Dimensionality reduction for full-waveform inversion

Felix J. Herrmann

SLIM
Seismic Laboratory for Imaging and Modeling
the University of British Columbia

Recent driver

HP and Shell Sensing System

HP and Shell are collaborating to develop a wireless sensing system to acquire extremely high-resolution seismic data on land. HP and Shell will use their complementary knowledge and experience to produce a groundbreaking solution that can sense, collect and store geophysical data.

- *1000.000 channel systems* (up from 40.000)
- will increase size data volumes by *orders of magnitude*
- aside from increasing # of cores no speedup on the horizon
- seismic data processing & *inversion* have become *challenging* because of *processor & IO limitations*

source http://www.hp.com/hpinfo/newsroom/press_kits/2010/sensingsolutions/index.html

[Tarantola, 84; Pratt, '98; Plessix, '06]

SLIM

FWI formulation

Multiexperiment unconstrained optimization problem:

$$\min_{\mathbf{m}\in\mathcal{M}} \sum_{i=1}^{N=n_s\times n_f} \frac{1}{2}\|\mathbf{D}_i - \mathcal{F}[\mathbf{m};\mathbf{Q}_i]\|_2^2 \quad \text{with} \quad \mathcal{F}[\mathbf{m};\mathbf{Q_i}] := \mathbf{P}_i\mathbf{H}_i^{-1}[\mathbf{m}]\mathbf{Q}_i$$

$$\begin{aligned}
\mathbf{D}_i &= \text{Monochromatic single-source data}\\
\mathbf{P}_i &= \text{Detection operator for each source experiment}\\
\mathbf{H}_i &= \text{Inverse of time-harmonic Helmholtz}\\
\mathbf{Q}_i &= \text{Monochromatic source}\\
\mathbf{m} &= \text{Unknown model, e.g. } c^{-2}(x)
\end{aligned}$$

[Pratt et. al., '98]
[Plessix '06]

SLIM

Adjoint state

Implicit solves of Helmholtz system for each experiment

$$\mathbf{H}[\mathbf{m}]\mathbf{u} = \mathbf{q} \quad \text{and} \quad \mathbf{H}^*[\mathbf{m}]\mathbf{v} = \mathbf{r}$$

with

$$\mathbf{r} = \mathbf{P}^*(\mathbf{d} - \mathcal{F}[\mathbf{m}, \mathbf{q}])$$

and compute gradient via

$$\delta\mathbf{m} = \Re\left(\sum_{\omega} \omega^2 \sum_{s} (\bar{\mathbf{u}} \odot \mathbf{v})_{s,\omega}\right)$$

FWI formulation
[complete data]

Multiexperiment unconstrained optimization problem:

$$\min_{\mathbf{m}\in\mathcal{M}} \sum_{i=1}^{N=n_s\times n_f} \frac{1}{2}\|\mathbf{D}_i - \mathcal{F}[\mathbf{m};\mathbf{Q}_i]\|_2^2 \quad \text{with} \quad \mathcal{F}[\mathbf{m};\mathbf{Q_i}] := \mathbf{P}\mathbf{H}^{-1}[\mathbf{m}]\mathbf{Q_i}$$

$\mathbf{D}$ = Multi-source and multi-frequency data volume

$\mathbf{P}$ = Single detection operator

$\mathbf{Q}$ = Seismic sources

$\mathbf{m}$ = Unknown model, e.g. $c^{-2}(x)$

[Tarantola, 84; Pratt, '98; Plessix, '06]

FWI formulation

[equivalent]

Multiexperiment unconstrained optimization problem:

$$\min_{\mathbf{m}\in\boldsymbol{\mathcal{M}}} \frac{1}{2}\|\mathbf{D}-\boldsymbol{\mathcal{F}}[\mathbf{m};\mathbf{Q}]\|_{2,2}^{2} \quad \text{with} \quad \boldsymbol{\mathcal{F}}[\mathbf{m};\mathbf{Q}] := \mathbf{P}\mathbf{H}^{-1}[\mathbf{m}]\mathbf{Q}$$

- requires large number of PDE solves
- linear in the sources
- apply *randomized* dimensionality reduction

[Tarantola, 84; Pratt, ’98; Plessix, ‘06]

[FJH et.al., '08-10', Krebs et.al., '09, Operto et. al., '09]
[Haber, Chung, and FJH, '10]

SLIM

Reduced FWI formulation

Multiexperiment unconstrained optimization problem:

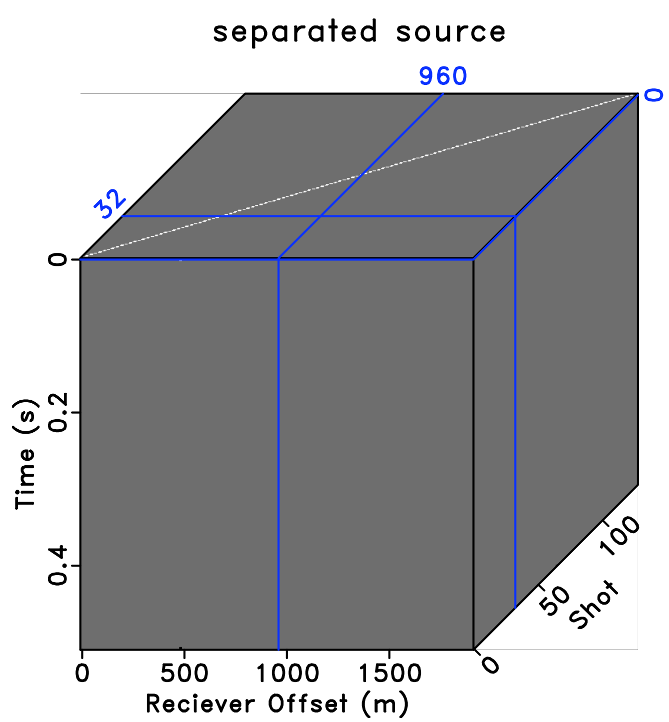
$$\min_{\mathbf{m}\in\mathcal{M}} \frac{1}{2}\|\underline{\mathbf{D}} - \mathcal{F}[\mathbf{m};\underline{\mathbf{Q}}]\|_{2,2}^2 \quad \text{with} \quad \mathcal{F}[\mathbf{m};\underline{\mathbf{Q}}] := \mathbf{P}\underline{\mathbf{H}}^{-1}\underline{\mathbf{Q}}$$

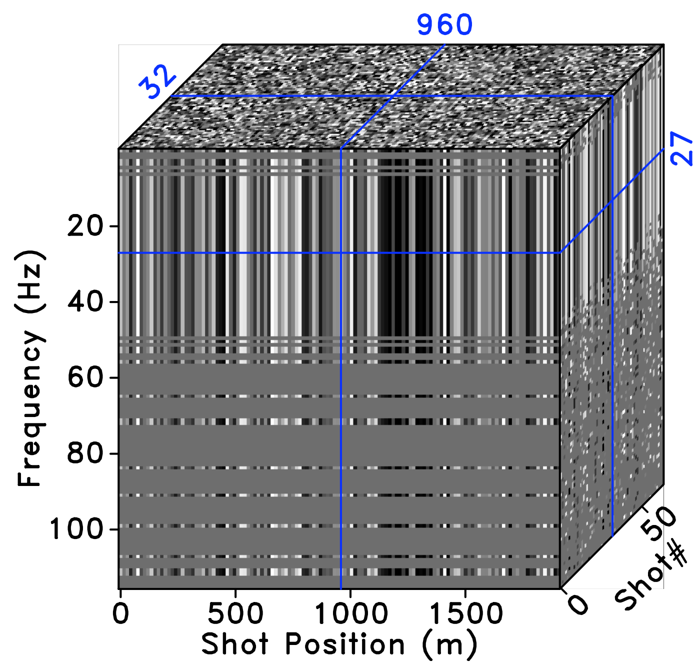
- requires *smaller* number of PDE solves
- explores *linearity* in the *sources* & *block-diagonal* structure of the *Helmholtz system*
- uses *randomized* frequency *selection* and *phase encoding*

[FJH et. al. '08-'10]

Batch/mini experiment

adapted from FJH et. al. ,09





$$\mathbf{Q} \qquad\qquad \underline{\mathbf{Q}} = \mathbf{RMQ}$$

Collection of *K simultaneous-source* experiments (*supershots*) with Batch size $K = n_f' \times n_s' \ll n_f \times n_s$

Math [Romberg, '07, FJH, '08-'10]

Compressive-sampling operator

$$\mathbf{RM} = \mathrm{vec}^{-1}\mathrm{blockdiag}\left[(\mathbf{RM})_{1\ldots n'_s}\right]\mathrm{vec}$$

with
$$(\mathbf{RM})_k = (\mathbf{R^\Sigma}_k\mathbf{M^\Sigma} \otimes \mathbf{I} \otimes \mathbf{R^\Omega}_k)$$

and
$$\mathbf{M^\Sigma} = \overbrace{\mathrm{sign}(\eta) \odot \mathbf{F_\Sigma}^H e^{j\theta}\mathbf{F_\Sigma}}^{\text{'Gaussian matrix'}}$$

where $\theta \in \mathrm{Uniform}(-\pi,\, \pi]$, and $\eta \in \mathrm{Normal}(0,1)$

Interpretations

Consider *randomized* dimensionality reduction as instances of

- *stochastic optimization & machine learning* [Haber, Chung, and FJH, '10]
- *compressive sensing* [FJH et. al, '08-'10]

Stochastic optimization

Replace *deterministic*-optimization problem

$$\min_{\mathbf{m}\in\boldsymbol{\mathcal{M}}} f(\mathbf{m}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\|\mathbf{d}_i - \mathcal{F}[\mathbf{m};\mathbf{q}_i]\|_2^2$$

with *sum* cycling over *different sources* & *corresponding monochromatic shot records* (columns of D & Q)

[Natterer, '01]

Stochastic *average* approximation

[Haber, Chung, and FJH, '10]

by *a stochastic-optimization* problem

$$\min_{\mathbf{m}\in\mathcal{M}} \mathbf{E}_{\mathbf{w}}\{f(\mathbf{m},\mathbf{w}) = \frac{1}{2}\|\mathbf{D}\mathbf{w}-\mathcal{F}[\mathbf{m};\mathbf{Q}\mathbf{w}]\|_2^2\}$$

$$\approx \frac{1}{K}\sum_{j=1}^{K}\frac{1}{2}\|\underline{\mathbf{d}}_j-\mathcal{F}[\mathbf{m};\underline{\mathbf{q}}_j]\|_2^2$$

with $\mathbf{w}\in N(0,1)$ and $\mathbf{E}_{\mathbf{w}}\{\mathbf{w}\mathbf{w}^H\}=\mathbf{I}$

and $\underline{\mathbf{d}}_j=\mathbf{D}\mathbf{w}_j,\ \underline{\mathbf{q}}_j=\mathbf{Q}\mathbf{w}_j$

Stochastic *average* approximation

In the *limit* $K \to \infty$, *stochastic* & *deterministic* formulations are *identical*

We *gain* as long as $K \ll N$ ...

Since the error in *Monte-Carlo* methods decays only slowly $(\mathcal{O}(K^{-1/2}))$

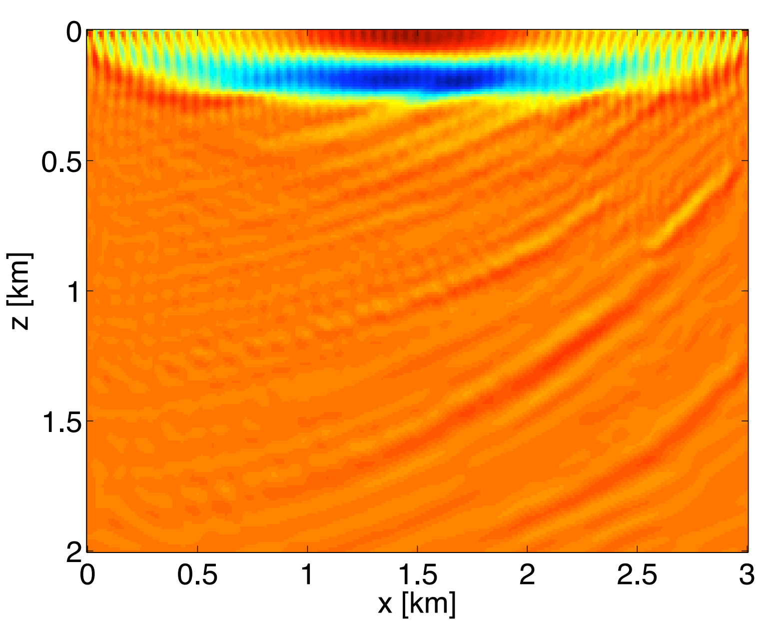
this approach may be problematic...

However, the location for the *minimum* of the *misfit* may be relatively *robust*...

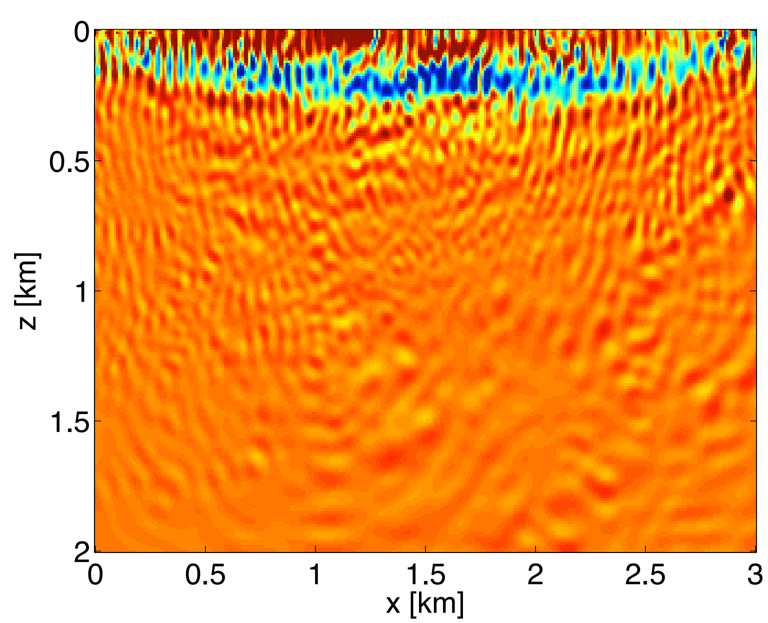
Stylized example

Search direction for batch size K:

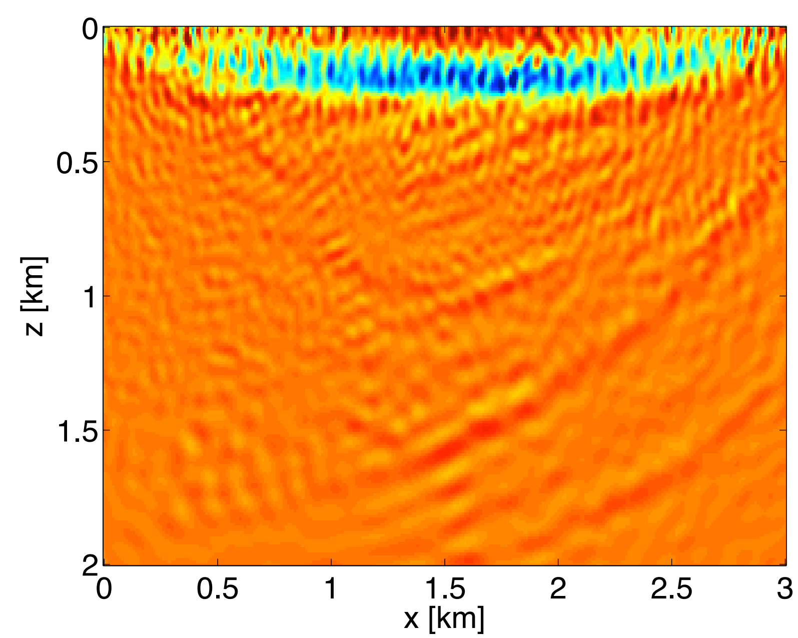
$$\mathbf{g}_K \approx \frac{1}{K}\sum_{j=1}^{K} \nabla\mathcal{F}^*[\mathbf{m};\underline{\mathbf{q}}_j]\delta\underline{\mathbf{d}}_j$$



full

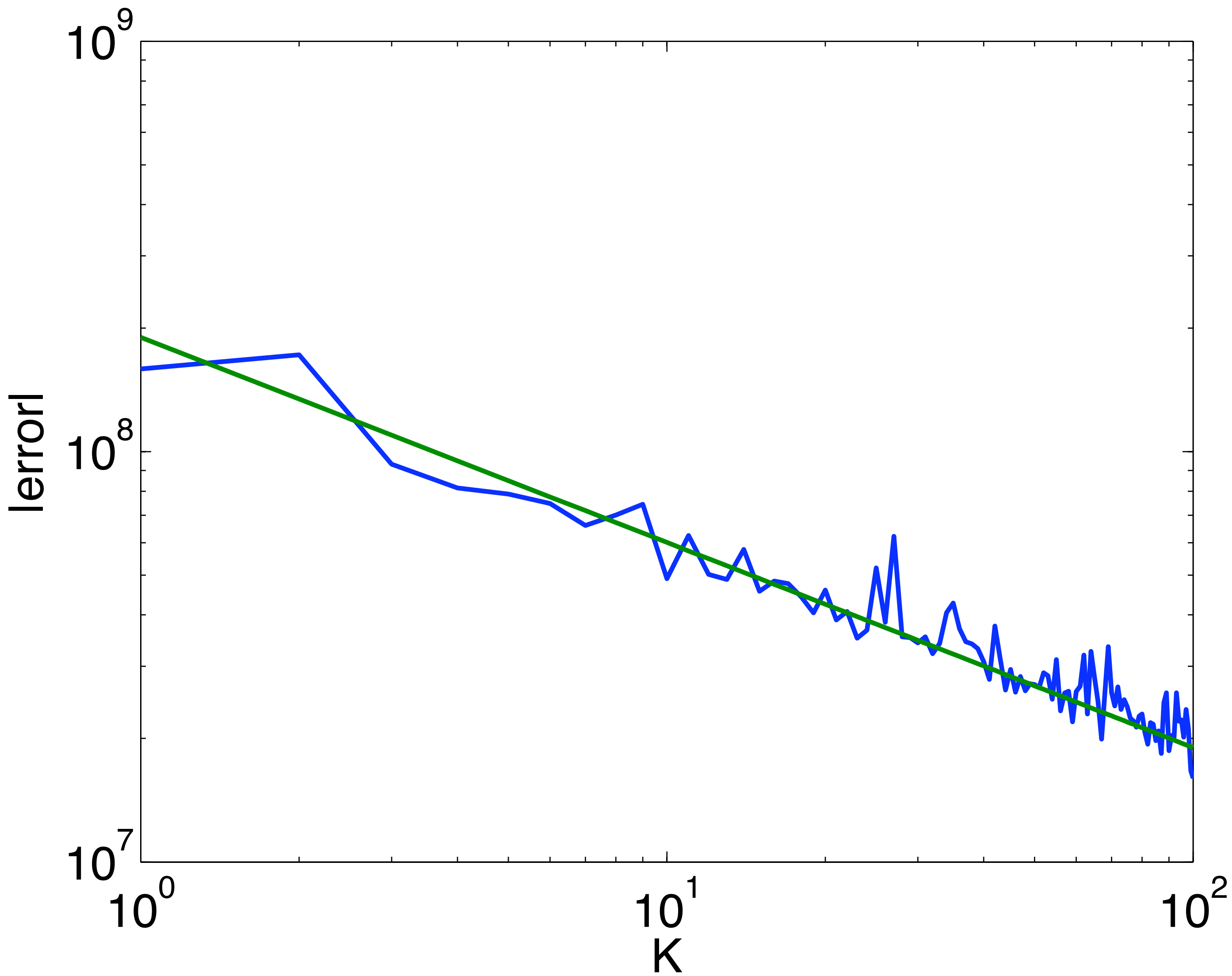


K=1



K=5

Decay

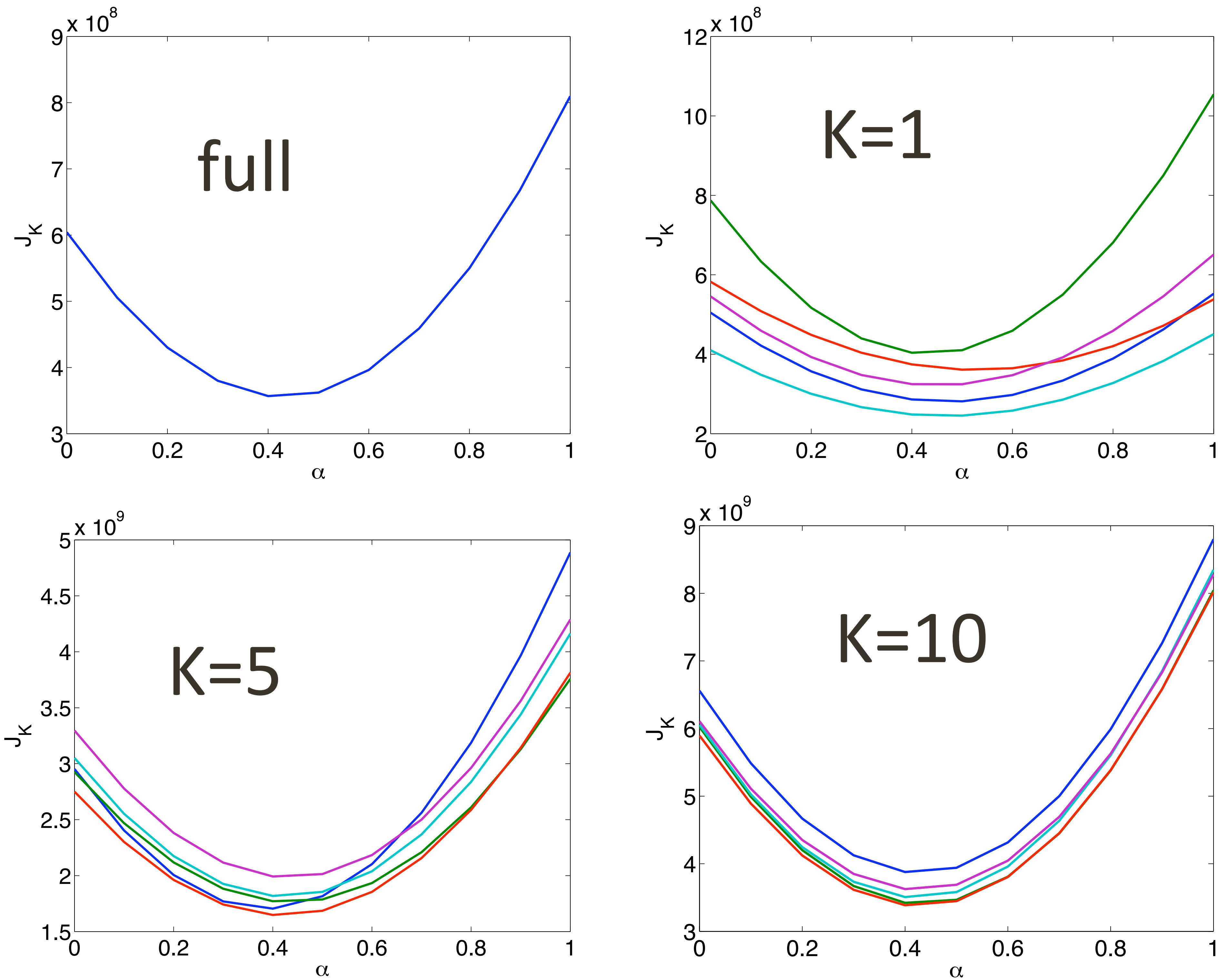


error between full and sampled gradient

Misfit functional

[adapted from Haber, Chung, and FJH, '10]

$$f_K(\mathbf{g}_K) = \frac{1}{K}\sum_{j=1}^{K}\frac{1}{2}\|\underline{\mathbf{d}}_j - \mathcal{F}[\mathbf{m}+\alpha\mathbf{g}_K;\underline{\mathbf{q}}_j]\|_2^2$$



Randomized trace estimates

FWI relies on computation of

$$\| \overbrace{\mathbf{D} - \mathcal{F}[\mathbf{m}; \mathbf{Q}]}^{\mathbf{S}} \|_F^2$$

which corresponds to computing the trace

$$\text{trace}(\mathbf{S}^*\mathbf{S}) = \|\mathbf{S}\|_F^2$$

Approximate this trace stochastically.

[Hutchinson., '90, Avron and S. Toledo, '10]

Randomized trace estimates

Use

$$H_K = \frac{1}{K}\sum_{j=1}^{K} \mathbf{w}_j^* \mathbf{B} \mathbf{w}_j \ \text{with} \ \mathbf{w}_j \ \ i.i.d.$$

and $\mathbf{B} = \mathbf{S}^*\mathbf{S}$.

Corresponds to SA via 'source encoding' for monochromatic experiments

- how to choose K and **w**'s such that

$$\Pr\left(|H_k - \text{trace}(\mathbf{B})| \leq \epsilon \ \text{trace}(\mathbf{B})\right) \geq 1 - \delta$$

for some $(\epsilon,\ \delta)$.

[Hutchinson., '90, Avron and S. Toledo, '10]

Randomized trace estimates

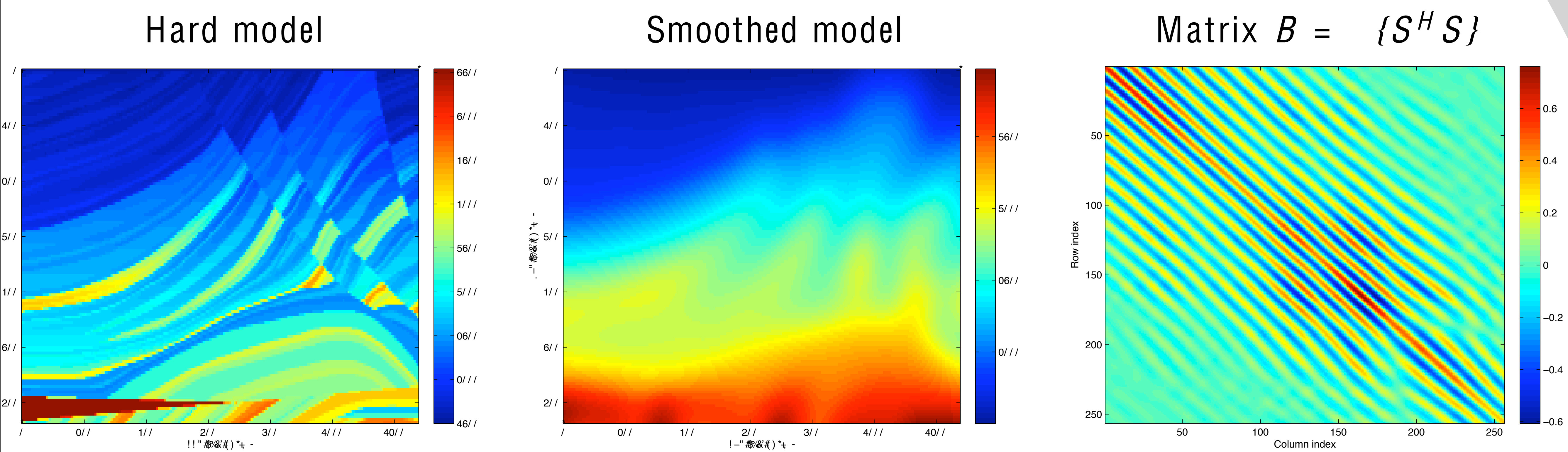
Set $(\epsilon, \delta) = (0.2, 0.1)$, yielding a possible error in the estimate of 25.

Estimates for K from the table are

$$K = (15, 12, 13, 100) \times 10^3$$

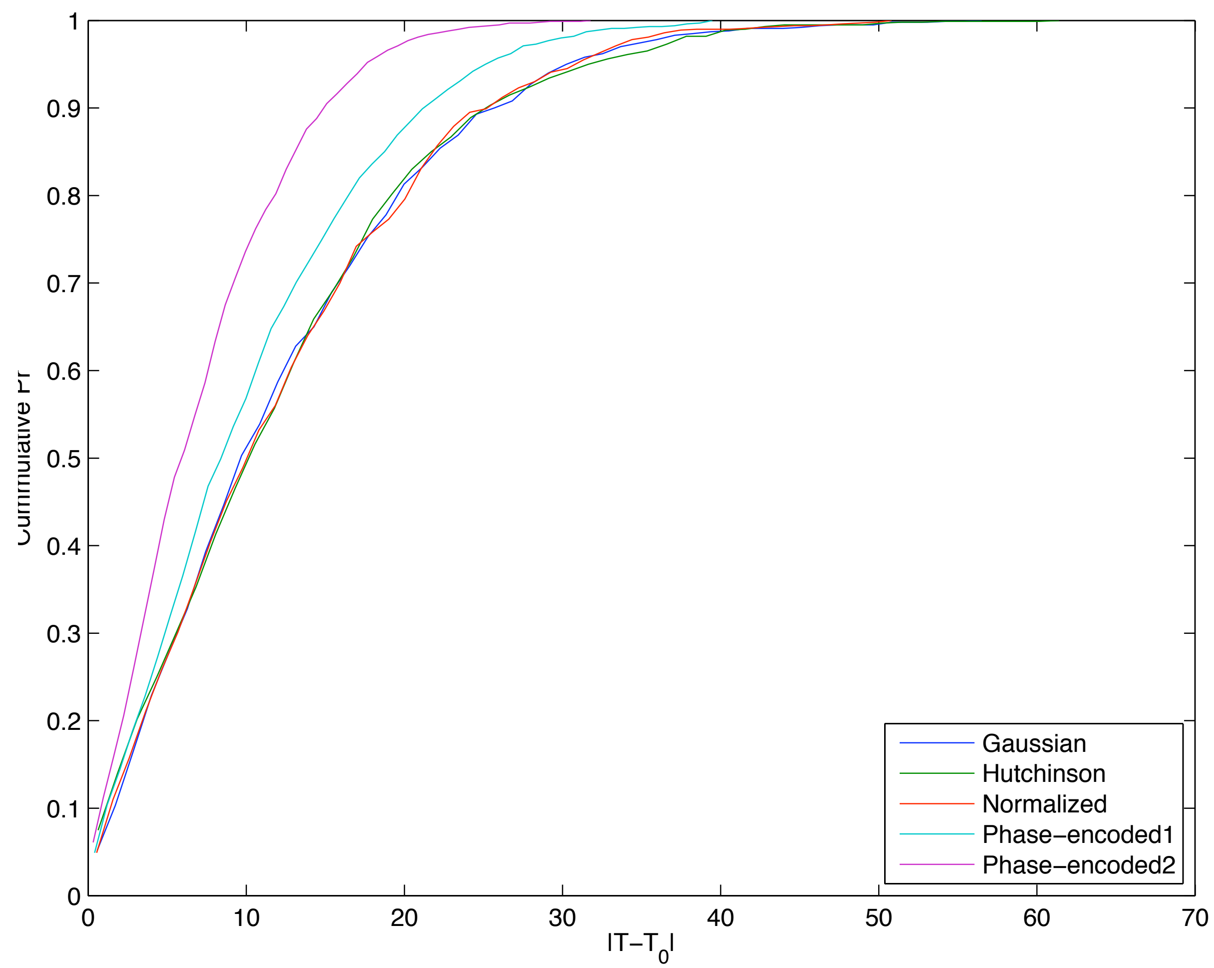
- pessimistic
- cross-over at $N = 15 \times 10^3$
- can we do better as seen with CS?

Example



- Modeled at 20 Hz with 256 sources
- significant amount of off-diagonal energy

Example
[for fixed N/K=16]



- different method perform similarly except for the phase encoding, which is better
- order of magnitude speedup

Stochastic approximation [Bertsekas,' '96; Nemirovski, '09]

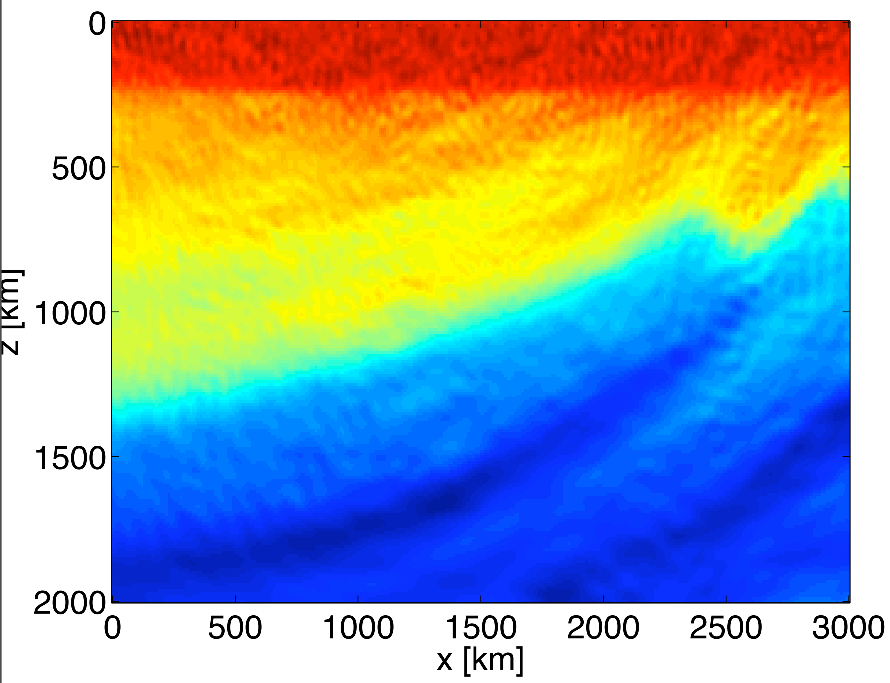
Use *different* simultaneous shots for each *subproblem*, i.e.,

$$\underline{\mathbf{Q}} \mapsto \underline{\mathbf{Q}}^k$$

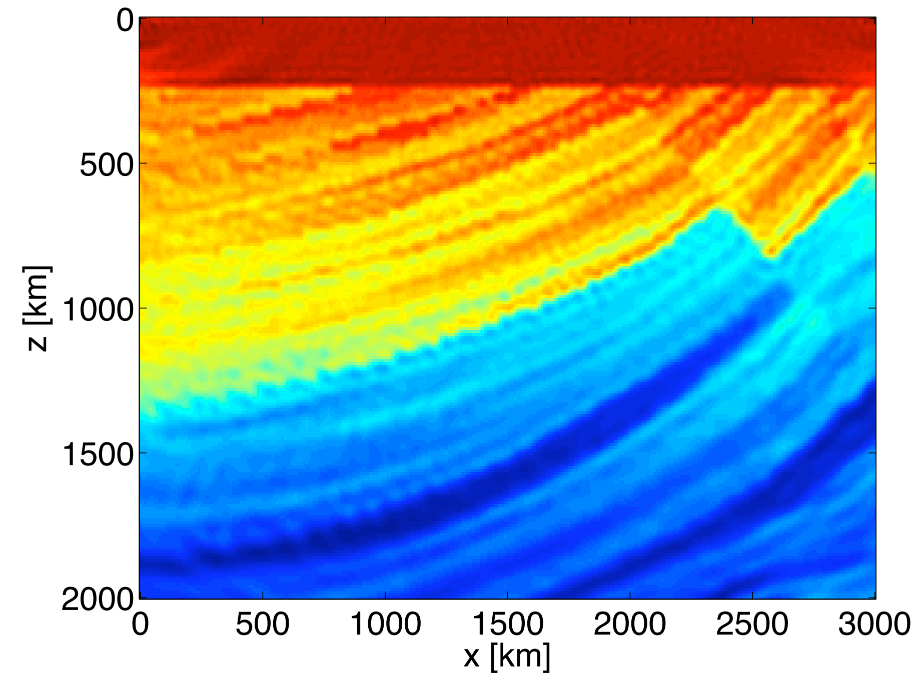
Requires *fewer* PDE solves for each GN *subproblem*...

- corresponds to *stochastic approximation* [Nemirovski, '09]
- related to Kaczmarz ('37) method applied by Natterer, '01
- *supersedes ad hoc* approach by Krebs *et.al.*, '09

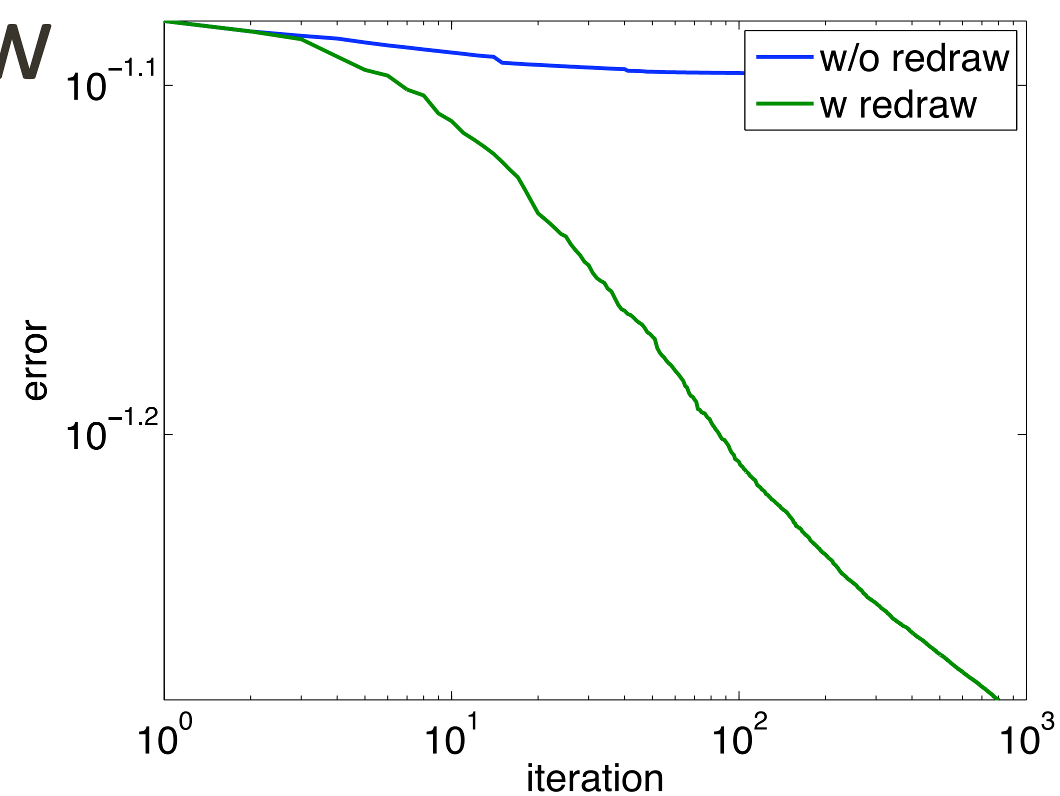
K=1 w and w/o redraw [noise-free case]



w/o redraw



w redraw



model error K=1, no averaging

Observations

SAA:

- Random trace estimates insightful but unclear how they relate to estimates for the model

SA

- Renewals improve convergence *significantly*
- *Averaging* removes noise *instability* but is *detrimental* to the *convergence*

Both produce '*noisy*' results ... Sounds *familiar?*

Combined approach

Leverage findings from *sparse recovery* & *compressive sensing*

- consider *phase-encoded* Gauss-Newton updates as separate *compressive-sensing* experiments
- remove *interferences* by *curvelet-domain sparsity* promotion
- exploit properties of Pareto curves in combination with stochastic optimization
- turn ‘overdetermined’ problems into ‘undetermined’ ones via *randomization*

SLIM

Rationale

Wavefields are *compressible* in curvelet *frames*

- *correlations* between *source* & *residual wavefields* are *compressible*
- *velocity* distributions of *sedimentary basins* are *also compressible*

Linearized subproblems are *convex*

Assume proximity Pareto *curves* for *successive linearizations*

Gauss-Newton

Algorithm 1: Gauss Newton

Result: Output estimate for the model $\mathbf{m}$

$\mathbf{m} \longleftarrow \mathbf{m}_0$; $k \longleftarrow 0$; `// initial model`

while not converged **do**

$\mathbf{p}^k \longleftarrow \arg\min_{\mathbf{p}} \frac{1}{2}\|\delta\mathbf{d} - \nabla\boldsymbol{\mathcal{F}}[\mathbf{m}^k;\mathbf{Q}]\mathbf{p}\|_2^2 + \lambda^k\|\mathbf{p}\|_2^2$; `// search dir.`

$\mathbf{m}^{k+1} \longleftarrow \mathbf{m}^k + \gamma^k\mathbf{p}^k$; `// update with linesearch`

$k \longleftarrow k+1$;

end

SLIM

Phase encoding

Algorithm 1: Gauss Newton with renewed phase encodings

Result: Output estimate for the model $\mathbf{m}$

$\mathbf{m} \longleftarrow \mathbf{m}_0$; $k \longleftarrow 0$; `// initial model`

while not converged **do**

$\quad \mathbf{p}^k \longleftarrow \arg\min_{\mathbf{p}} \frac{1}{2}\|\delta\underline{\mathbf{d}}^k - \nabla\boldsymbol{\mathcal{F}}[\mathbf{m}^k;\underline{\mathbf{Q}}^k]\mathbf{p}\|_2^2 + \lambda^k\|\mathbf{p}\|_2^2$; `// search dir.`

$\quad \mathbf{m}^{k+1} \longleftarrow \mathbf{m}^k + \gamma^k\mathbf{p}^k$; `// update with linesearch`

$\quad k \longleftarrow k+1$;

end

[Wang & Sacchi, '07]

Sparse recovery

Least-squares migration with *sparsity* promotion

$$\delta\widetilde{\mathbf{m}} = \mathbf{S}^* \arg\min_{\delta\mathbf{x}} \frac{1}{2}\|\delta\mathbf{x}\|_{\ell_1} \quad \text{subject to} \quad \|\delta\underline{\mathbf{d}} - \nabla\boldsymbol{\mathcal{F}}[\mathbf{m}_0; \underline{\mathbf{Q}}]\mathbf{S}^*\delta\mathbf{x}\|_2 \leq \sigma$$

$$\delta\mathbf{x} = \text{Sparse curvelet-coefficient vector}$$

$$\mathbf{S}^* = \text{Curvelet synthesis}$$

leads to *significant* speedup as long as

$$n_{PDE}^{\ell_1} \times K \ll n_{PDE}^{\ell_2} \times n_f \times n_s$$

Compressive updates

Algorithm 1: Gauss Newton with sparse updates

Result: Output estimate for the model $\mathbf{m}$

$\mathbf{m} \longleftarrow \mathbf{m}_0;\ k \longleftarrow 0$; `// initial model`

while not converged **do**

$\mathbf{p}^k \longleftarrow \mathbf{S}^* \arg\min_{\mathbf{x}} \frac{1}{2}\|\delta\underline{\mathbf{d}}^k - \nabla\mathcal{F}[\mathbf{m}^k;\underline{\mathbf{Q}}^k]\mathbf{S}^*\mathbf{x}\|_2^2 \quad \text{s.t.} \quad \|\mathbf{x}\|_1 \leq \tau^k$

$\mathbf{m}^{k+1} \longleftarrow \mathbf{m}^k + \gamma^k\mathbf{p}^k$; `// update with linesearch`

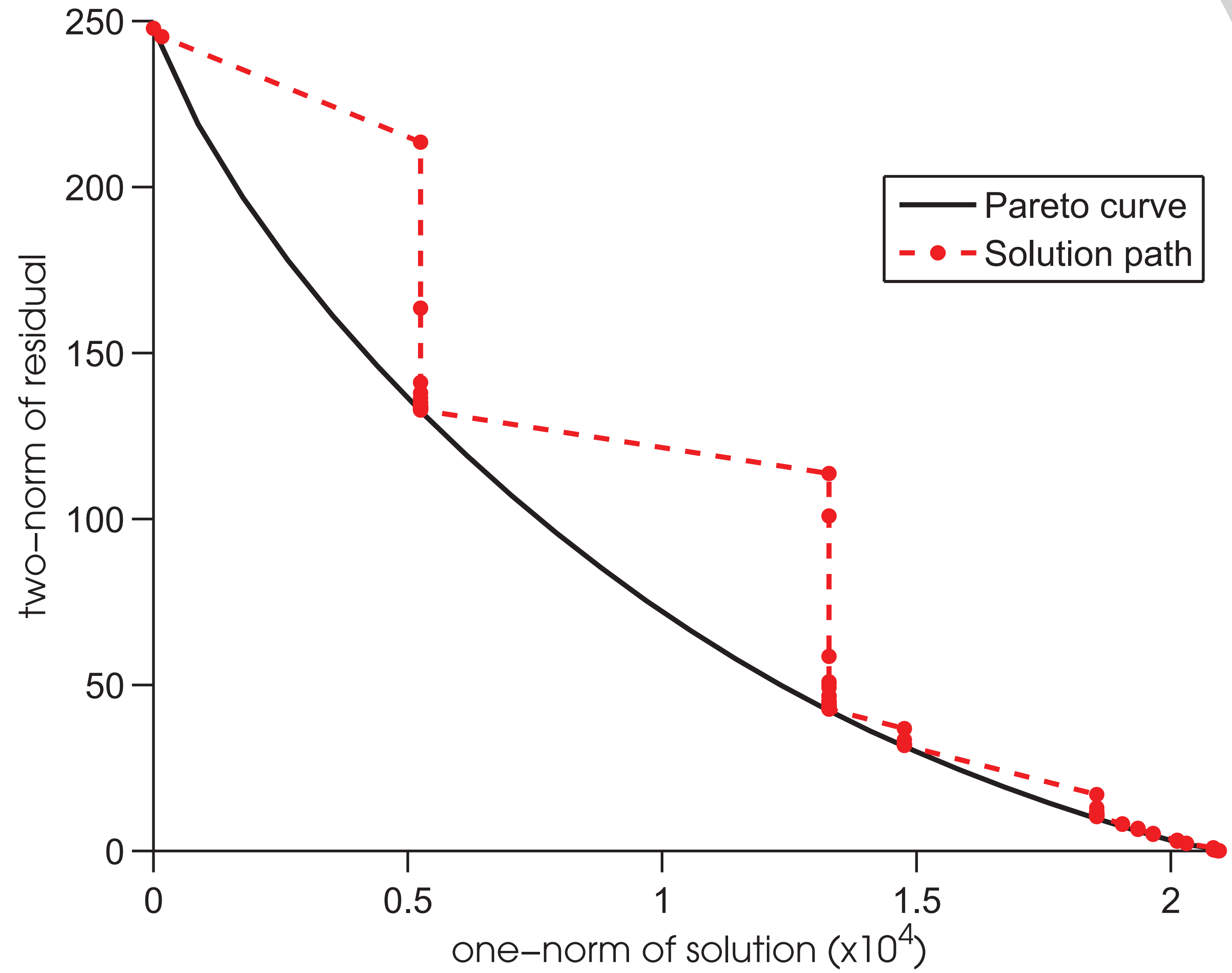
$k \longleftarrow k+1;$

end

[van den Berg & Friedlander, ’08]

Solution strategy

- Draw *new* CS *experiment* when *Pareto curve* is reached
- Do *new linearization*
- *Sweep* from low to hight frequencies



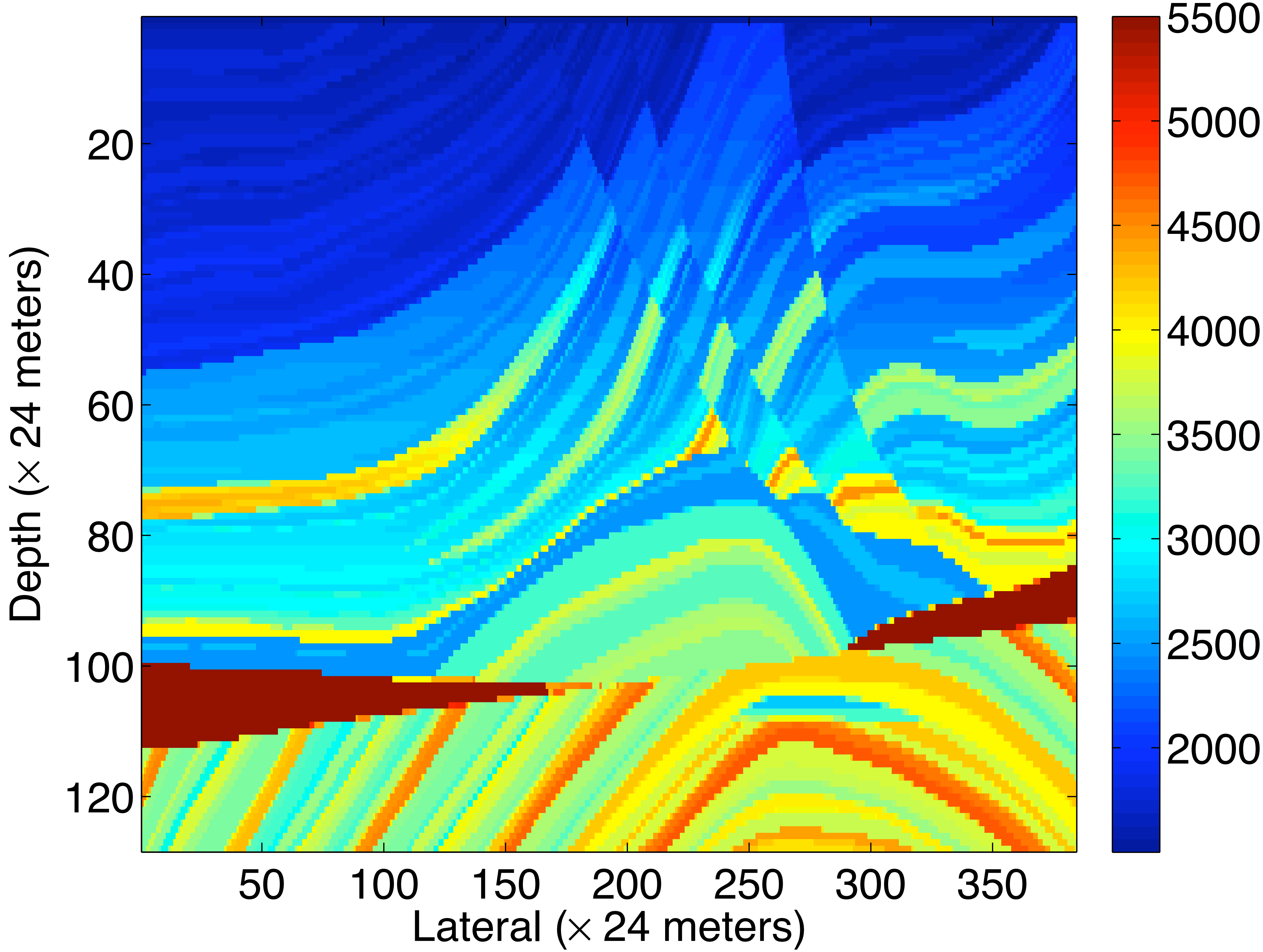
Example

FWI specs:

- Committed *inversion crime*
- Frequency continuation over 10 bands
- 15 *simultaneous* shots with 10 *frequencies* each

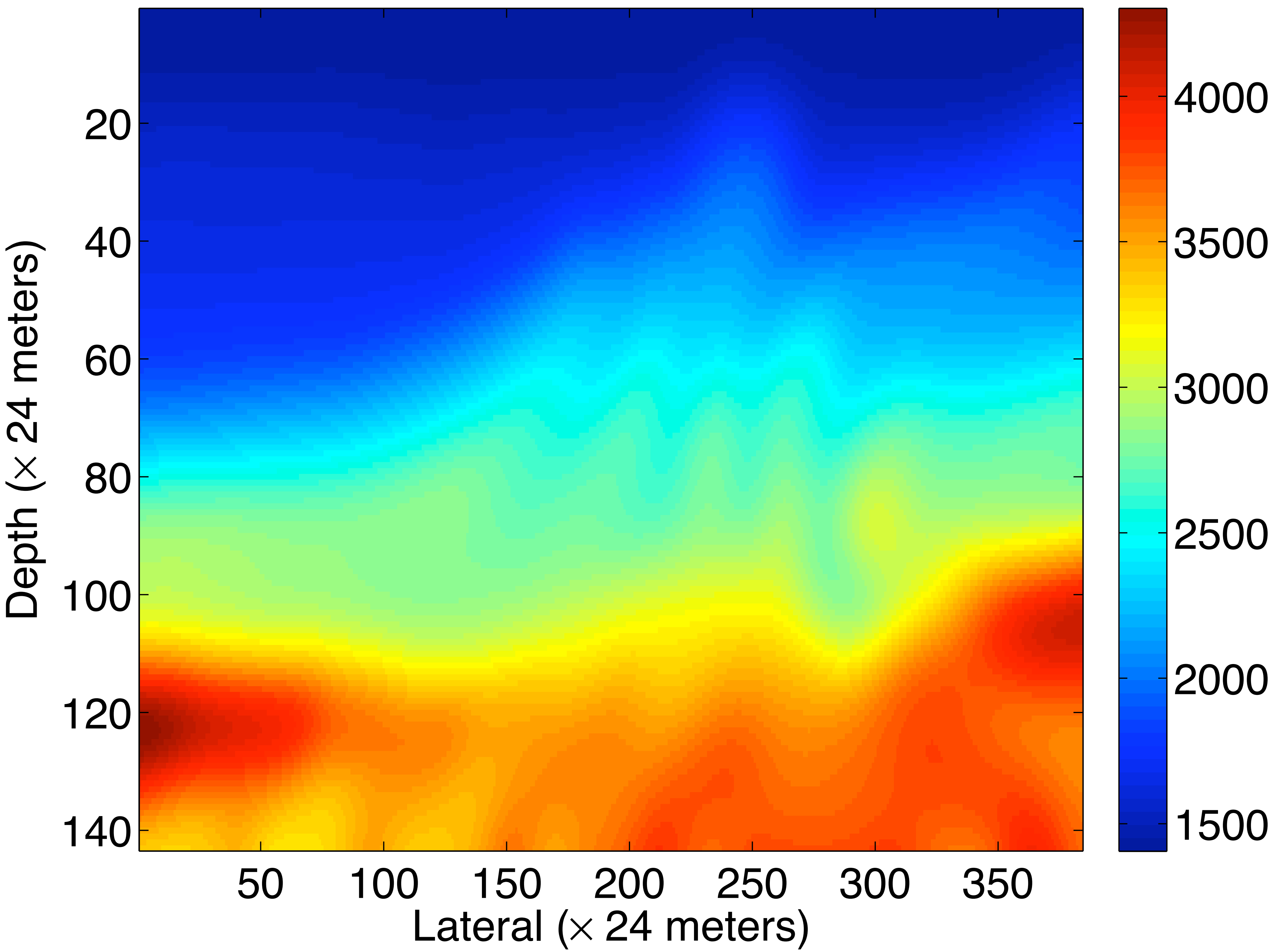
$$K = 10 \times 15 \ll 100 \times 384$$

True model

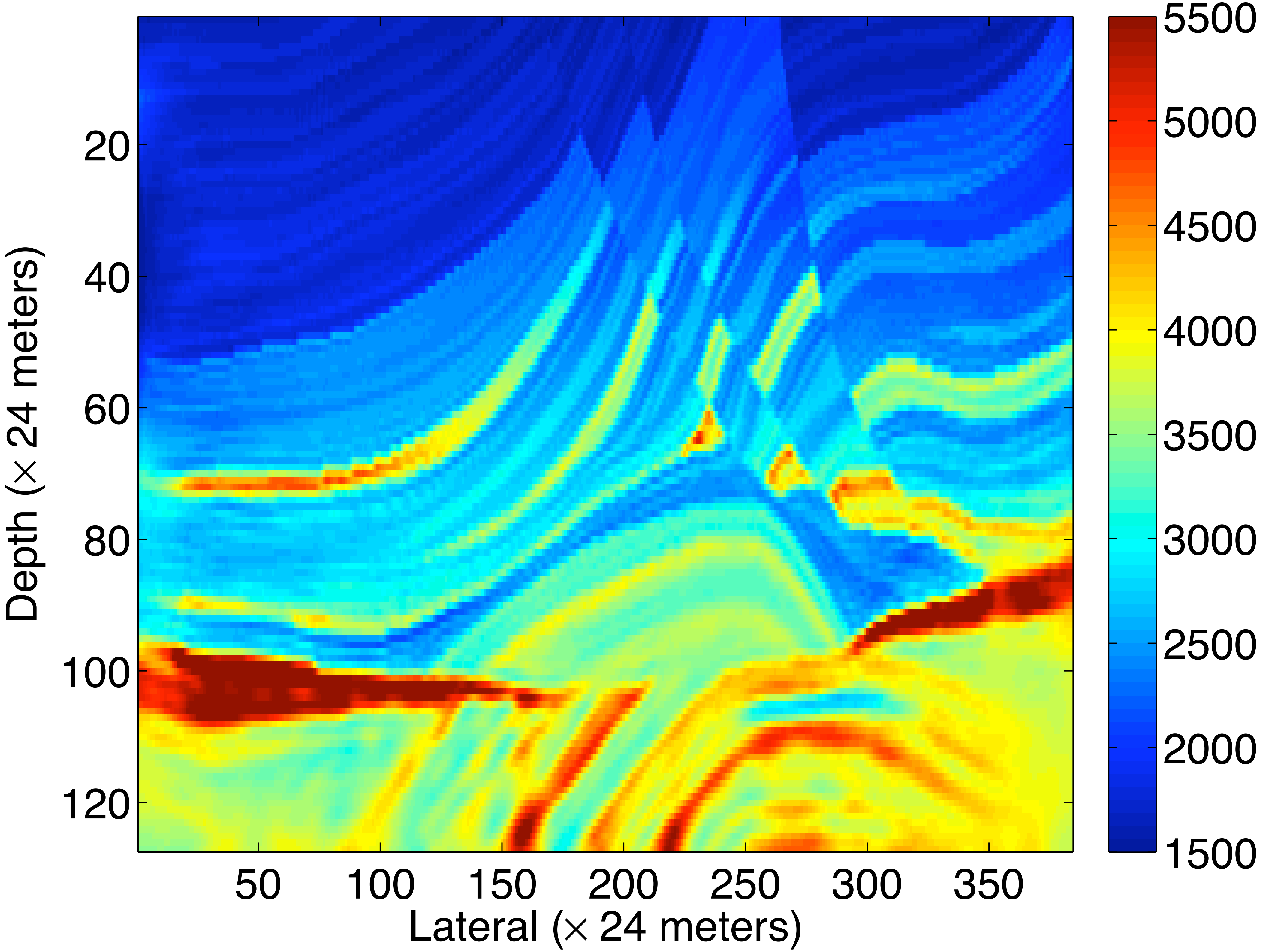


Depth (× 24 meters)
20
40
60
80
100
120
50
100
150
200
250
300
350
Lateral (× 24 meters)
5500
5000
4500
4000
3500
3000
2500
2000

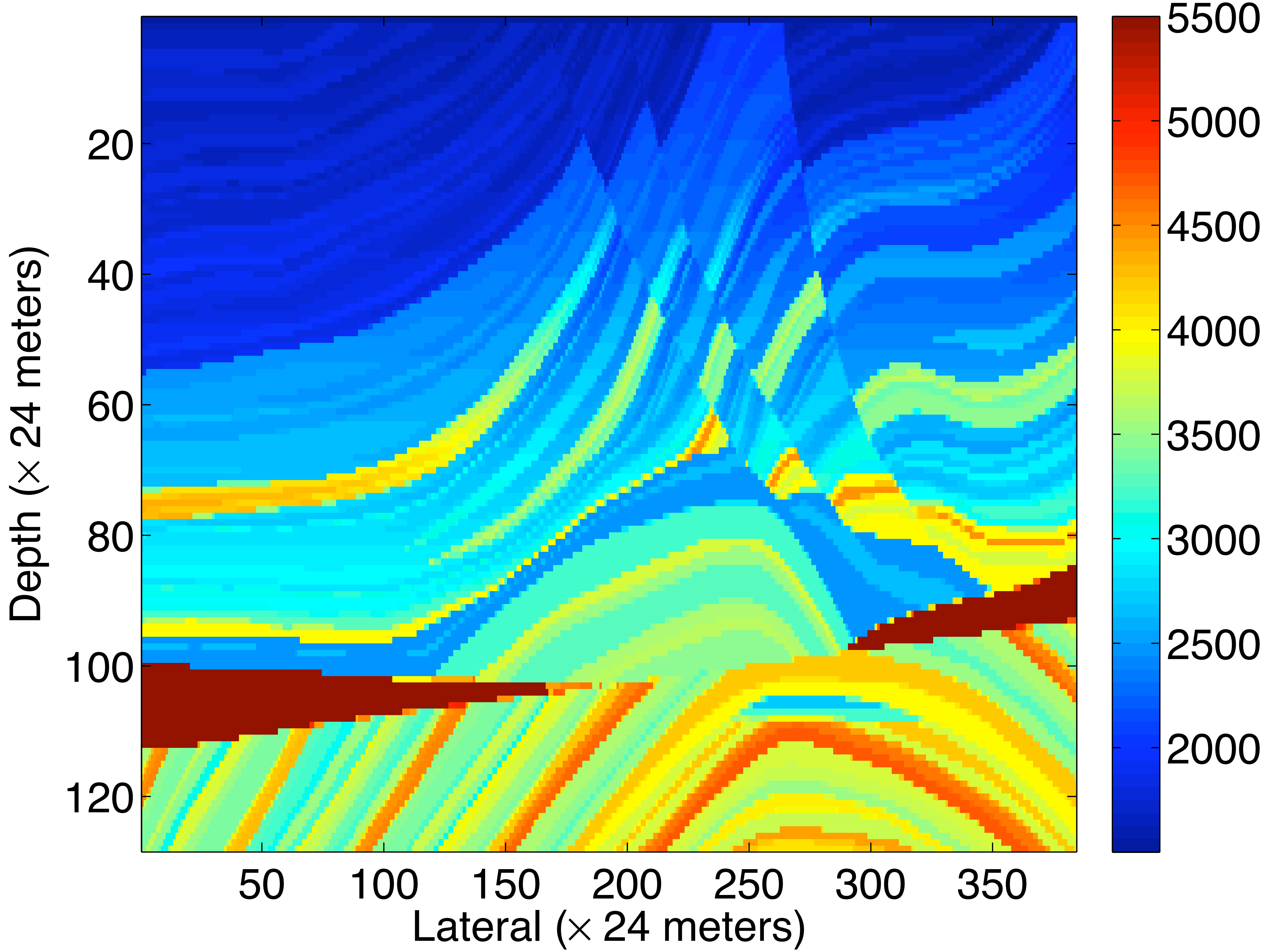
Initial model



Inverted model

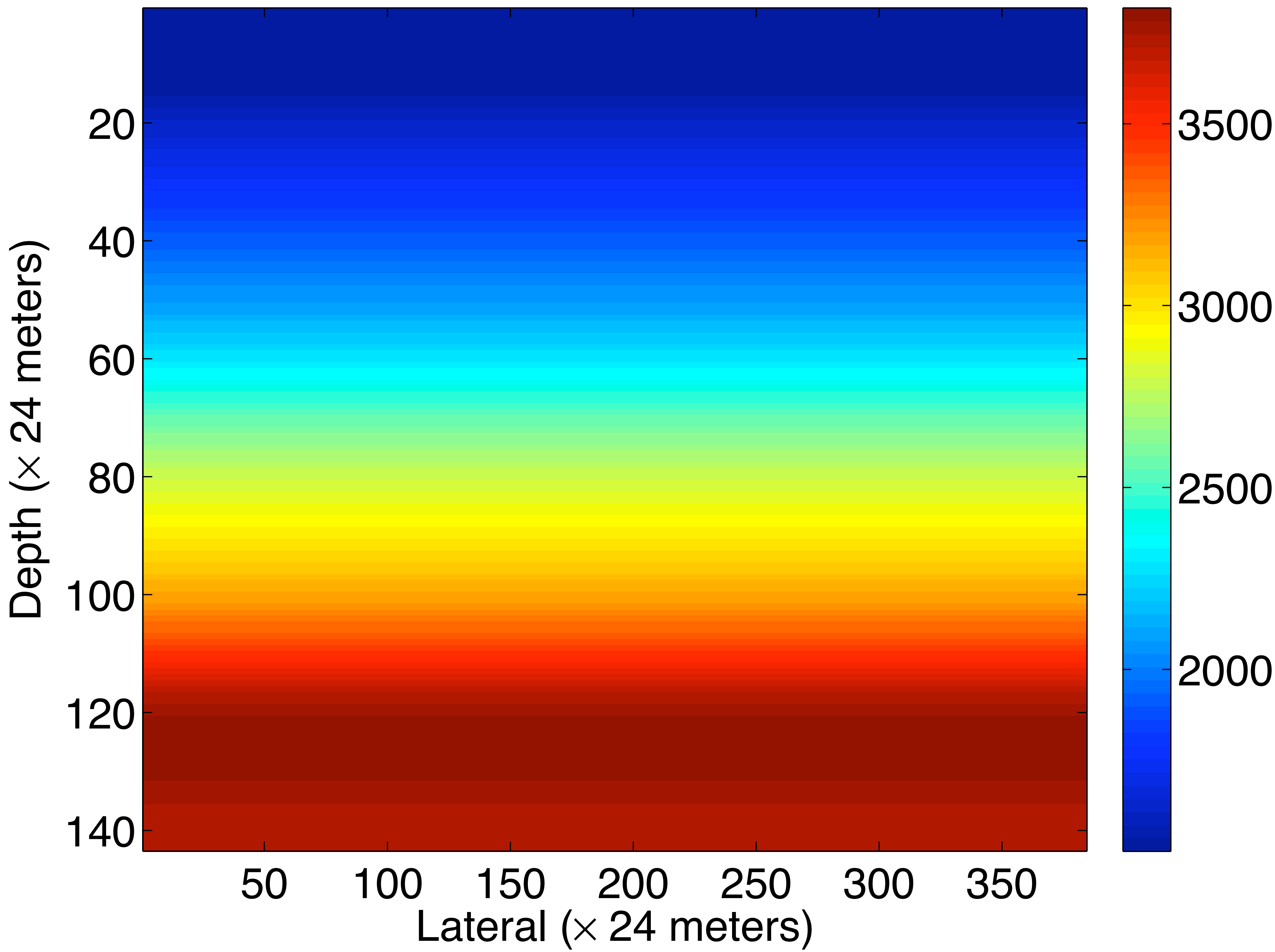


True model

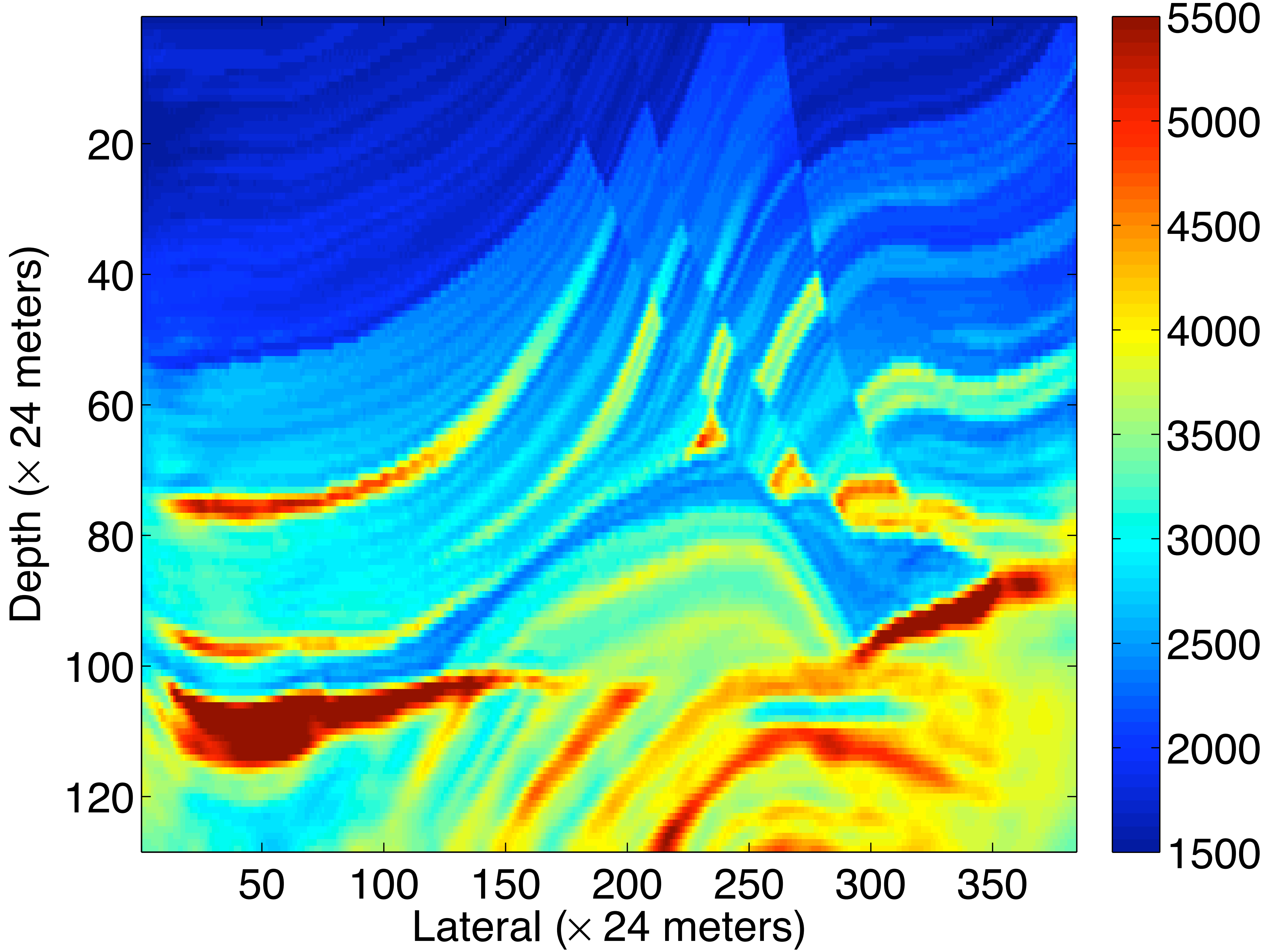


Depth (× 24 meters)
20
40
60
80
100
120
Lateral (× 24 meters)
50
100
150
200
250
300
350
5500
5000
4500
4000
3500
3000
2500
2000

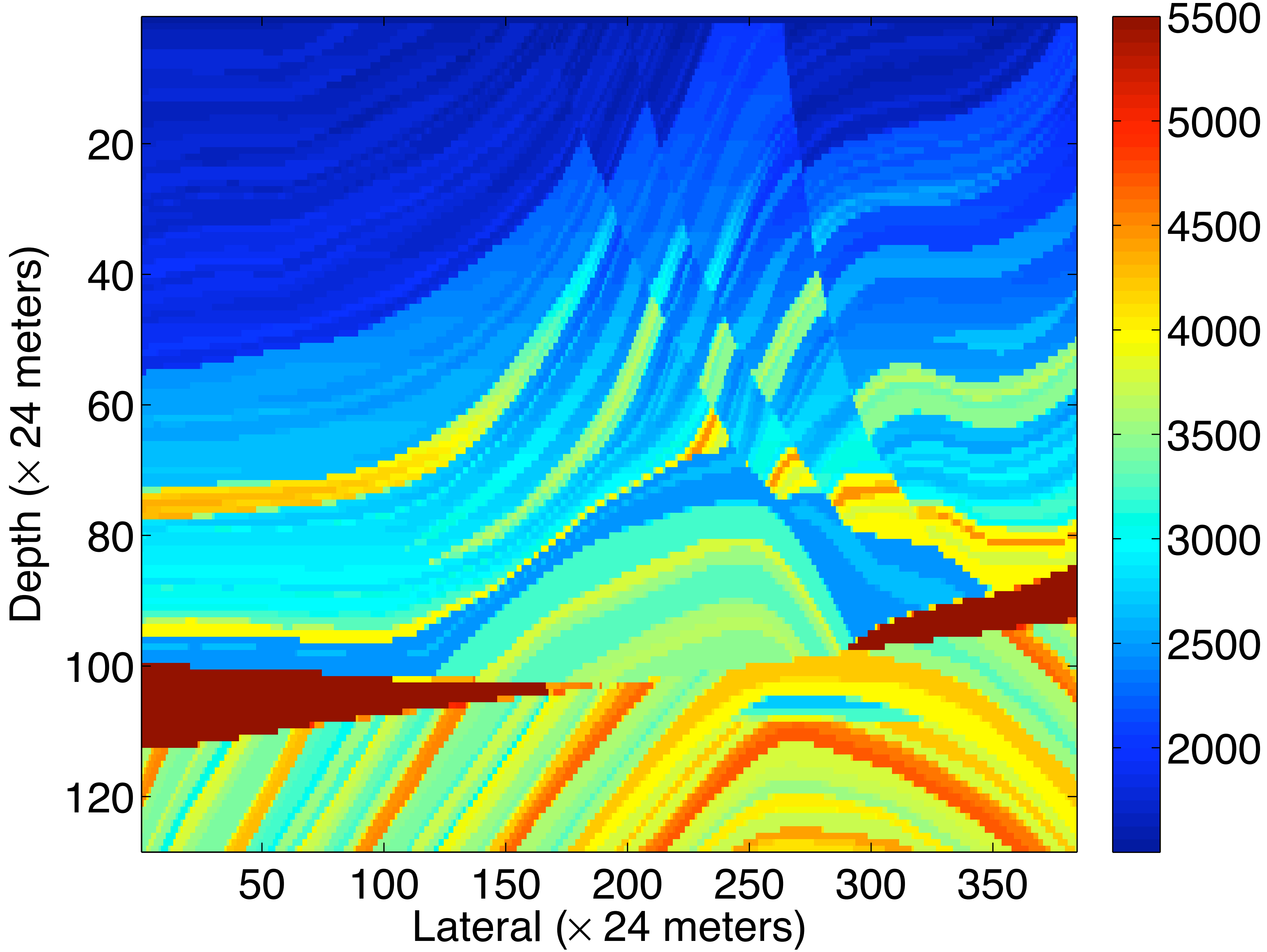
Initial model



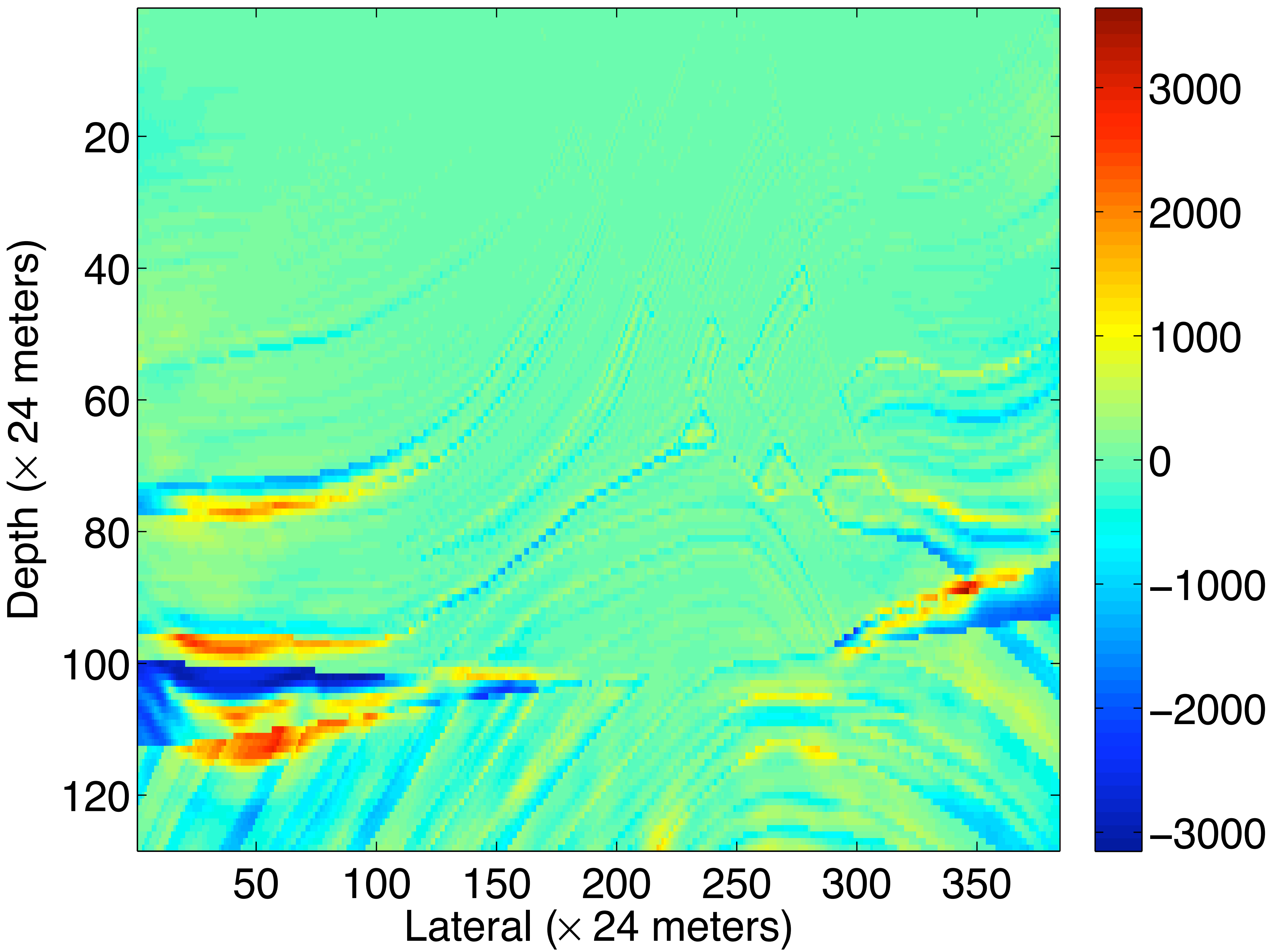
Inverted model



True model


Depth (× 24 meters)
20
40
60
80
100
120
50
100
150
200
250
300
350
Lateral (× 24 meters)
5500
5000
4500
4000
3500
3000
2500
2000

Difference


20
40
60
80
100
120
Depth (× 24 meters)
50
100
150
200
250
300
350
Lateral (× 24 meters)
3000
2000
1000
0
−1000
−2000
−3000

Performance

Remember per *subproblem*

$$n_{PDE}^{\ell_1} \times K \ll n_{PDE}^{\ell_2} \times n_f \times n_s$$

$$\begin{aligned} n_{PDE}^{\ell_1} &\approx 200 \\ K &= 150 \end{aligned}$$

versus

$$\begin{aligned} n_{PDE}^{\ell_2} &\approx 10 \\ K &= 38400 \end{aligned}$$

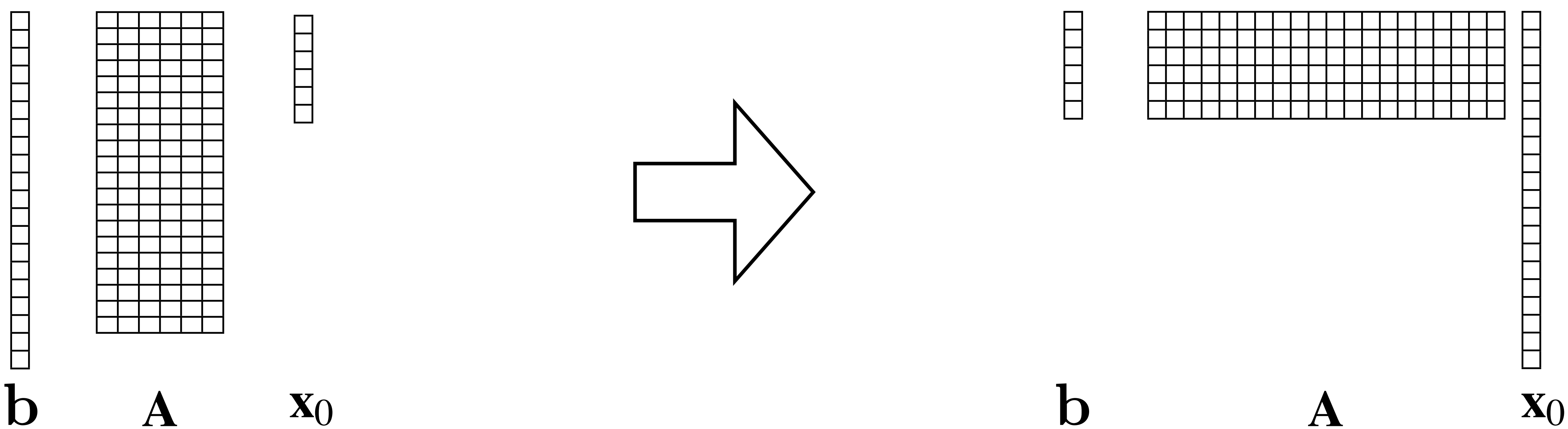
SPEEDUP of 13 X

Recap

$$\min_{\mathbf{x}} \|\mathbf{b} - \mathbf{A}_{\text{total}}\|_2 \qquad\qquad \min_{\mathbf{x}} \|\mathbf{b} - \mathbf{A}_k\mathbf{x}\|_2 \quad \text{s.t.} \quad \|\mathbf{x}\|_1 \leq \tau_k$$

overdetermined
expensive

underdetermined
cheap

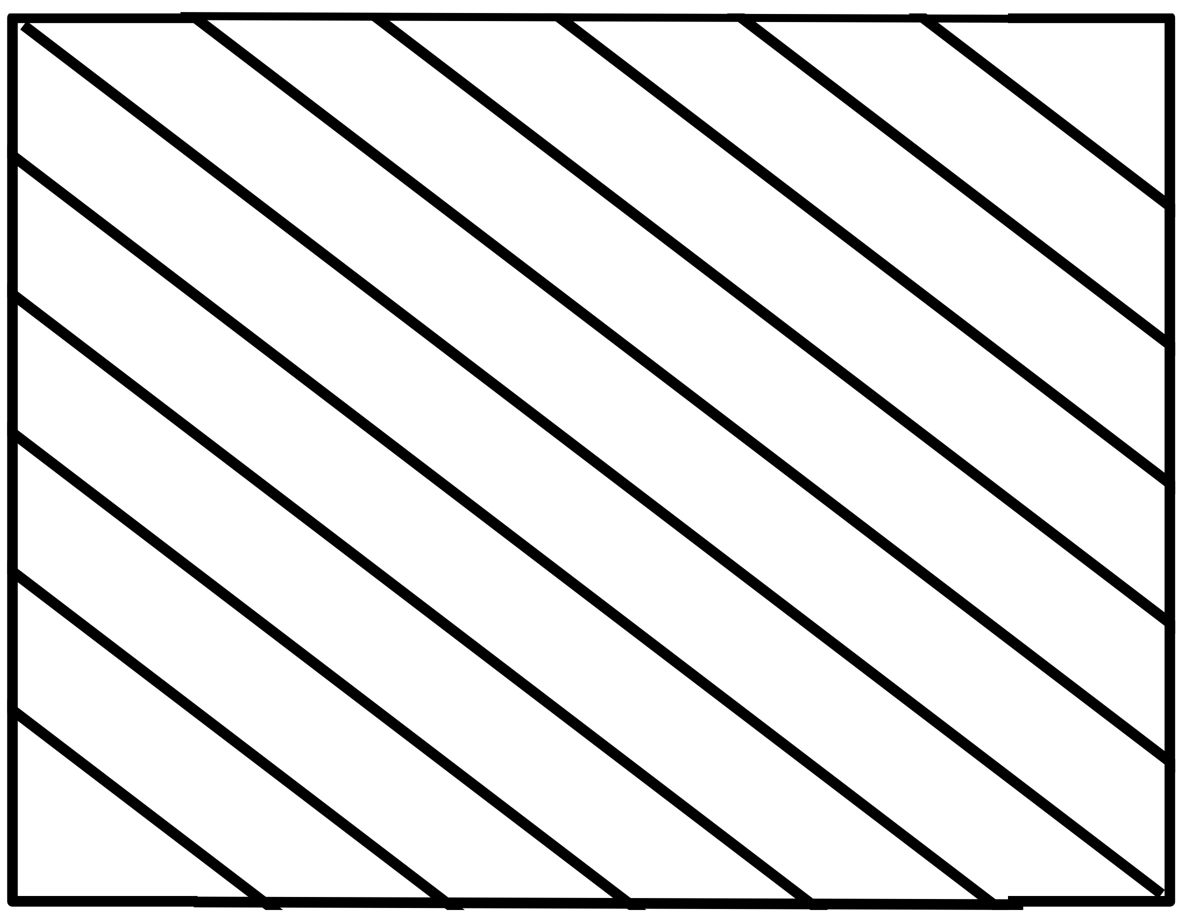


Choose a new set of *simultaneous* sources after each ‘GN’ subproblem is solved

SLIM

Carry home message ...

Seismic inversion involves very large *full matrices*

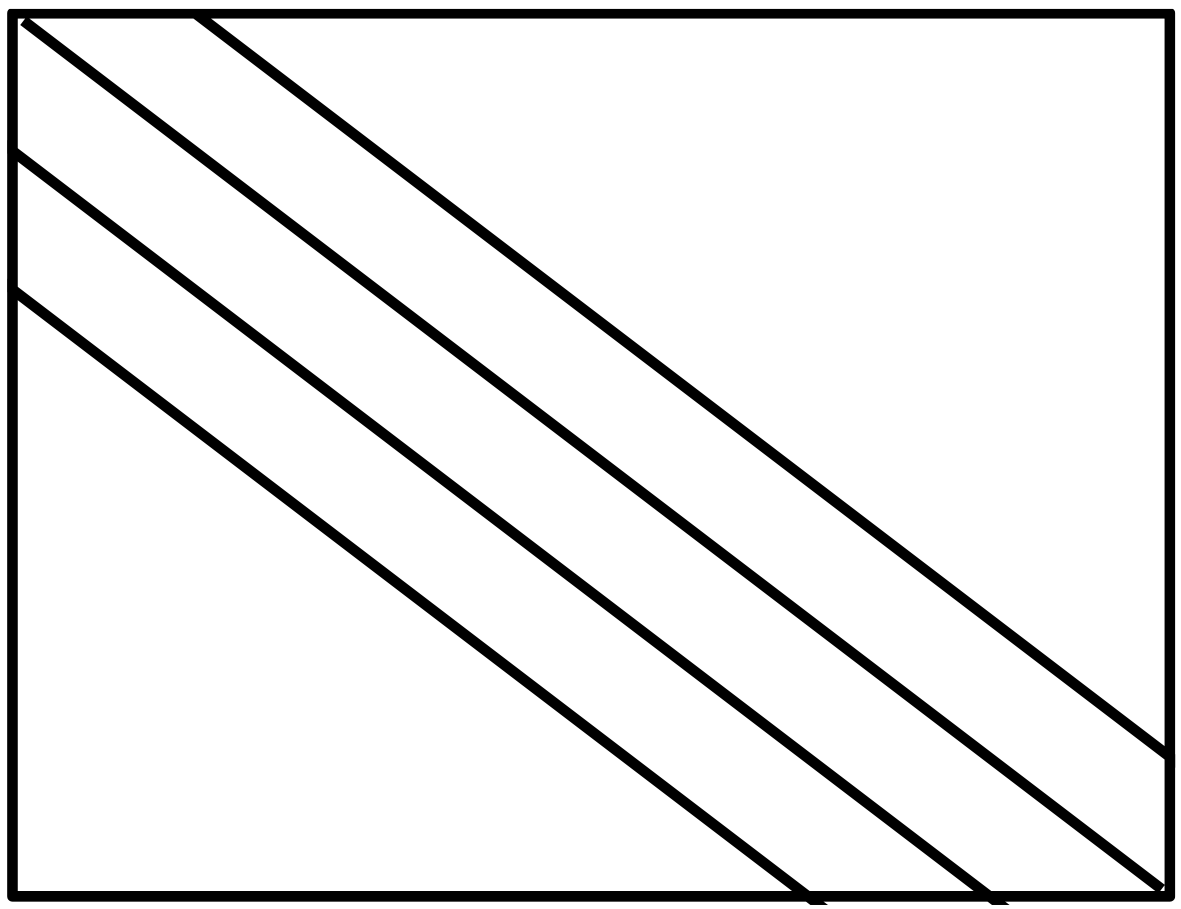


[de Hoop et al., ‘08-’09]
[Smit et. al., ‘09]

SLIM

“Holy grail”

Find a representation to “diagonalize”



"Holy grail"

Major engineering effort to keep track of matrix *permutations*

- killed by constants
- leaks to off diagonals

[FJH & Lin., ‘07]
[Demanet & Peyré, ‘08]

CS alternative

Model-size reduction by CS

SLIM

CS alternative

Leverage *invariance* under solution operators <=> preservation of *sparsity*

Sparsity promotion takes care of keeping track of the *permutations* ***implicitly*** **...!**

SLIM

Conclusions

Leveraged

- curvelet-domain sparsity on the model
- invariance under solution operators <=> preservation of sparsity

Indications that *compressive sensing* supersedes the *stochastic approximation by sparse recovery of dimensionality reduced* subproblems