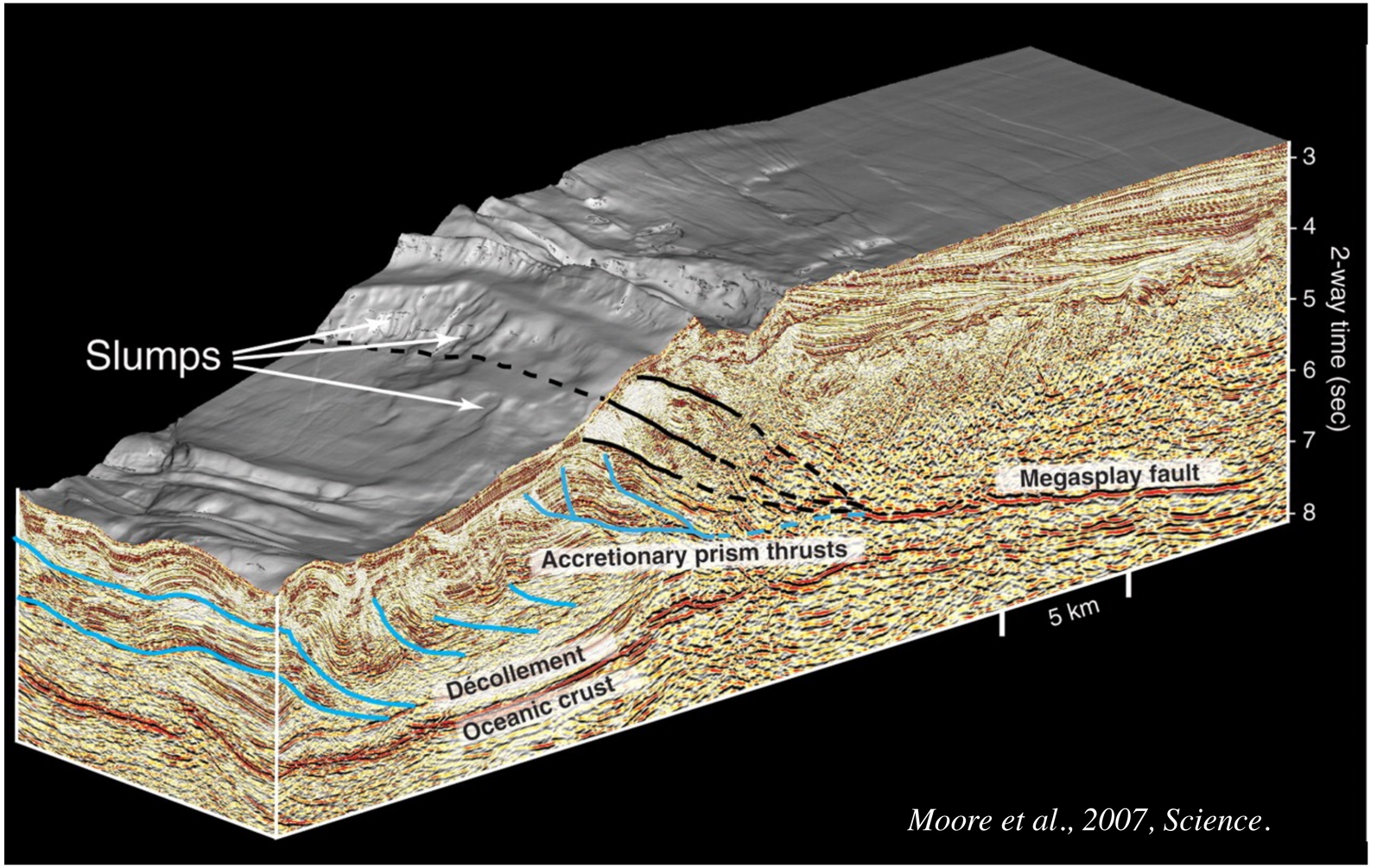
SIOG 231
GEOMAGNETISM AND ELECTROMAGNETISM

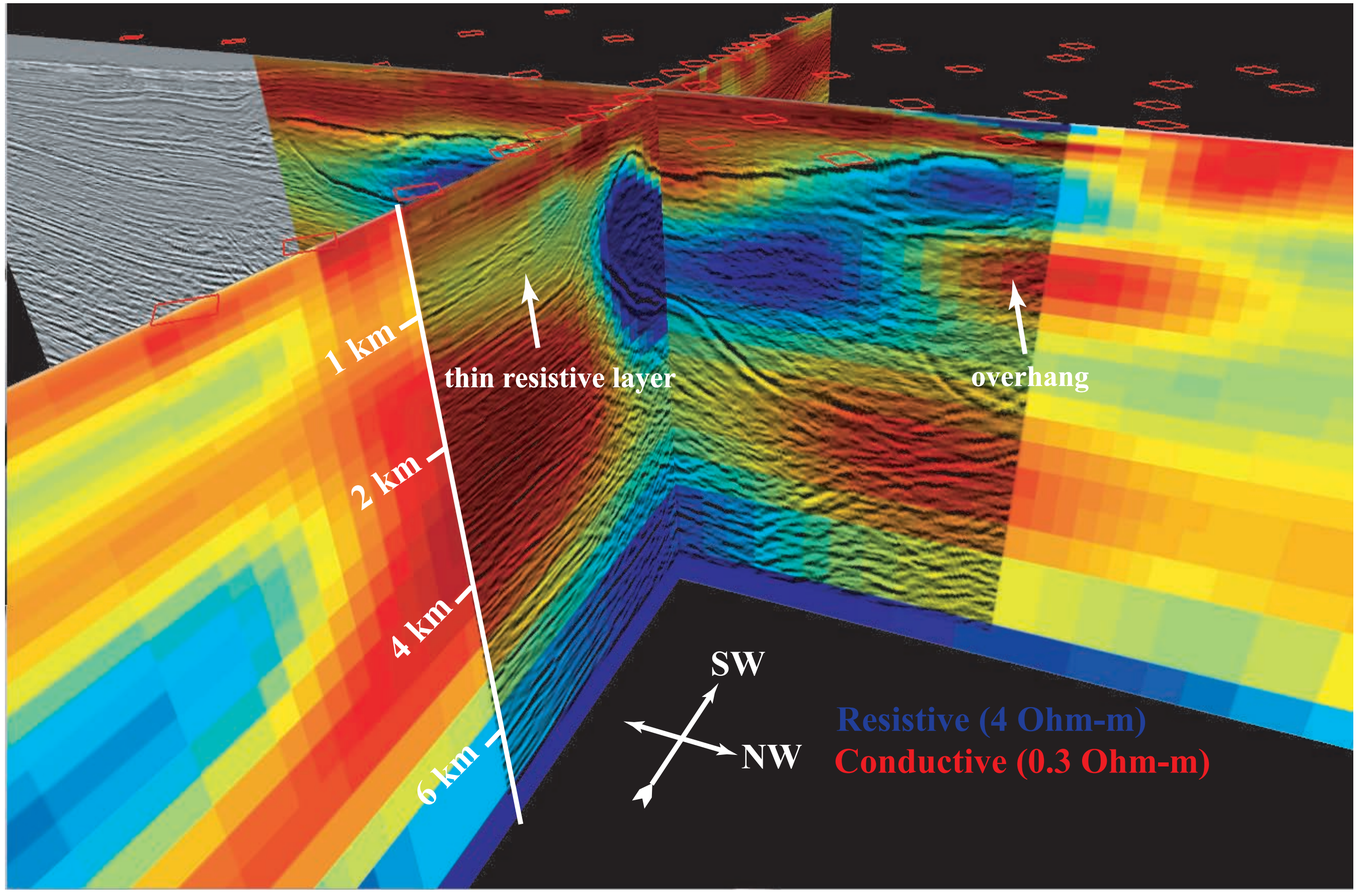
Lecture 17
Regularized Inversion
1/3/2022

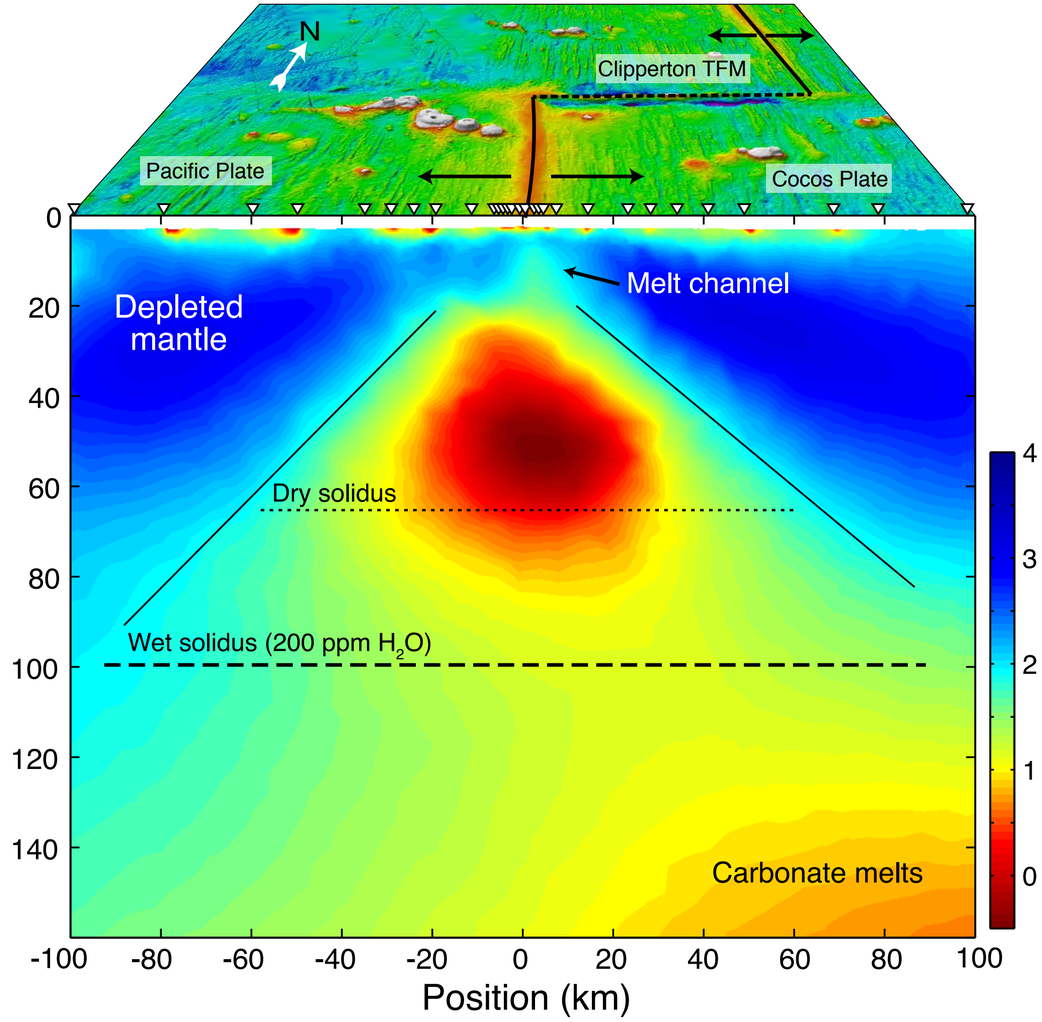
With seismic reflection images you can often see the geology in the data.

For EM and potential field methods you need inversion to recover something that can be interpreted as geology. (Same is true for seismic tomography and full waveform inversion, as well as potential field methods.)

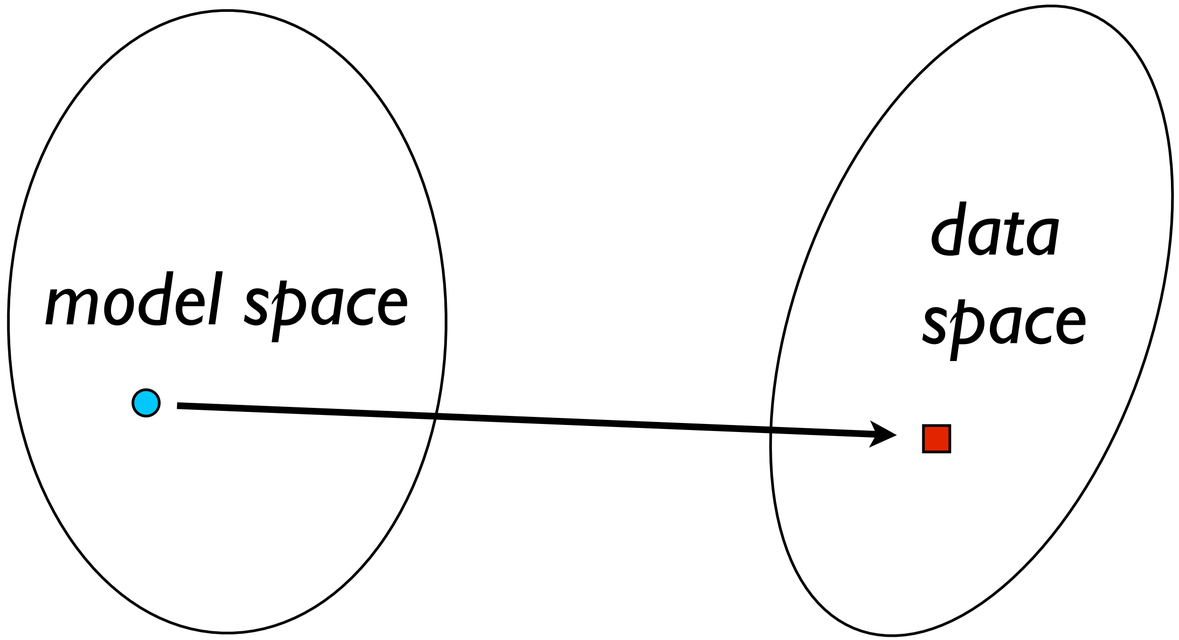


So how do we get images like these?





Forward modeling:



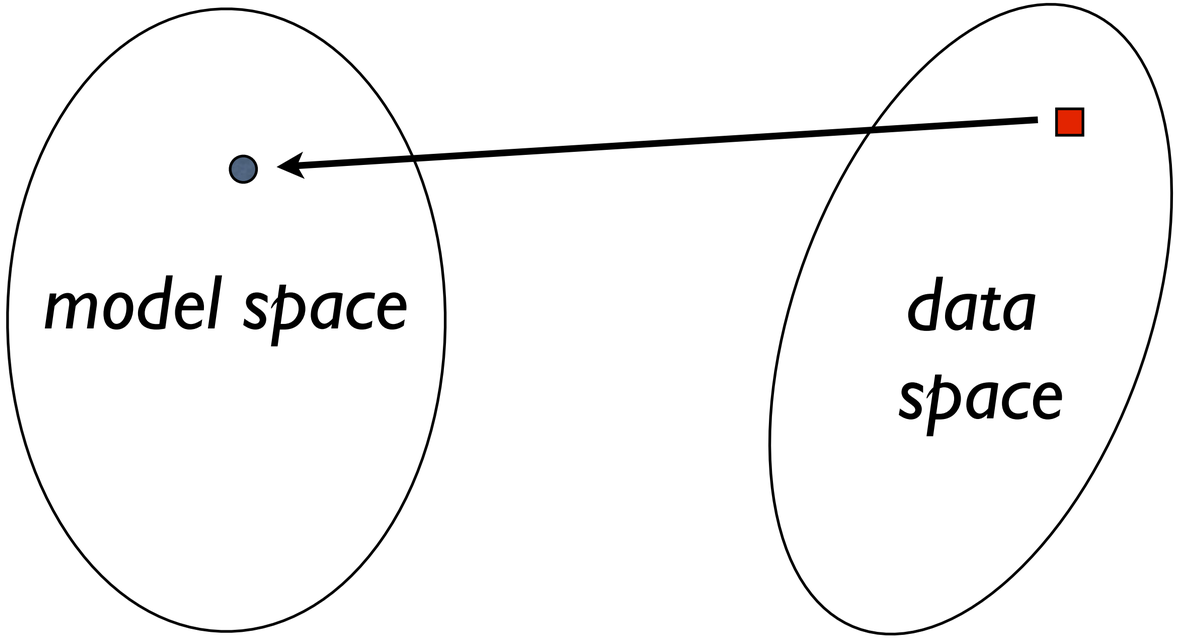
$$\hat{\mathbf{d}} = f(\mathbf{x}, \mathbf{m})$$ Some forward functional f

$$\mathbf{m} = (m_1, m_2, ..., m_N)$$ Model parameters (layers, blocks, ...)

$$\mathbf{x} = (x_1, x_2, x_3, ..., x_{kM})$$ Independent variables (freqs., locations, ...)

$$\hat{\mathbf{d}} = (\hat{d}_1, \hat{d}_2, \hat{d}_3, ..., \hat{d}_M)$$ Predicted data (gravity, magnetic, electric, ...)

Inverse modeling:



Given real (observed) data $\mathbf{d} = (d_1, d_2, d_3, ..., d_M)$

with errors $\sigma = (\sigma_1, \sigma_2, ..., \sigma_M)$

find an $\mathbf{m}$

There are various approaches to inversion:

Trial and error modeling

Stochastic

Monte Carlo, Markov Chains
Genetic Algorithms
Simulated annealing, etc.
(Bayesian Searches)

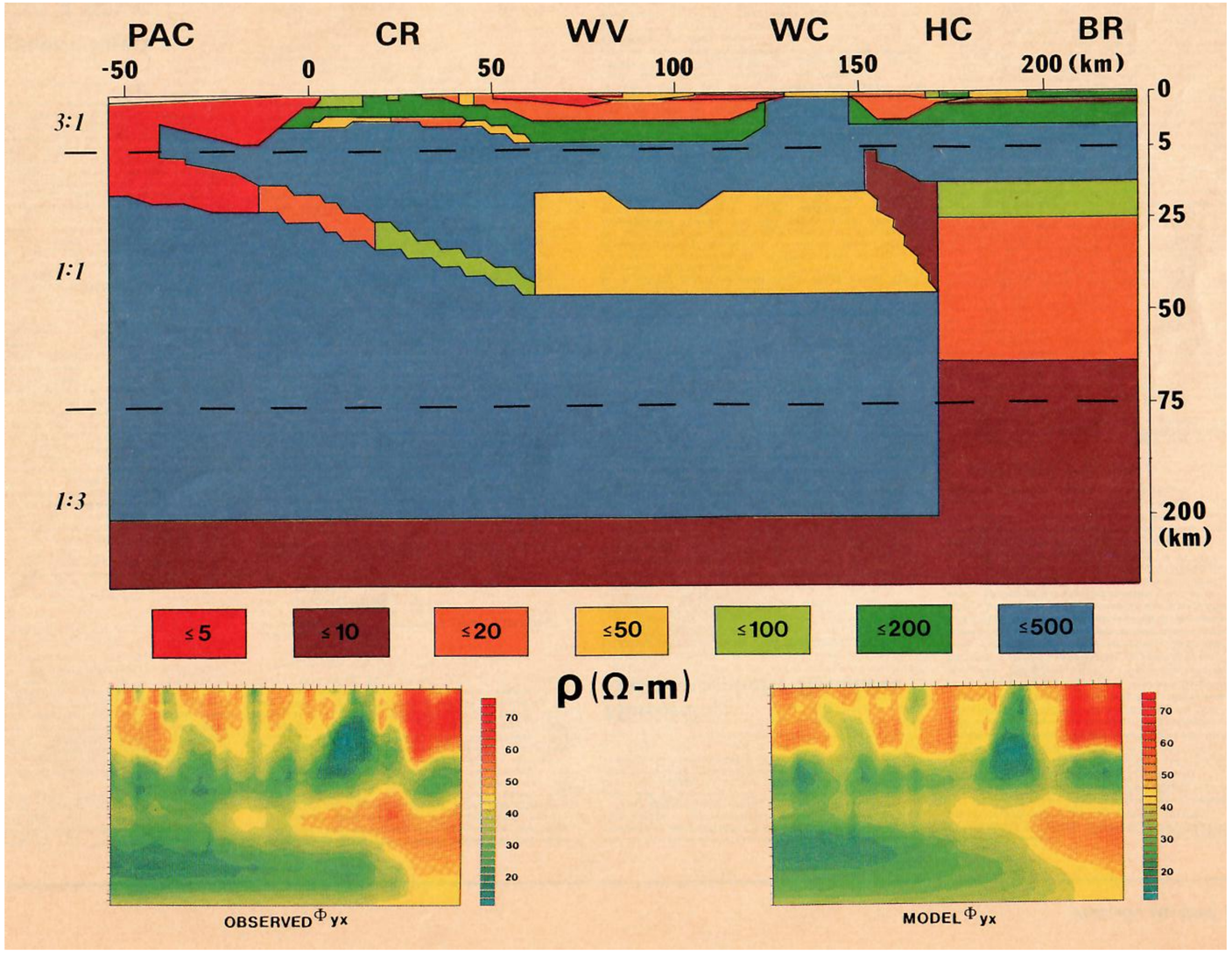
Deterministic

Newton Algorithms
Steepest descent
Conjugate Gradients
Quadratic (and Linear) Programming, etc.

Analytical

D+ (1D MT)
Bilayer (1D resistivity)
Ideal body theory in gravity and magnetism

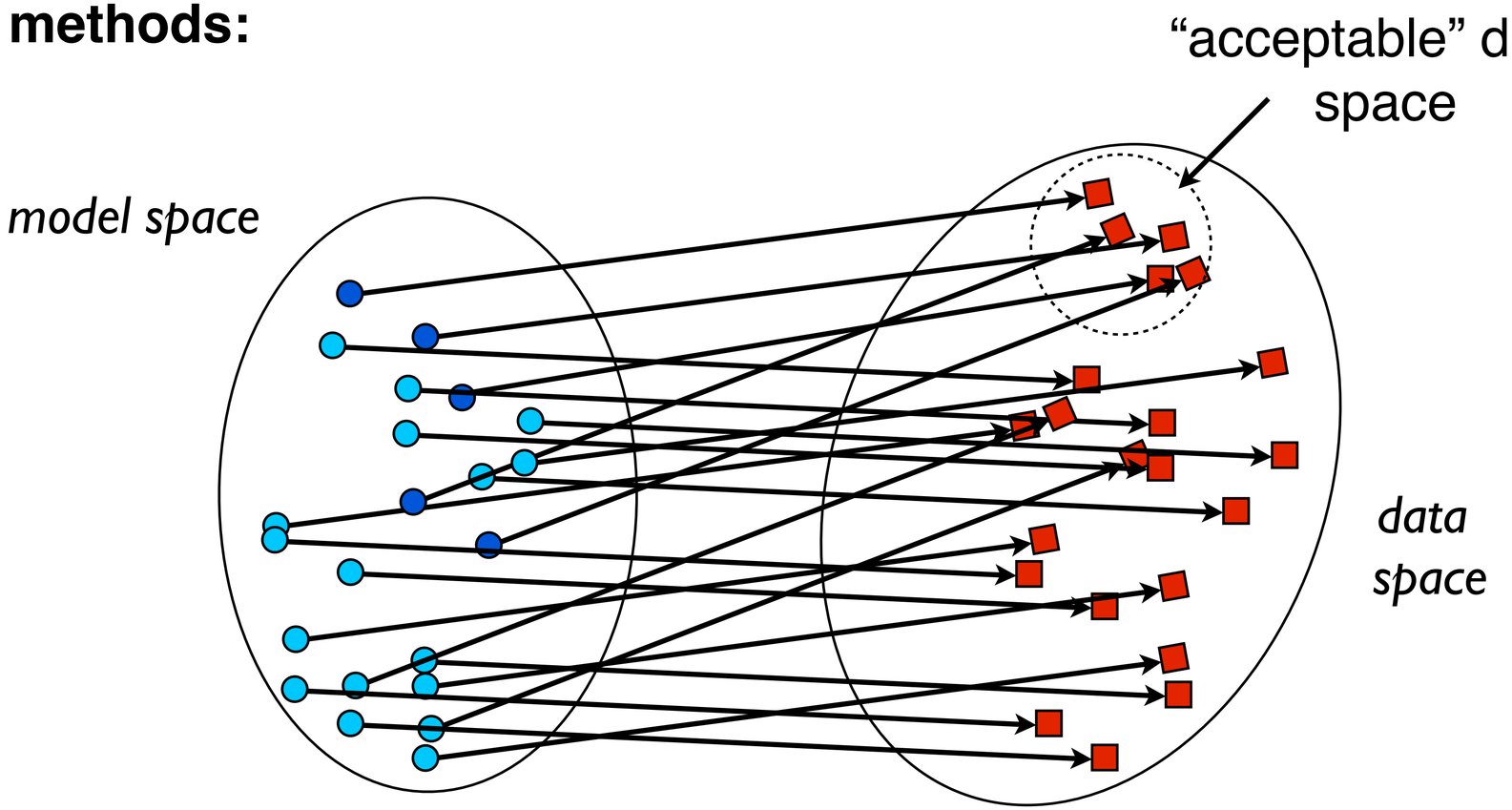
Trial and error modeling



(from EOS, Feb 1988)

The first 2D regularized inversion wasn't available until about 1990.

Stochastic methods:

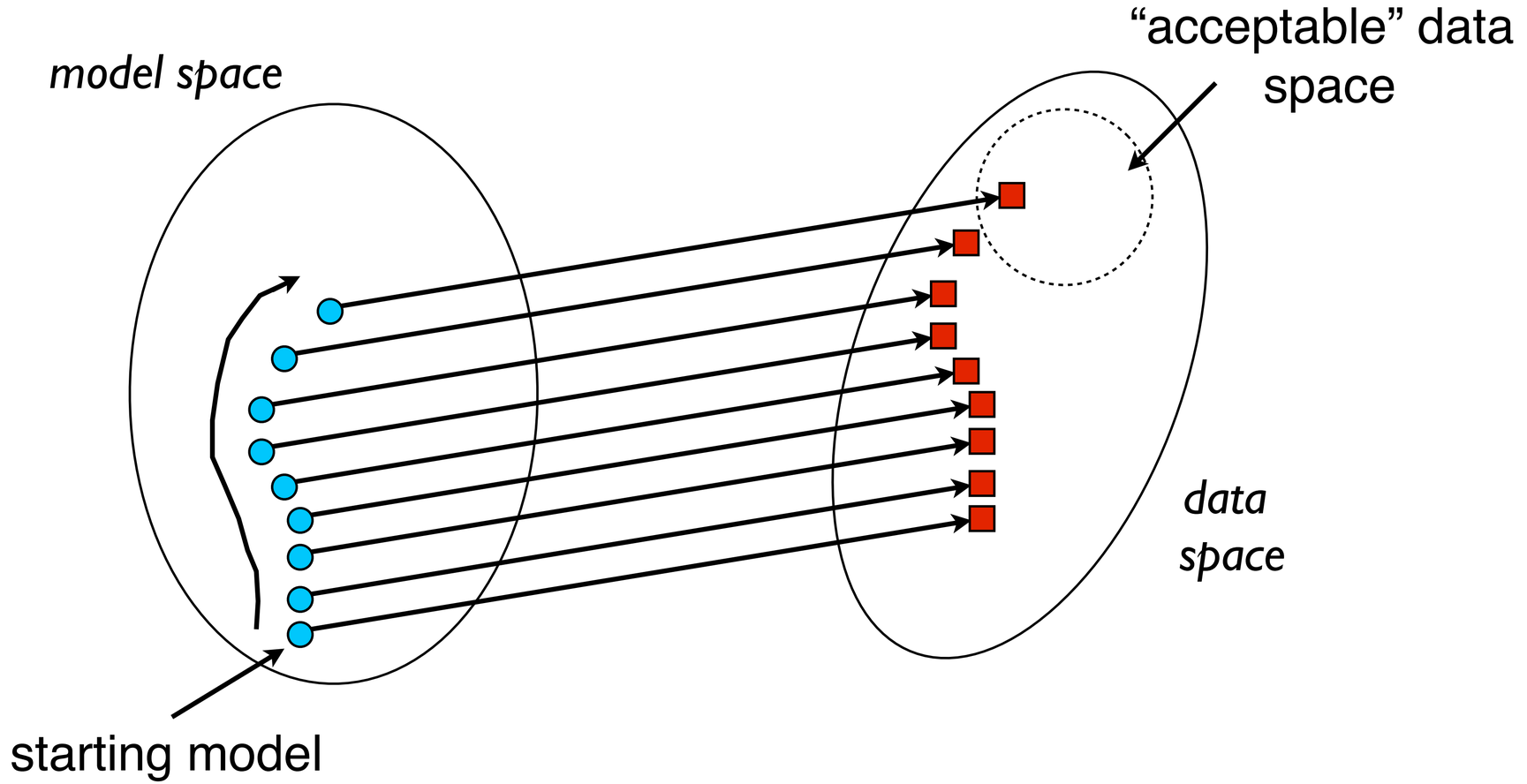


Often called **Bayesian inversion**. A useful approach, largely restricted to simple problems (because millions of models required), with most of the subtlety in model generation methods.

The advantages are that (i) only forward calculations are made and (ii) some statistics can be obtained on model parameters. Best for sparsely parameterized models. One needs to be careful that bounds on explored model space don't unduly influence the outcome.

Deterministic

Newton Algorithms
Steepest descent
Conjugate Gradients



The direction of the search is determined by how changing parts of the model affects the fit to the data. This usually involves using the Jacobian matrix

$$J_{ij} = \frac{\partial f(x_i, \mathbf{m}_0)}{\partial m_j}$$

Analytical

e.g. D+ (1D MT) and Bilayer (DC resistivity)

my data

and across the insulating interval $z_k < z < z_{k+1}$ we find

$$E_{k+1} = E_k + (z_{k+1} - z_k) D_k^+ \tag{52}$$

$$= E_k + (z_{k+1} - z_k) D_{k+1}^- \tag{53}$$

Define the admittance just above the k-th conductor in the usual way

$$C_k = -E_k / D_k^- . \tag{54}$$

Then by means of equations (50), (51) and (53) we can eliminate the E_k and $D_k^\pm$ as we did for uniform layers (although C_k is not continuous):

$$C_k = \frac{E_k}{-D_k^-} = \frac{E_k}{\mathrm{i}\omega\mu_0\tau_k E_k - D_k^+} = \frac{1}{\mathrm{i}\omega\mu_0\tau_k - D_k^+/E_k} \tag{55}$$

$$= \frac{1}{\mathrm{i}\omega\mu_0\tau_k - D_{k+1}^-/E_k} = \frac{1}{\mathrm{i}\omega\mu_0\tau_k - \dfrac{D_{k+1}^-}{E_{k+1} - (z_{k+1} - z_k) D_{k+1}^-}} . \tag{56}$$

Finally, dividing by D_{k+1}^- in the bottom tier we find the connection between the admittance at one level to the one above:

$$C_k = \frac{1}{\mathrm{i}\omega\mu_0\tau_k + \dfrac{1}{z_{k+1} - z_k + C_{k+1}}} . \tag{57}$$

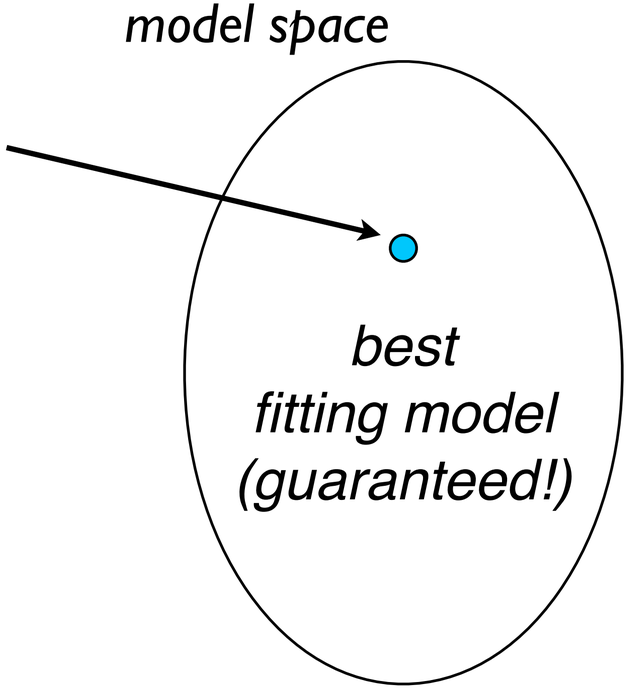
We could solve (48) by recurring upwards in the familiar way, starting with $E(H) = 0 = C_{K+1}$, to get the value of $E(0)$ and hence of $C_1 = c(\omega)$. But now we do something different: we substitute repeatedly from the top, and we get a magnificent **continued fraction** for the admittance:

$$c(\omega) = z_1 + \cfrac{1}{\mathrm{i}\omega\mu_0\tau_1 + \cfrac{1}{z_2 - z_1 + \cfrac{1}{\mathrm{i}\omega\mu_0\tau_2 + \cfrac{1}{z_3 - z_2 + \cfrac{1}{\mathrm{i}\omega\mu_0\tau_3 + \cdots \cfrac{1}{H - z_K}}}}}} . \tag{58}$$

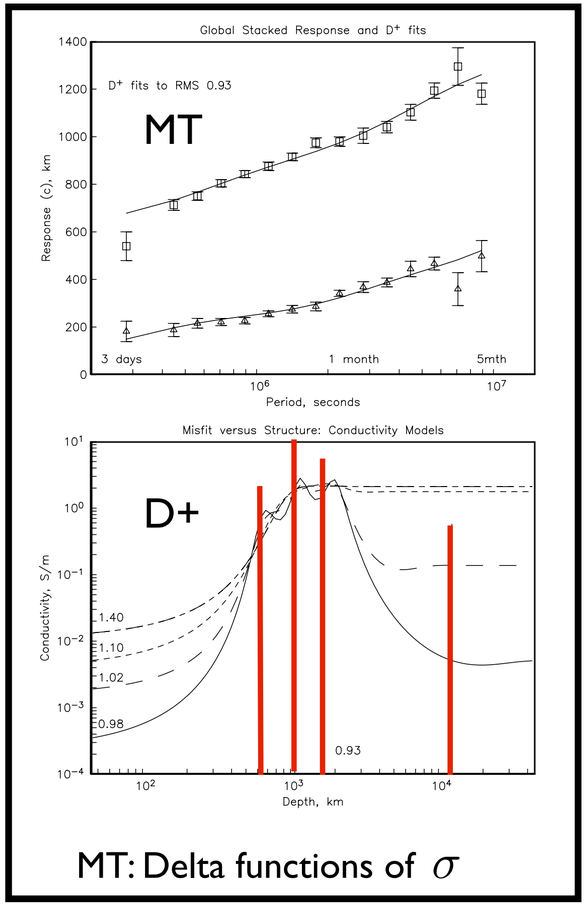
The initial z_1 allows us to put an insulator at $z = 0$, rather than a conducting sheet at the surface. While not exactly the same as the continued fractions described in the introduction, (58) can be rearranged by similar elementary algebra to be a *finite* partial fraction expansion:

$$c(\omega) = z_1 + \sum_{k=1}^{K} \frac{\alpha_k}{\lambda_k + \mathrm{i}\omega} . \tag{59}$$

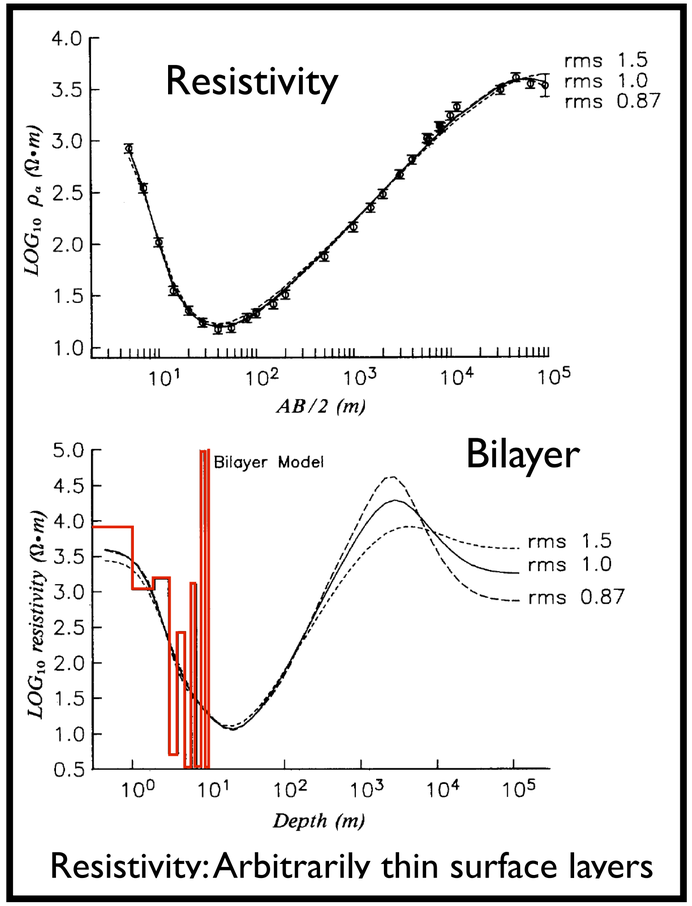
model space



These solutions are guaranteed best fitting but pathologically rough.



MT: Delta functions of σ

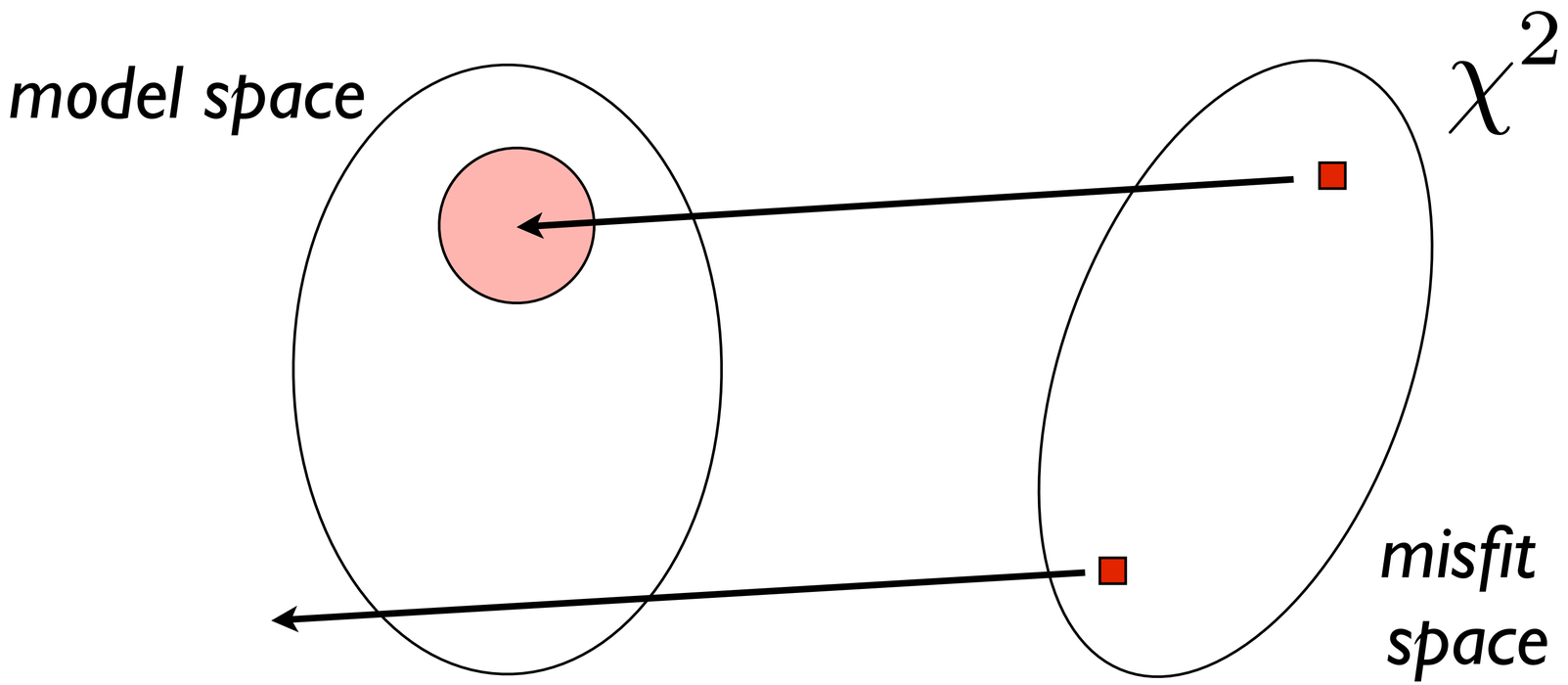


Resistivity: Arbitrarily thin surface layers

We don't know for sure, but least squares (LS) fits to higher dimensional models are probably also pathological.

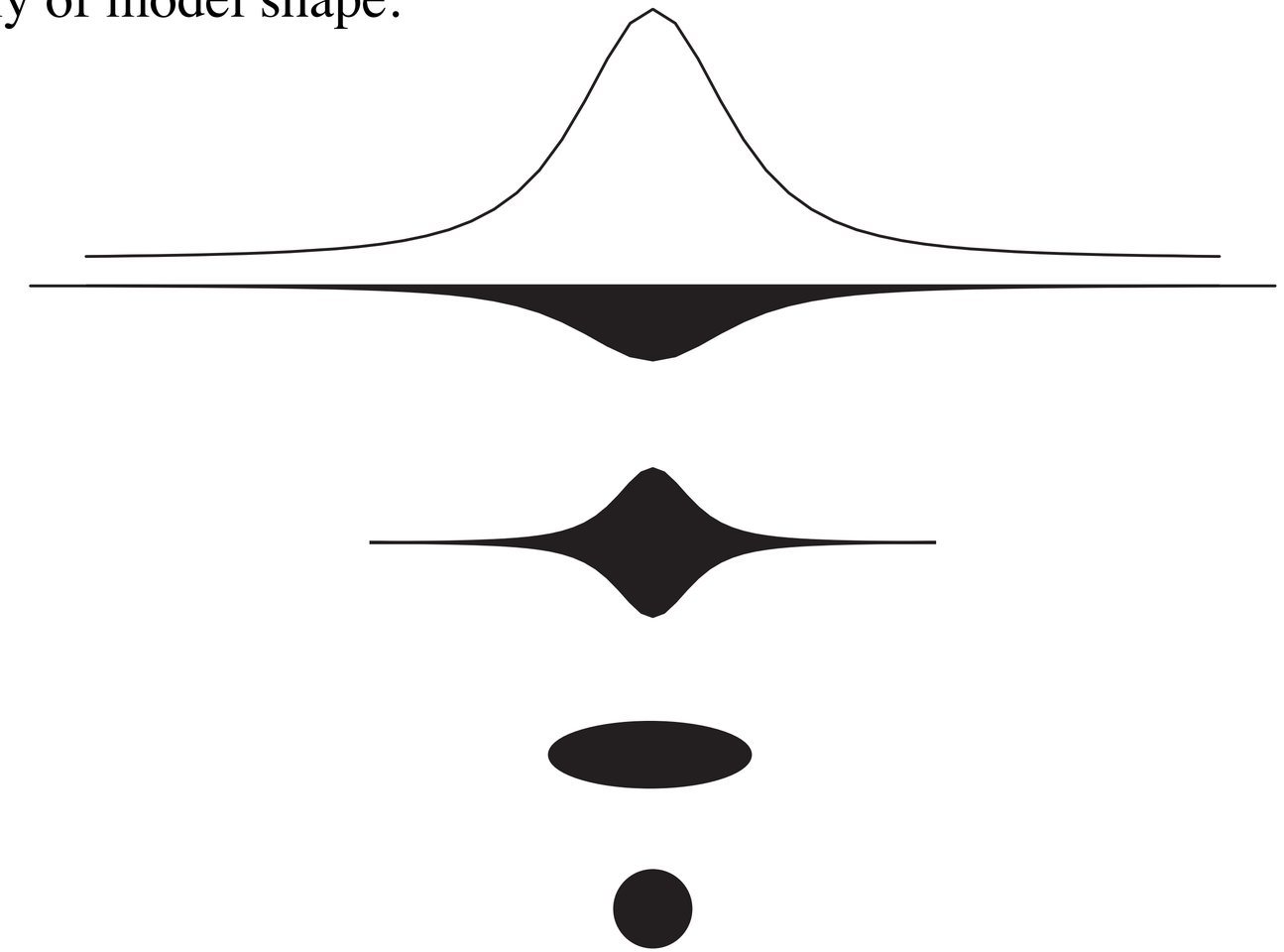
But we are pretty sure that true LS solutions are maximally "rough".

Geophysical inversion is non-unique:

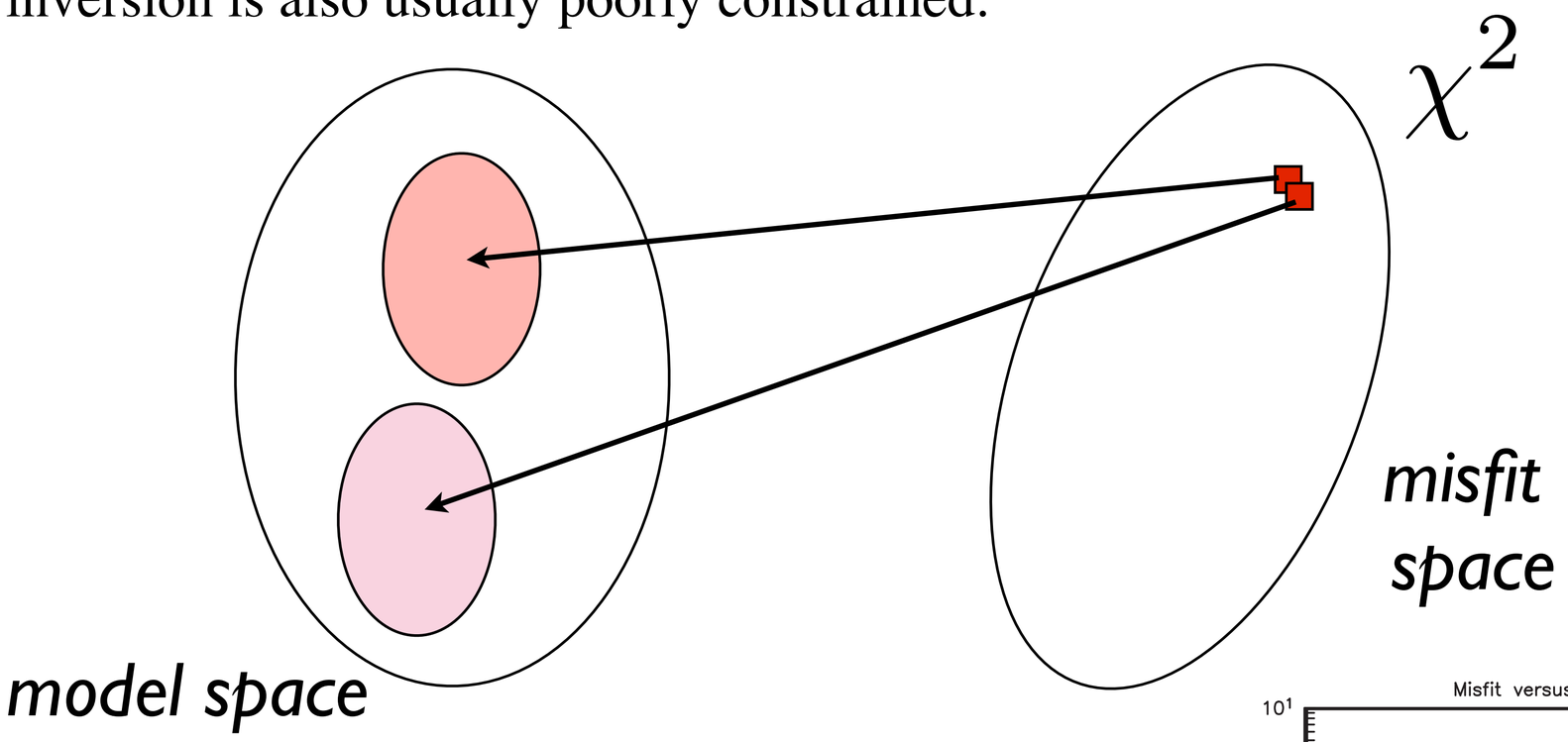


A single misfit will map into an infinite number of models (or none at all!).

What we are talking abou is **model construction**. For a great many geophysicists this is what they think of when inversion is mentioned. More rigorous approaches try to obtain bounds on model properties - something that is true of all models. The classic example is total mass from gravity: inversion of gravity data is terribly non-unique, but total excess mass and maximum depth can both be estimated independently of model shape.

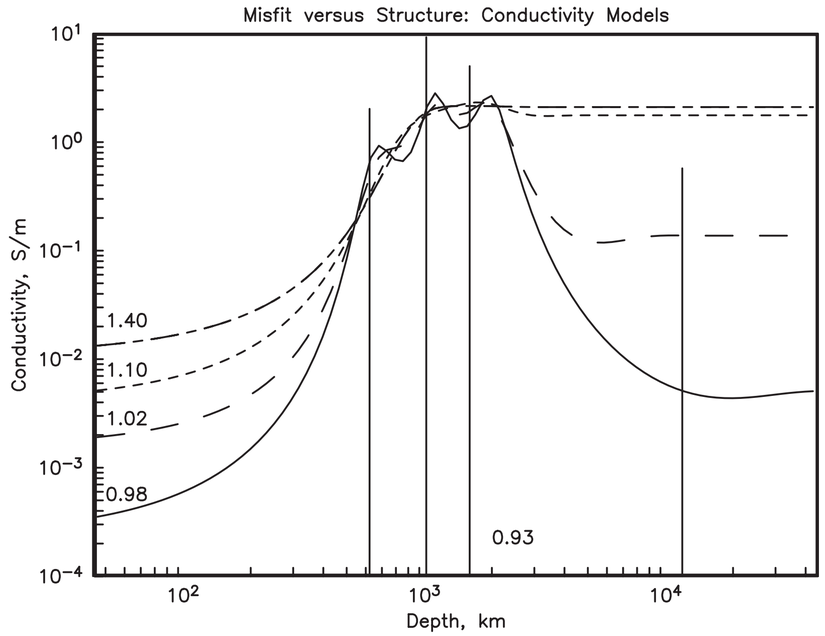


Geophysical inversion is also usually poorly constrained:



A small distance in χ^2 corresponds to a large distance in **m**

(And don't forget: the minimum χ^2 is likely outside your model parameterization).



We introduced the sum-squared misfit measure:

$$\chi^2 = \sum_{i=1}^{M} \frac{1}{\sigma_i^2} \left[d_i - f(x_i, \mathbf{m}) \right]^2$$

Our misfit measure can be written in matrix notation

$$\chi^2 = ||\mathbf{W}(\mathbf{d} - \hat{\mathbf{d}})||^2 = ||\mathbf{W}\mathbf{d} - \mathbf{W}f(\mathbf{m})||^2$$

$$\mathbf{W} = \mathrm{diag}(1/\sigma_1, 1/\sigma_2, ...1/\sigma_M) \quad .$$

Least squares minimizes this misfit with respect to all model parameters simultaneously. If the errors are zero-mean, independent, and normally distributed, χ^2 is Chi-squared distributed with *M-N* degrees of freedom, and least squares gives a maximum likelihood and unbiased estimate of **m**.

In practice the errors are rarely so well behaved, but least squares if fairly tolerant.

$$\mathrm{RMS} = \sqrt{\chi^2/M} \quad .$$

For a least squares solution we solve in the usual way by differentiating and setting to zero:

$$\nabla\chi^2 = -2(\mathbf{WJ})^T[\mathbf{W}(\mathbf{d} - f(\mathbf{m_0})) - \mathbf{WJ}\Delta\mathbf{m}] = 0$$

where we get a linear system

$$\beta = \alpha\Delta\mathbf{m}$$

$$\beta = (\mathbf{WJ})^T\mathbf{W}(\mathbf{d} - f(\mathbf{m_0}))$$
$$\alpha = (\mathbf{WJ})^T\mathbf{WJ} \quad .$$

So, given a starting model $\mathbf{m}_0$ we can find an update $\Delta\mathbf{m}$ and iterate until we converge. (This is Gauss-Newton.)

$$\Delta\mathbf{m} = \alpha^{-1}\beta$$

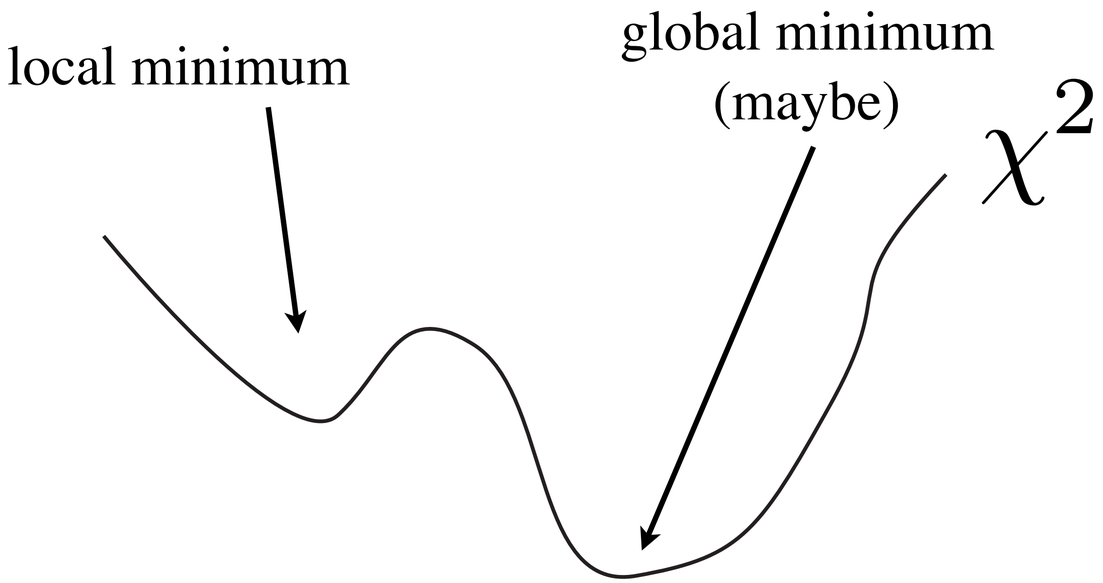
But this only works for small *N* (it isn't even defined for $N > M$). If *N* is big then the solutions become unstable, oscillatory, and generally useless (they are probably trying to converge to D+ type solutions).

Global versus local minima:

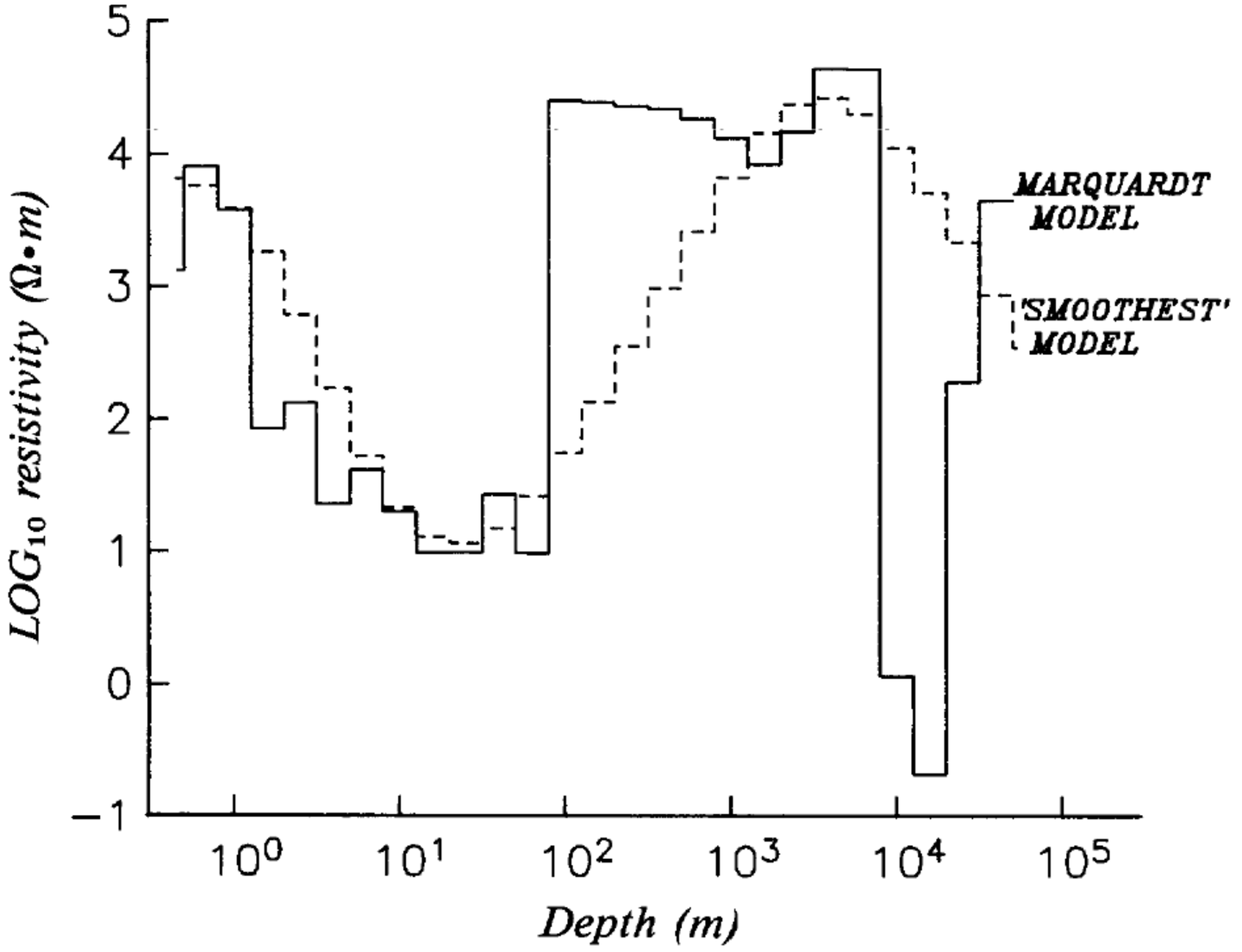
For nonlinear problems, there are no guarantees that Gauss-Newton will converge.

There are no guarantees that if it does converge the solution is a global one.

The solution might well depend on the starting model.



Here is an example of a 27-layer Marquardt model (thicknesses are fixed).



Constable, Parker, and Constable, 1987

Almost all inversion today incorporates some type of regularization, which minimizes some aspect of the model as well as fit to data:

$$U = \left(||\mathbf{Wd} - \mathbf{W}f(\mathbf{m})||^2 \right) + \mu ||\mathbf{Rm}||^2$$

where $\mathbf{Rm}$ is some measure of the model and μ is a trade-off parameter or Lagrange multiplier. In 1D a typical $\mathbf{R}$ might be:

$$\mathbf{R}_1 = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & -1 & 1 & 0 & \dots & 0 \\ & & \ddots & & & \ddots & \\ & & & & & & \\ & & & & & -1 & 1 \end{pmatrix}$$

m_1	-1						
m_2	+1	-1					
m_3		+1	-1				
m_4			+1	-1			
m_5				+1	-1		
m_6					+1	-1	
m_7						+1	-1
m_8							+1

which extracts a measure of slope. This stabilizes the inversion, creates a single solution, allows N > M, and manufactures models with useful properties.

This is easily extended to 2D and 3D modeling.

The trade-off between roughness and misfit:

$$U = \left(||\mathbf{W}\mathbf{d} - \mathbf{W}f(\mathbf{m})||^2\right) + \mu||\mathbf{R}\mathbf{m}||^2$$

When μ is small, model roughness is ignored and we try to fit the data. When μ is large, we smooth the model at the expense of data fit.

The trade-off between roughness and misfit:

$$U = \left(||\mathbf{W}\mathbf{d} - \mathbf{W}f(\mathbf{m})||^2\right) + \mu||\mathbf{R}\mathbf{m}||^2$$

When μ is small, model roughness is ignored and we try to fit the data. When μ is large, we smooth the model at the expense of data fit.

One approach is to choose μ and minimize U by least squares, but picking μ *a priori* is simply choosing how rough your model is.

The trade-off between roughness and misfit:

$$U = \left(||\mathbf{Wd} - \mathbf{W}f(\mathbf{m})||^2\right) + \mu||\mathbf{Rm}||^2$$

When μ is small, model roughness is ignored and we try to fit the data. When μ is large, we smooth the model at the expense of data fit.

One approach is to choose μ and minimize U by least squares, but picking μ *a priori* is simply choosing how rough your model is.

We ought to have a decent idea of how well our data can be fit. This forms the basis of the "Occam" approach, where a target data misfit χ^2_* is chosen:

$$U = \left(||\mathbf{Wd} - \mathbf{W}f(\mathbf{m})||^2 - \chi^2_*\right) + \mu||\mathbf{Rm}||^2$$

Note that if we achieve the target misfit, all we are doing is minimizing the model roughness

$$U = \left(||\mathbf{Wd} - \mathbf{W}f(\mathbf{m})||^2 - \chi_*^2\right) + \mu||\mathbf{Rm}||^2$$

For linearized, iterative inversion we follow the same approach as before:

$$\hat{\mathbf{d}} = f(\mathbf{m}_1) = f(\mathbf{m}_0 + \Delta\mathbf{m}) \approx f(\mathbf{m}_0) + \mathbf{J}\Delta\mathbf{m}$$

$$J_{ij} = \frac{\partial f(x_i, \mathbf{m}_0)}{\partial m_j}$$

and our unconstrained functional becomes

$$U = ||\mathbf{Rm}_1||^2 + \mu^{-1}\left(||\mathbf{Wd} - \mathbf{W}\left(f(\mathbf{m}_0) + \mathbf{J}(\mathbf{m}_1 - \mathbf{m}_0)\right)||^2 - \chi_*^2\right)$$

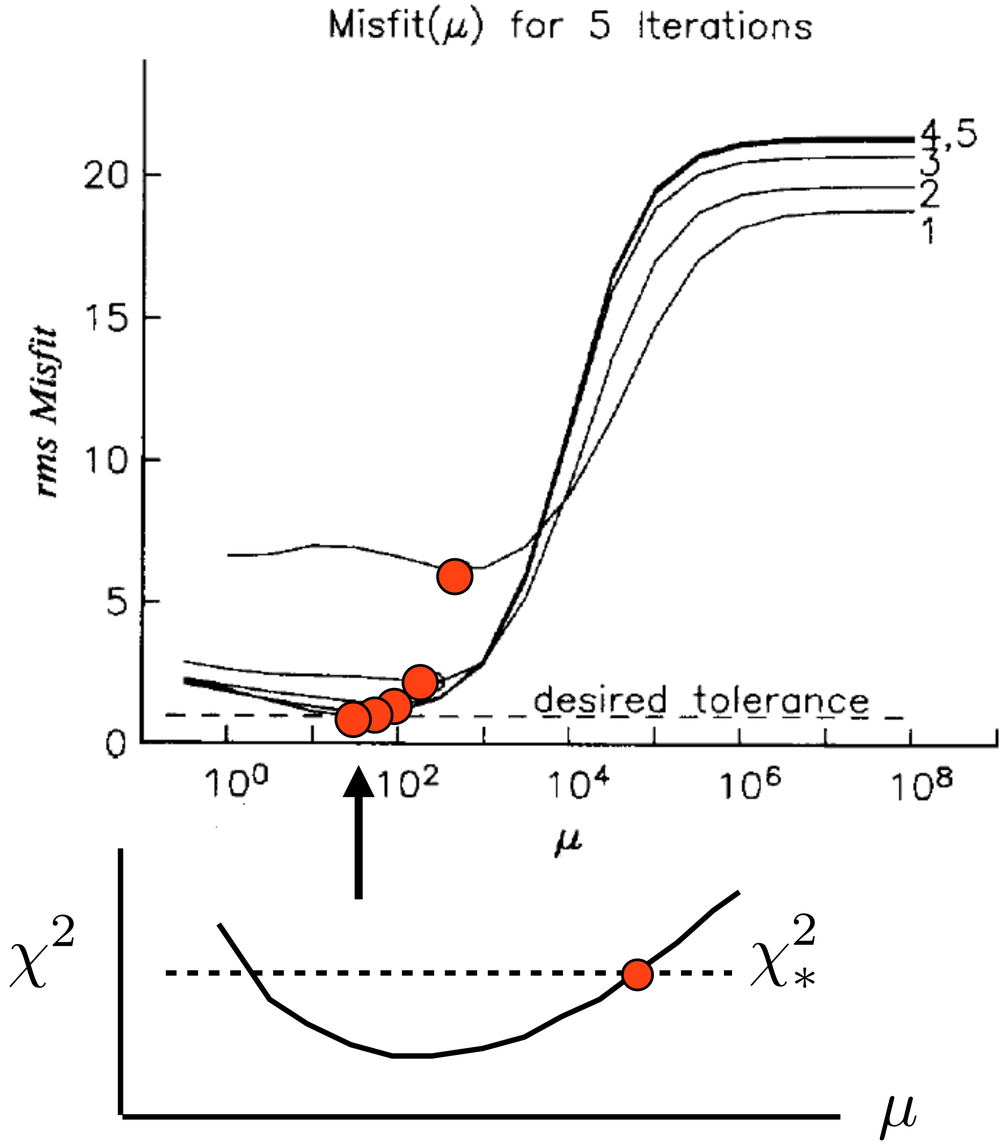
After differentiation and setting to zero we get an expression for a new model:

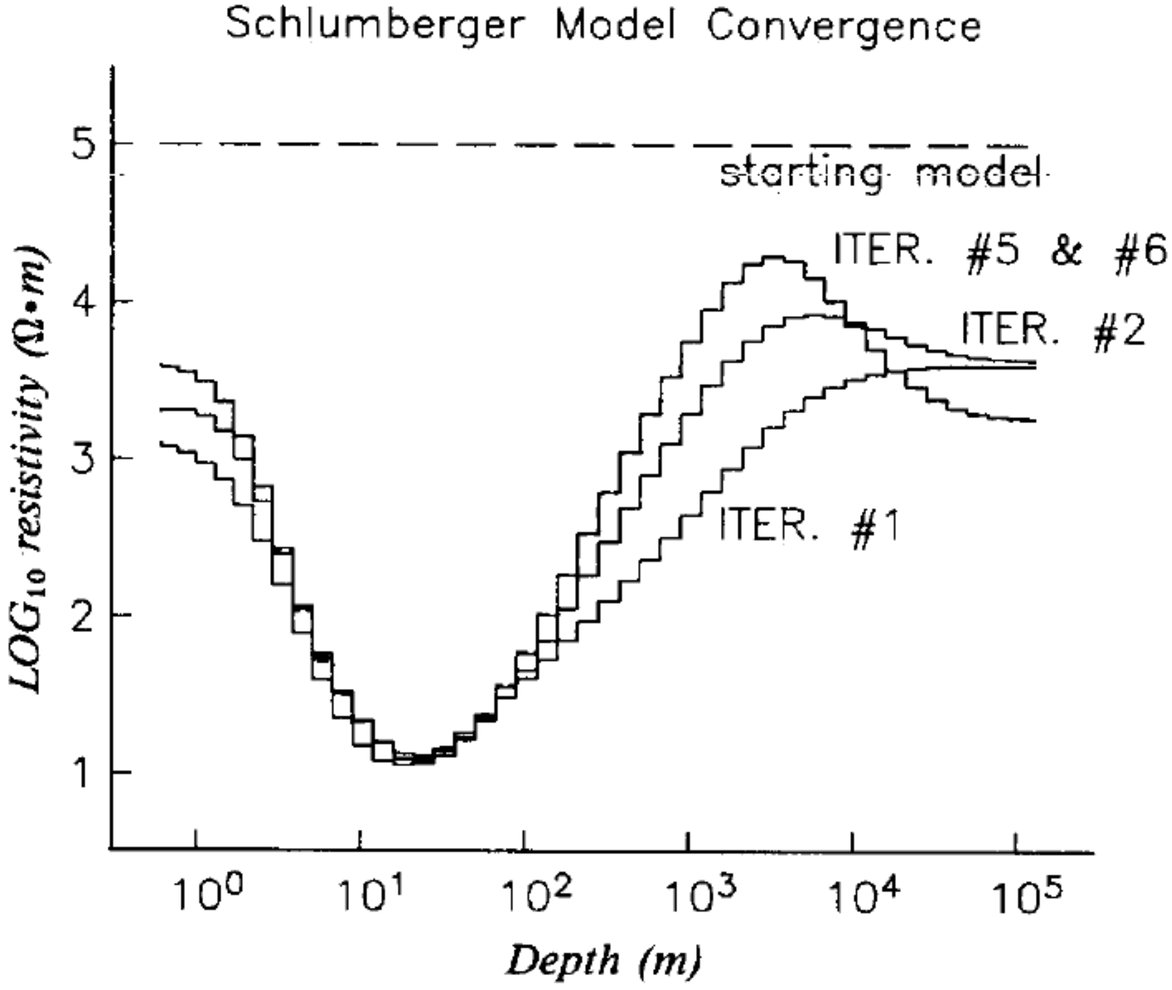
$$\mathbf{m}_1 = \left[\mu\mathbf{R}^T\mathbf{R} + (\mathbf{WJ})^T\mathbf{WJ}\right]^{-1}(\mathbf{WJ})^T\mathbf{W}(\mathbf{d} - f(\mathbf{m_0}) + \mathbf{J}\mathbf{m}_0) \quad .$$

Note that we are solving for the next model directly, and not a model update $\Delta\mathbf{m}$

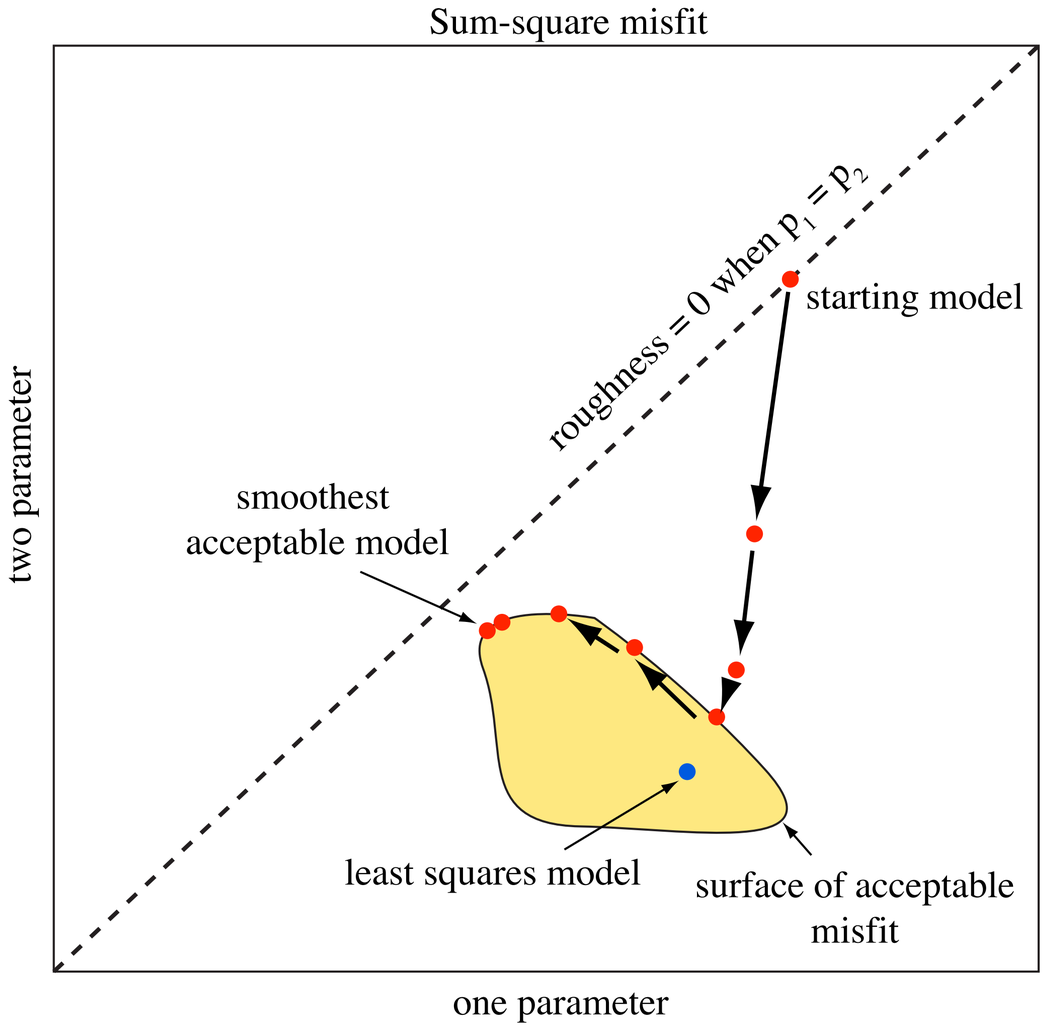
All we need is a way to find the correct μ (the one that makes $\chi^2 = \chi_*^2$)

Occam does this by carrying out a line search to find the ideal μ. Before χ^2_* is reached, we minimize χ^2. After χ^2_* is reached we choose the μ which gives us exactly χ^2_*.



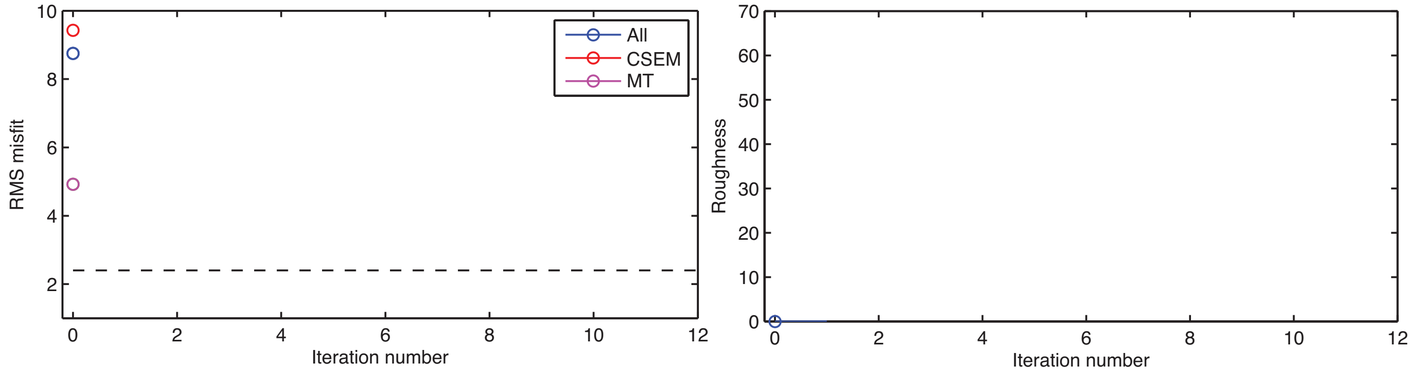


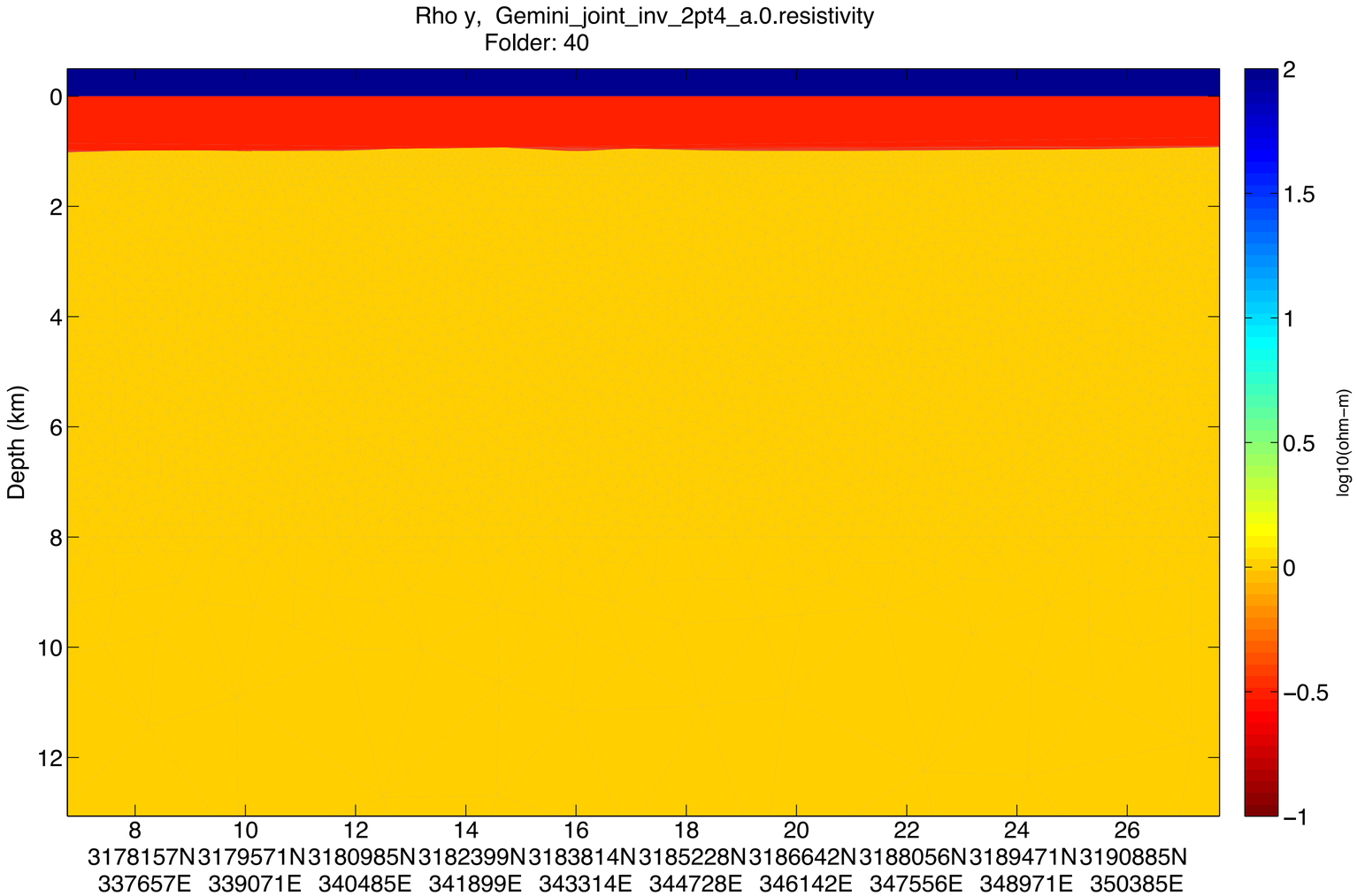
If the Occam algorithm does not get hung up in a local minimum, it will converge to the smoothest model for a given misfit.

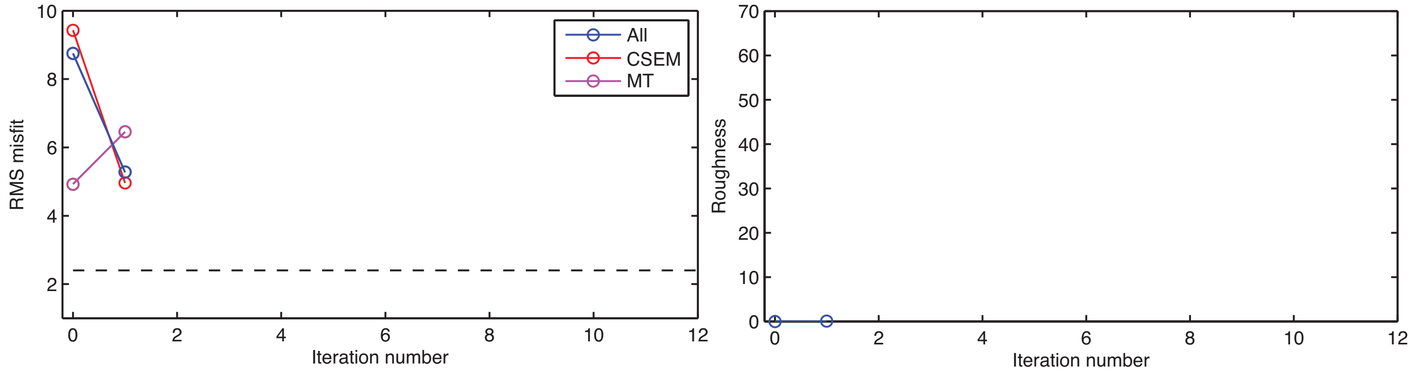


It is important to run the inversion to convergence, and not stop as soon as the target misfit is achieved.

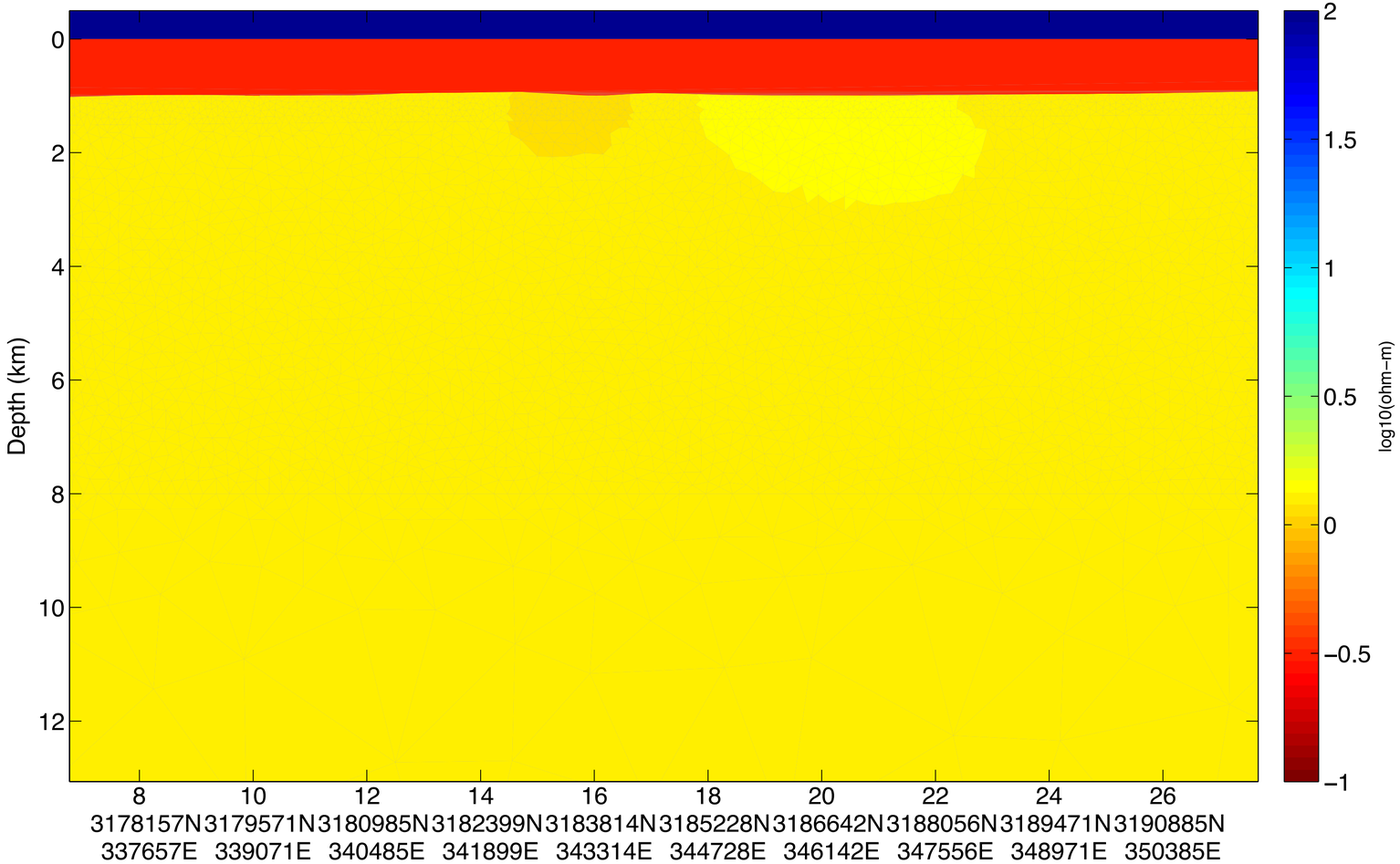
Occam demo

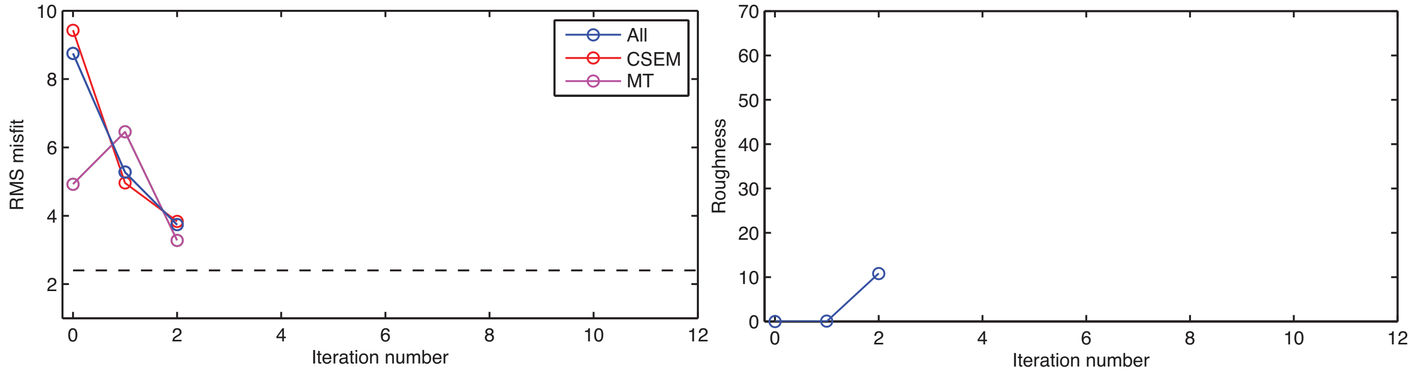

All
CSEM
MT
RMS misfit
Iteration number
Roughness
Iteration number


Rho y, Gemini_joint_inv_2pt4_a.0.resistivity
Folder: 40
Depth (km)
log10(ohm-m)
3178157N 3179571N 3180985N 3182399N 3183814N 3185228N 3186642N 3188056N 3189471N 3190885N
337657E 339071E 340485E 341899E 343314E 344728E 346142E 347556E 348971E 350385E

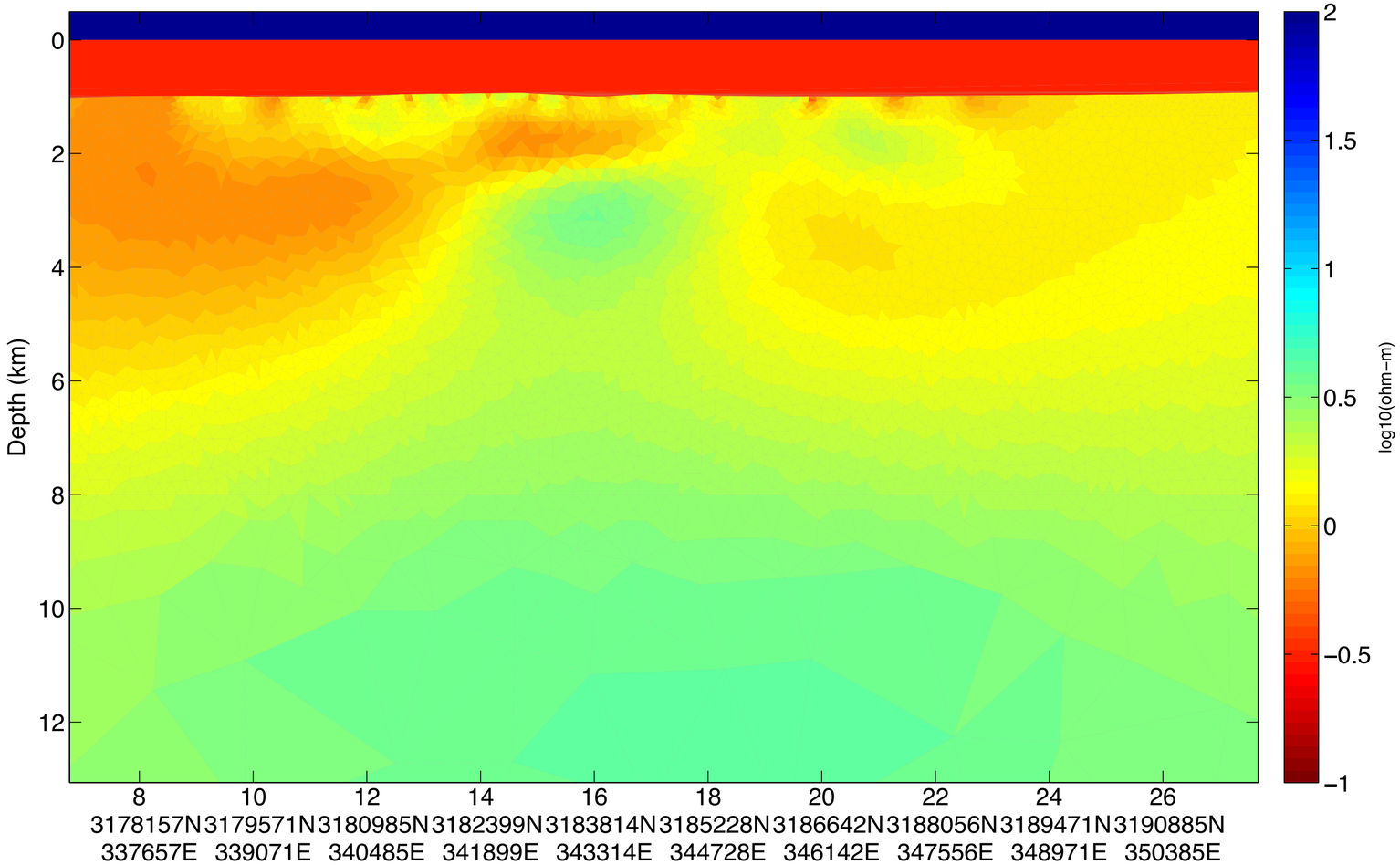


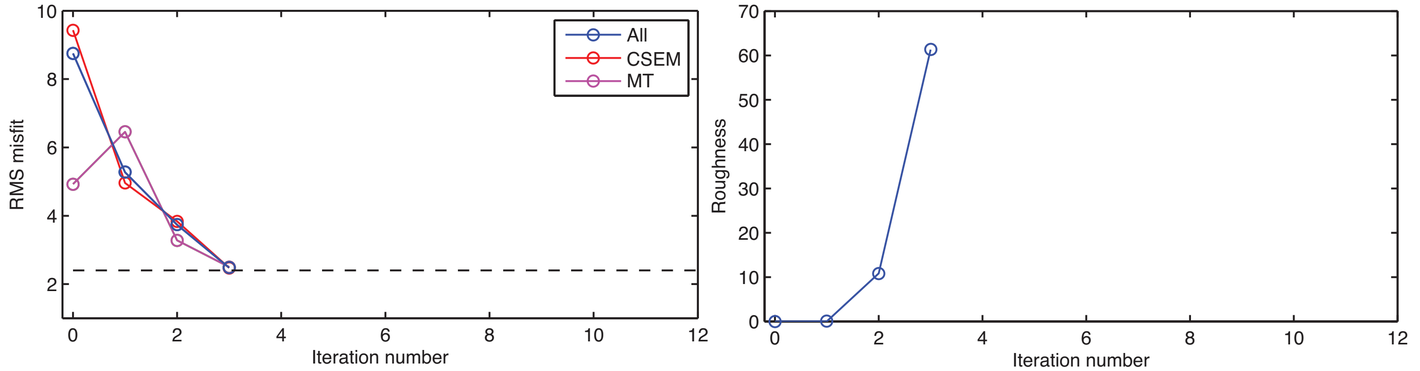
Rho y, RMS: 5.2768 Gemini_joint_inv_2pt4_a.1.resistivity
Folder: 40



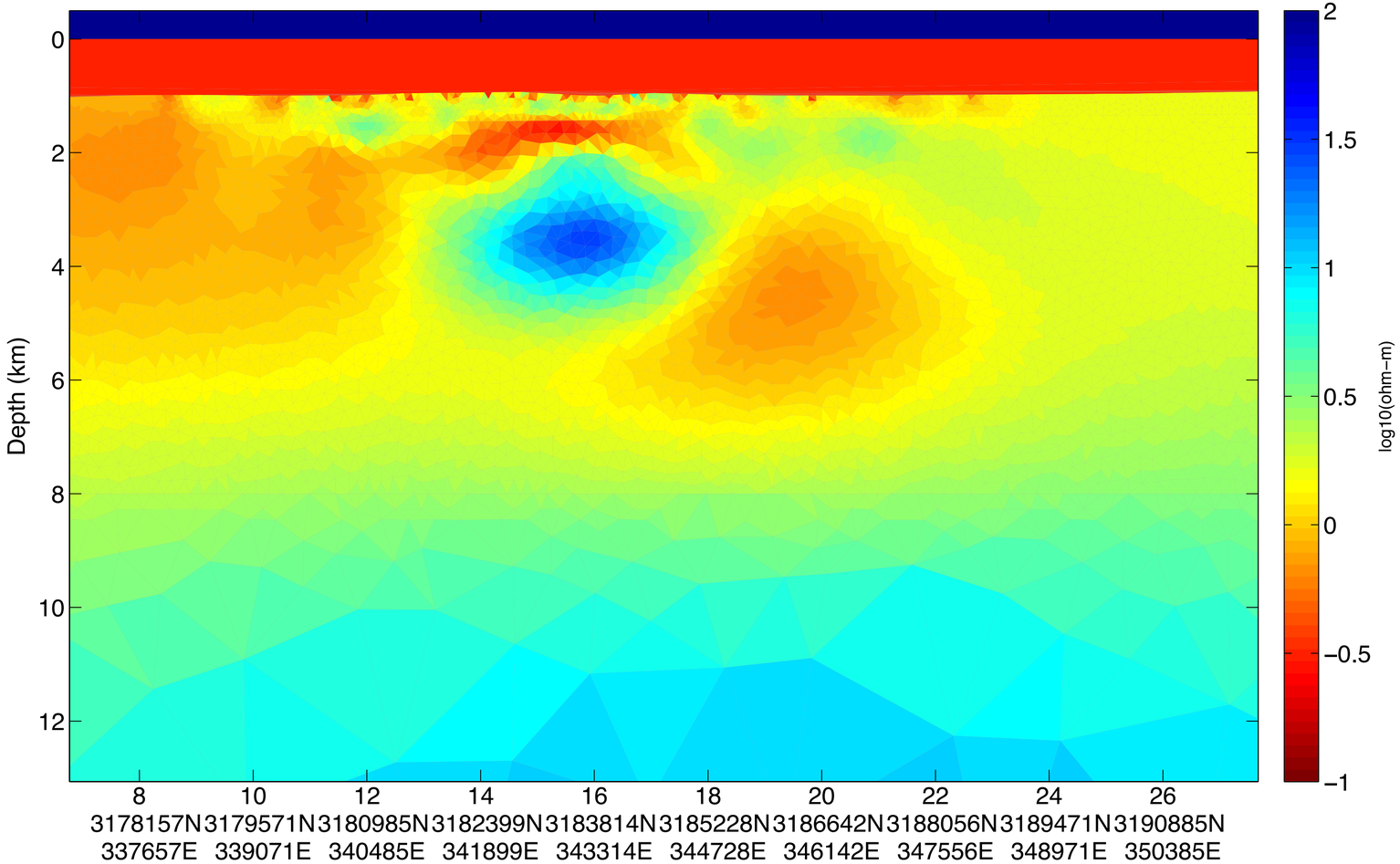


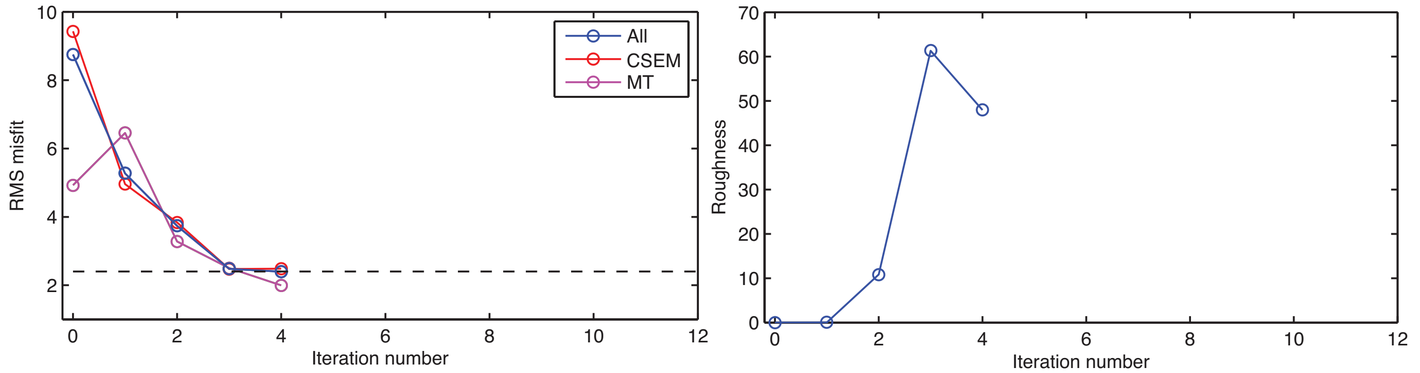
Rho y, RMS: 3.7362 Gemini_joint_inv_2pt4_a.2.resistivity
Folder: 40



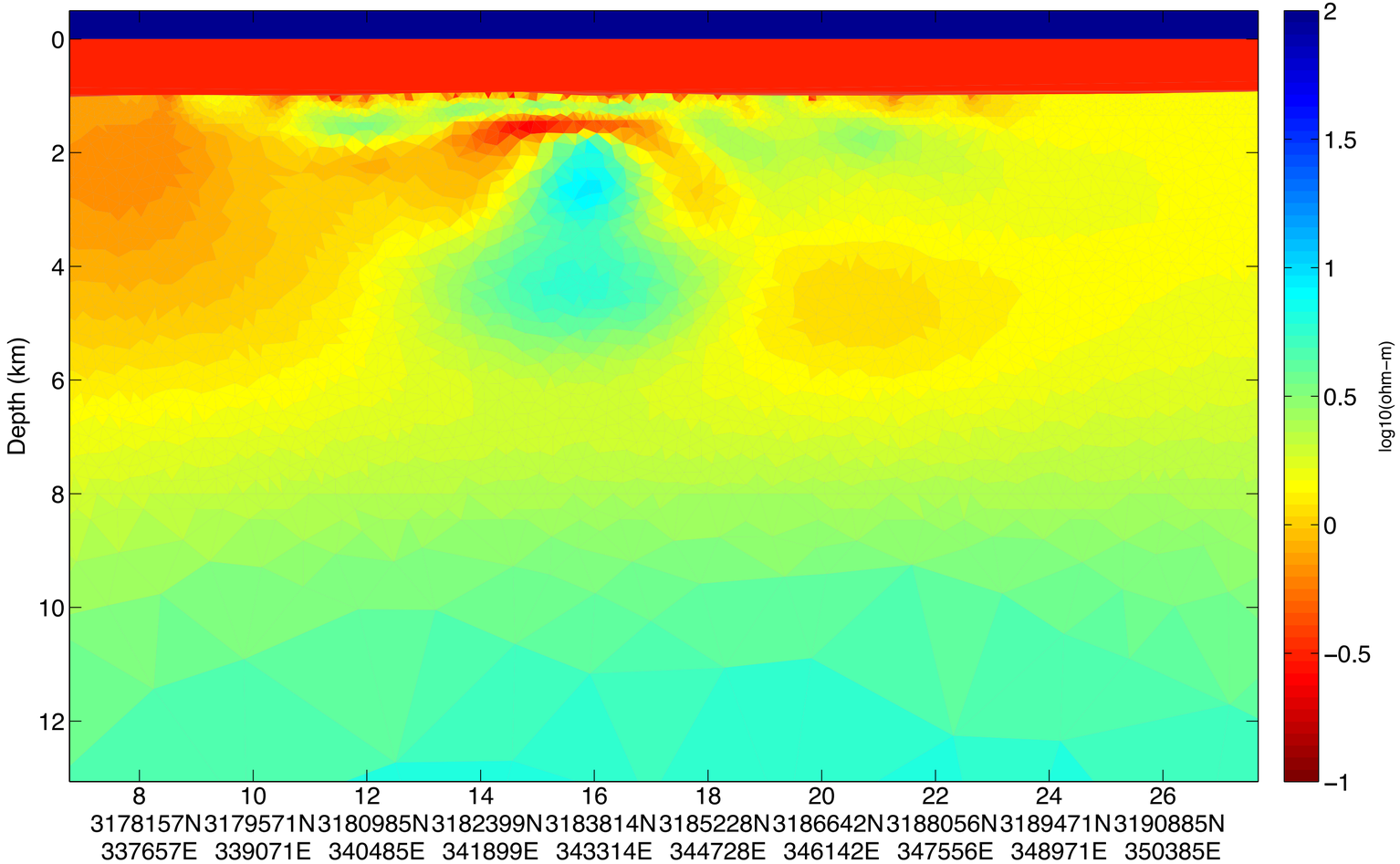


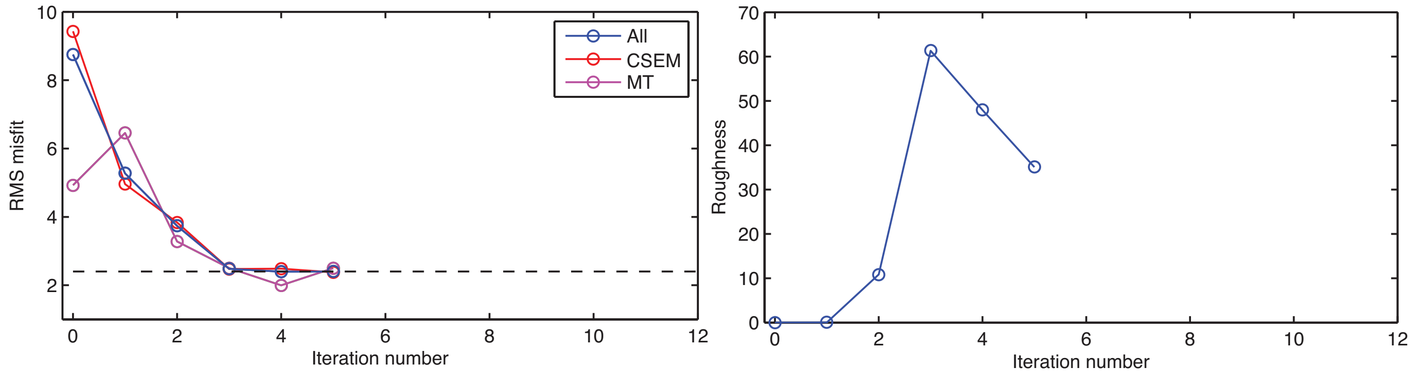
Rho y, RMS: 2.4814 Gemini_joint_inv_2pt4_a.3.resistivity
Folder: 40



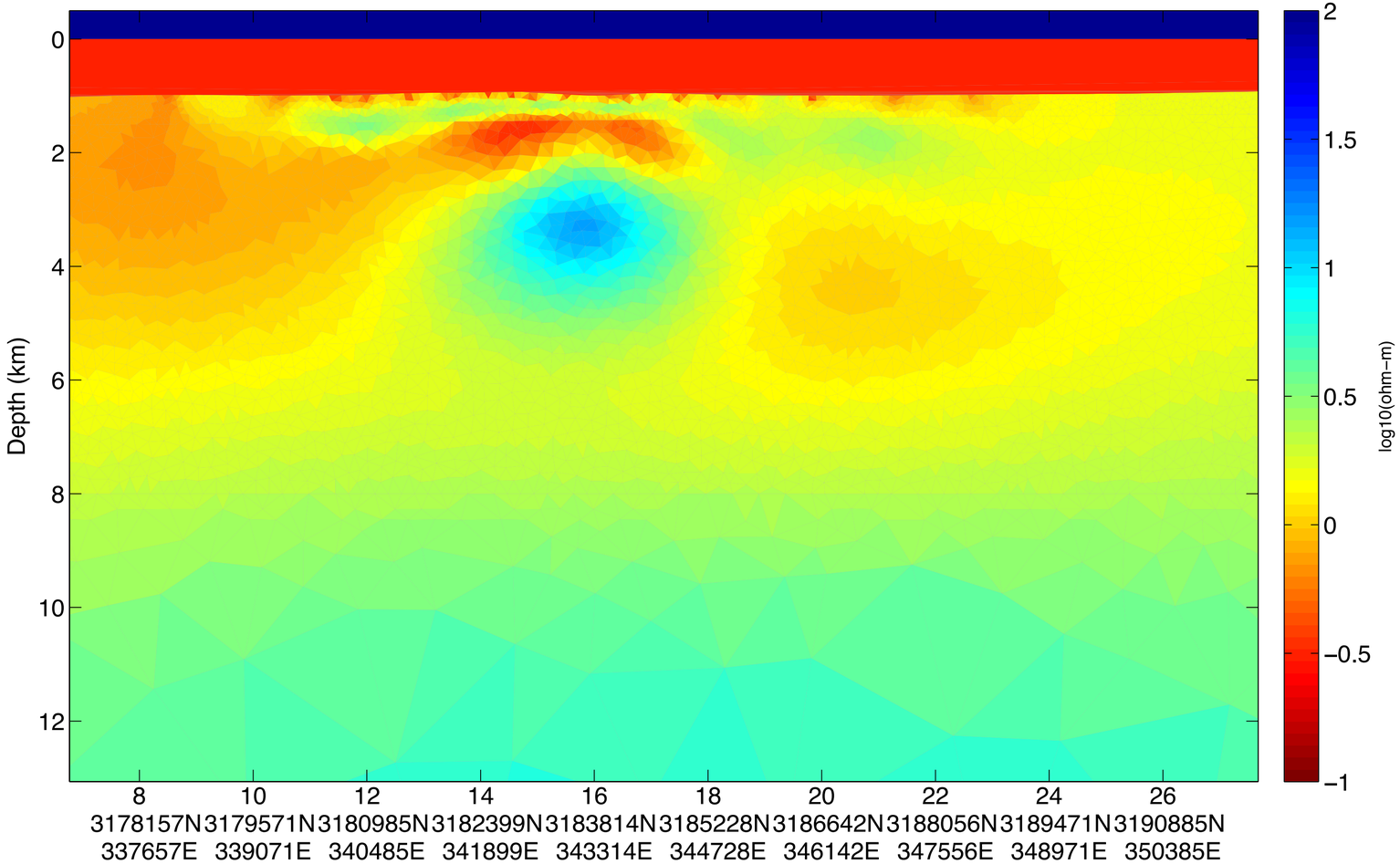


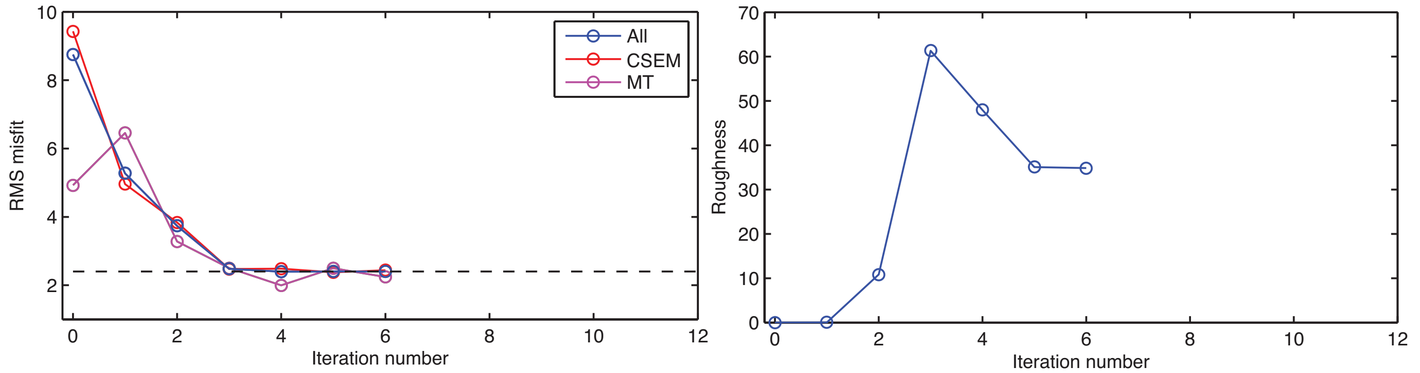
Rho y, RMS: 2.3978 Gemini_joint_inv_2pt4_a.4.resistivity
Folder: 40



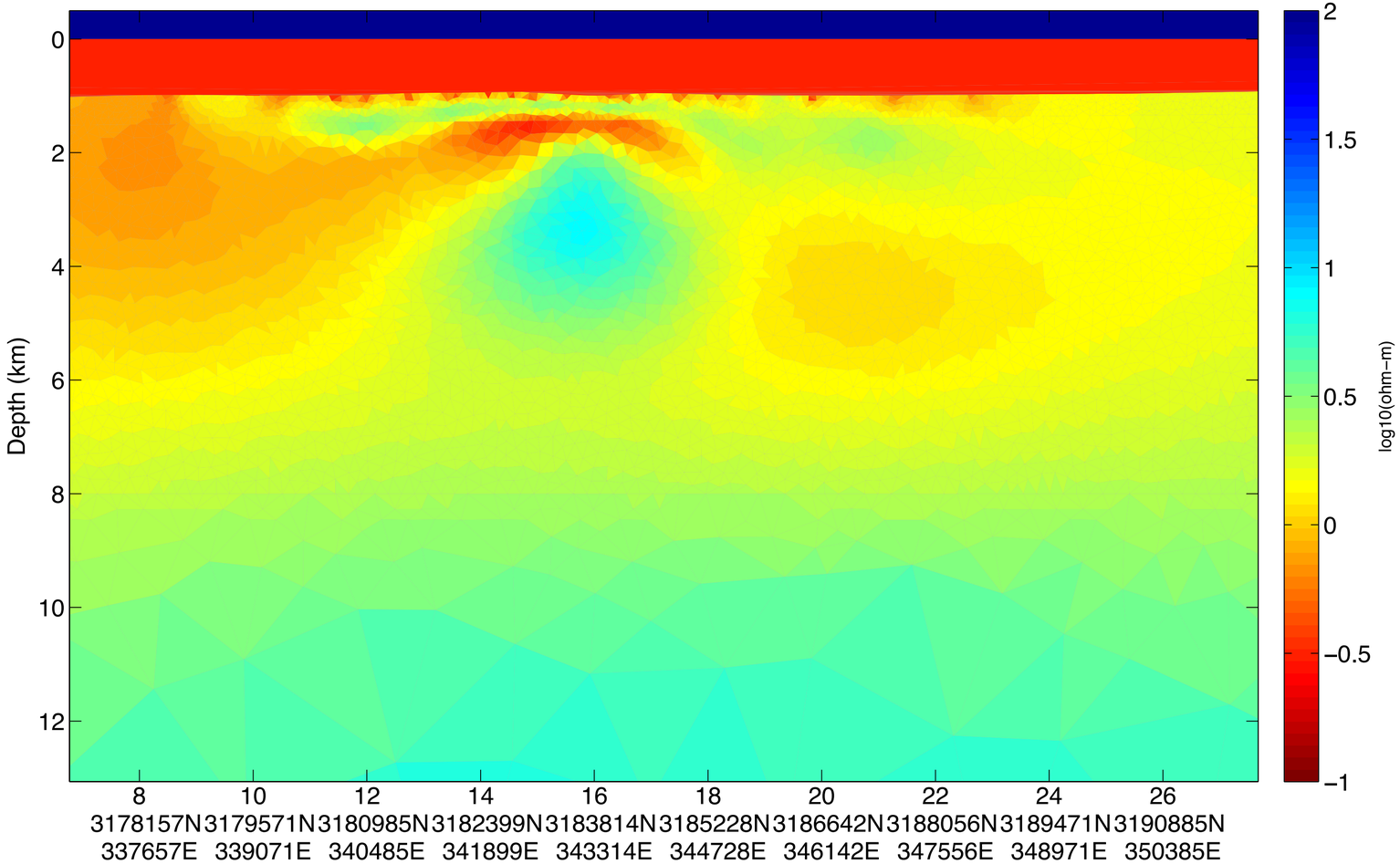


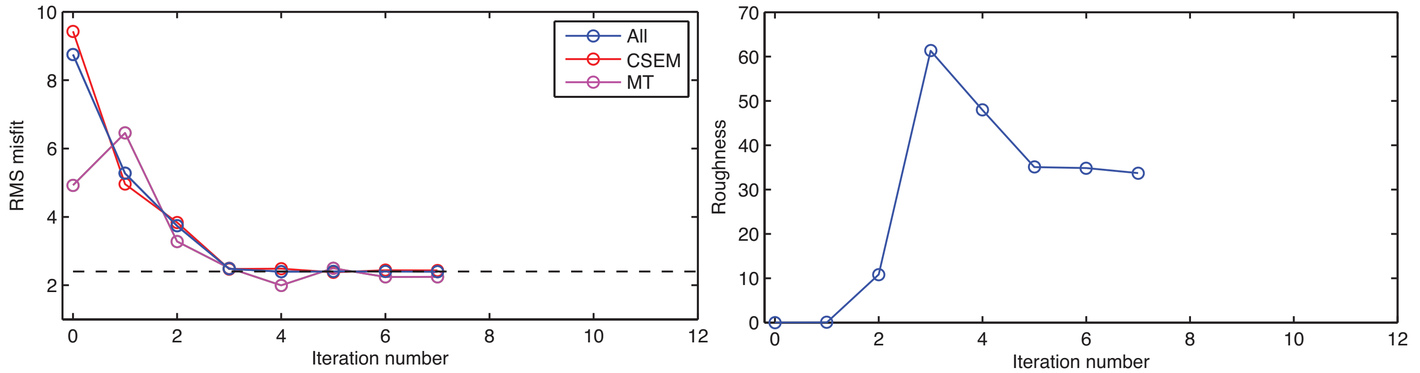
Rho y, RMS: 2.4027 Gemini_joint_inv_2pt4_a.5.resistivity
Folder: 40



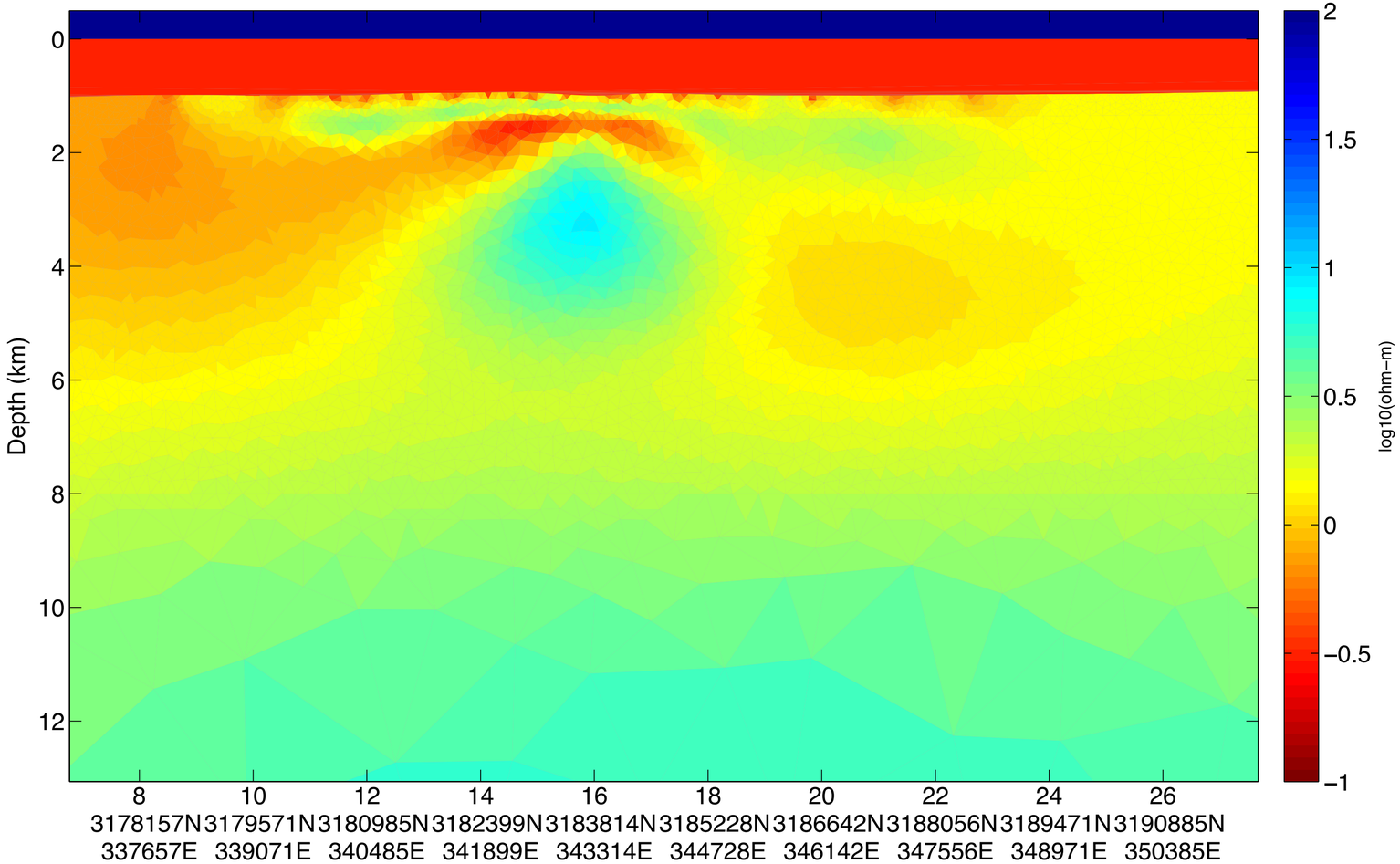


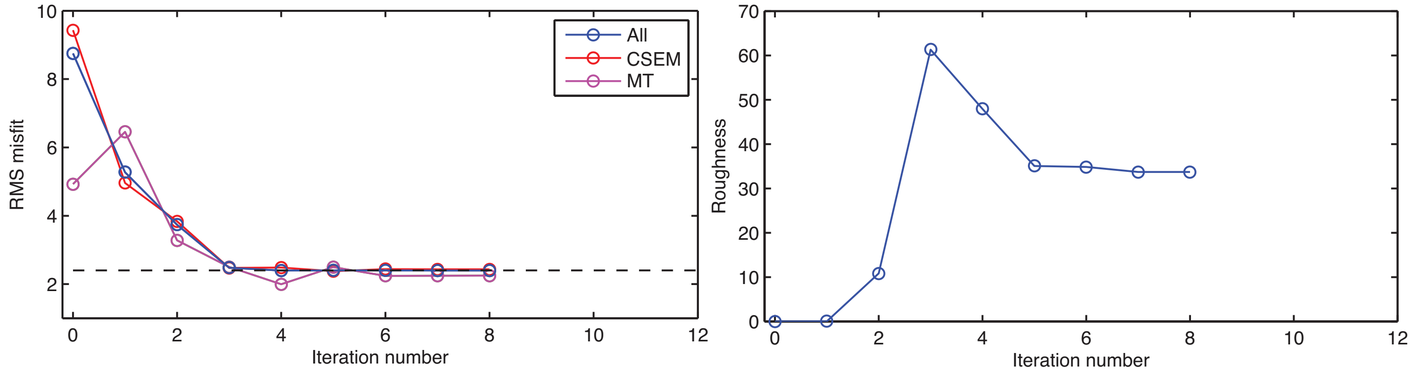
Rho y, RMS: 2.3993 Gemini_joint_inv_2pt4_a.6.resistivity
Folder: 40



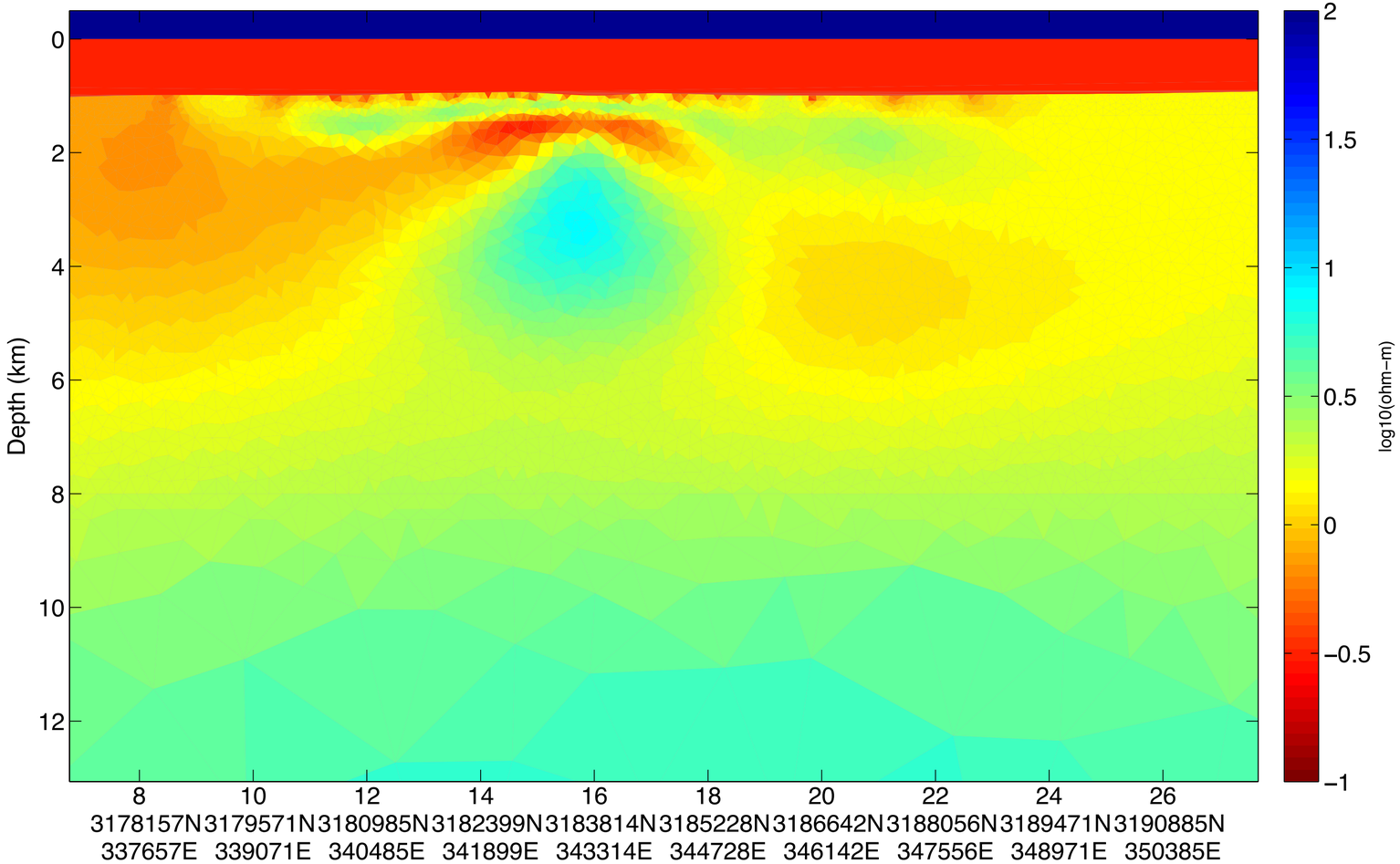


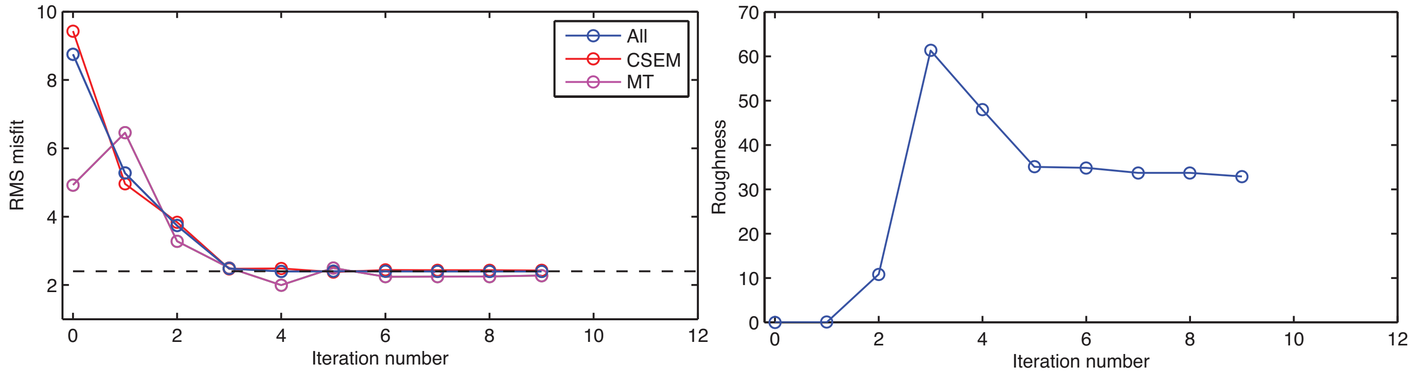
Rho y, RMS: 2.3982 Gemini_joint_inv_2pt4_a.7.resistivity
Folder: 40



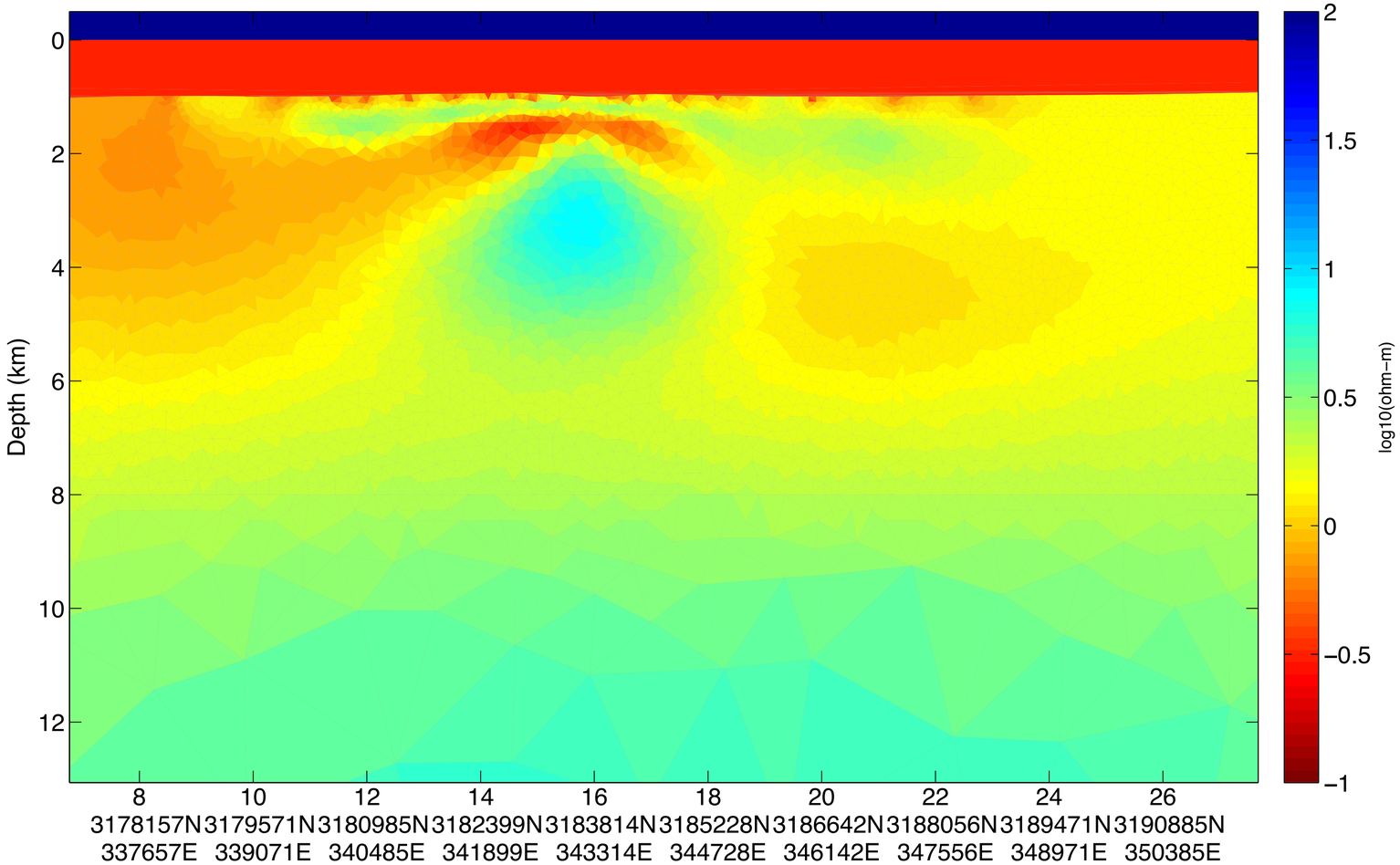


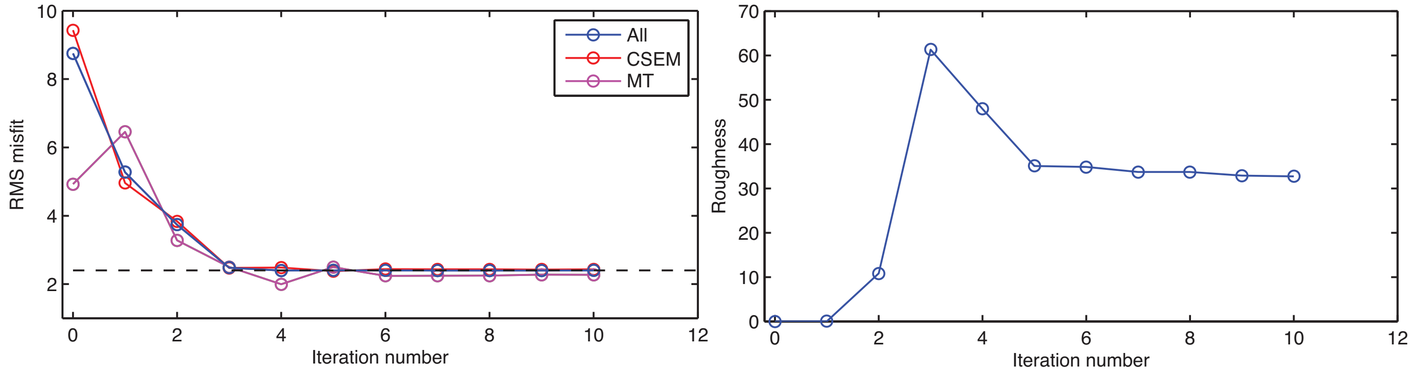
Rho y, RMS: 2.3974 Gemini_joint_inv_2pt4_a.8.resistivity
Folder: 40



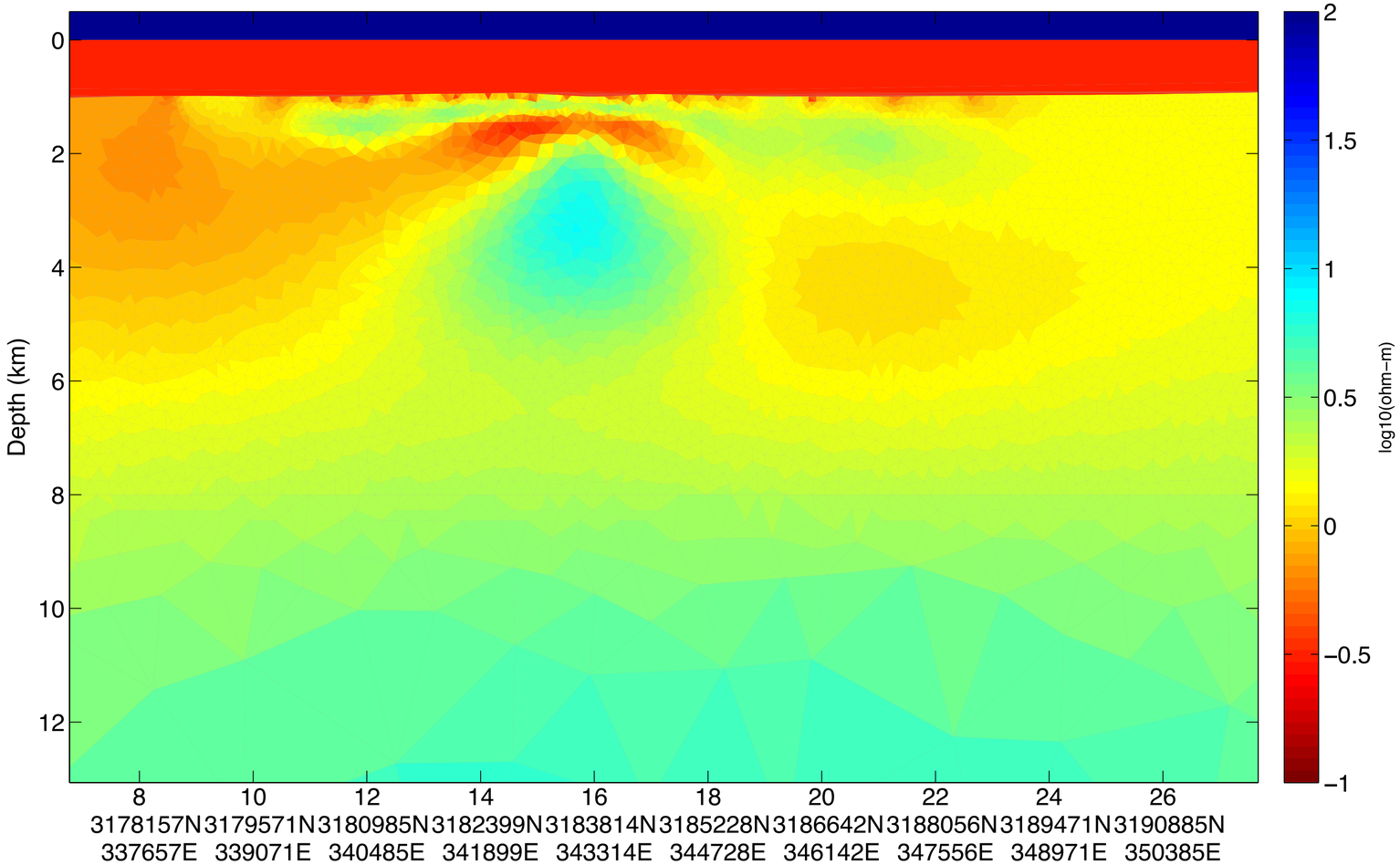


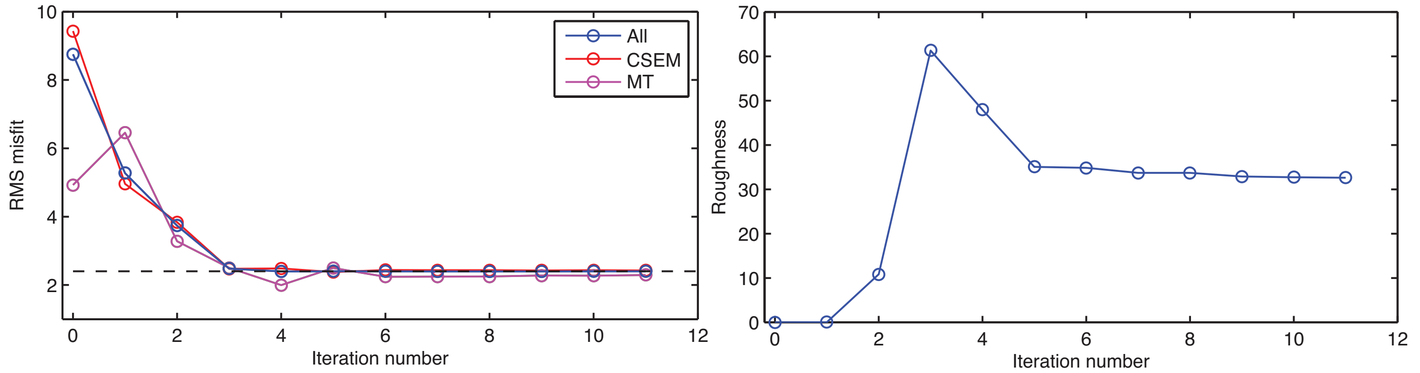
Rho y, RMS: 2.398 Gemini_joint_inv_2pt4_a.9.resistivity
Folder: 40



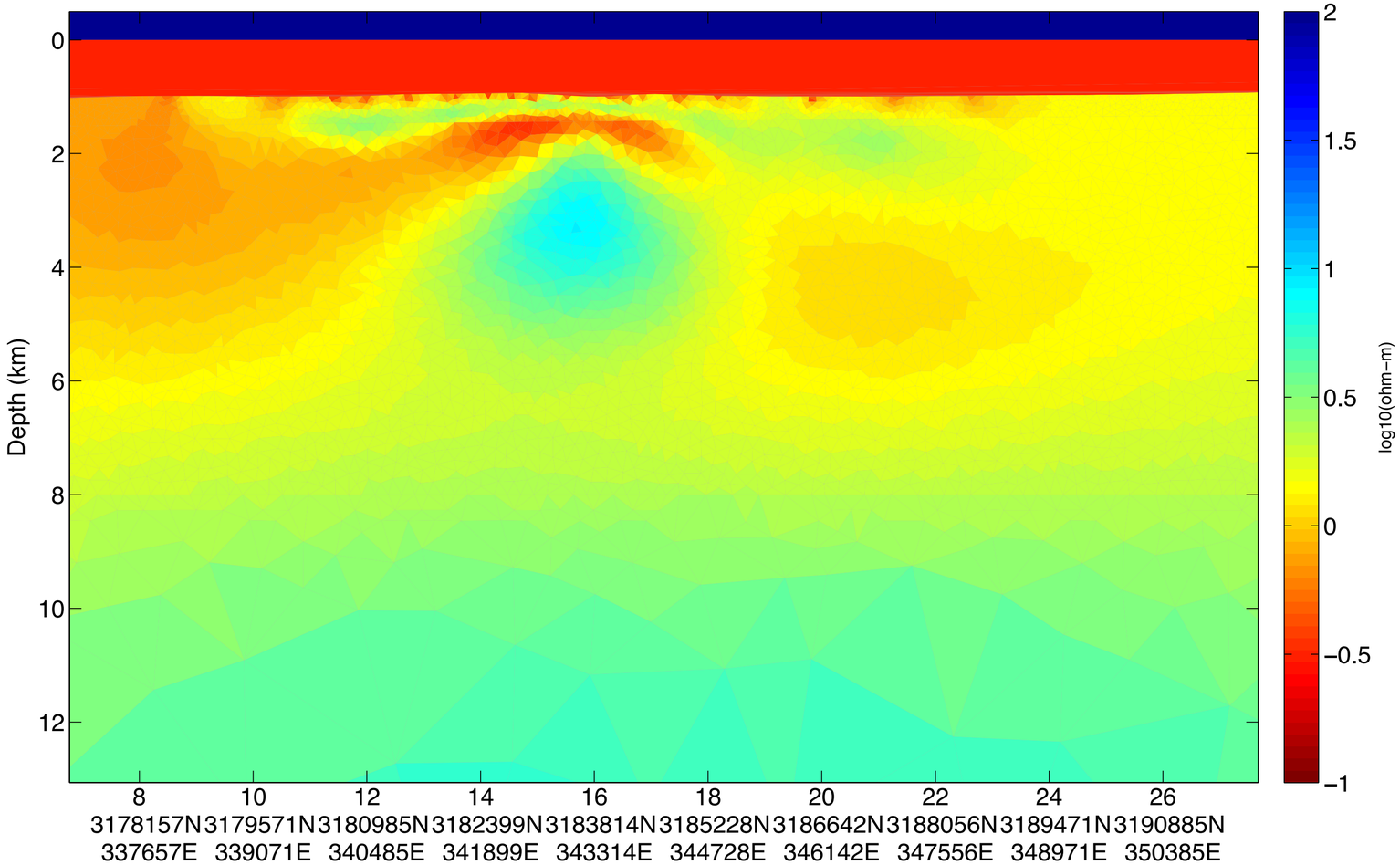


Rho y, RMS: 2.402 Gemini_joint_inv_2pt4_a.10.resistivity
Folder: 40

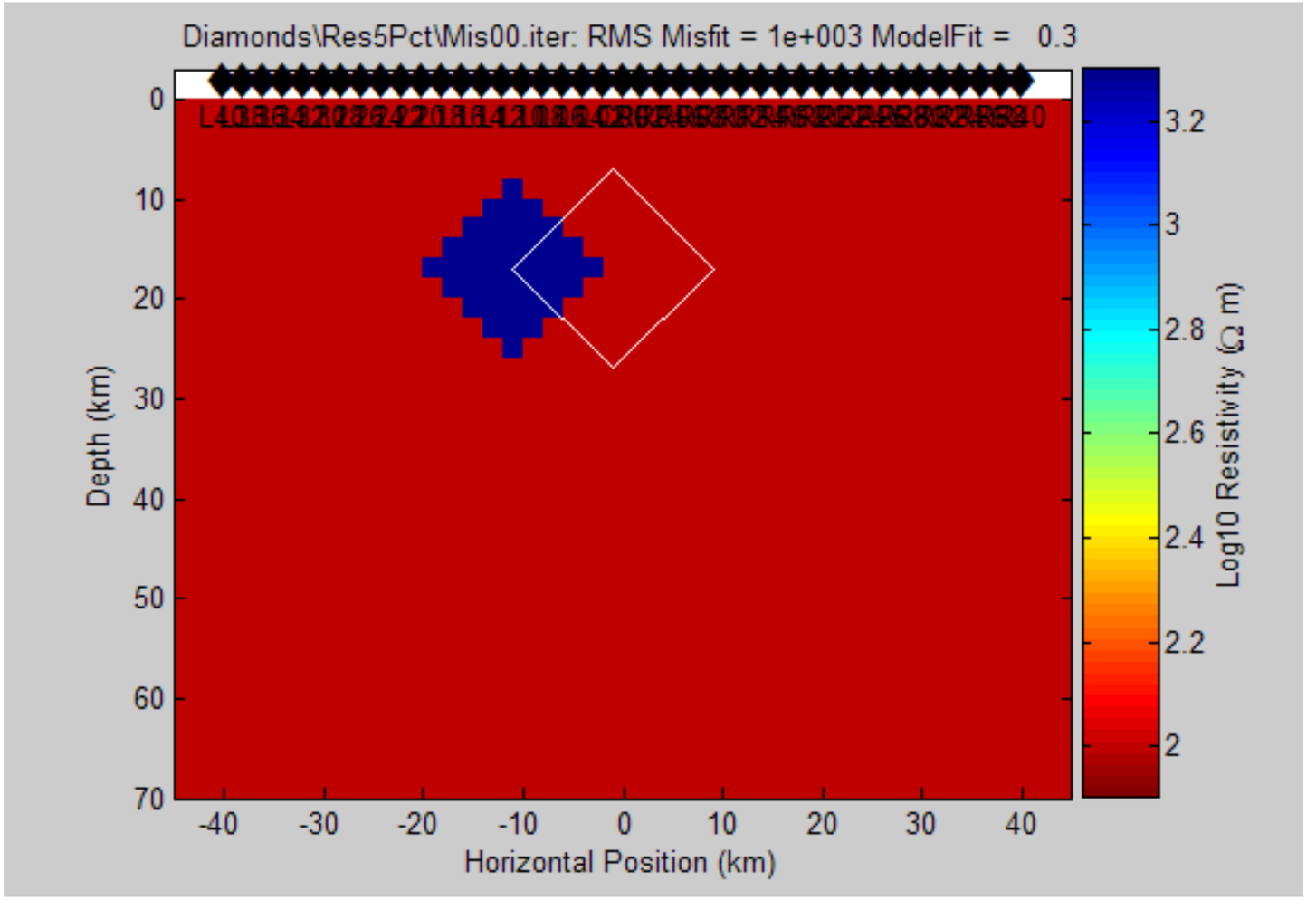




Rho y, RMS: 2.4008 Gemini_joint_inv_2pt4_a.11.resistivity
Folder: 40



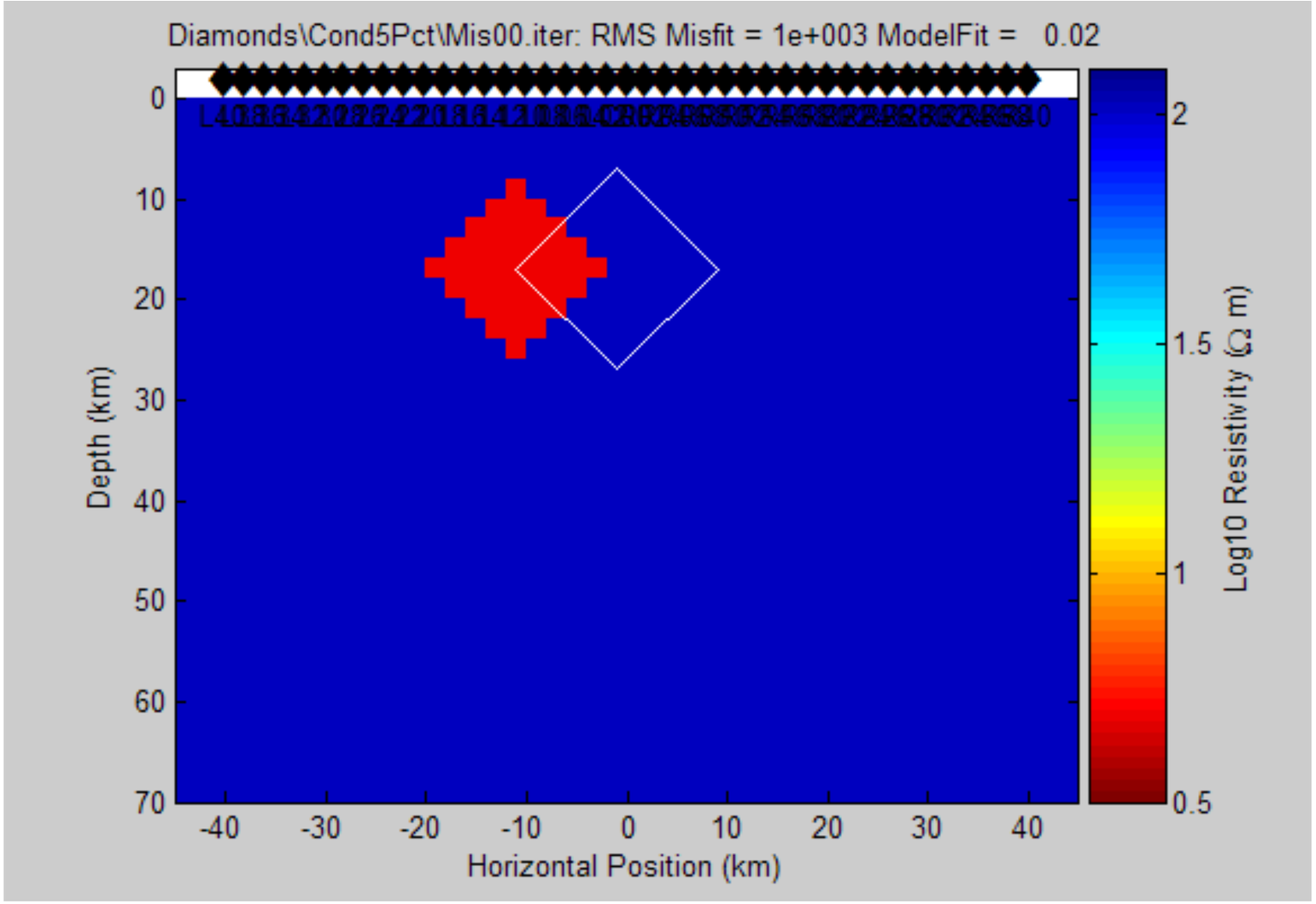
Because **J** depends on **m**, it is best start inversions from a half-space



MT: misaligned starting resistor - no harm done

Courtesy David Myer.

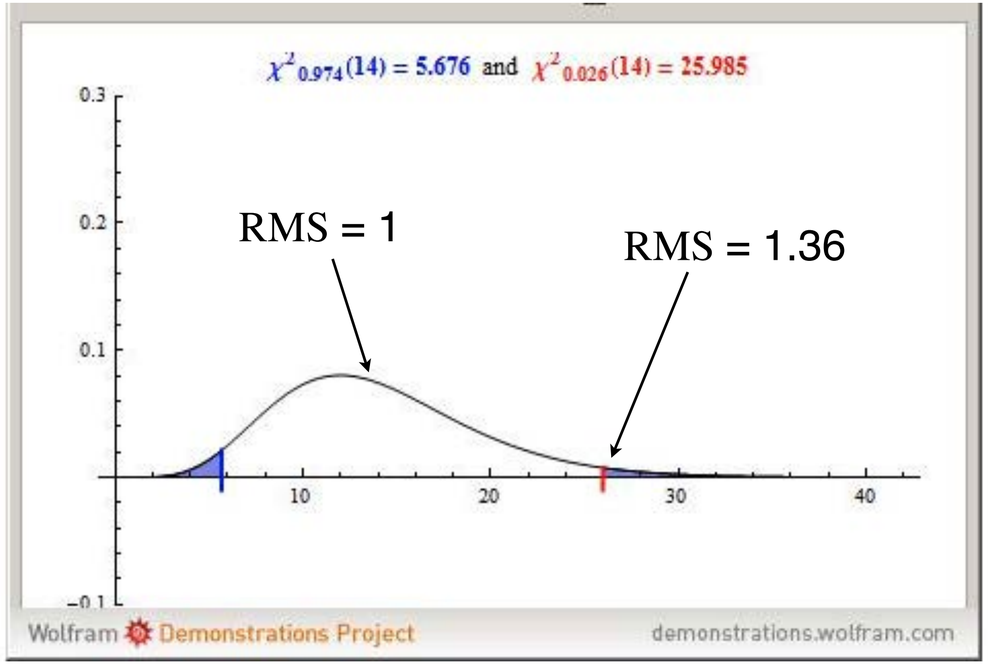
Misaligned starting conductor - forever trapped by **J**



Courtesy David Myer.

So what constitutes an adequate misfit?

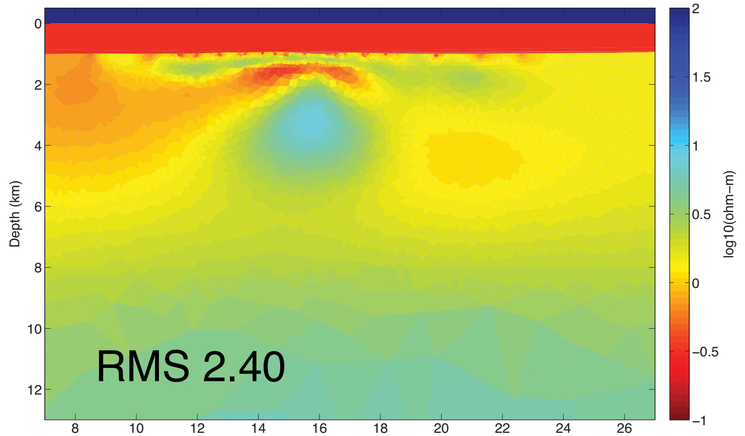
For zero-mean, Gaussian, independent errors, χ^2 is chi-squared distributed with M degrees of freedom. The expectation value is just M, which corresponds to RMS=1, and so this could be a reasonable target misfit. Or, one could look up the 95% (or other) confidence interval for chi-squared M.

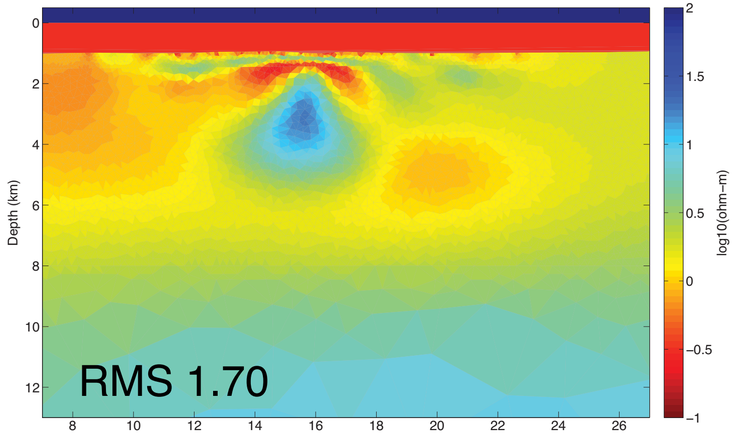


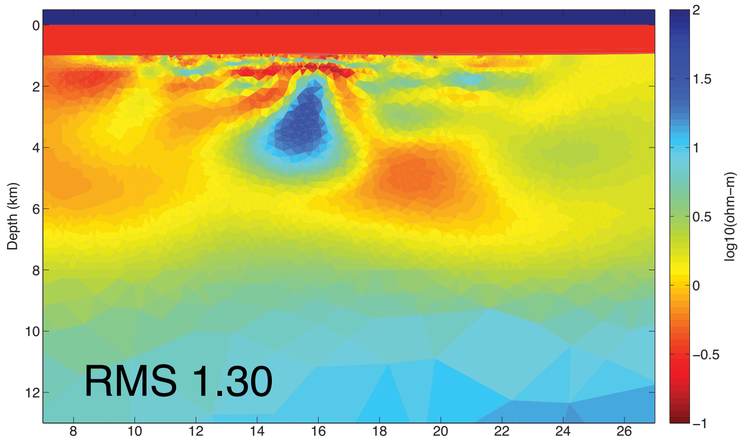
χ^2 for 14 data. For large data sets, RMS=1 and $RMS_{95\%}$ are very much the same.

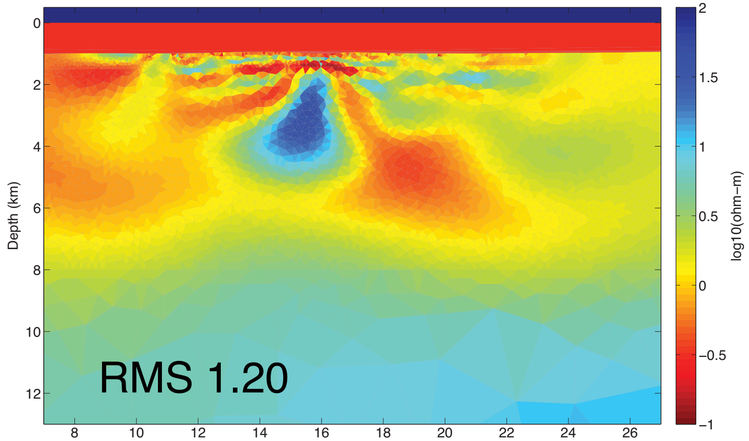
We could use other measures of fit, but the quadratic measure works with the mathematics of minimization, and for Gaussian errors has nice statistical properties (unbiased, maximum likelihood, minimum variance).

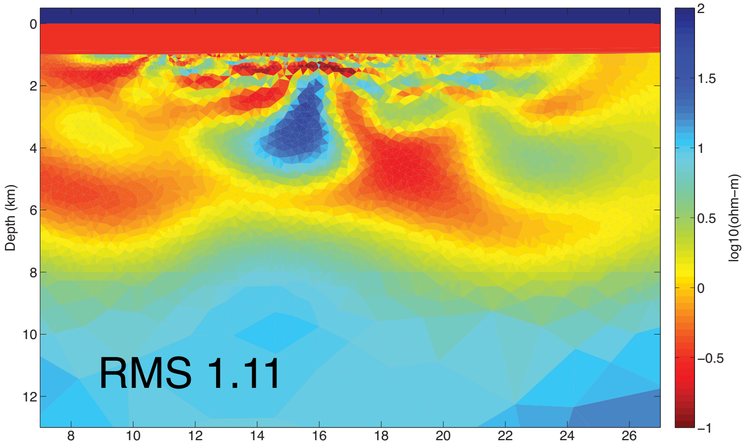
Even with well-estimated errors, choice of misfit can still be somewhat subjective.

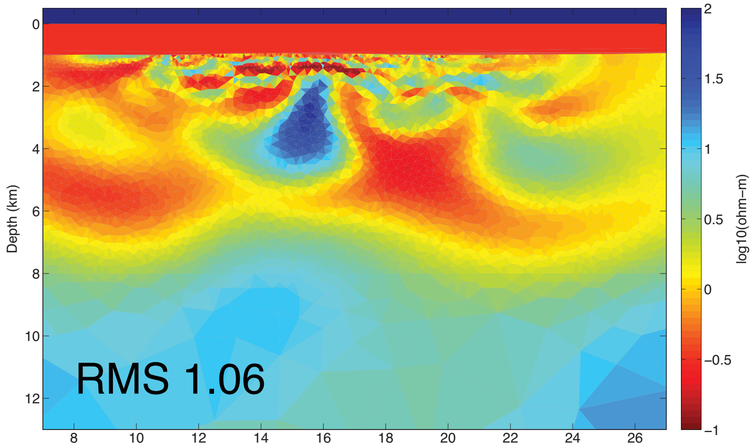






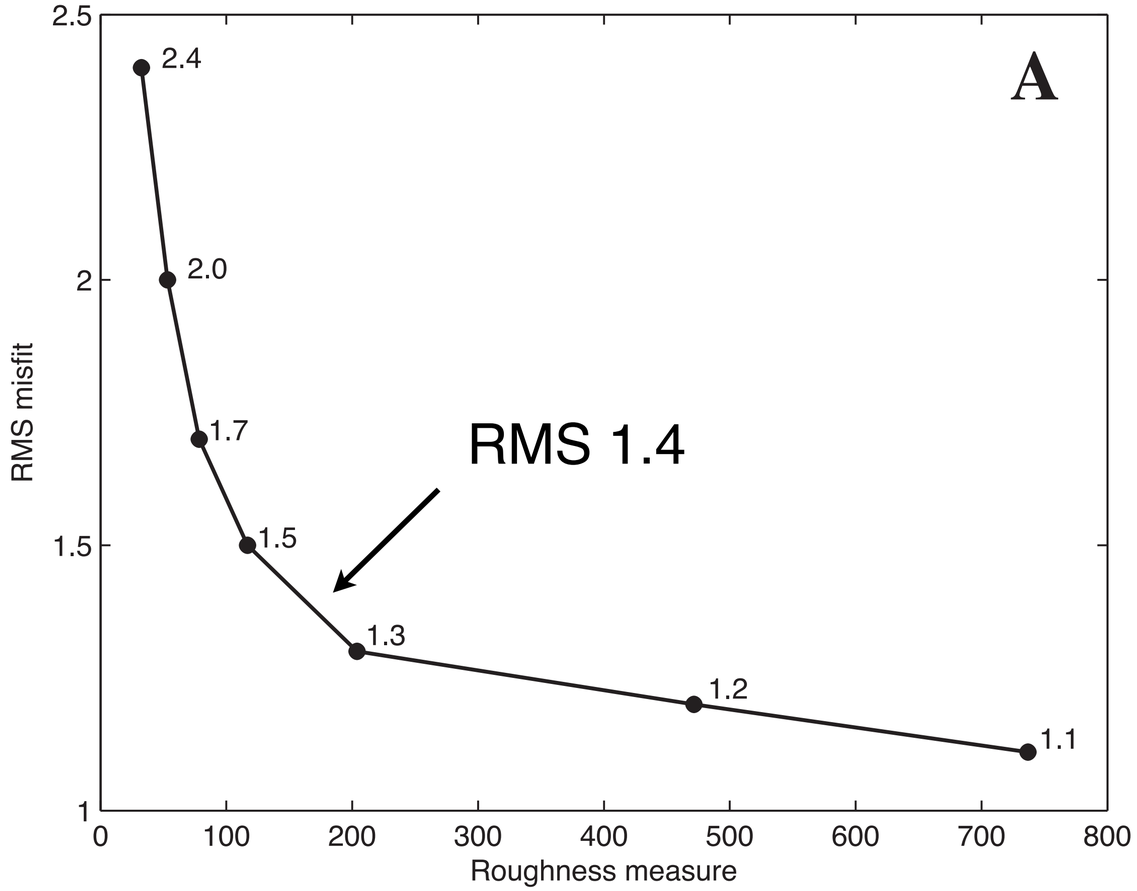




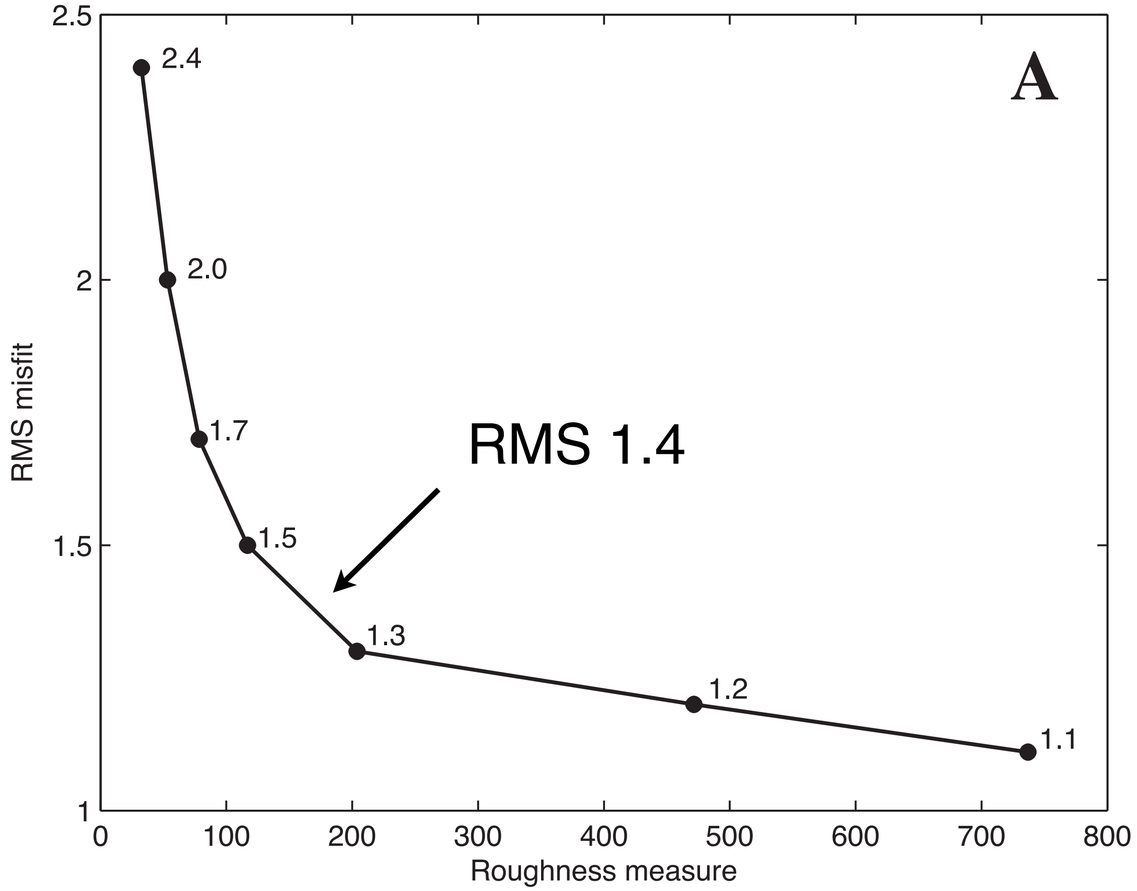


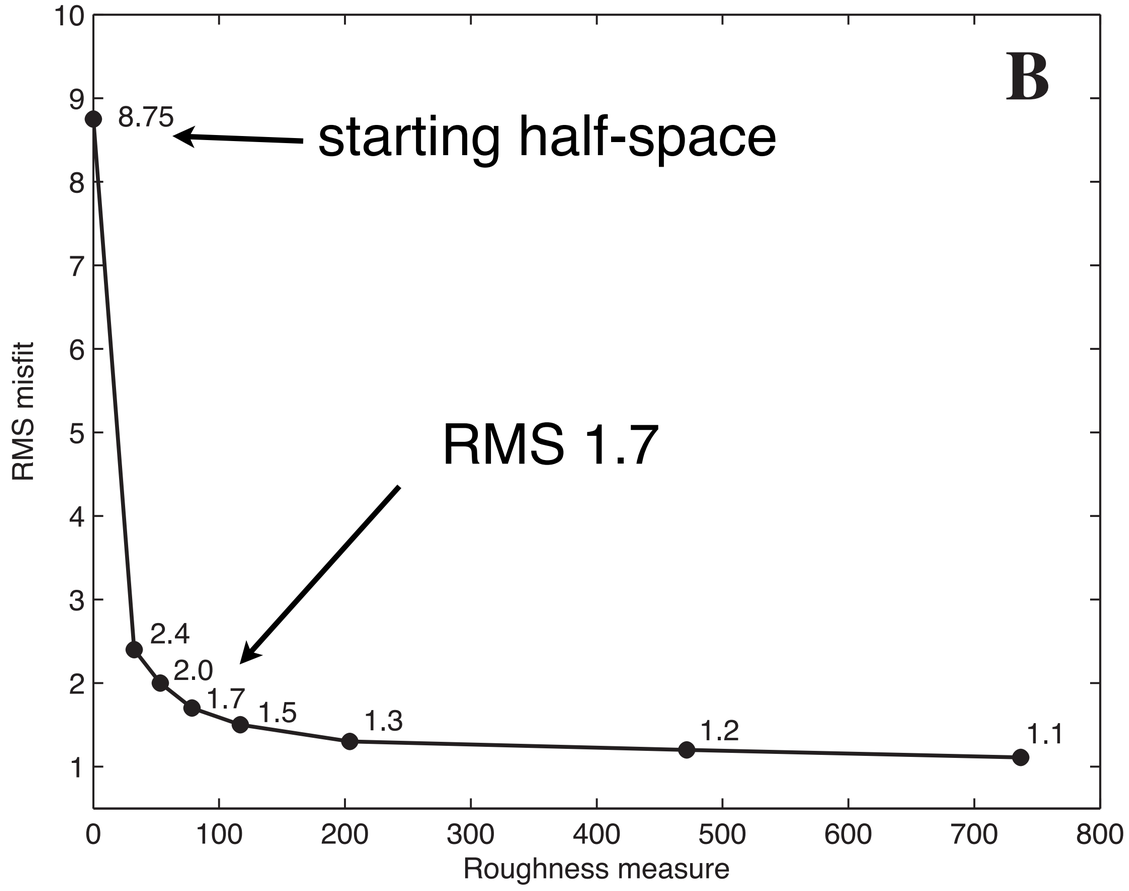
Constable, Orange, and Key, 2015

So we are often left without statistical guidance and have to use judgement in determining an adequate fit. Some people like trade-off, or “L”-curves...

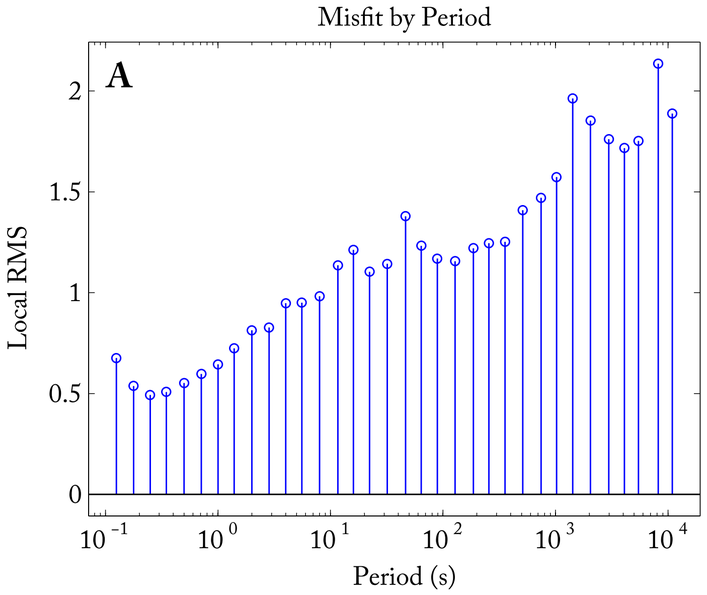


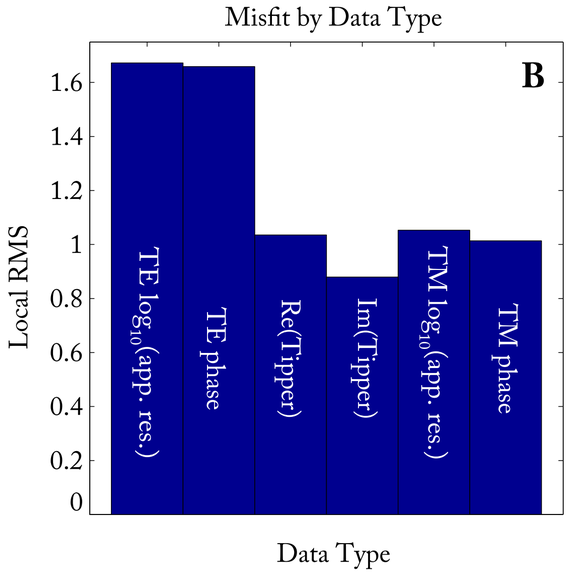
… but I am not one of them. They depend on the scope of the data and the scaling of the axes. The best way to choose a misfit level is to have a good understanding of the data and their errors.

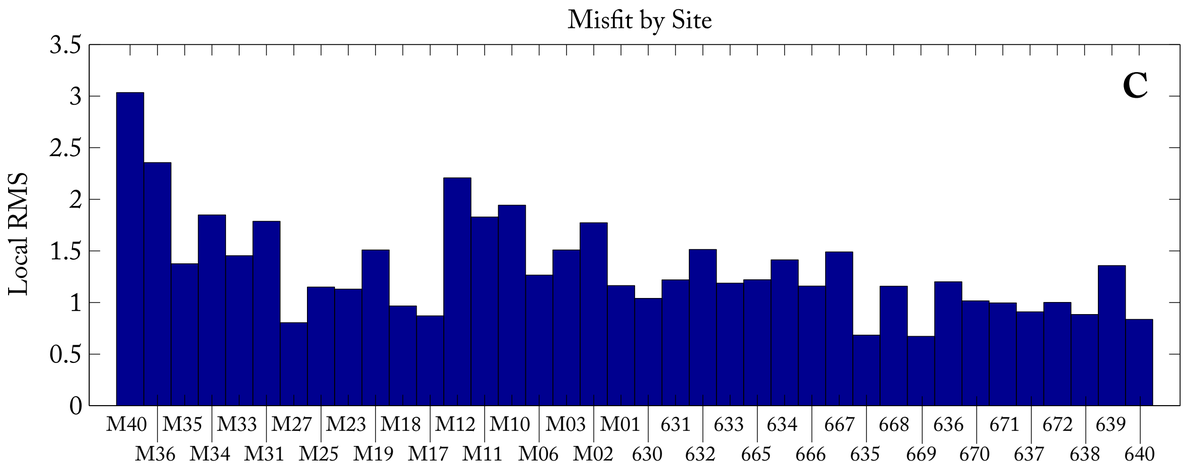




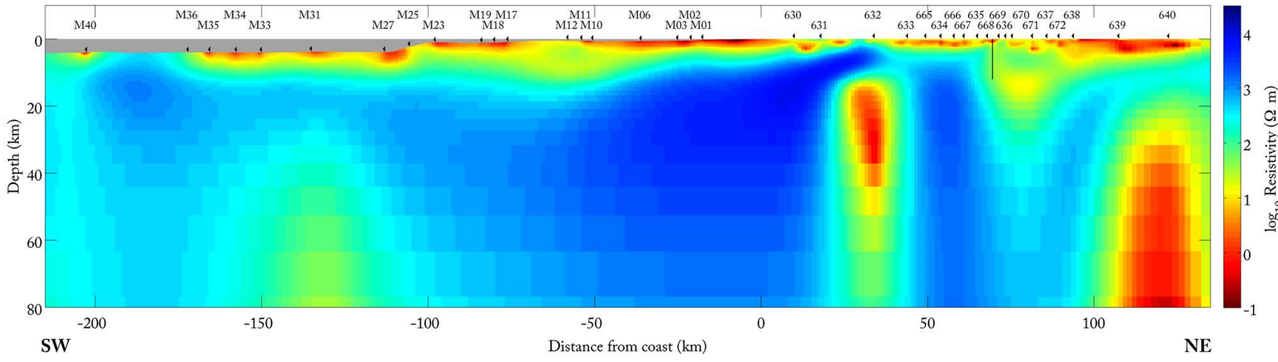
It is also a good idea to look at how the misfit is partitioned across the data: Ideally it should be random, but in practice very rarely is. Example is from MT data.



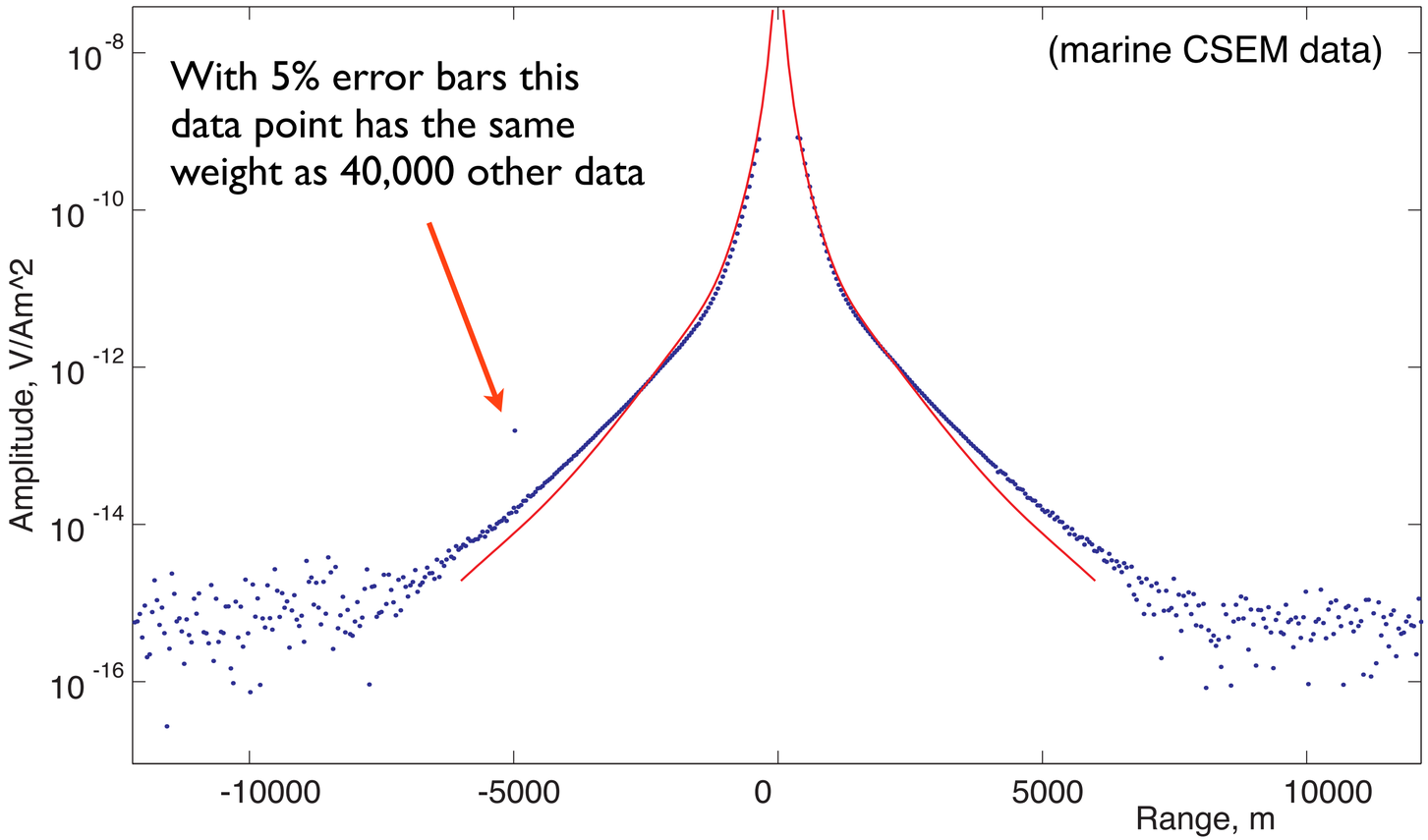




PhD thesis, Brent Wheelock, 2012.



Sum-squared misfit measures are unforgiving of outliers:



With Gaussian noise, the probability of a data point being misfit by 6 error bars is about one in a billion.

All through any inversion process you should monitor weighted residuals to ensure that there are no bad guys out there.

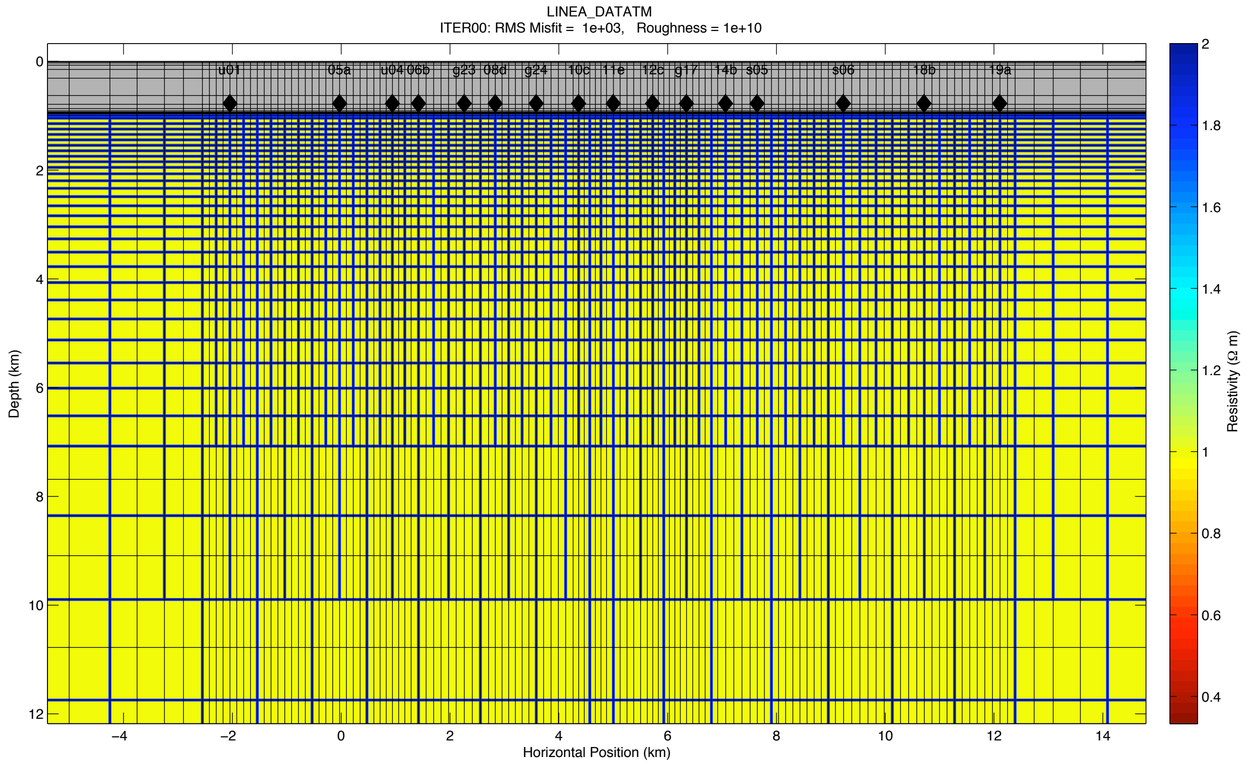
Computing **J** :

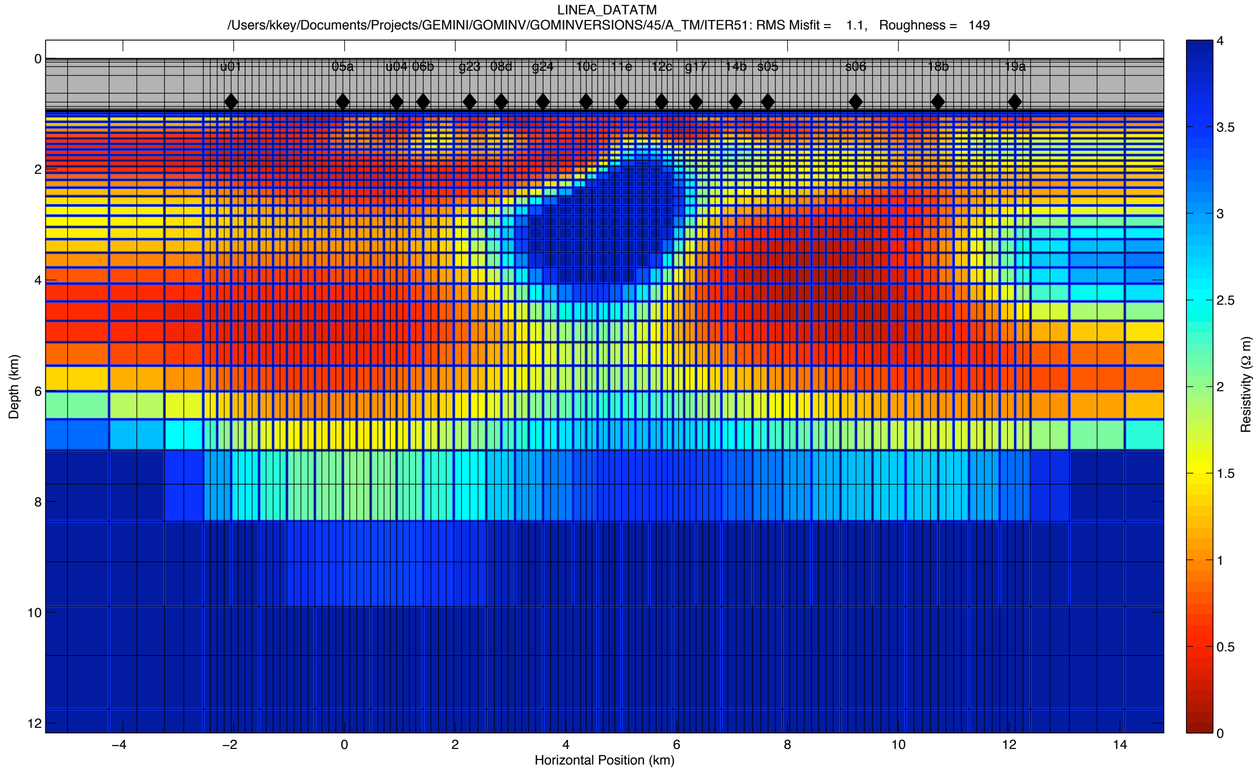
Analytical formula (works for 1D)

Forward or central differencing (easy, but can be expensive)

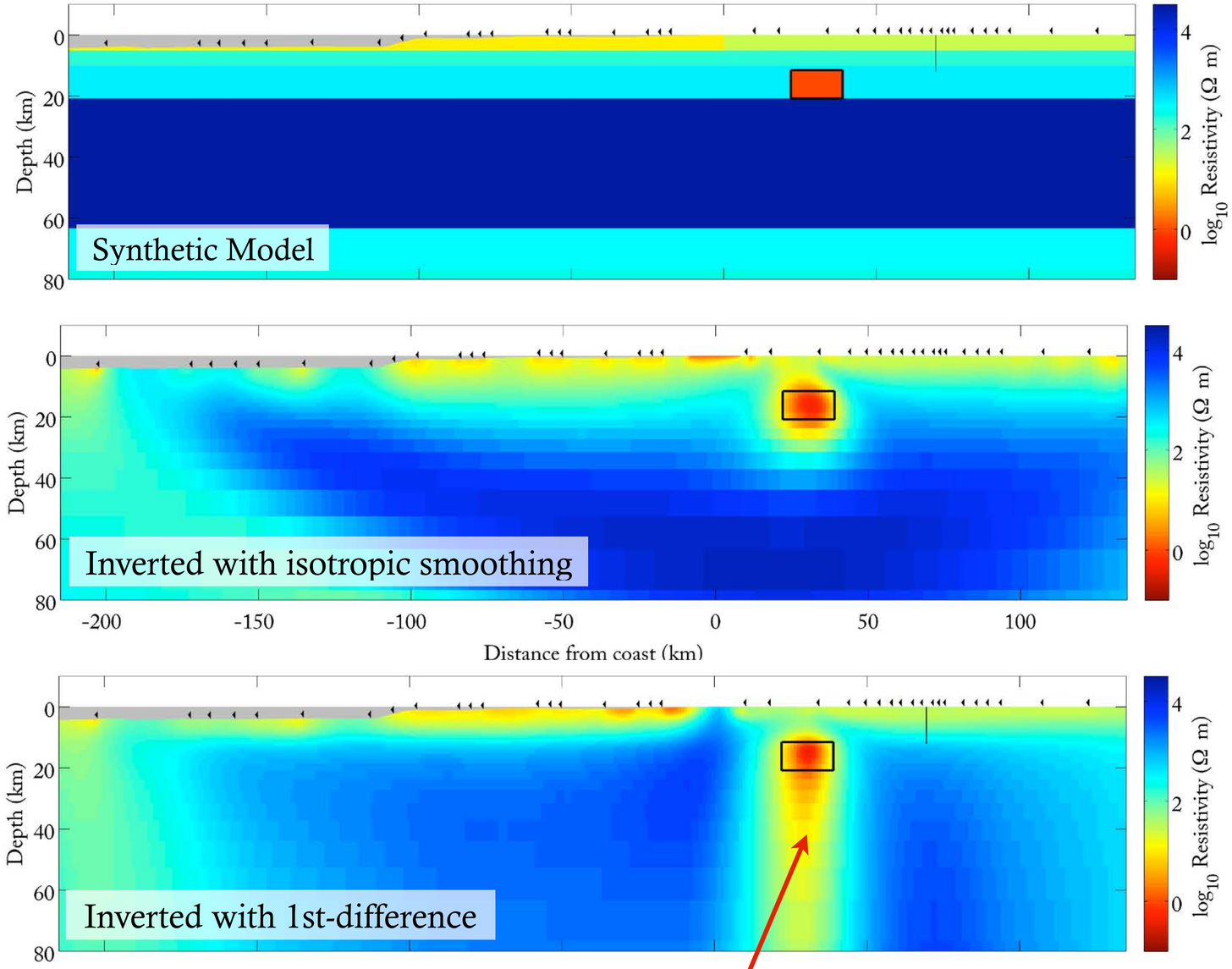
Adjoint method (computes **J** using a few forward calculations)

We often use a “dual grid” to subsample the computational mesh and allow model blocks to grow with depth. This also keeps the inversion matrices smaller. For codes such as MARE2DEM, which use adaptive mesh refinement to keep the forward calculation accurate, this is built into the code.





The regularization determines what the model looks like, just as much as the data does:



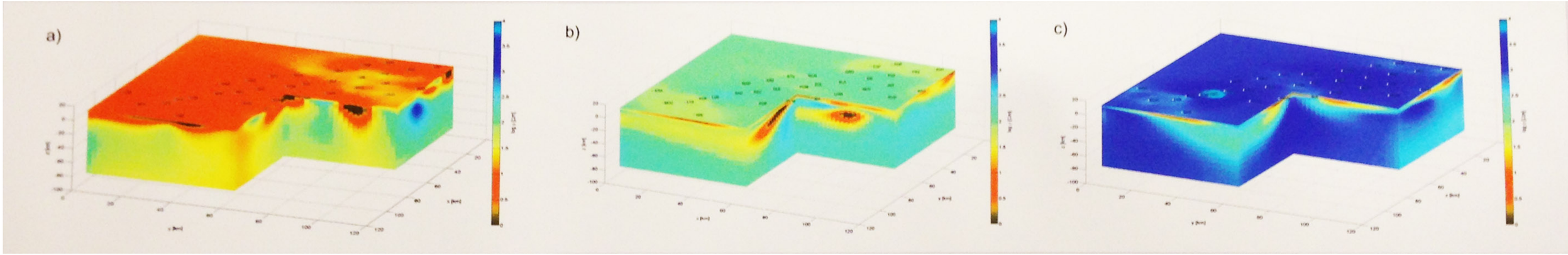
default regularization generates
a conductive artifact

Courtesy Brent Wheelock.

Be aware that many codes regularize against a "prior" model $\mathbf{m_o}$, often set to the starting model without the consent of the user.

$$U = \left(||\mathbf{Wd} - \mathbf{W}f(\mathbf{m})||^2\right) + \mu||\mathbf{R}(\mathbf{m} - \mathbf{m_o})||^2$$

Here are three inversions of the same data using ModEM, a popular 3D MT inversion package, with 10 Ωm, 100 Ωm, and 1,000 Ωm starting models:



From: Slezak, Jozwiak, Nowozynski, and Brasse, 2016 EM Induction Workshop, Thailand.

In EM, both MT apparent resistivities and CSEM amplitudes can vary by many orders of magnitude. This suggests that one should use error floors that are percentages.

One might also parameterize the data as logs. For small ϵ:

$$d' \pm 0.434\epsilon = \log_{10}(d \pm \epsilon d)$$

log data data — linear data — fractional error

$$0.434 = 1/\ln(10)$$

It ought not to matter how you parameterize the data (so long as the errors are properly scaled and the appropriate chain rule is applied to the Jacobian):

$$\mathbf{m}_1 = \left[\mu\mathbf{R}^T\mathbf{R} + (\mathbf{WJ})^T\mathbf{WJ}\right]^{-1}(\mathbf{WJ})^T\mathbf{W}(\mathbf{d} - f(\mathbf{m_0}) + \mathbf{Jm}_0) \quad .$$

make data log

take the log of the forward model

use the chain rule to convert $\partial\mathbf{d}/\partial\mathbf{m}$ to $\partial\log(\mathbf{d})/\partial\mathbf{m}$

But …

... it does.

Here we consider MT data over a half-space, varying only half-space resistivity R.

Recall that MT impedance (Z) is:

$$E_x = Z H_y$$

$$\rho = \frac{1}{2\pi f \mu} |Z|^2$$

For small R, misfit flattens for linear ρ

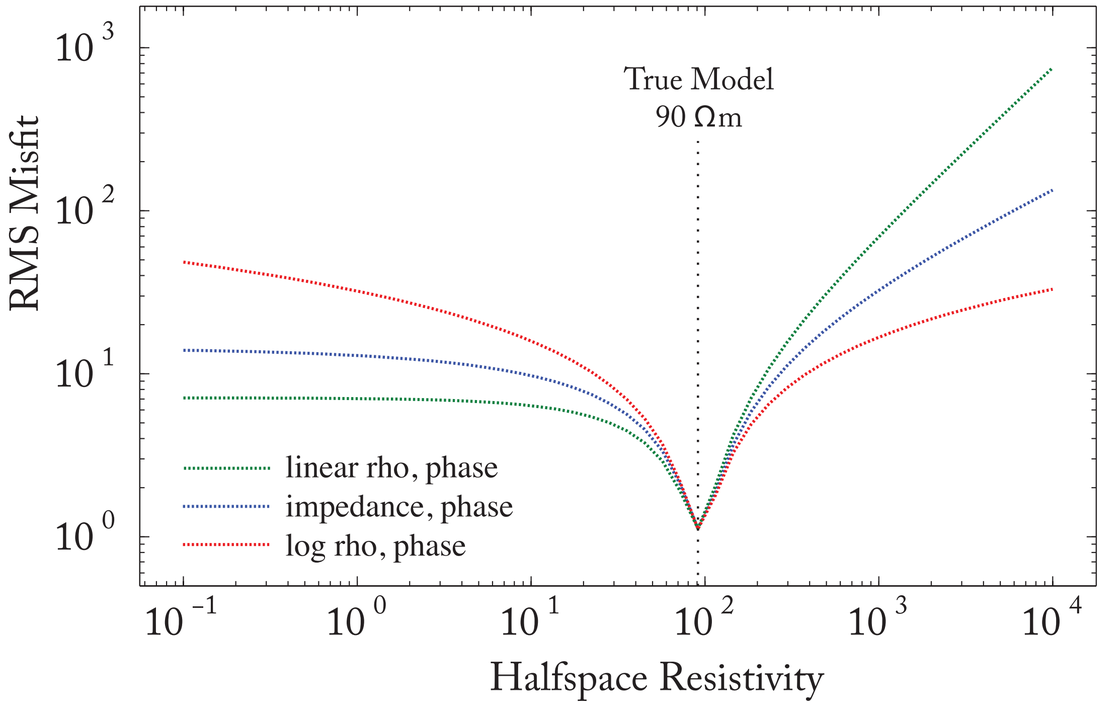
This is because

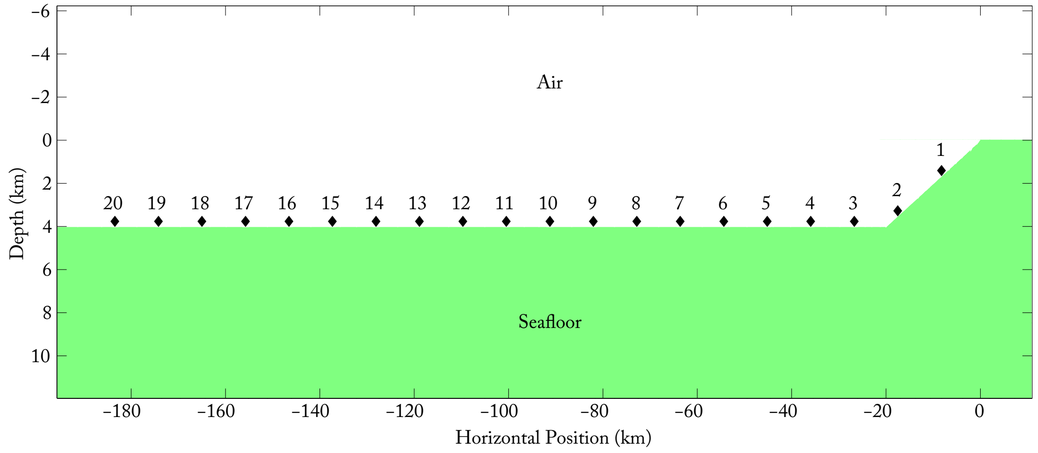
$$(\mathbf{d} - f(\mathbf{m})) \rightarrow \text{const.}$$

as $f(\mathbf{m}) \rightarrow 0$ but

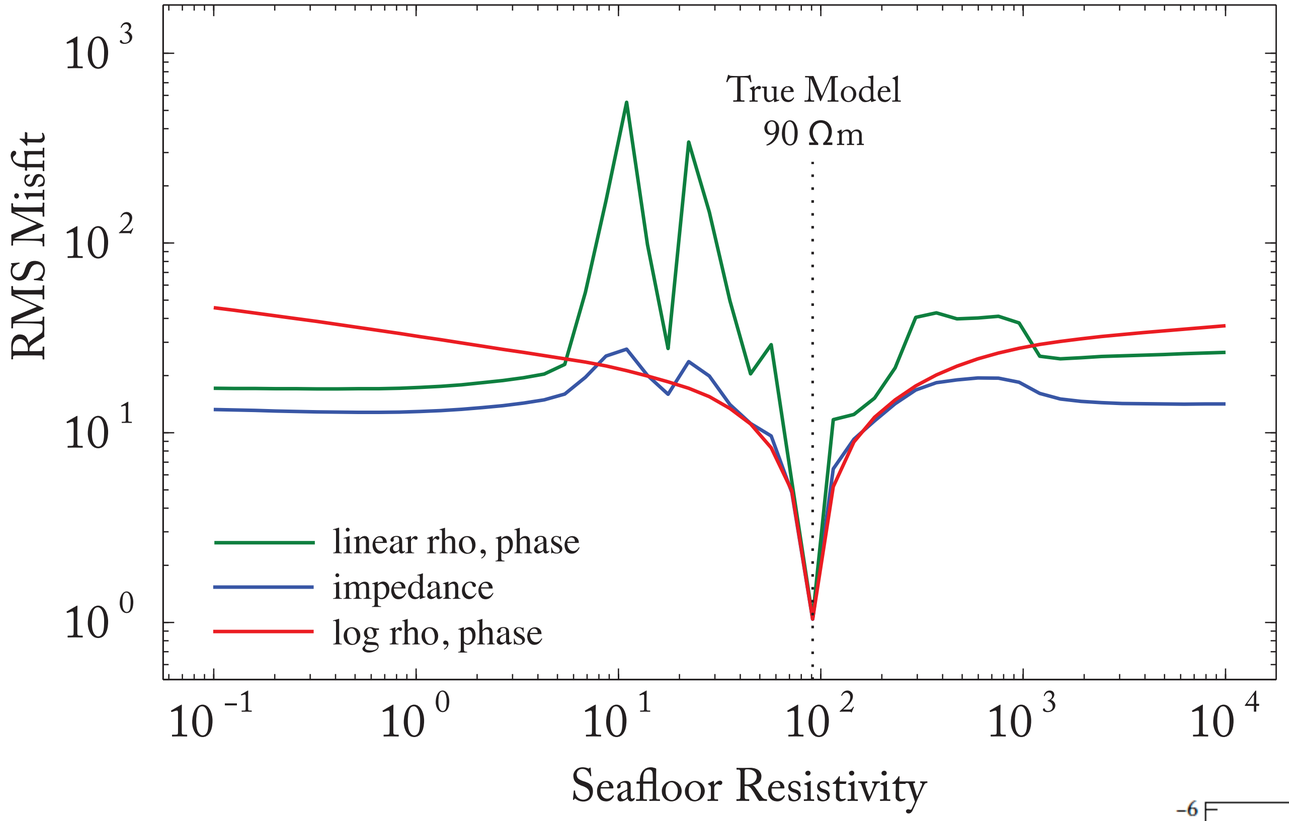
$$(\log \mathbf{d} - \log f(\mathbf{m}))$$

does not.



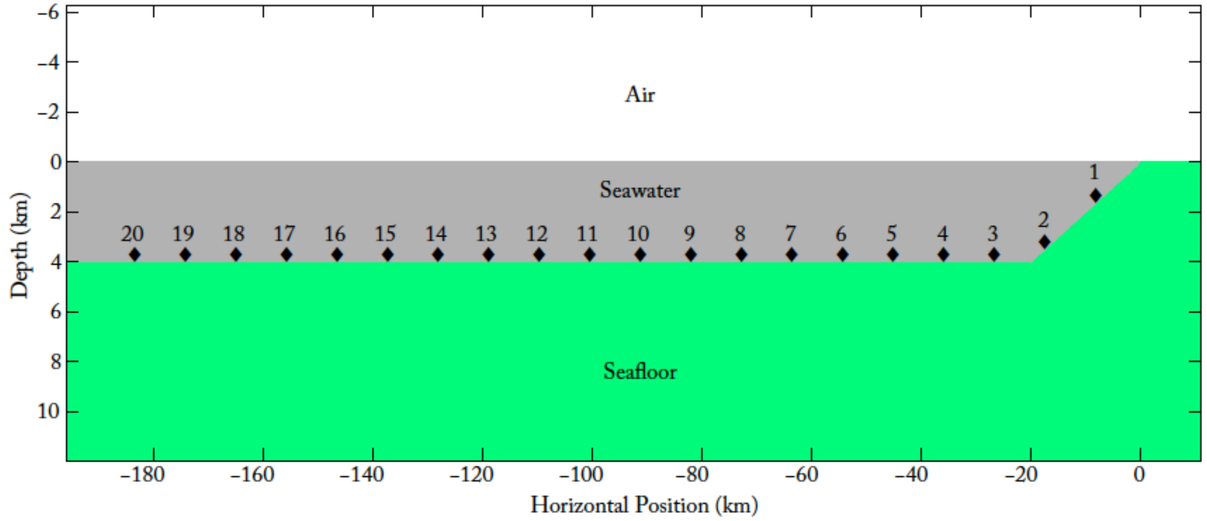


(modified from Wheelock et al., 2015)



This effect can be even worse for marine MT data affected by bathymetry.

Local minima develop, and misfit flatlines at low R.



(modified from Wheelock et al., 2015)

It is important to remember that models from geophysical inversion depend on much more than the data alone: