{\text{freq}}} \mathbf{u}_{i,l}^\top \frac{\partial \mathbf{A}_{i,l}^\top}{\partial \mathbf{m}} \mathbf{v}_{i,l}$$

$$\mathbf{u}_{i,l} = \mathbf{A}_{i,l}(\mathbf{m})^{-1} \mathbf{q}_{i,l}$$

$$\mathbf{v}_{i,l} = -\mathbf{A}_{i,l}^{-\top}(\mathbf{m}) \mathbf{P}_i^\top \mathbf{r}_{i,l}$$

$$\mathbf{r}_{i,l} = \mathbf{P}_i \mathbf{A}_{i,l}(\mathbf{m})^{-1} \mathbf{q}_{i,l} - \mathbf{d}_{i,l}$$

2 PDE solves are required !

[van Leeuwen, T and Herrmann, F J , 2013]
[Peters, B, Herrmann, F J and van Leeuwen, T, 2014]
[Golub, G H and Pereyra, V , 2003]

Wavefield-reconstruction inversion

Penalty method:

$$\min_{\mathbf{u},\mathbf{m}} \frac{1}{2N} \sum_{i=1}^{N_{\text{src}}} \sum_{l=1}^{N_{\text{freq}}} \|\mathbf{P}_i \mathbf{u}_{i,l} - \mathbf{d}_{i,l}\|_2^2 + \lambda^2 \|\mathbf{A}_{i,l}(\mathbf{m})\mathbf{u}_{i,l} - \mathbf{q}_{i,l}\|_2^2$$

Eliminating $\mathbf{u}$ w/ variable projection:

$$\bar{\mathbf{u}} = \arg\min_{\mathbf{u}} \frac{1}{2N} \sum_{i=1}^{N_{\text{src}}} \sum_{l=1}^{N_{\text{freq}}} \|\mathbf{P}_i \mathbf{u}_{i,l} - \mathbf{d}_{i,l}\|_2^2 + \lambda^2 \|\mathbf{A}_{i,l}(\mathbf{m})\mathbf{u}_{i,l} - \mathbf{q}_{i,l}\|_2^2$$

[van Leeuwen, T and Herrmann, F J , 2013]

[Golub, G and Pereyra, V, 2003]

SLIM

Wavefield-reconstruction inversion

Corresponds to solving the following augmented system:

$$\begin{pmatrix} \lambda \mathbf{A}_{i,l} \\ \mathbf{P}_i \end{pmatrix} \overline{\mathbf{u}}_{i,l} = \begin{pmatrix} \lambda \mathbf{q}_{i,l} \\ \mathbf{d}_{i,l} \end{pmatrix}$$

with the gradient

1 augmented system solves is required !

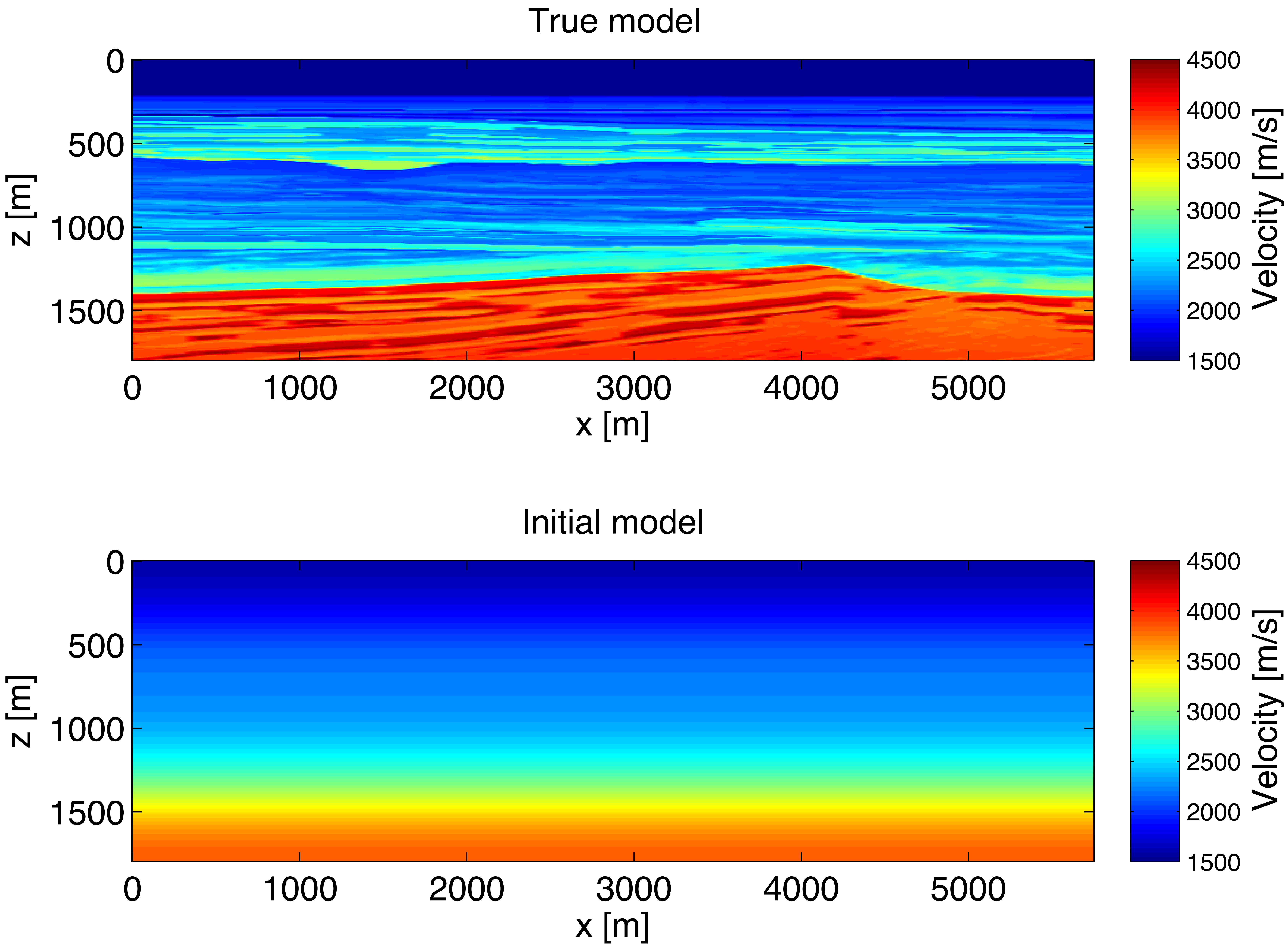
$$\mathbf{g} = \frac{1}{N} \sum_{i=1}^{N_{\mathrm{src}}} \sum_{l=1}^{N_{\mathrm{freq}}} \overline{\mathbf{u}}_{i,l}^{\top} \frac{\partial \mathbf{A}_{i,l}^{\top}}{\partial \mathbf{m}} \overline{\mathbf{v}}_{i,l}$$

$$\overline{\mathbf{v}}_{i,l} = \mathbf{A}_{i,l}(\mathbf{m}) \overline{\mathbf{u}}_{i,l} - \mathbf{q}_{i,l}$$

True & initial model

[van Leeuwen, T and Herrmann, F J , 2013]

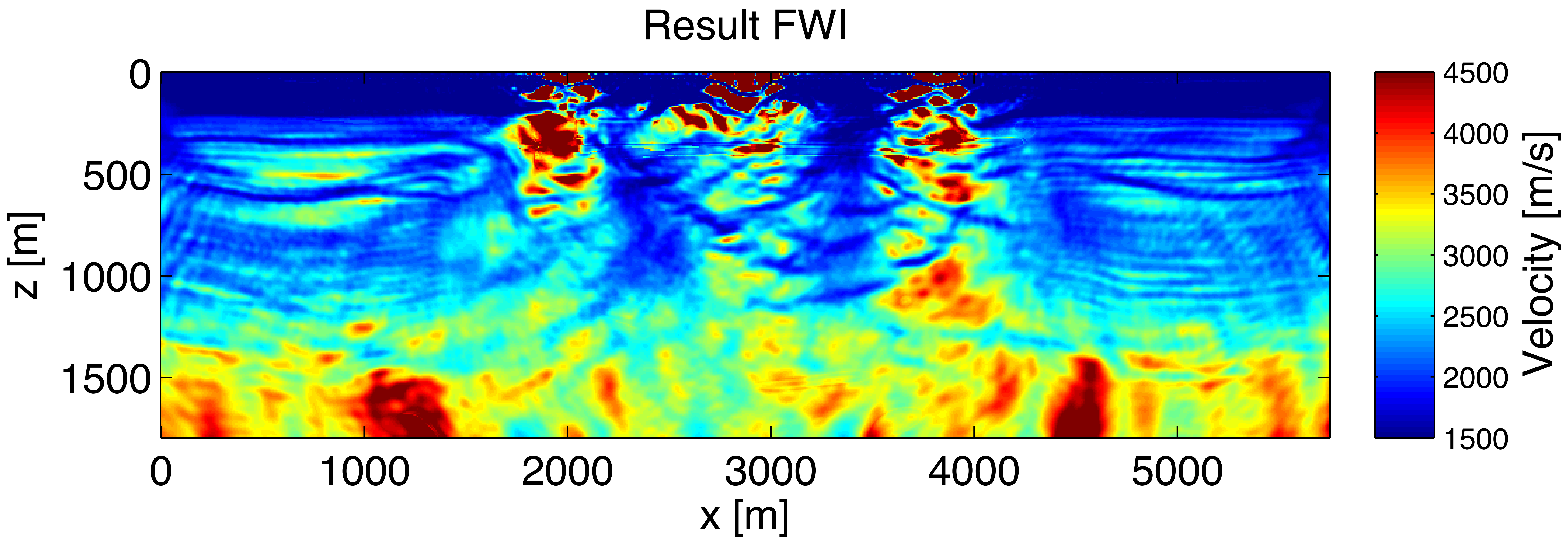
[Peters, B, Herrmann, F J and van Leeuwen, T, 2014]

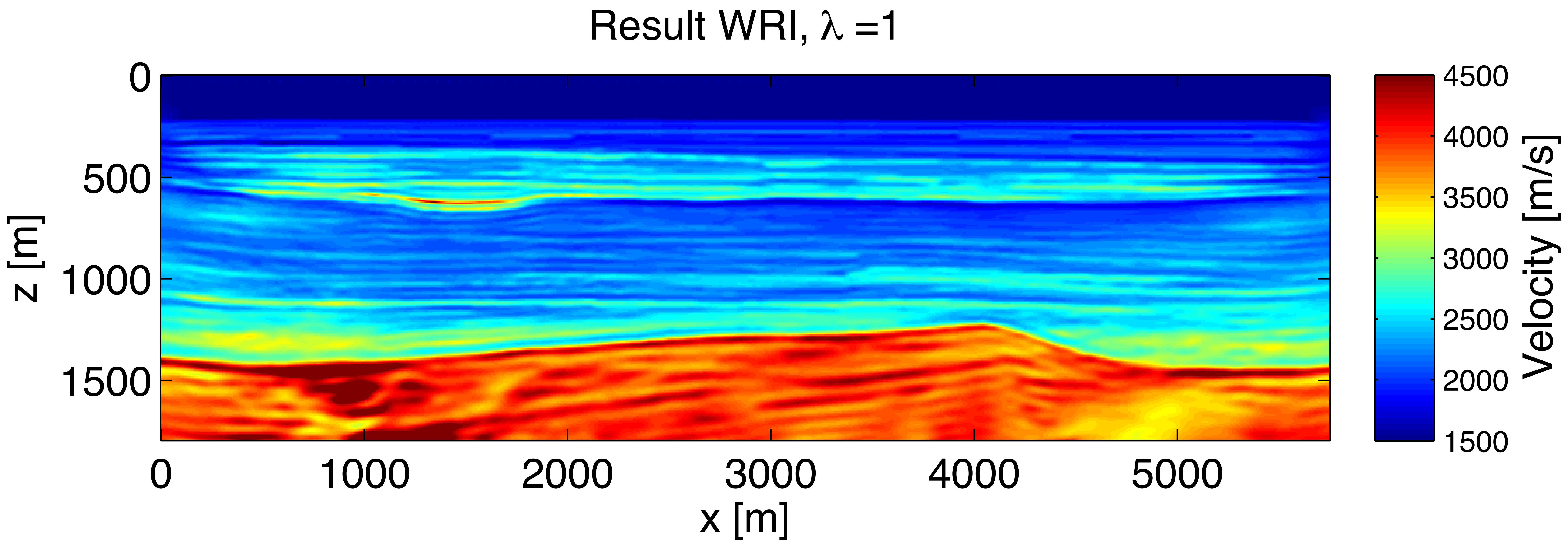


FWI vs WRI

[van Leeuwen, T and Herrmann, F J , 2013]

[Peters, B, Herrmann, F J and van Leeuwen, T, 2014]





WRI with source estimation

Triple parameters optimization problem:

$$\min_{\mathbf{u},\mathbf{m},\alpha} \frac{1}{2N} \sum_{i=1}^{N_{\mathrm{src}}} \sum_{l=1}^{N_{\mathrm{freq}}} \|\mathbf{P}_i \mathbf{u}_{i,l} - \mathbf{d}_{i,l}\|_2^2 + \lambda^2 \|\mathbf{A}_{i,l}(\mathbf{m})\mathbf{u}_{i,l} - \alpha_{i,l}\mathbf{e}_{i,l}\|_2^2$$

SLIM

WRI with source estimation

Triple parameters optimization problem:

$$\min_{\mathbf{u},\mathbf{m},\alpha} \frac{1}{2N} \sum_{i=1}^{N_{\text{src}}} \sum_{l=1}^{N_{\text{freq}}} \|\mathbf{P}_i \mathbf{u}_{i,l} - \mathbf{d}_{i,l}\|_2^2 + \lambda^2 \|\mathbf{A}_{i,l}(\mathbf{m})\mathbf{u}_{i,l} - \alpha_{i,l}\mathbf{e}_{i,l}\|_2^2$$

Eliminate $\mathbf{u}$ and α jointly w/ variable projection:

$$[\overline{\mathbf{u}}, \overline{\alpha}] =$$

$$\arg\min_{\mathbf{u},\alpha} \frac{1}{2N} \sum_{i=1}^{N_{\text{src}}} \sum_{l=1}^{N_{\text{freq}}} \|\mathbf{P}_i \mathbf{u}_{i,l} - \mathbf{d}_{i,l}\|_2^2 + \lambda^2 \|\mathbf{A}_{i,l}(\mathbf{m})\mathbf{u}_{i,l} - \alpha_{i,l}\mathbf{e}_{i,l}\|_2^2$$

WRI with source estimation

Corresponds to solving the following augmented system:

$$\begin{pmatrix} \lambda \mathbf{A}_{i,l} & -\lambda \mathbf{e}_{i,l} \\ \mathbf{P}_i & 0 \end{pmatrix} \begin{pmatrix} \overline{\mathbf{u}}_{i,l} \\ \overline{\alpha}_{i,l} \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{d}_{i,l} \end{pmatrix}$$

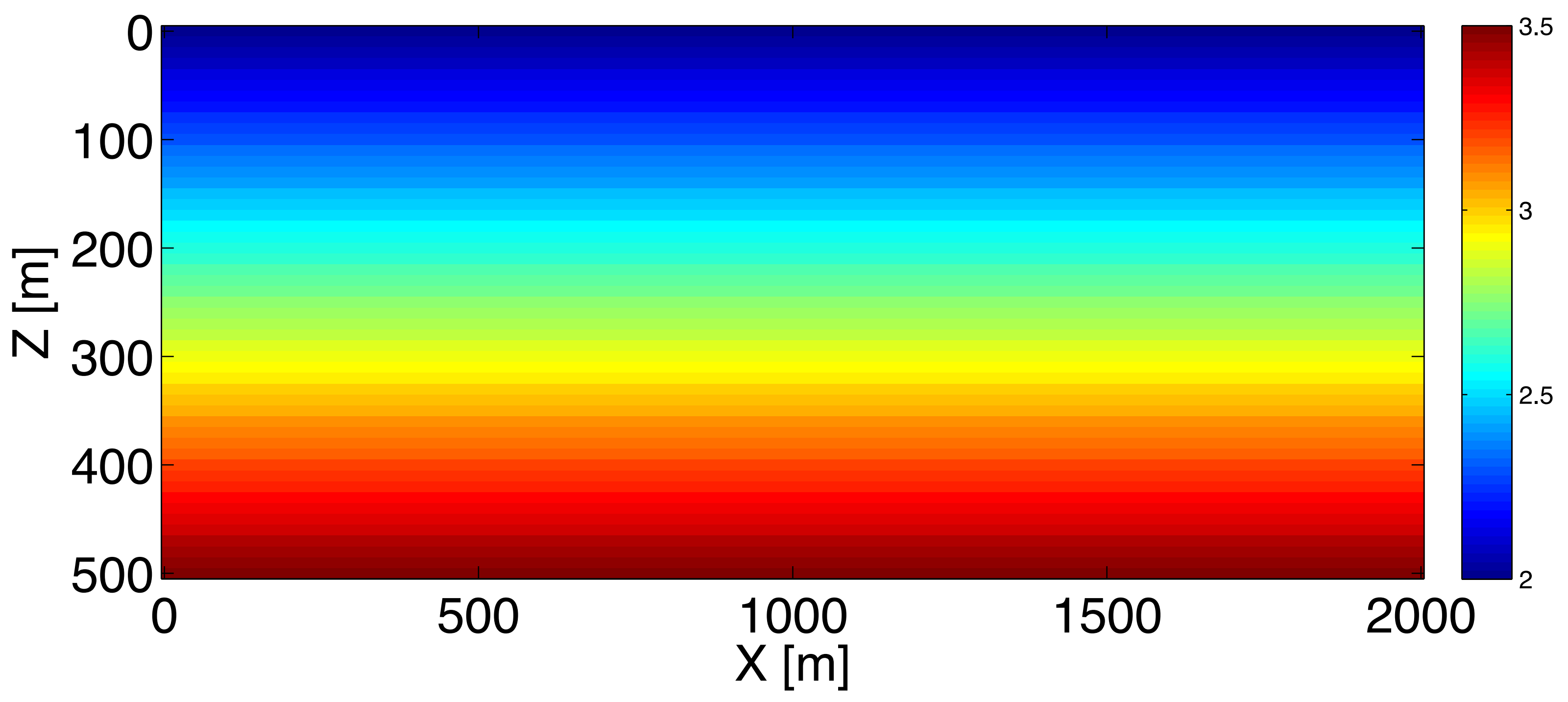
Cf. original augmented system:

Full column rank!

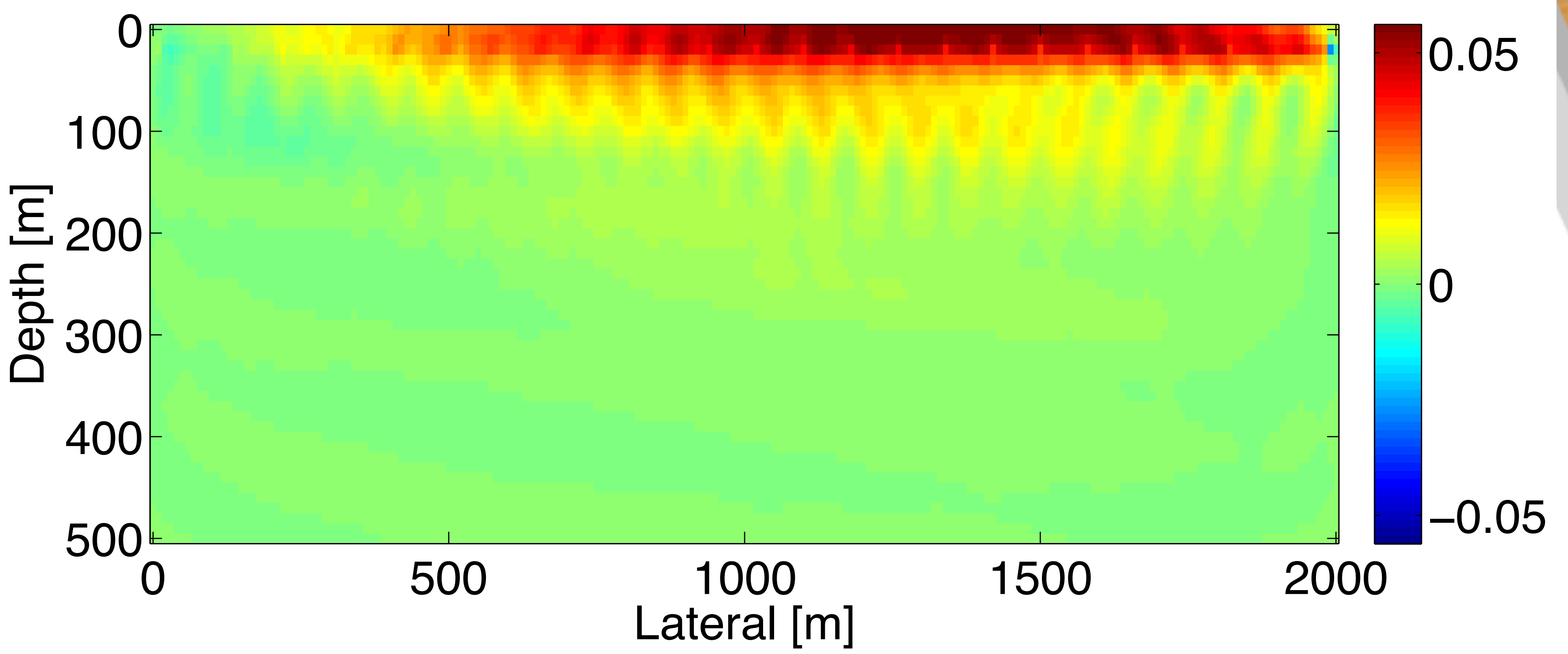
No additional computational cost!

$$\begin{pmatrix} \lambda \mathbf{A}_{i,l} \\ \mathbf{P}_i \end{pmatrix} \overline{\mathbf{u}}_{i,l} = \begin{pmatrix} \lambda \mathbf{q}_{i,l} \\ \mathbf{d}_{i,l} \end{pmatrix}$$

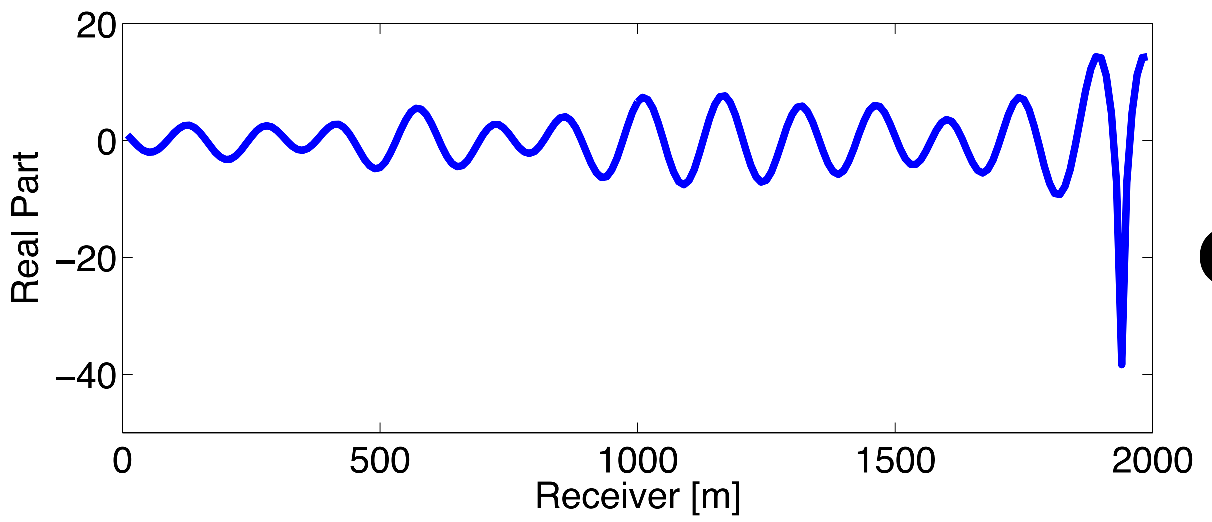
WRI with source estimation



$\mathbf{m}$



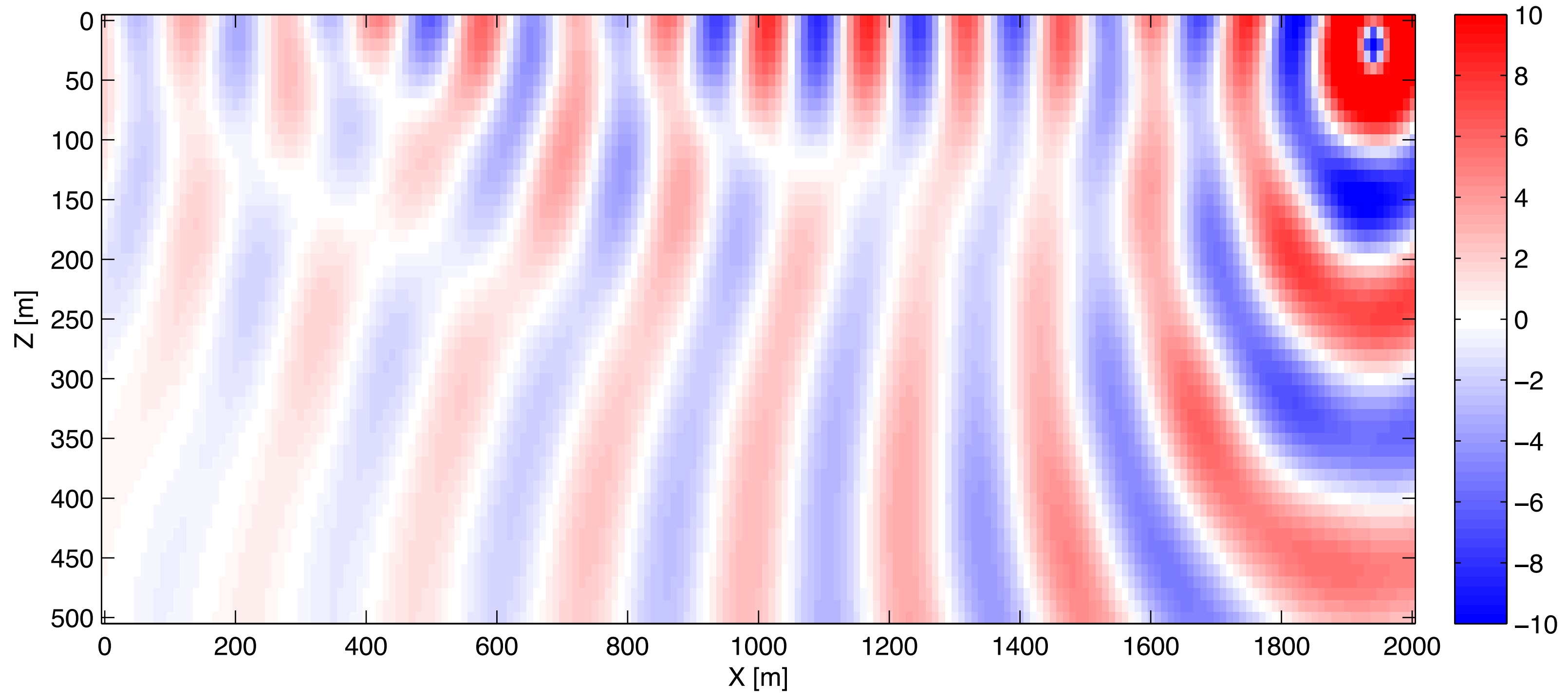
$\mathbf{g}$



$\mathbf{d_{obs}}$

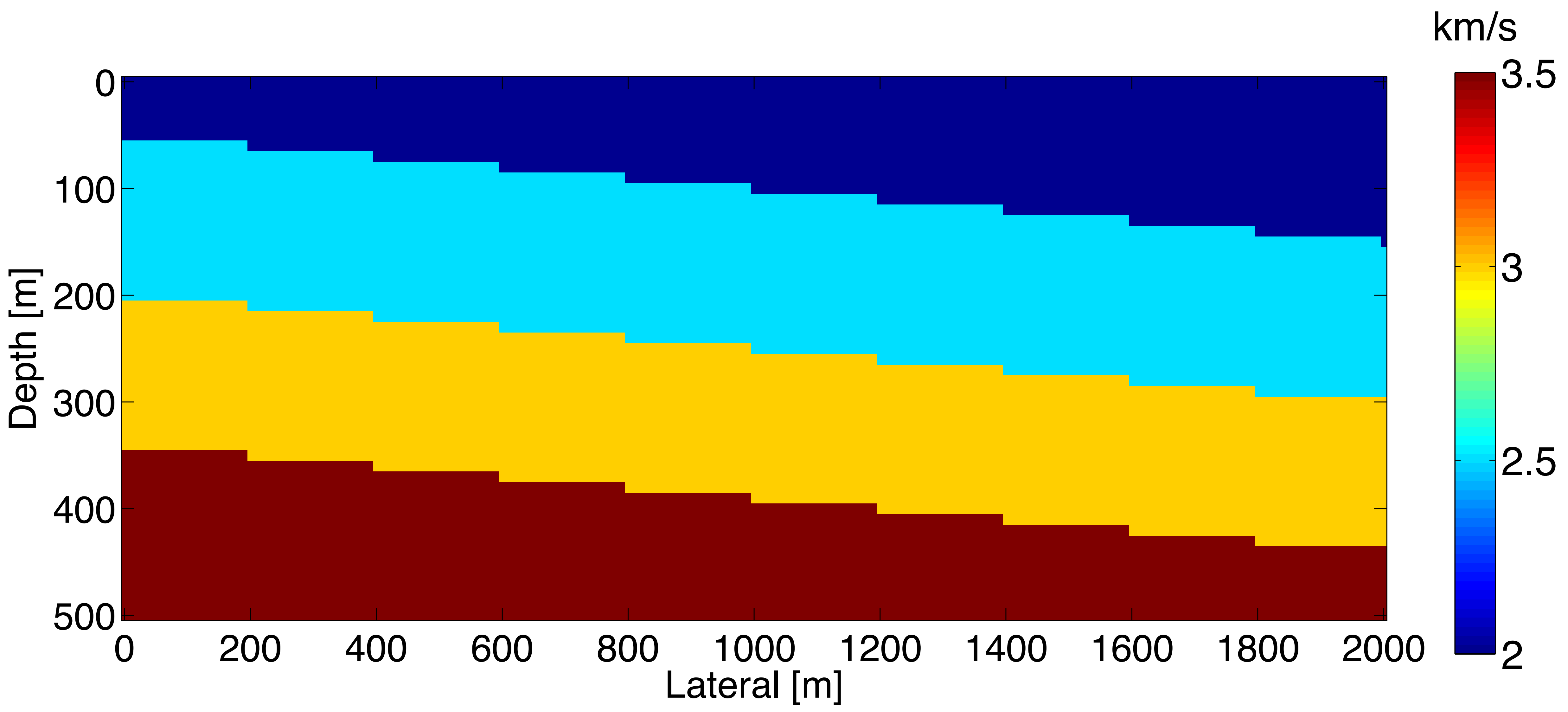
$$\begin{pmatrix} \lambda\mathbf{A} & -\lambda\mathbf{e} \\ \mathbf{P} & 0 \end{pmatrix} \begin{pmatrix} \overline{\mathbf{u}} \\ \overline{\alpha} \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{d_{obs}} \end{pmatrix}$$

$\overline{\alpha}$ and

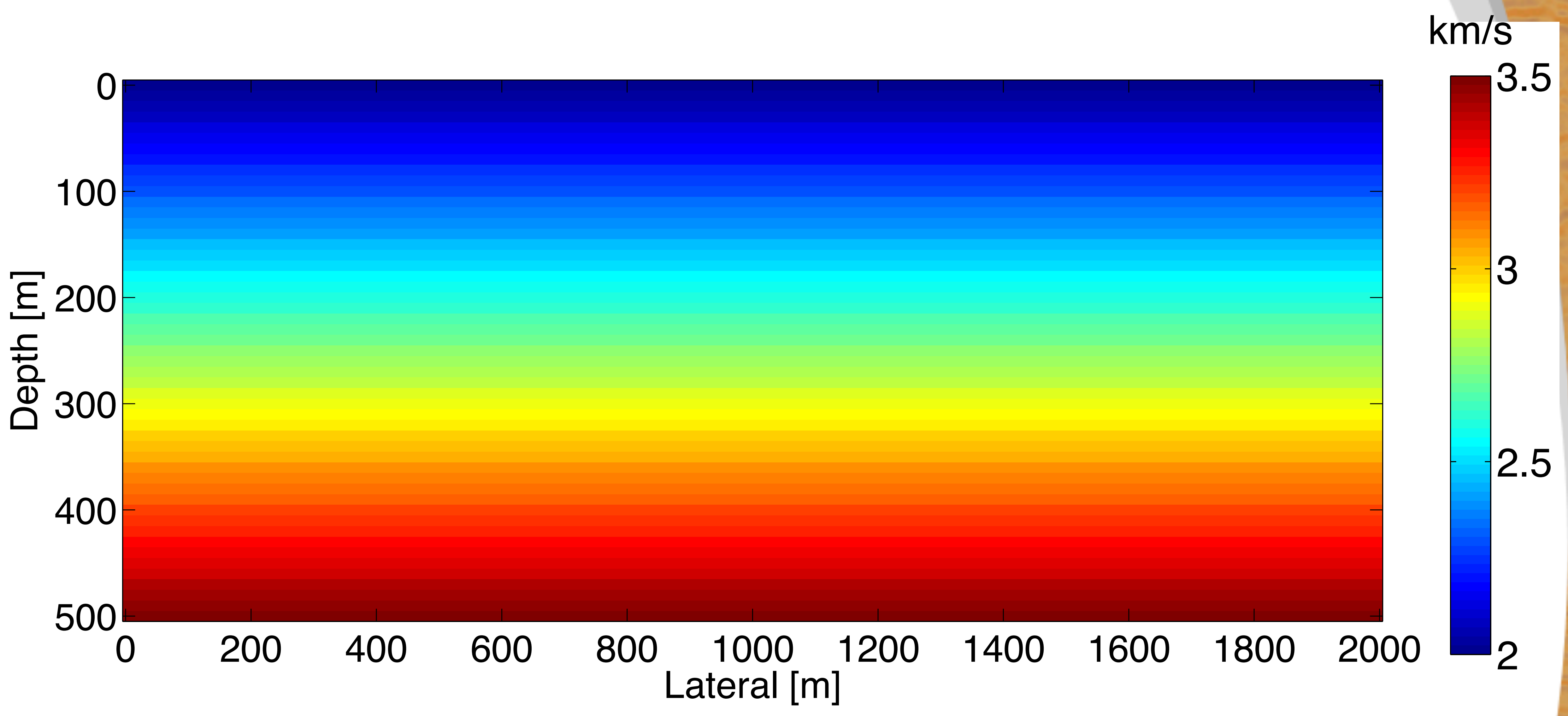


$\overline{\mathbf{u}}$

Synthetic example

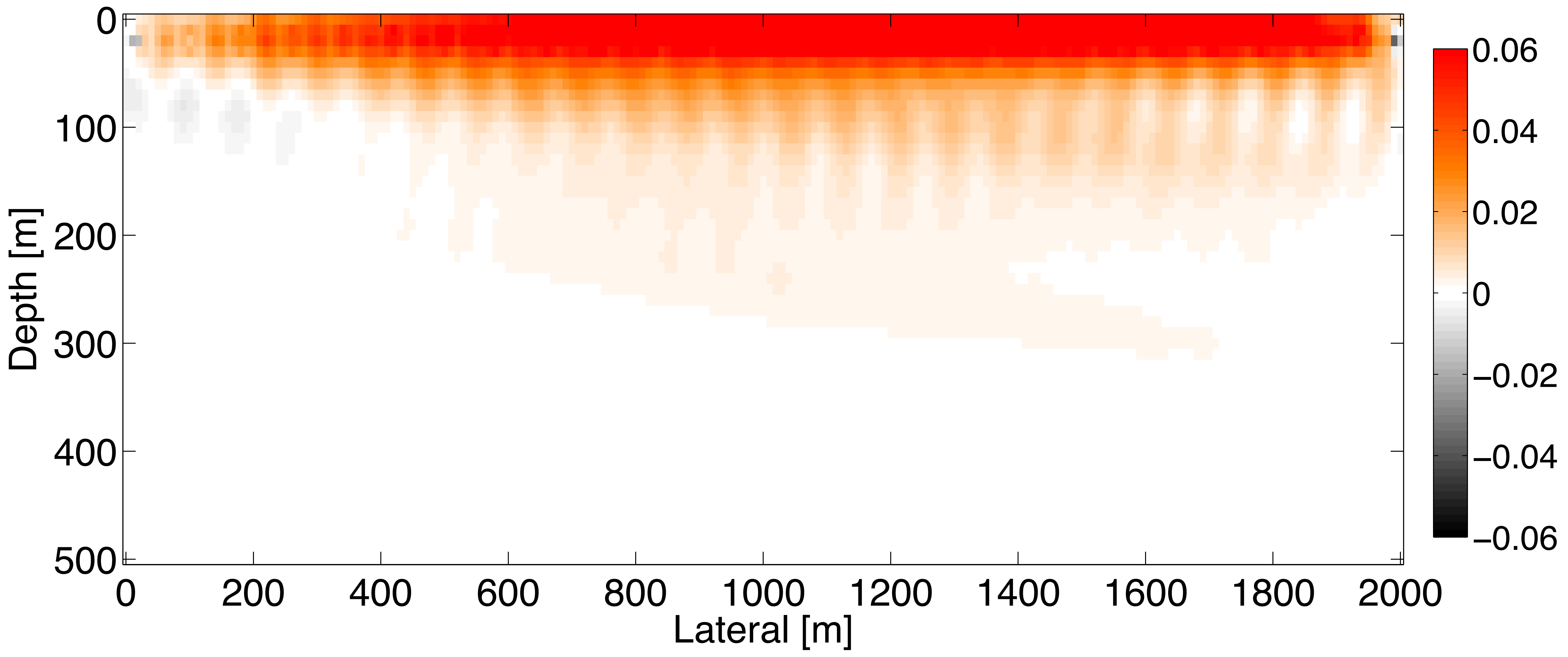


True Model

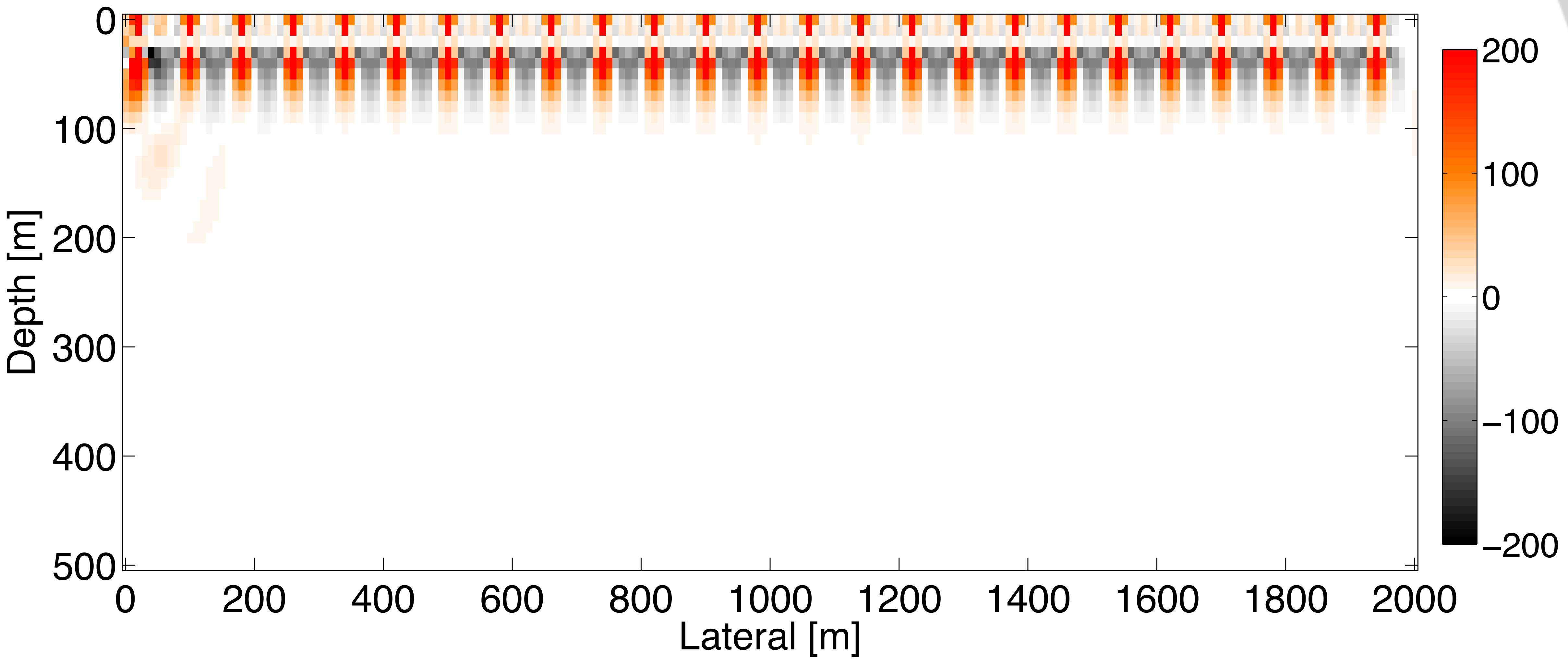


Initial Model

Gradient comparison

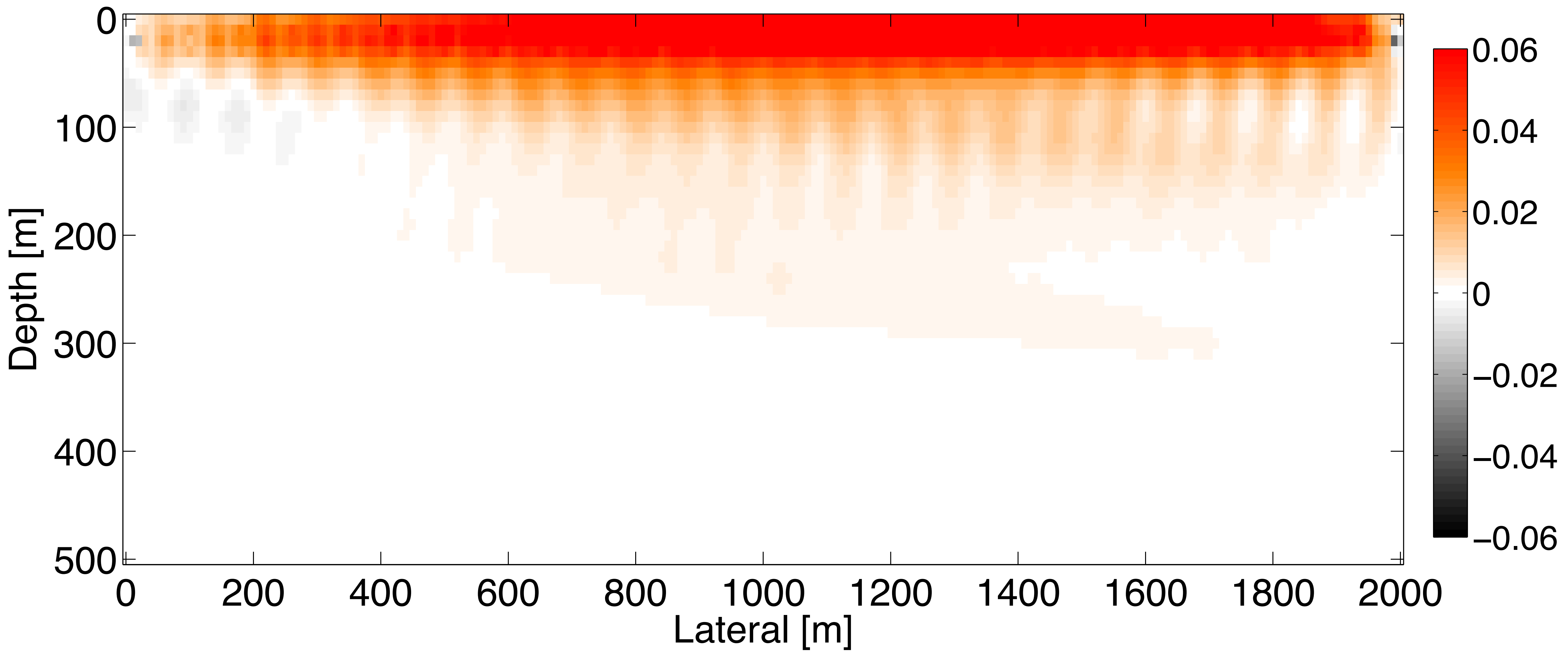


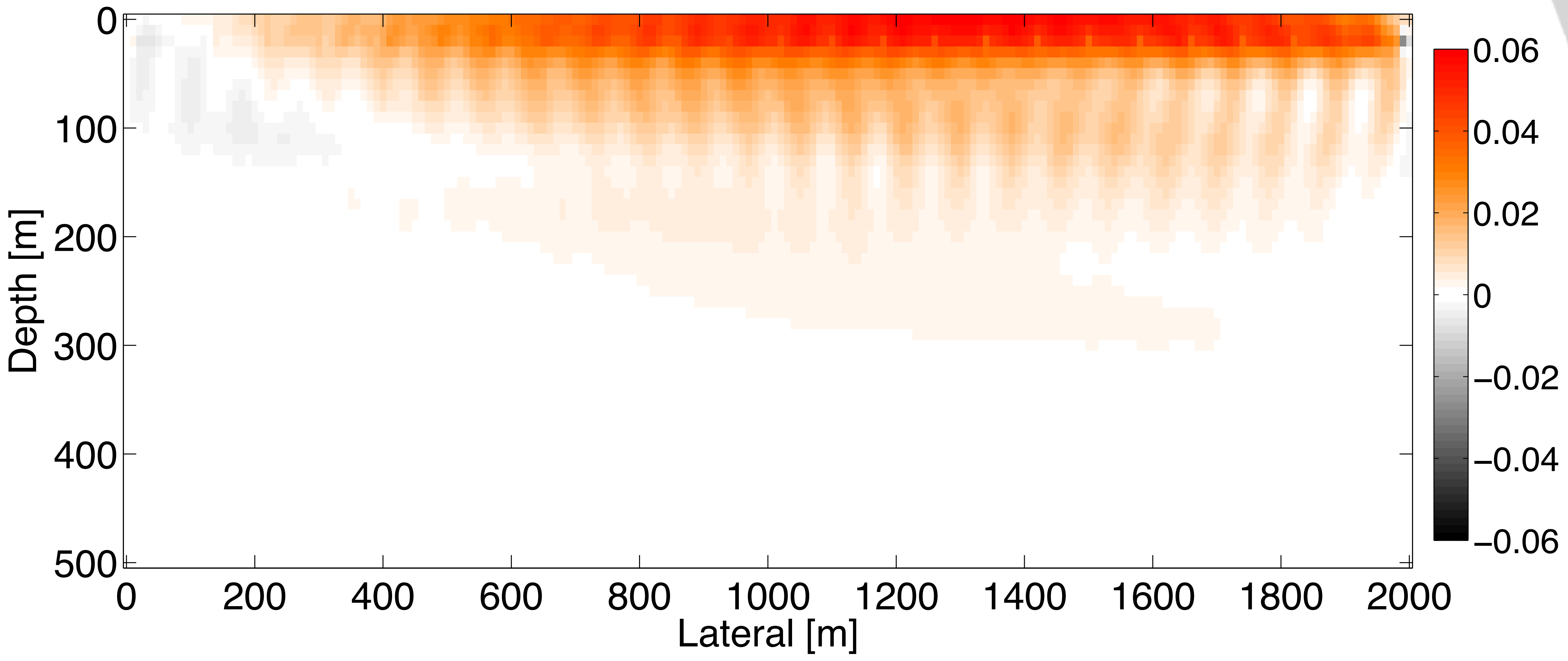
Gradient with true source wavelet



Gradient with wrong source wavelet

Gradient comparison





Gradient with true source wavelet

Gradient with estimated source wavelet

WRI with source estimation

Objective function:

$$\varphi(\mathbf{m}) = \frac{1}{2N} \sum_{i=1}^{N_{\text{src}}} \sum_{l=1}^{N_{\text{freq}}} \|\mathbf{P}_i \overline{\mathbf{u}}_{i,l}(\mathbf{m}) - \mathbf{d}_{i,l}\|_2^2 + \lambda^2 \|\mathbf{A}_{i,l}(\mathbf{m}) \overline{\mathbf{u}}_{i,l}(\mathbf{m}) - \alpha_{i,l}(\mathbf{m}) \mathbf{e}_{i,l}\|_2^2$$

with

$$\begin{pmatrix} \lambda \mathbf{A}_{i,l} & -\lambda \mathbf{e}_{i,l} \\ \mathbf{P}_i & 0 \end{pmatrix} \begin{pmatrix} \overline{\mathbf{u}}_{i,l} \\ \overline{\alpha}_{i,l} \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{d}_{i,l} \end{pmatrix}$$

[Nesterov, 2004]

SLIM

WRI with source estimation

Objective function:

$$\varphi(\mathbf{m}) = \frac{1}{2N}\sum_{i=1}^{N_{\mathrm{src}}}\sum_{l=1}^{N_{\mathrm{freq}}} \|\mathbf{P}_i\overline{\mathbf{u}}_{i,l}(\mathbf{m}) - \mathbf{d}_{i,l}\|_2^2 + \lambda^2\|\mathbf{A}_{i,l}(\mathbf{m})\overline{\mathbf{u}}_{i,l}(\mathbf{m}) - \alpha_{i,l}(\mathbf{m})\mathbf{e}_{i,l}\|_2^2$$

with

$$\begin{pmatrix} \lambda\mathbf{A}_{i,l} & -\lambda\mathbf{e}_{i,l} \\ \mathbf{P}_i & 0 \end{pmatrix} \begin{pmatrix} \overline{\mathbf{u}}_{i,l} \\ \overline{\alpha}_{i,l} \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{d}_{i,l} \end{pmatrix}$$

expensive ~ one PDE solve

Stochastic optimization

Full objective function:

$$\min_{\mathbf{m}} \varphi(\mathbf{m}) = \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{m})$$

Full gradient (FG):

$$\min_{\mathbf{m}} G(\mathbf{m}) = \frac{1}{N}\sum_{i=1}^{N} g_i(\mathbf{m})$$

Challenge:

- N is large
- computing $\{f_i(\mathbf{m})\}_{1\leq i\leq N}$ and $\{g_i(\mathbf{m})\}_{1\leq i\leq N}$ are expensive

[Nesterov, 2004]

SLIM

Stochastic optimization

Full gradient method:

$$\mathbf{m}^{k+1} = \mathbf{m}^k - \alpha_k G(\mathbf{m}^k) = \mathbf{m}^k - \frac{\alpha_k}{N}\sum_{i=1}^{N} g_i(\mathbf{m}^k)$$

Linear convergence rate:

$$\varphi(\mathbf{m}^k) - \varphi(\mathbf{m}^*) = \mathcal{O}(\rho^k)$$

for some $\rho < 1$

[Nemirovski and Yudin, 1983]
[Nemirovski *et al.*, 2009]
[Agarwal *et al.*, 2012]

Stochastic optimization

Stochastic objective function:

$$\min_{\mathbf{m}} \overline{\varphi}_k(\mathbf{m}) = \frac{1}{n_{\mathcal{I}_k}} \sum_{i \in \mathcal{I}_k} f_i(\mathbf{m})$$

Stochastic gradient (SG):

$$\overline{G}_k(\mathbf{m}) = \frac{1}{n_{\mathcal{I}_k}} \sum_{i \in \mathcal{I}_k} g_i(\mathbf{m})$$

[Nemirovski and Yudin, 1983]
[Nemirovski *et al*., 2009]
[Agarwal *et al*., 2012]

Stochastic optimization

Stochastic gradient method:

$$\mathbf{m}^{k+1} = \mathbf{m}^k - \alpha_k \overline{G}_k(\mathbf{m}^k)$$

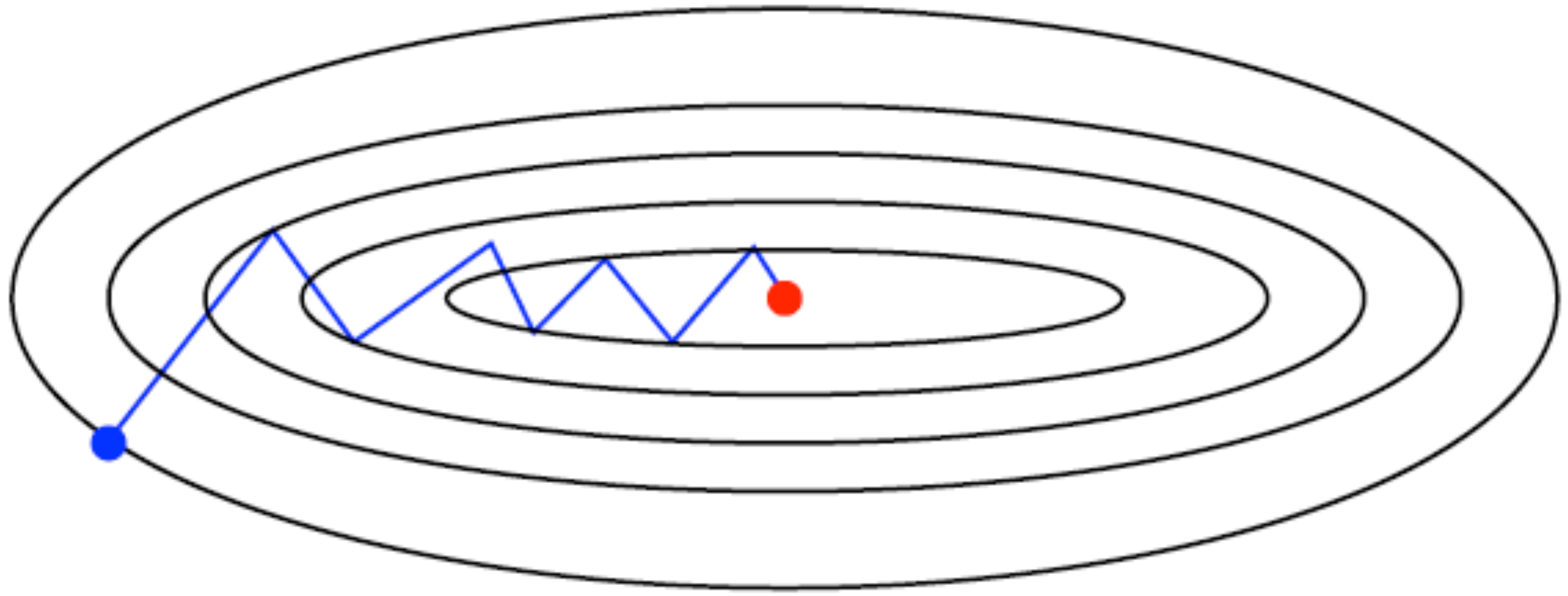
Sublinear convergence rate:

$$\mathbb{E}[\varphi(\mathbf{m}^k)] - \varphi(\mathbf{m}^*) = \mathcal{O}(1/k)$$

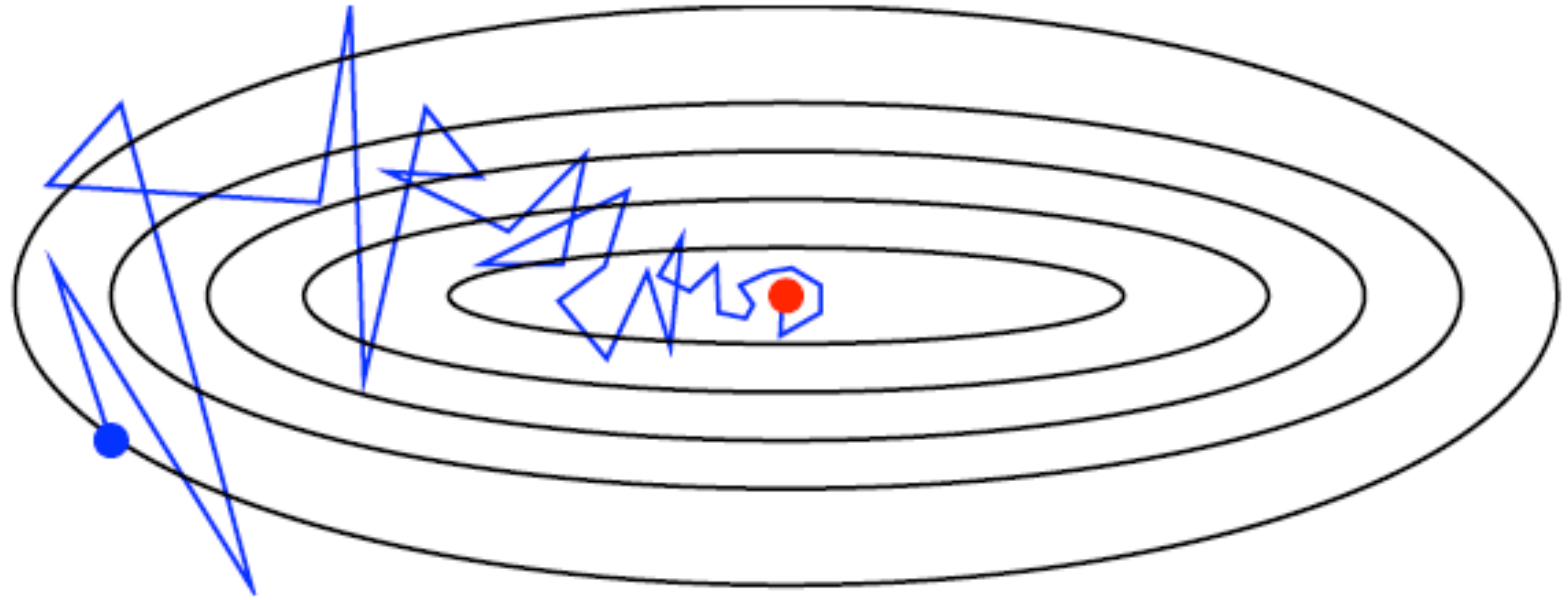
[Schmidt 2013]

FG vs SG

Full gradient method:



Stochastic gradient method:

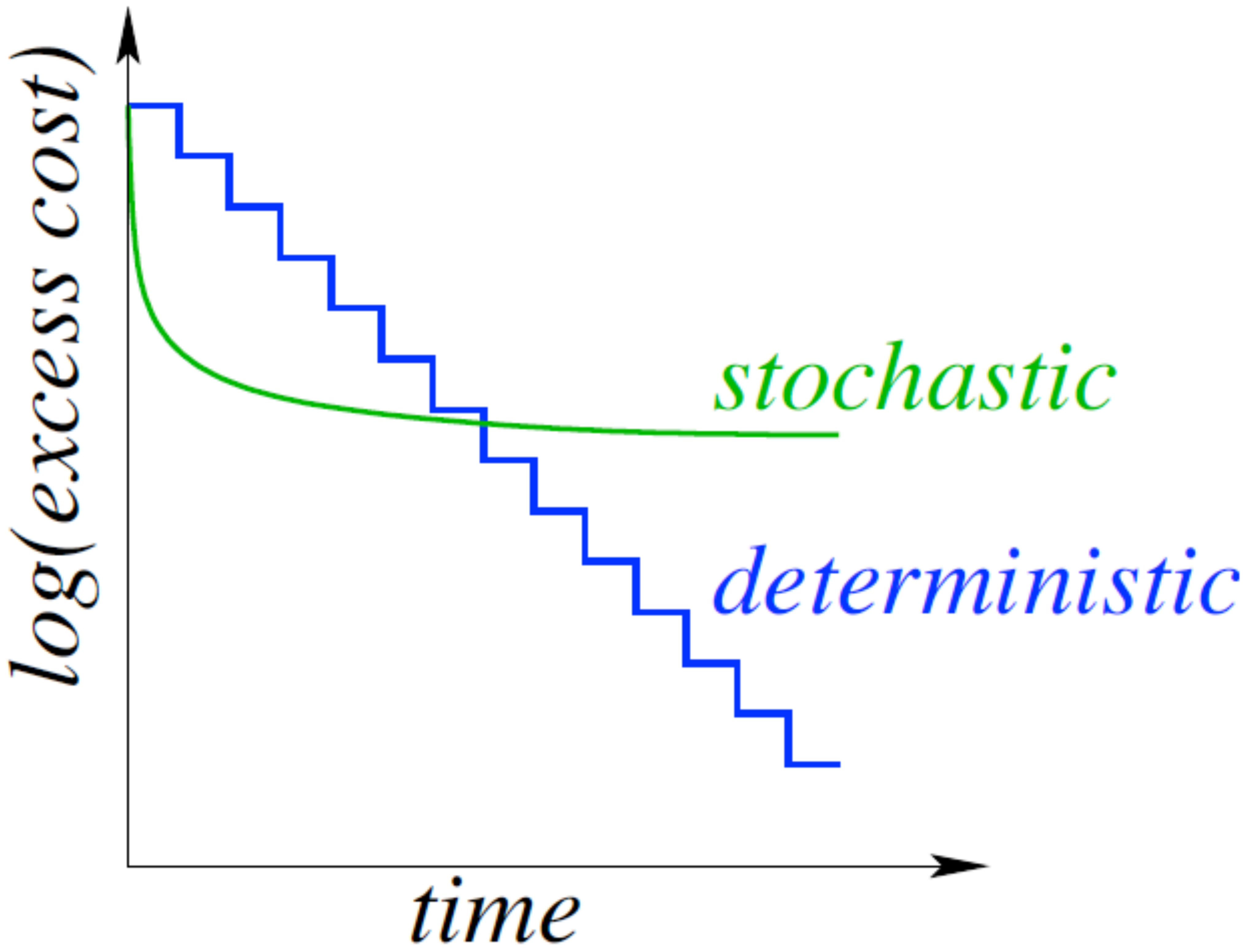


[Schmidt 2013]

FG vs SG

Convergence comparison:

- FG method has O(N) cost with $\mathcal{O}(\rho^k)$ rate;
- SG method has O(1) cost with O(1/k) rate;

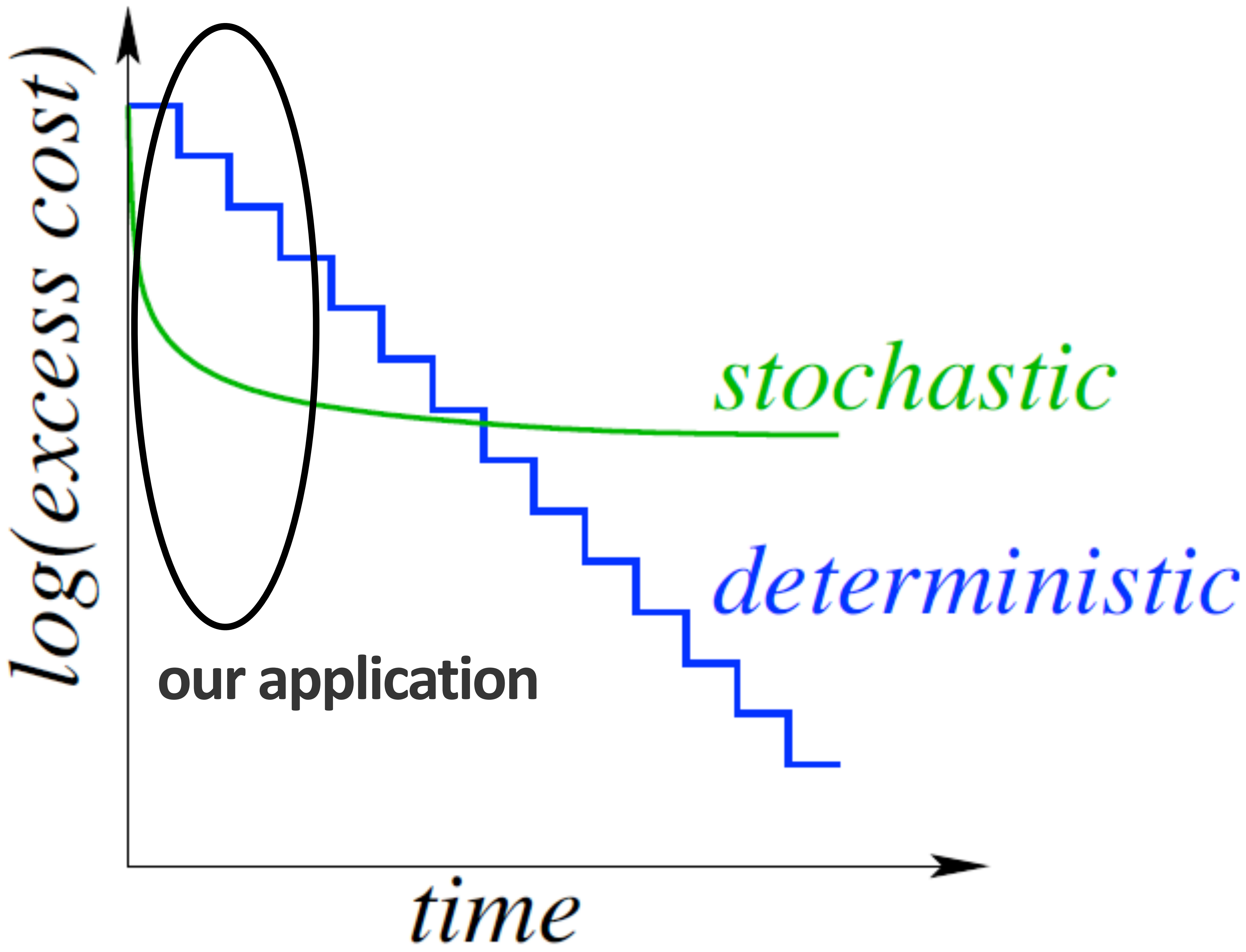


[Schmidt 2013]

FG vs SG

Convergence comparison:

- FG method has O(N) cost with $\mathcal{O}(\rho^k)$ rate;
- SG method has O(1) cost with O(1/t) rate;



Stochastic WRI

Stochastic objective function:

$$\overline{\varphi}(\mathbf{m}) = \frac{1}{n_{\mathcal{I}_k}} \sum_{i \in \mathcal{I}_k} f_i(\mathbf{m})$$

with

$$f_i(\mathbf{m}) = \frac{1}{2N_{\text{freq}}} \sum_{l=1}^{N_{\text{freq}}} \|\mathbf{P}_i \overline{\mathbf{u}}_{i,l}(\mathbf{m}) - \mathbf{d}_{i,l}\|_2^2 + \lambda^2 \|\mathbf{A}_{i,l}(\mathbf{m})\overline{\mathbf{u}}_{i,l}(\mathbf{m}) - \alpha_{i,l}(\mathbf{m})\mathbf{e}_{i,l}\|_2^2$$

Stochastic gradient:

$$\overline{G}_k(\mathbf{m}) = \frac{1}{n_{\mathcal{I}_k}} \sum_{i \in \mathcal{I}_k} g_i(\mathbf{m})$$

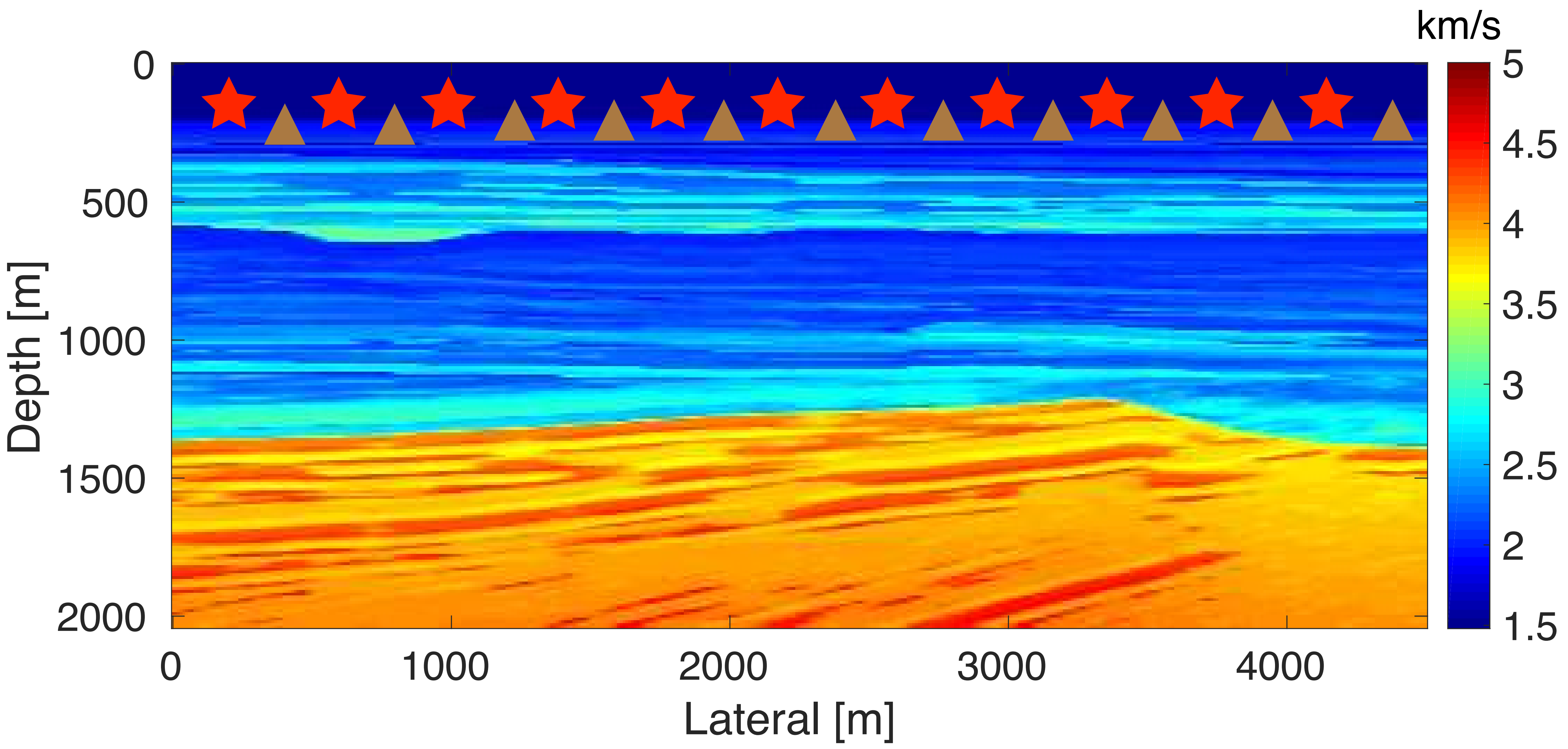
Stochastic WRI

Stochastic l-BFGS:

$$\mathbf{m}^{k+1} = \mathbf{m}^k - \alpha_k \overline{H}_k^{-1}(\mathbf{m}^k)\overline{G}_k(\mathbf{m}^k)$$

where $\overline{H}_k(\mathbf{m}^k)$ is the l-BFGS Hessian

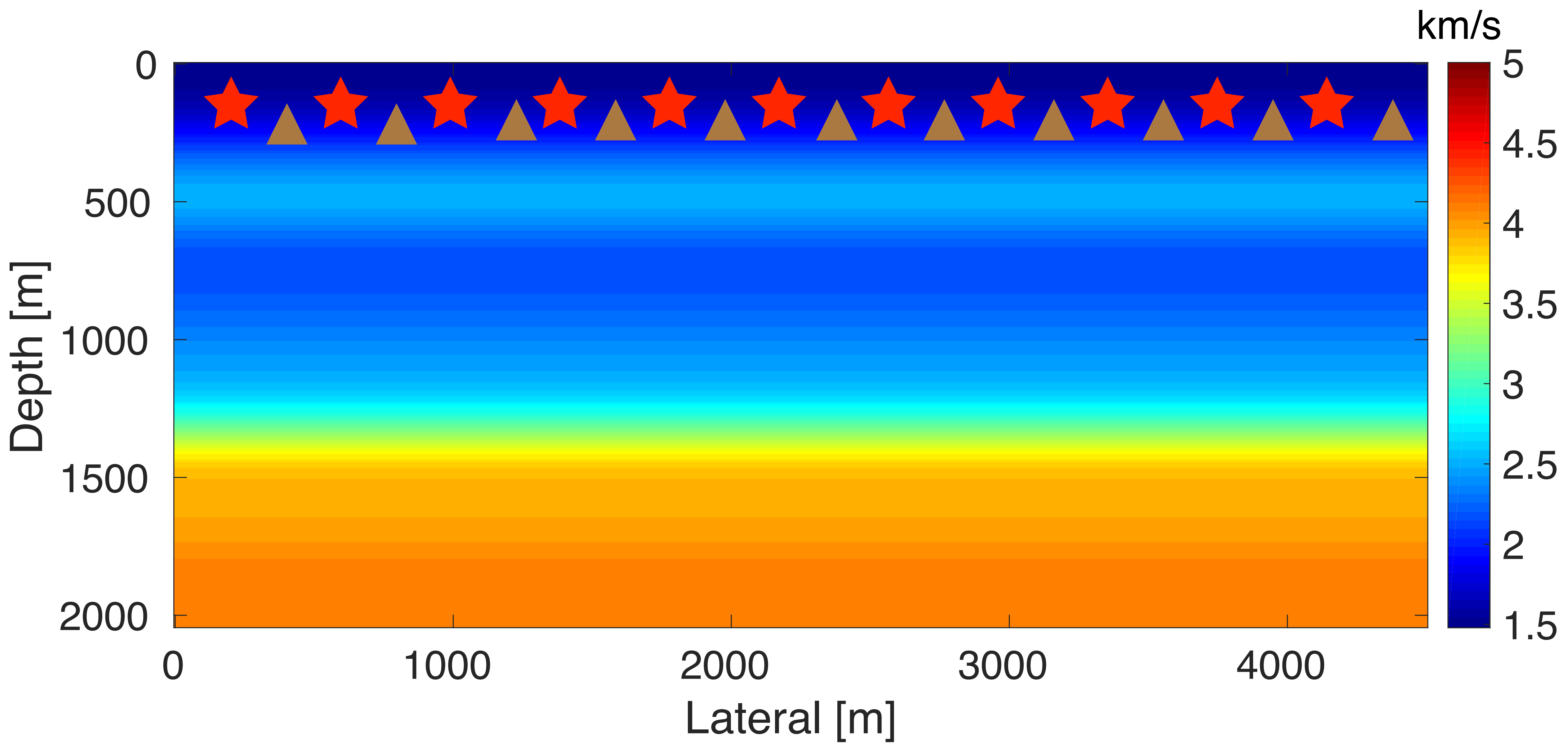
BG model



True model

Modeling information:
Model size: 2000m x 4500m
Source spacing: 50m
Receiver spacing: 10m
Fixed spread 4.5km
Frequency : 2~31 Hz

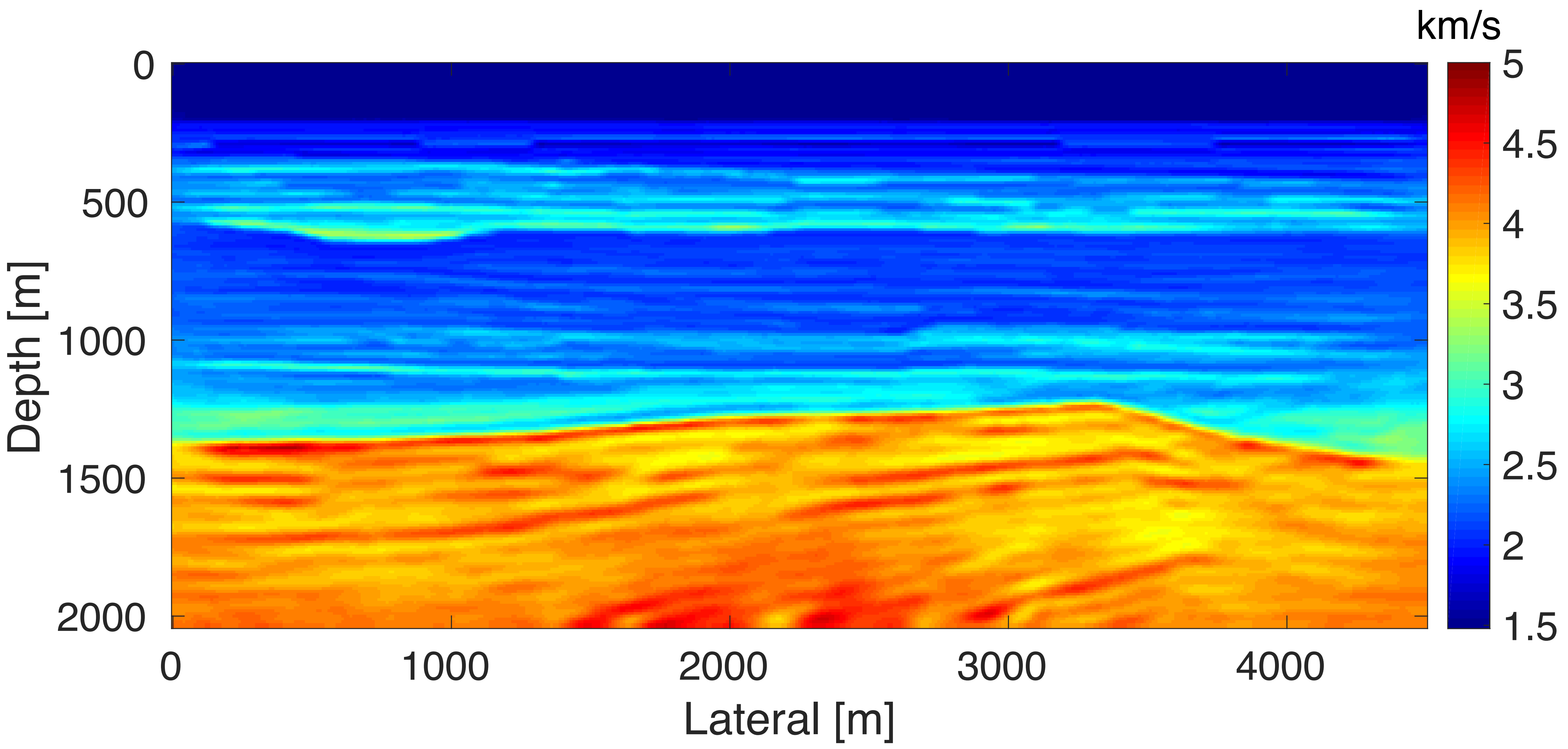
BG model



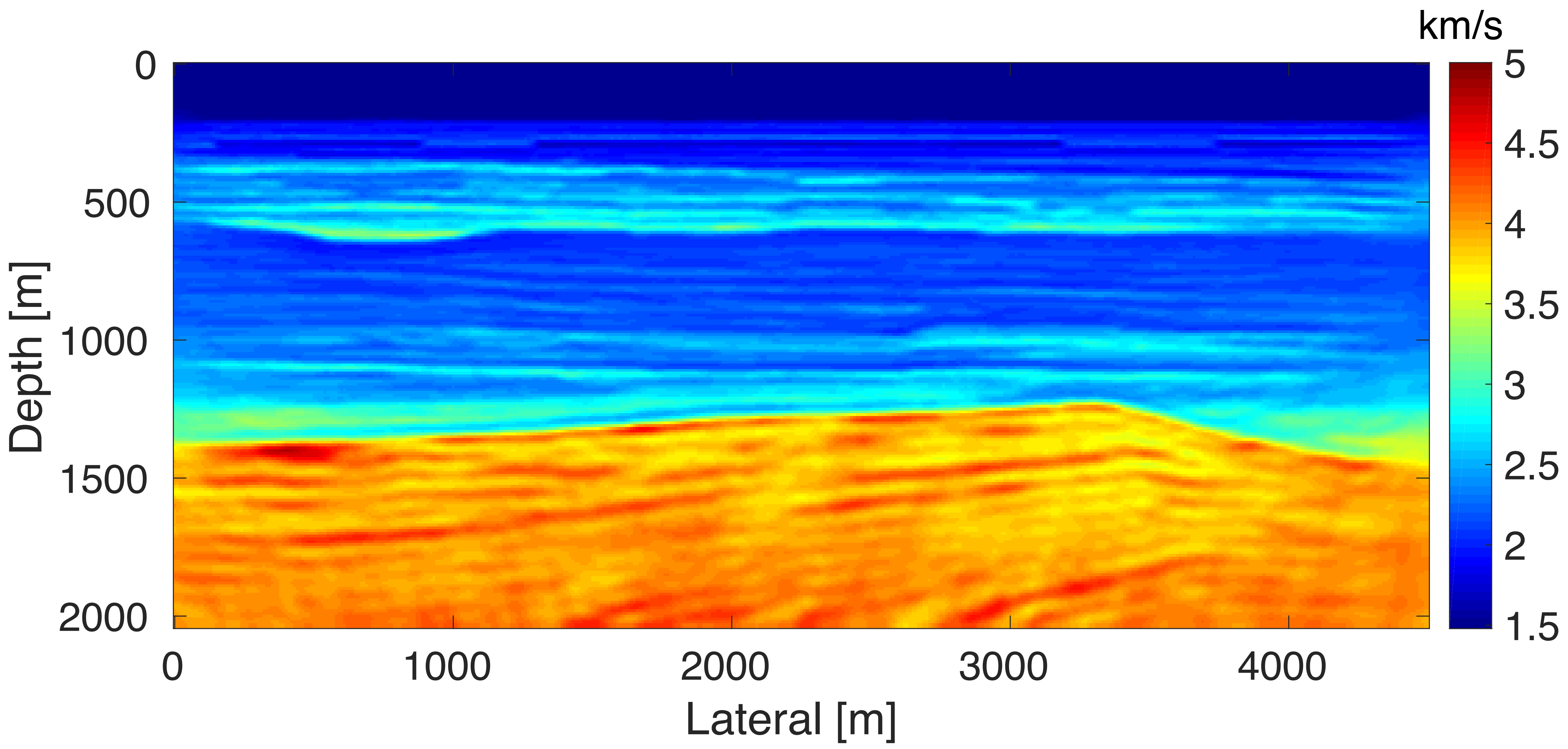
Initial model

Inversion information:
Optimization Solver: l-BFGS
Iterations per frequency band: 21
Batch size: 15

Inversion results

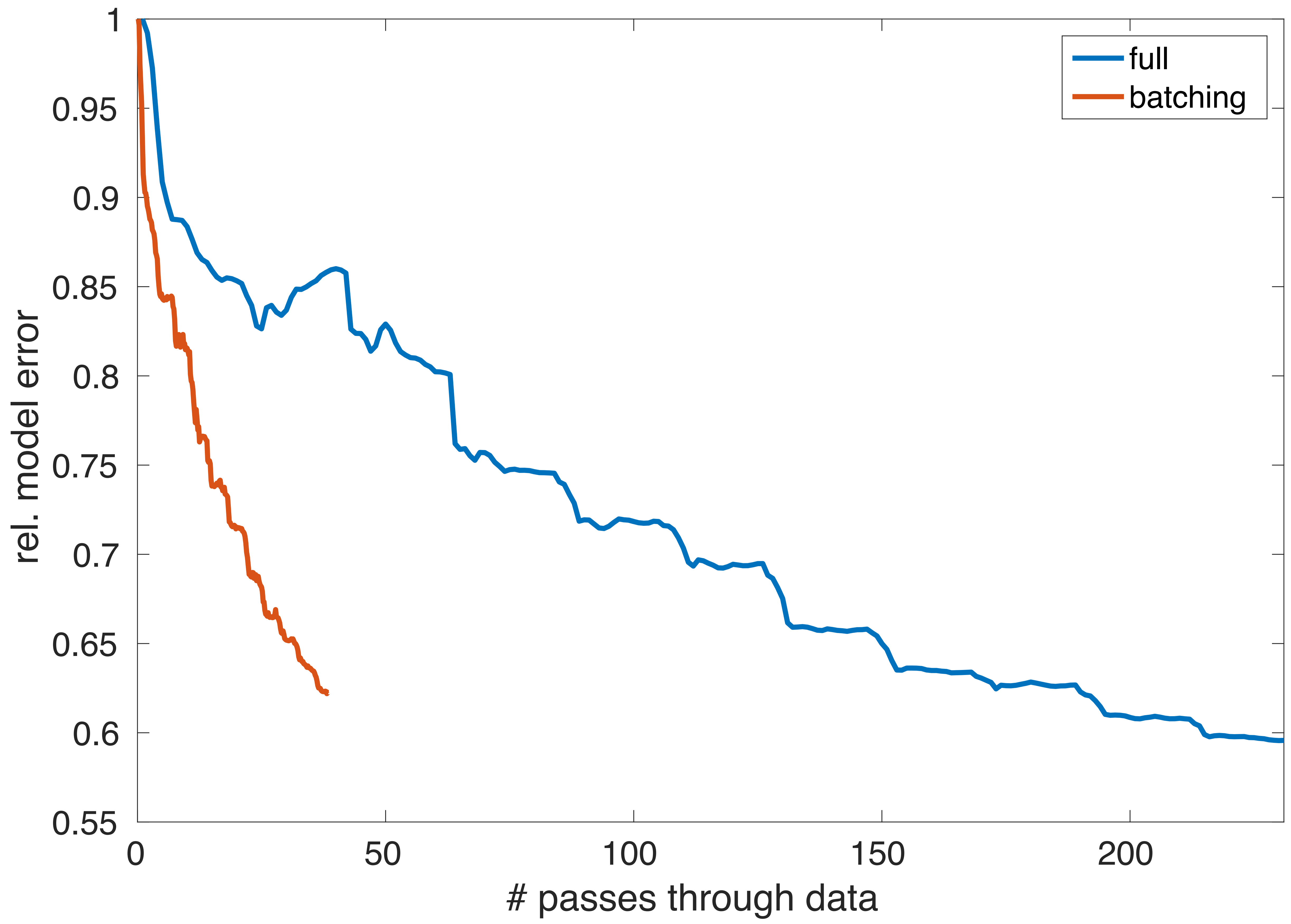


full data w/ true source wavelet



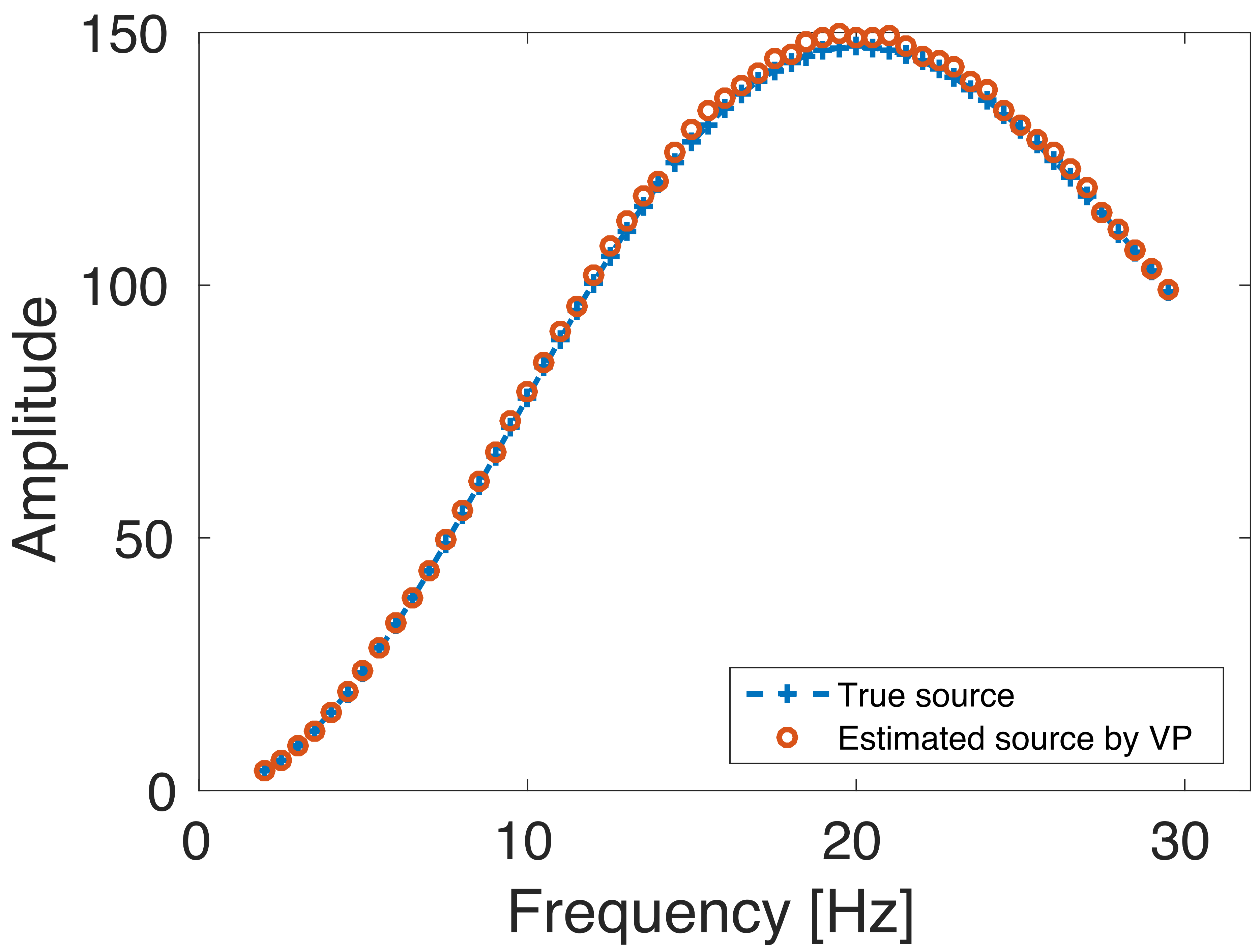
batch data w/ estimated source wavelet

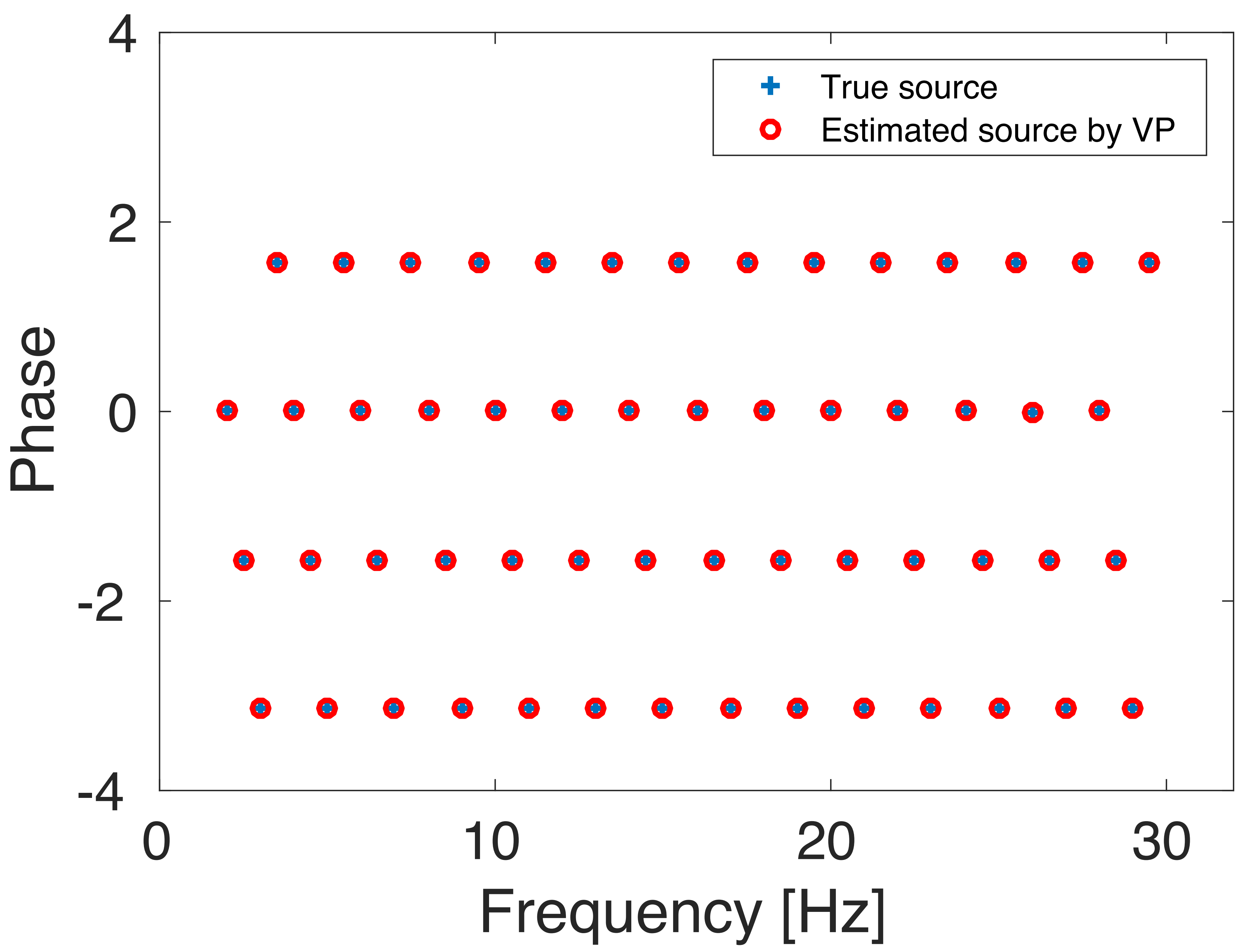
Convergence comparison



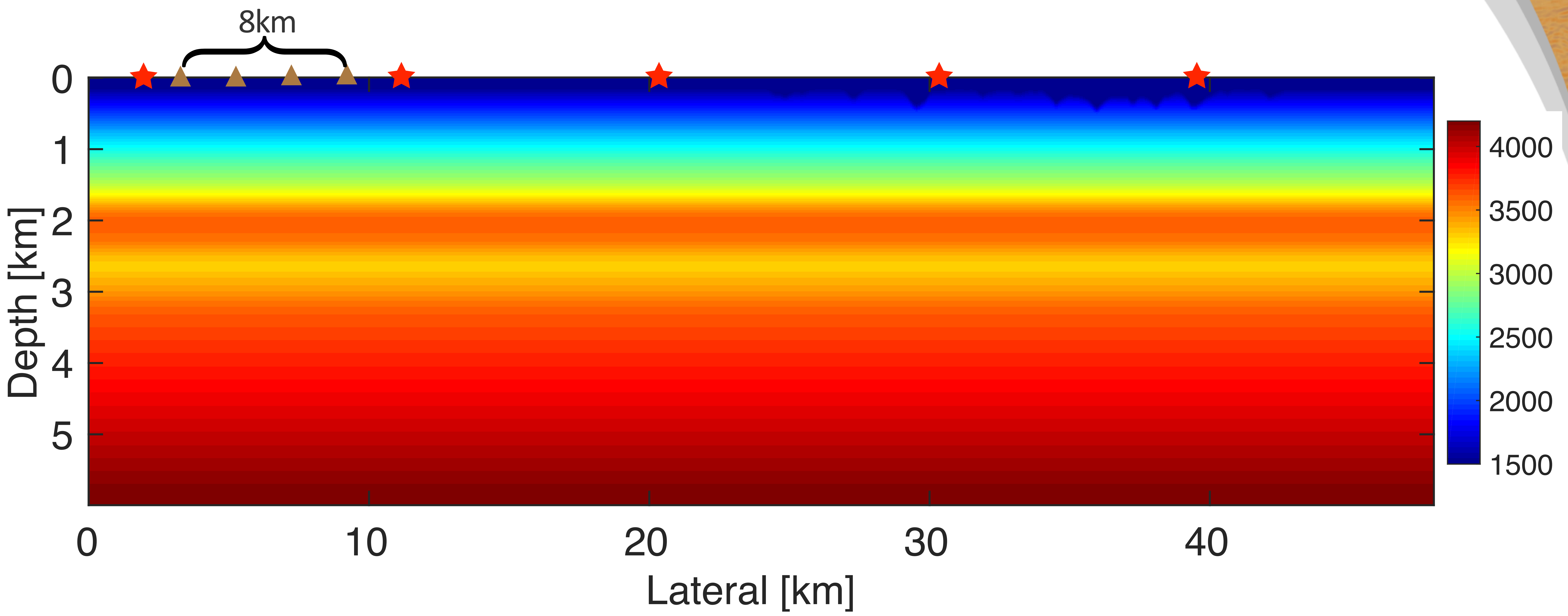
6x speed up

Source wavelet comparison

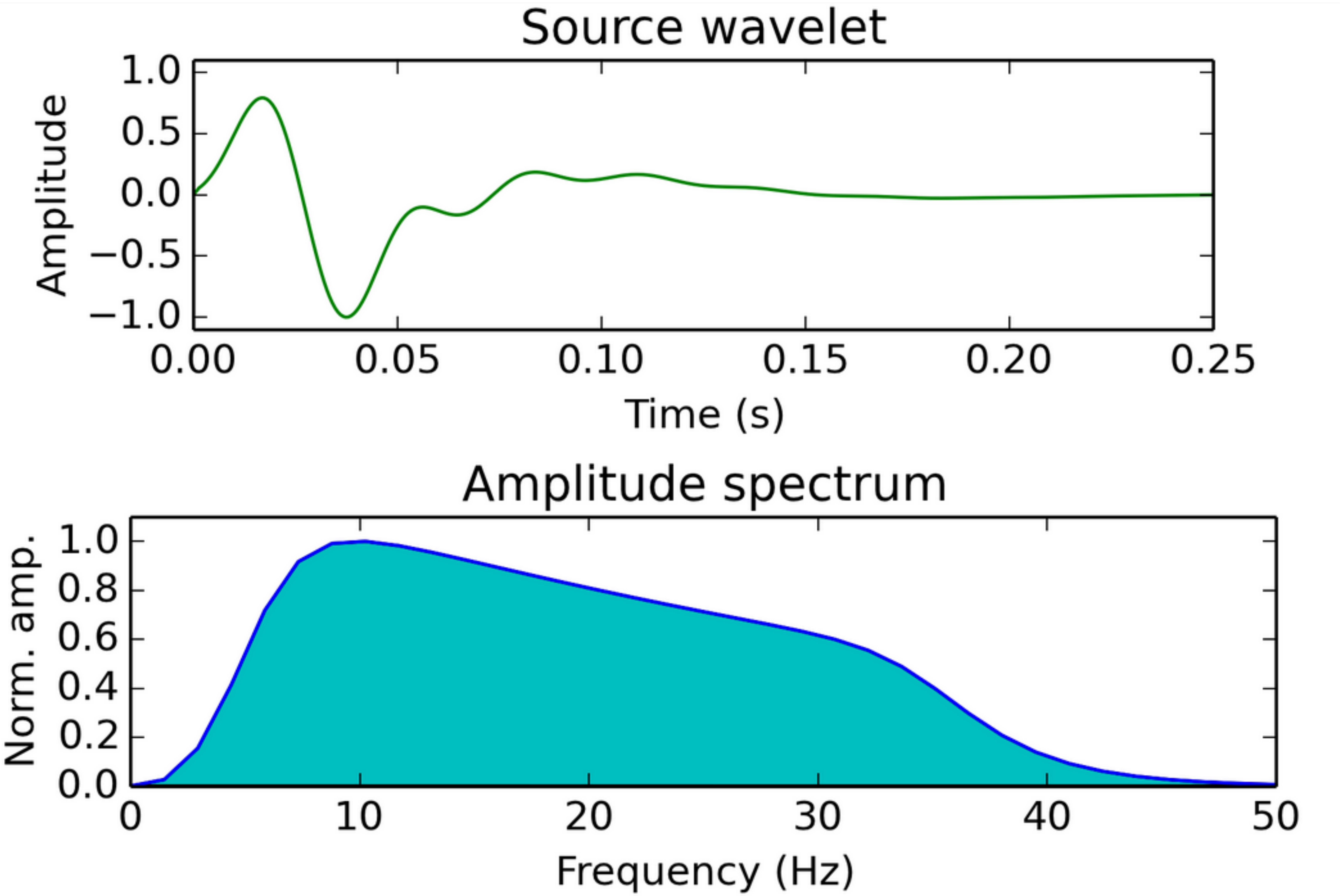




Chevron blind test data



Chevron blind test data



Chevron blind test data

Data-set information:

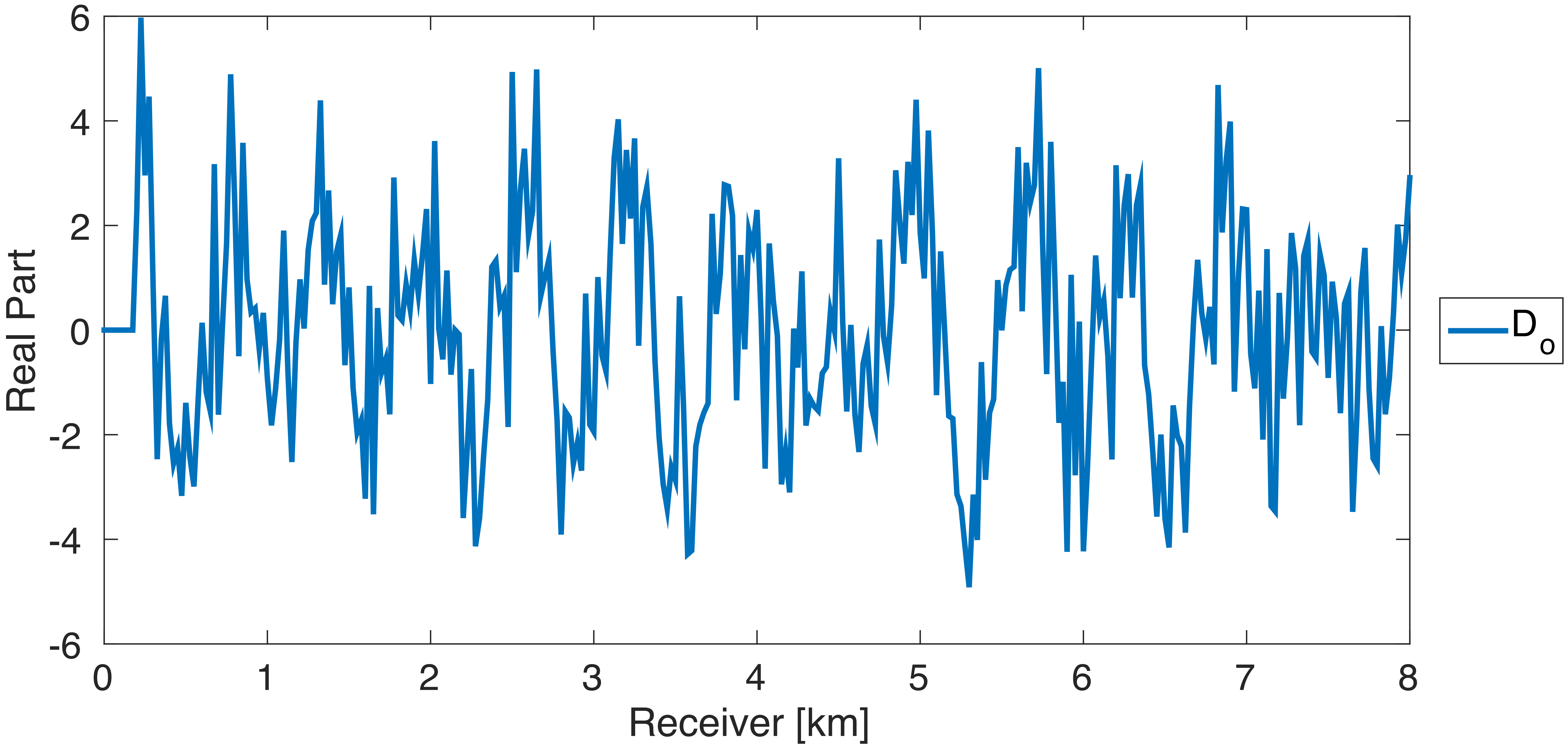
1. 1600 shots: ds = 25 m, Source depth = 15 m;
2. 321 hydrophone recs/shot: dr = 25 m, Receiver depth = 15 m;
3. Maximum offset = 8000 m;
4. Record time = 8.0 s, sample rate 4 ms;
5. Vp water = constant = 1510 m/s;
6. With free surface multiples present in the data;
7. Isotropic Elastic.

Chevron blind test data

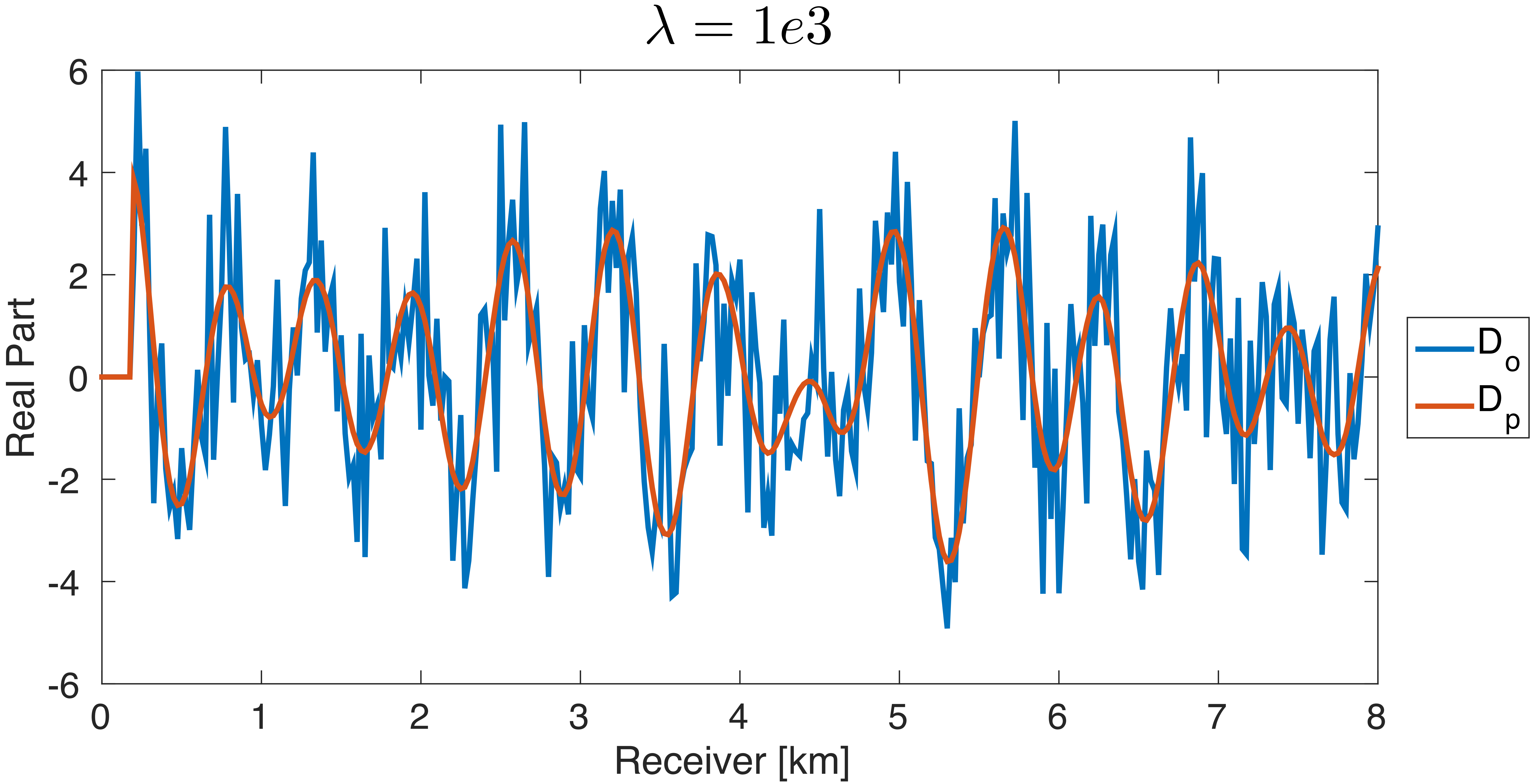
Inversion strategy:

1. Frequency domain WRI with Source estimation;
2. Frequency bands: [3:0.2:5]Hz, [3:0.2:7]Hz, [3:0.2:9]Hz, [3:0.2:11]Hz, [3:0.2:19]Hz;
3. Batch sizes of random frequency subsets: 3, 6, 10, 10, 15;
4. Batch size of random source subsets: 300;
5. Optimization solver: l-BFGS with 20 iterations per frequency band;
6. 4 passes of WRI at frequency 3-11 Hz and 1 pass to 19 Hz;
7. Grid size: 20m for 3-11Hz and 12m for 3-19Hz;
8. No pre-processing !!!

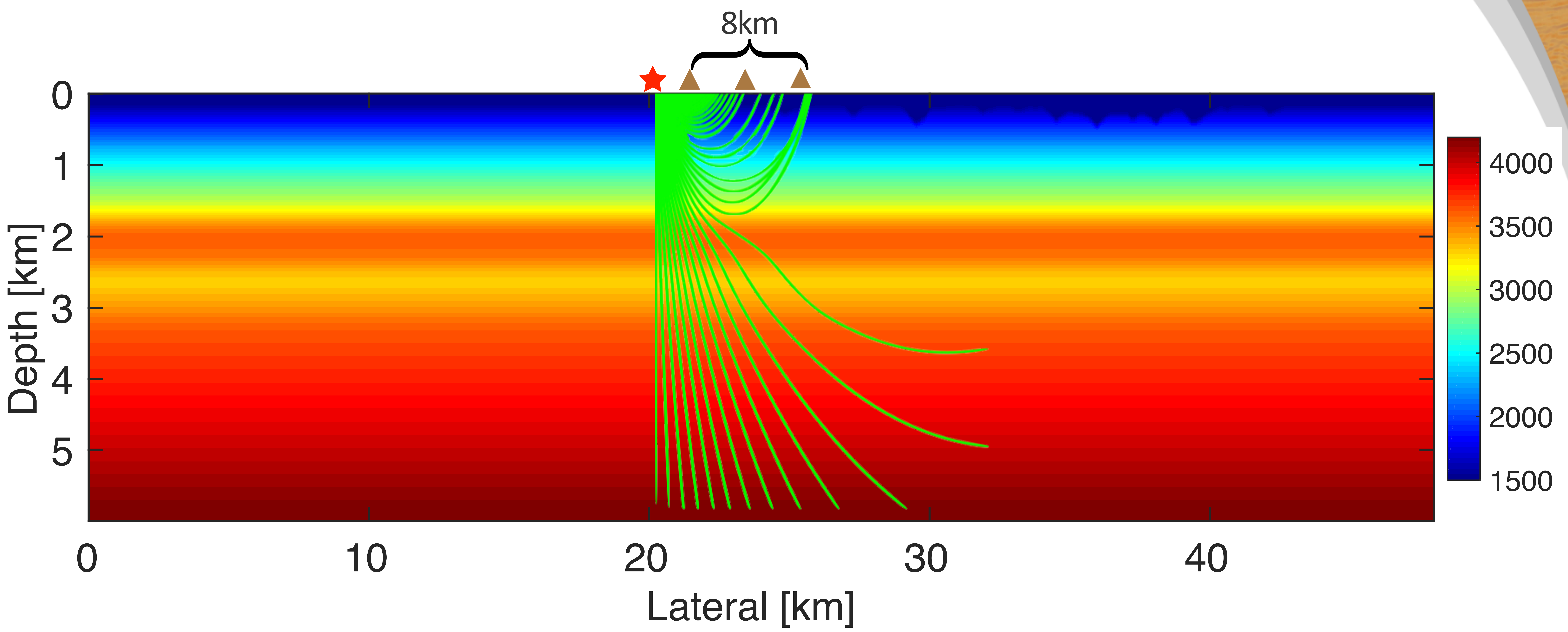
Data comparison
— 3 Hz Data of 800th shot



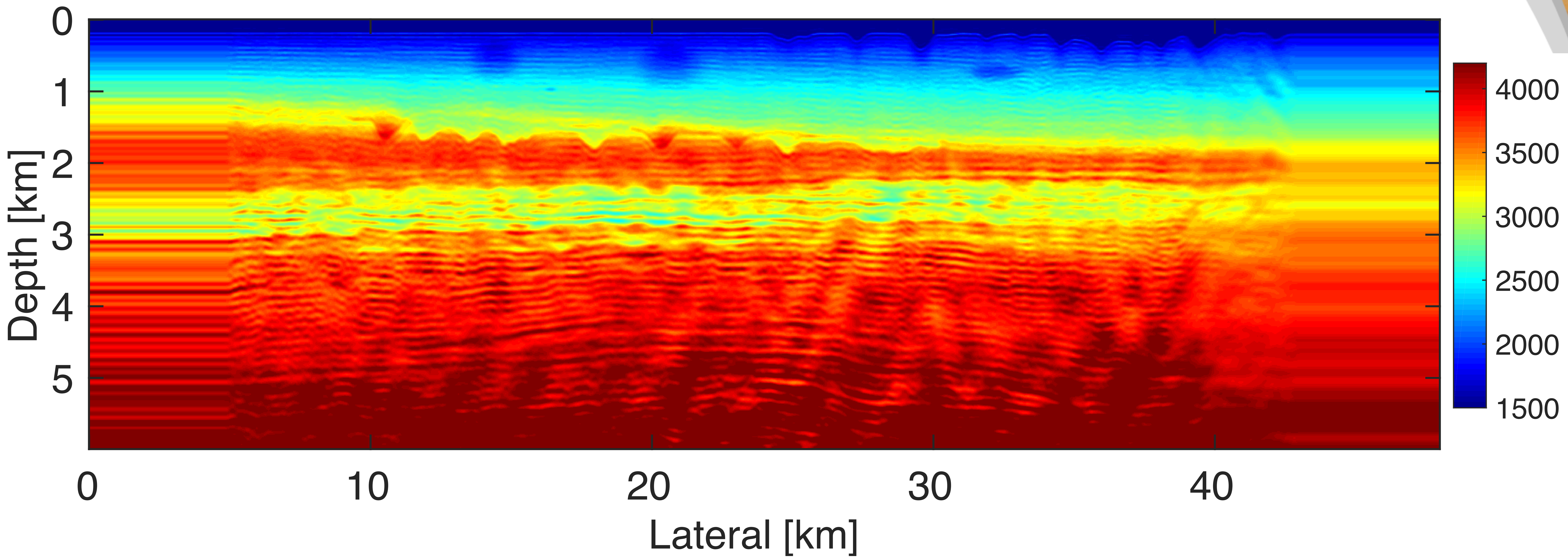
Data comparison
— 3 Hz Data of 800th shot



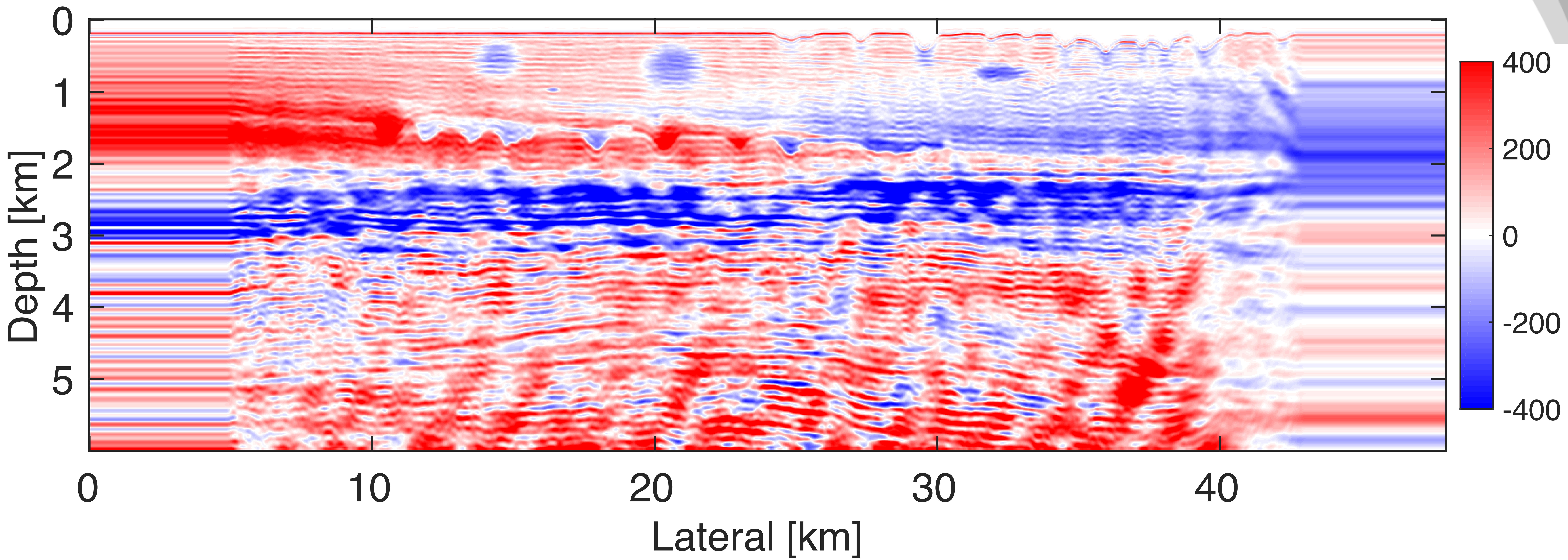
Initial model



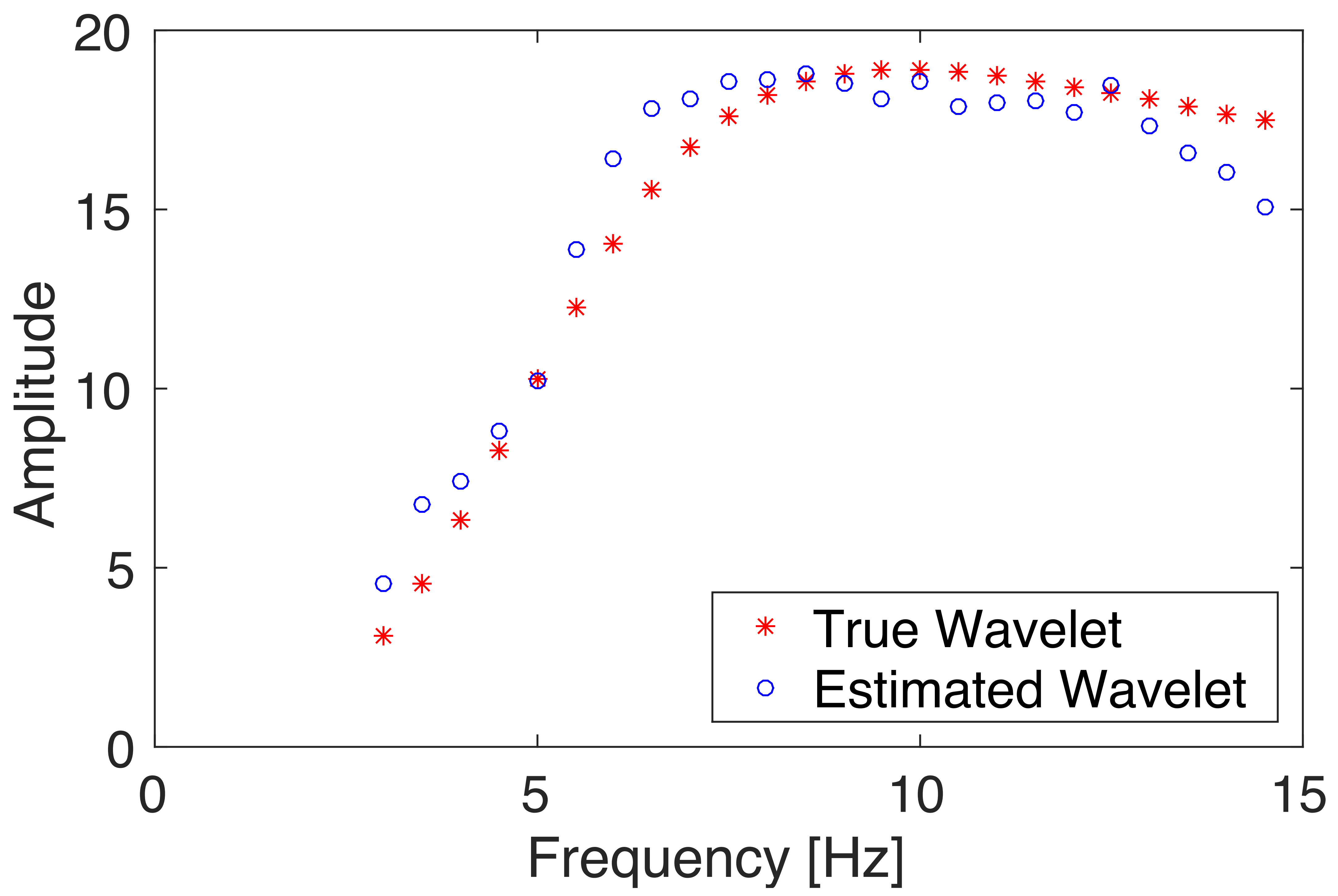
Inversion result



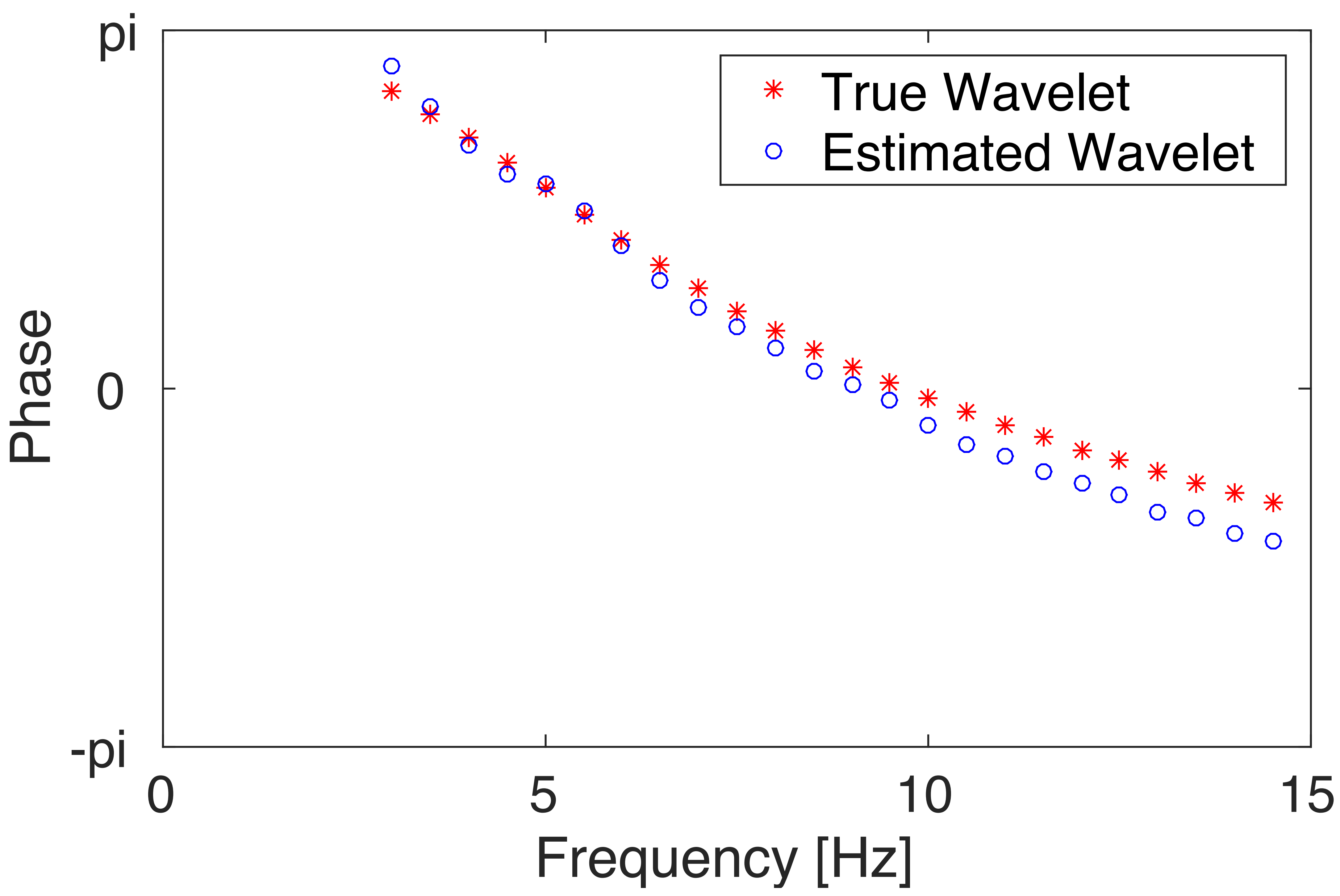
Model update



Source wavelet comparison

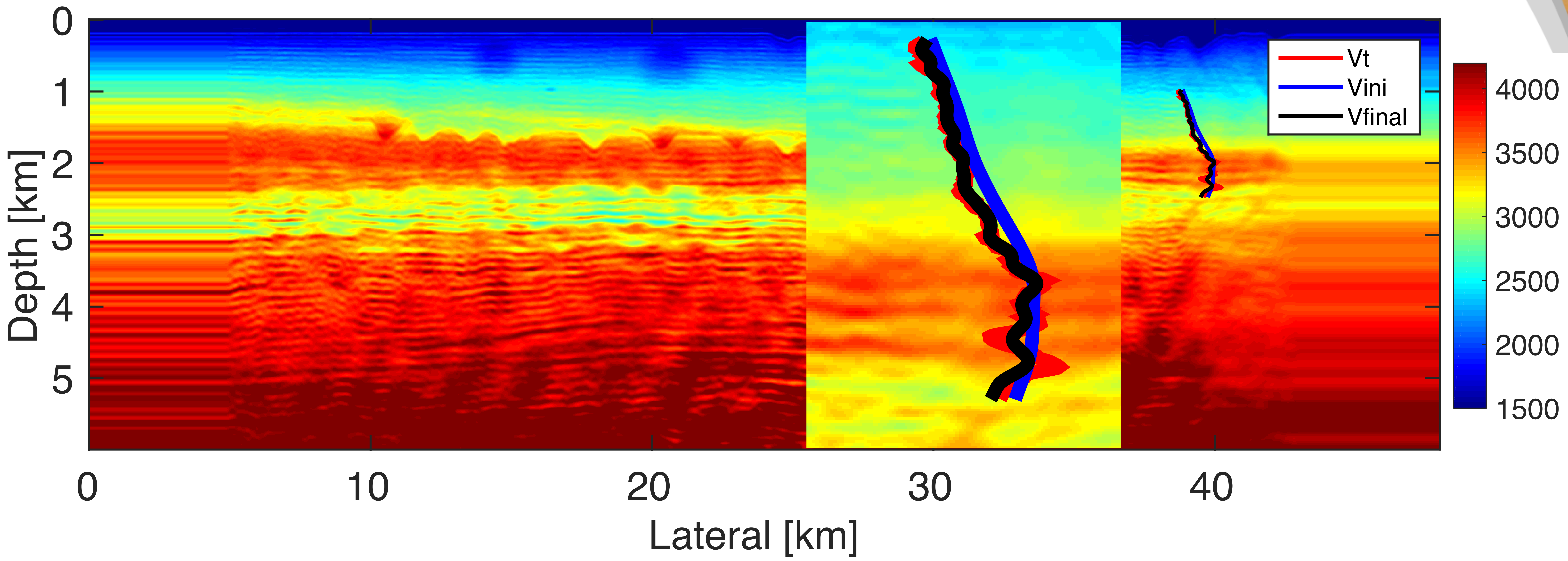


Amplitude

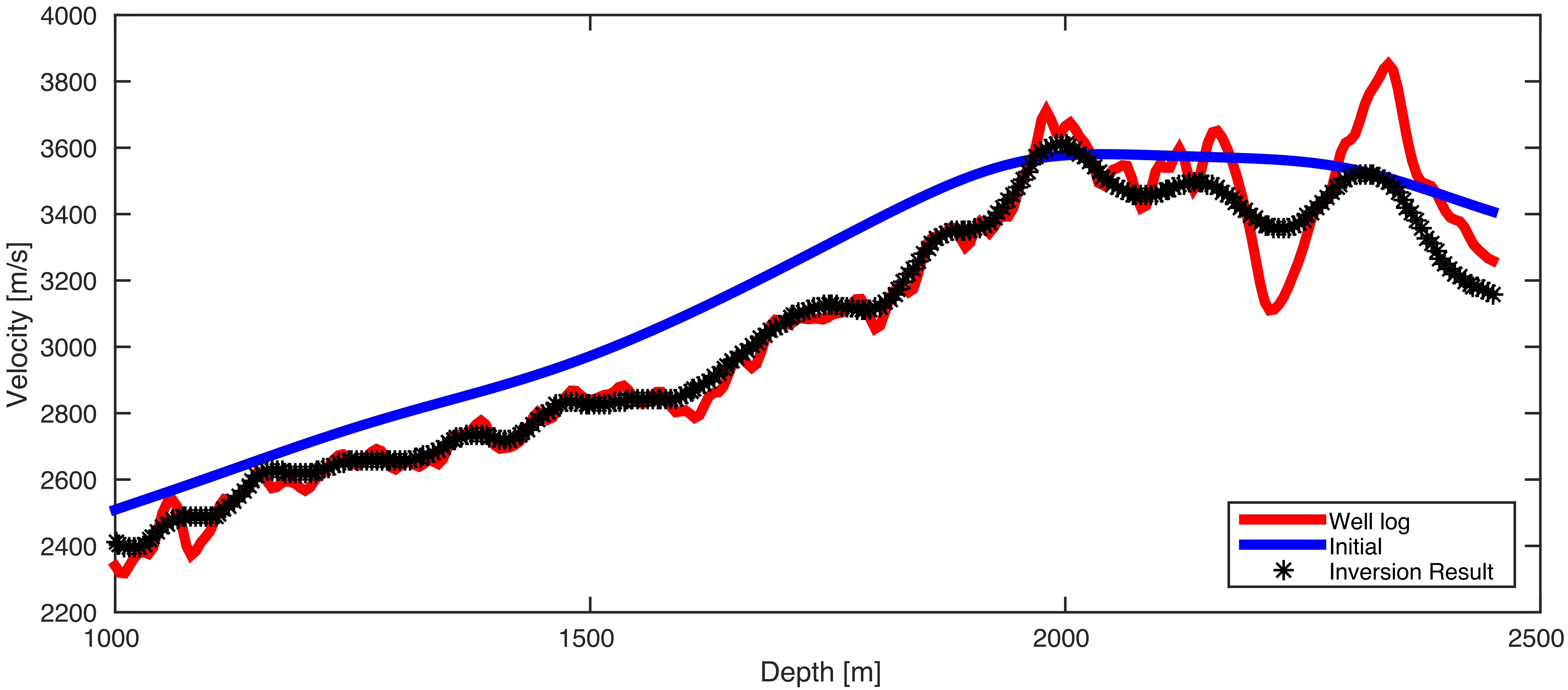


Phase

Well-log comparison



Well-log comparison



Conclusion

Stochastic WRI with source estimation:

- dimensionality reduction – select subsets of sources and frequencies randomly
- source estimation – reconstruct the wavefields and source wavelets by variable projection method
- numerical examples – speed up the inversion with a factor of at least 6 and invert the unknown model without additional prior information about the source wavelets

Reference

Albert Tarantola, and Bernard Valette. Generalized nonlinear inverse problems solved using the least squares criterion. *Reviews of Geophysics* 20 (2): 219-232, 1982.

Jean Virieux and Stéphane Operto. An overview of full-waveform inversion in exploration geophysics. *Geophysics*, 74(6):WCC1–WCC26, 2009.

Tristan van Leeuwen and Felix J. Herrmann. Mitigating local minima in full-waveform inversion by expanding the search space. *Geophysical Journal International*, 195:661–667, 10 2013.

Tristan van Leeuwen and Felix J. Herrmann. A penalty method for PDE-constrained optimization in inverse problems. *Inverse Problems*, 32(1):015007, 12 2015.

Bas Peters, Felix J. Herrmann and Tristan van Leeuwen. Wave-equation based inversion with the penalty method: adjoint-state versus wavefield-reconstruction inversion. *EAGE,* 2014.

Gene Golub and Victor Pereyra. Separable nonlinear least squares: the variable projection method and its applications. *Inverse problems,* 19 (2): R1, 2003.

Yurii Nesterov. Introductory lectures on convex optimization. *Springer,* 2004.

Arkadii Nemirovskii, David Borisovich Yudin, and Edgar Ronald Dawson. Problem complexity and method efficiency in optimization. 1983.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization,* 19 (4): 1574-1609, 2009.

Alekh Agarwal, Martin J. Wainwright, Peter L. Bartlett, and Pradeep K. Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. *Advances in Neural Information Processing Systems*: 1-9, 2009.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*: 1-30, 2013.

Acknowledgements

This research was carried out as part of the SINBAD project with the support of the member organizations of the SINBAD Consortium.