Estimating the posterior using prior sampling: The case of rapid seismic source inference.

Jeannot
Trampert

Universiteit Utrecht

2 routes to posterior sampling:

Posterior sampling: the samples are drawn to explain a certain set of observations.
---> new observation means new sampling

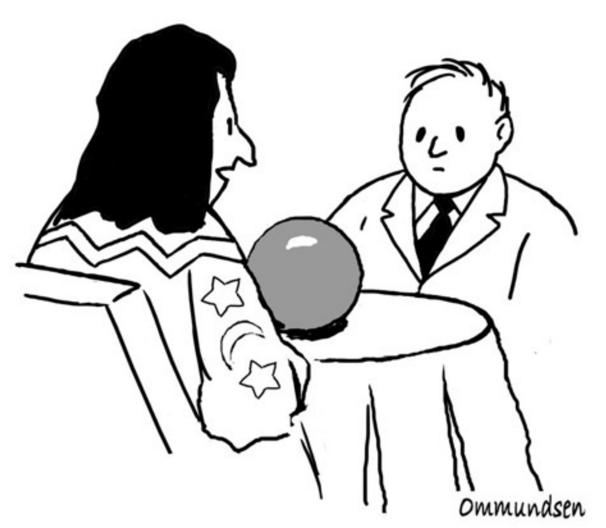
Prior sampling: independent of any set of observations and only relies on some prior information.
---> samples can be recycled to solve similar problems repeatedly

We use the concept of conjunction of information (Tarantola, 2005), which is a generalization of Bayes' theorem, to solve inference problems

Posterior = (Prior * Likelihood) / Evidence

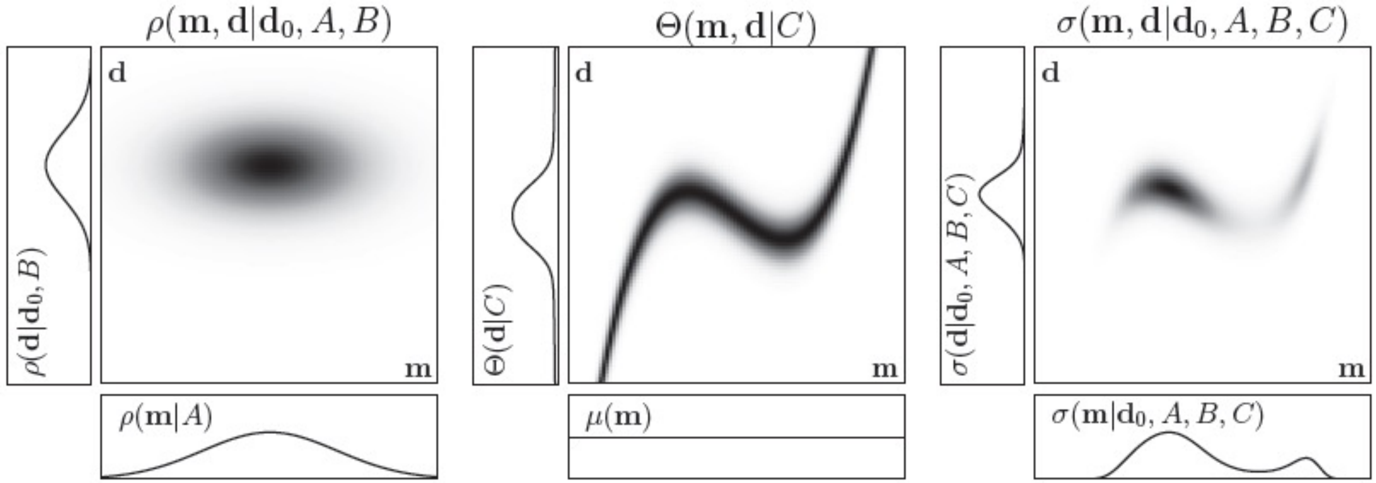
$$\sigma(\mathbf{m}, \mathbf{d}, \mathbf{d}_0 | A, B, C) = k \frac{\rho(\mathbf{m}|A)\rho(\mathbf{d}, \mathbf{d}_0|B)\Theta(\mathbf{m}, \mathbf{d}|C)}{\mu(\mathbf{m})\mu(\mathbf{d})\mu(\mathbf{d}_0)},$$

m is a vector of model parameters
d=g(m) is a data prediction based on m
d_0 is some observation
A,B,C are some assumptions



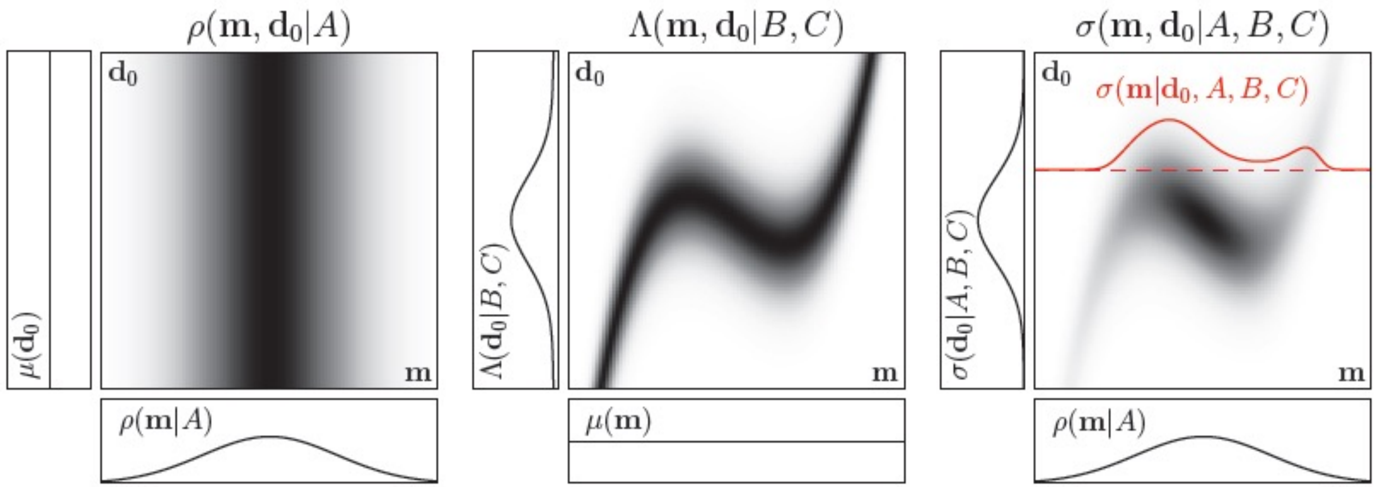
"Is this needed for a Bayesian analysis?"

Posterior route



(a) The fact that we have observed $\mathbf{d}_0$ enters the inference process in the form of prior information described by the pdf $\rho(\mathbf{d}|\mathbf{d}_0, B)$ (left hand panel). Theoretical predictions are represented by $\Theta(\mathbf{m}, \mathbf{d}|C)$, taking into account any assumptions about the relation between data and model parameters (middle panel). The posterior pdf is given by the marginal pdf $\int_{\mathbb{D}} \sigma(\mathbf{m}, \mathbf{d}|\mathbf{d}_0, A, B, C)\mathrm{d}\mathbf{d}$ (right hand panel).

Prior route: dataless sampling



(b) We assume that the observation $\mathbf{d}_0$ is available only at a later time and is thus treated as an unknown variable with prior distribution $\mu(\mathbf{d}_0)$ (left hand panel). The pdf $\Lambda(\mathbf{m}, \mathbf{d}_0|B, C)$ carries the combined assumptions on theoretical modelling and observational uncertainties (middle panel). The posterior pdf is given by the conditional pdf $\sigma(\mathbf{m}|\mathbf{d}_0, A, B, C)$ (red solid line, right hand panel).

In practice, due to non-Gaussian distributions and the non-linearity of g(m), we do not get closed form solutions
→ sampling

Comparing posterior versus prior sampling
→ different sample density!

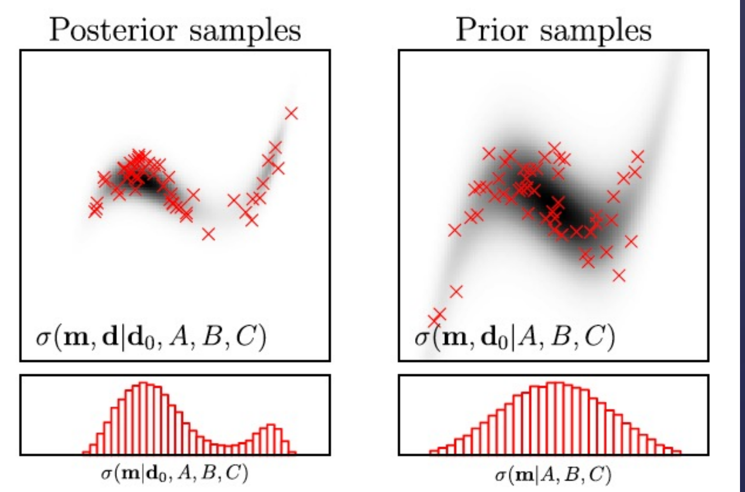
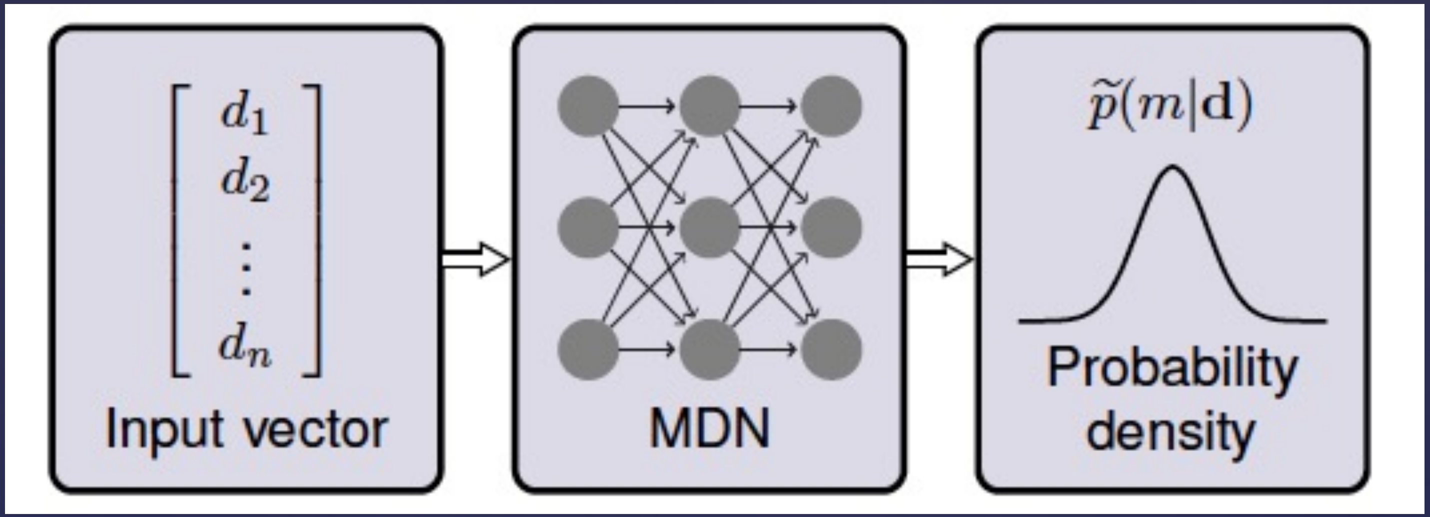


Figure 2. Two approaches to sampling the relationship shown in Fig. 1. A set of posterior samples $\mathcal{D}_{\text{post}}$ (left), whose density follows the distribution $\sigma(\mathbf{m}, \mathbf{d}|\mathbf{d}_0, A, B, C)$ and a set of prior samples $\mathcal{D}_{\text{prior}}$ (right) following $\sigma(\mathbf{m}, \mathbf{d}_0|A, B, C)$.

Solution obtained by marginalization versus conditioning.

We use neural networks as a regression engine for $\sigma(\mathbf{m}|\mathbf{d}_0, A, B, C)$ using the prior samples



A neural network requires

- Architecture
- Activity rule
- Training (we train on synthetic data)
- Assessment

A Toy Problem: locate a particle in c-dimensional space by knowing it's distance from the origin

g(m)=||m||$_2$ (L-2 norm)

Prior on m is uniform in the c-dimensional cube $[-1,1]^c$

Measurement d_0 subject to Gaussian noise
Analytical solution for d_0=0

Find $\sigma(m|d_0)$

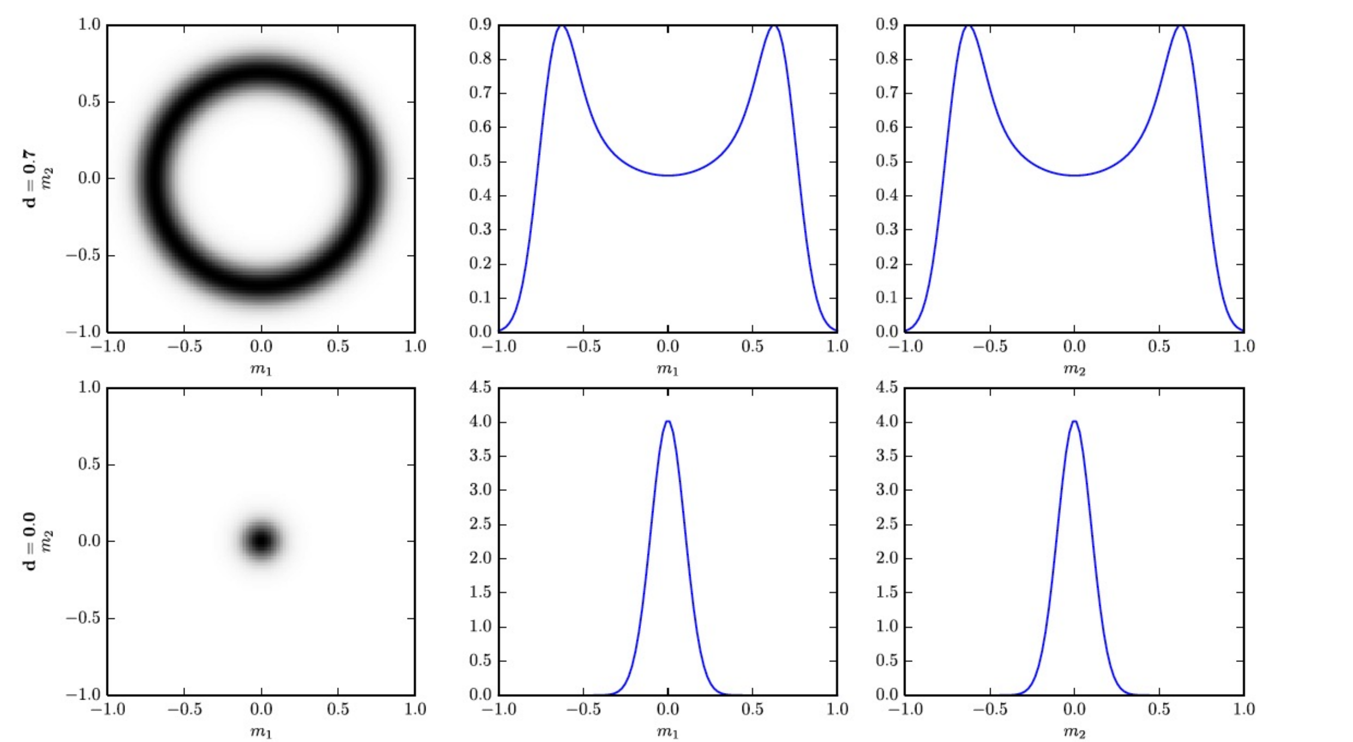


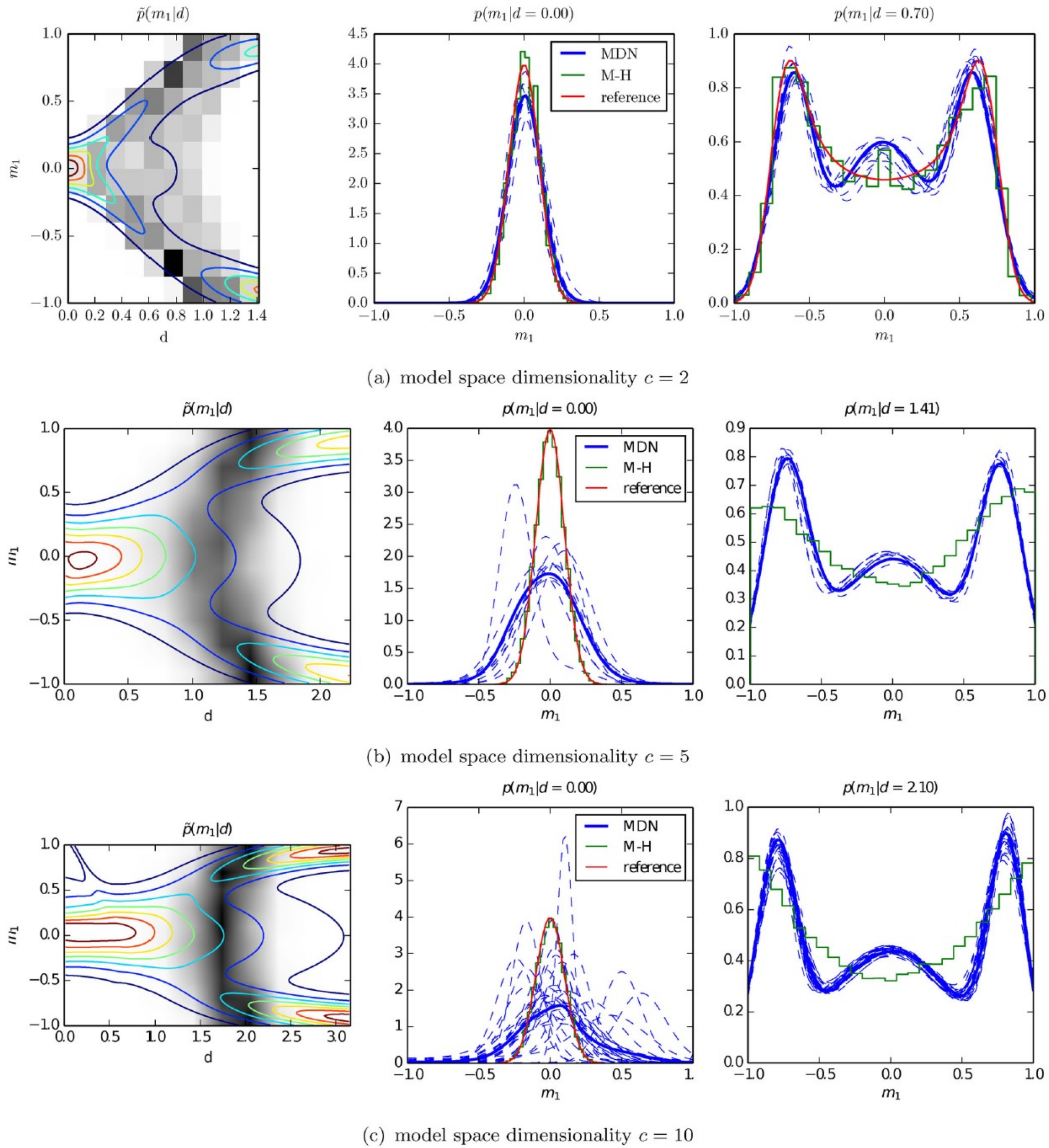
Figure 3. Top row, left: posterior $p(\mathbf{m}|d_0 = 0.7)$ for the case $n = 2$. Middle and right: marginals $p(m_1|d_0 = 0.7)$ and $p(m_2|d_0 = 0.7)$, respectively. Bottom row: the same, but for $d_0 = 0.0$. Darker colours correspond to higher probabilities.

Grey shading is prior sample density

Posterior sampling using M-H (run twice)

MDN's comprise several members each trained by different random initializations.

Dark blue line in the average used for inference (evaluated twice)



Advantages of prior sampling and NN regression:

- Computationally-expensive sampling performed separately from 'solving the inverse problem'
- No need to consider burn-in, chain thinning, etc
- Samples can be reused in conjunction with many different observations
- A single set of simulations (samples) may be processed and used in a variety of ways

Disadvantages:

- Only a few samples are 'close' to any given observation; the rest do not provide useful information
- Reliant on assumption of smoothness
- Training more difficult

Inference is conservative:

- Posterior depends on sample density as well as likelihood and prior

Ideal if:

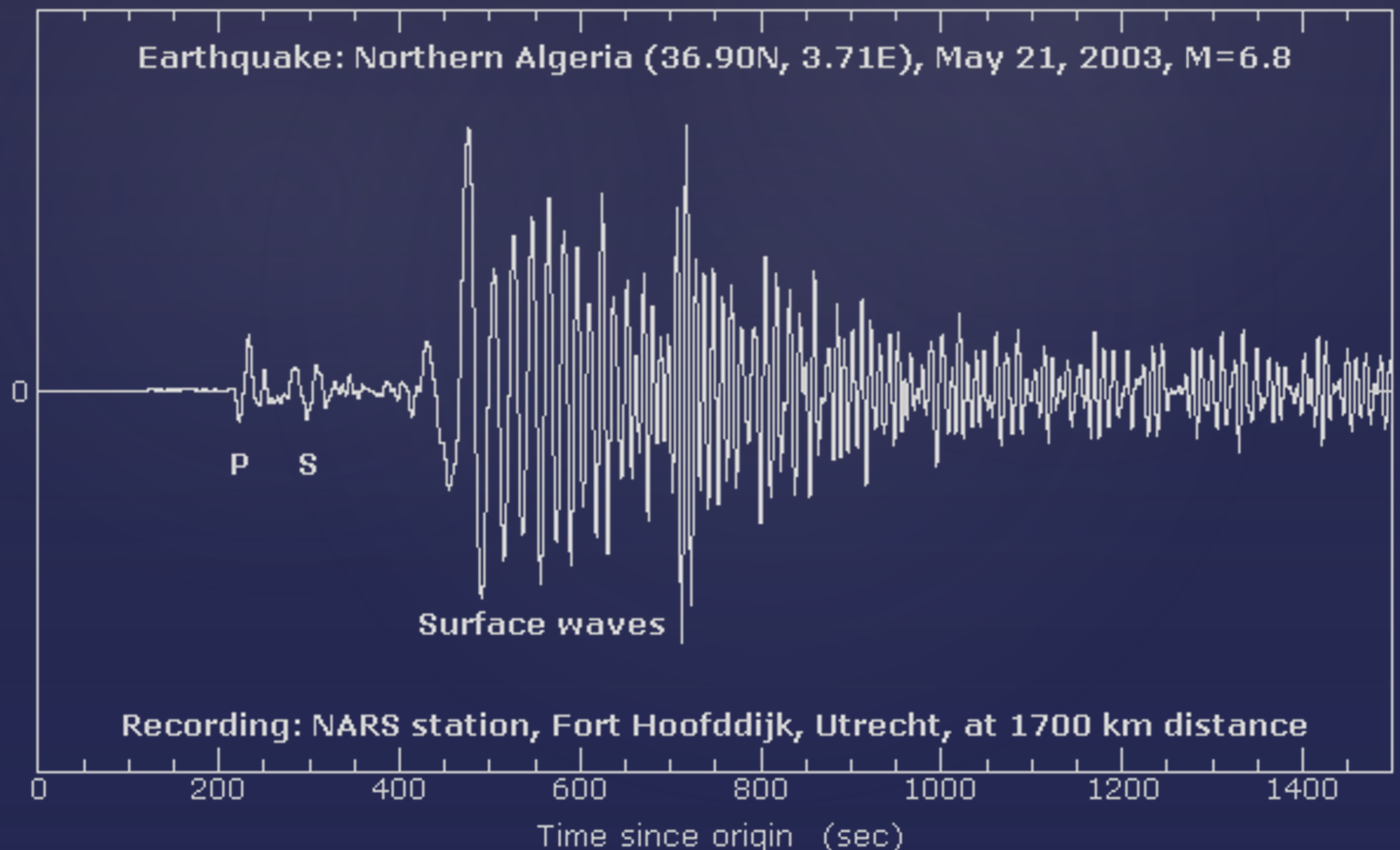
- The same inverse problem must be solved repeatedly
- It is necessary to minimise the time between observation and result

An example of earthquake early warning

→ Fast and repeated inversion of seismic waveforms for earthquake parameters

We use seismograms (records of ground displacement) to infer the physics of earthquake sources or the Earth's internal structure

Seismogram = Source * Green's function * Instrument



Workflow

1. Generate synthetic data (with exact physics)
2. Define target (parameter of interest)
3. Approximate posterior using a neural network
4. Test network prediction
5. Apply to real data

Area

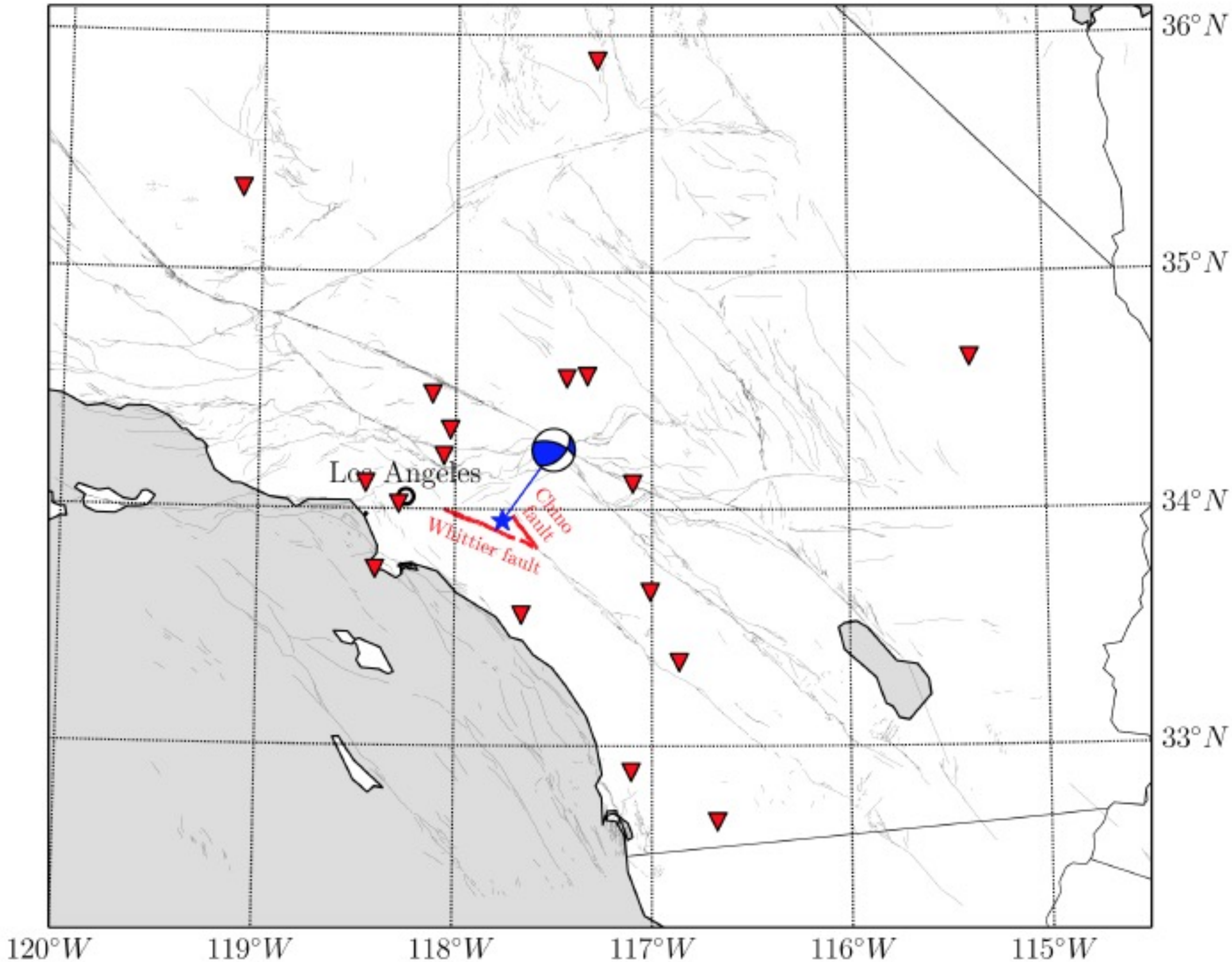


Figure 5.3: Surface traces of known faults in the study region (taken from http//earthquakes.usgs.gov/regional/qfaults/). The locations of 17 strong-motion instruments of the SCSN and ANZA Regional Network are shown as red triangles, the epicentral location of the 2008 M_w 5.4 Chino Hills event, located between the Whittier and Chino fault sections, is shown as a blue star, where the beach-ball corresponds to a moment tensor point source solution obtained by Hauksson et al. (2008).

Training data: earthquake locations

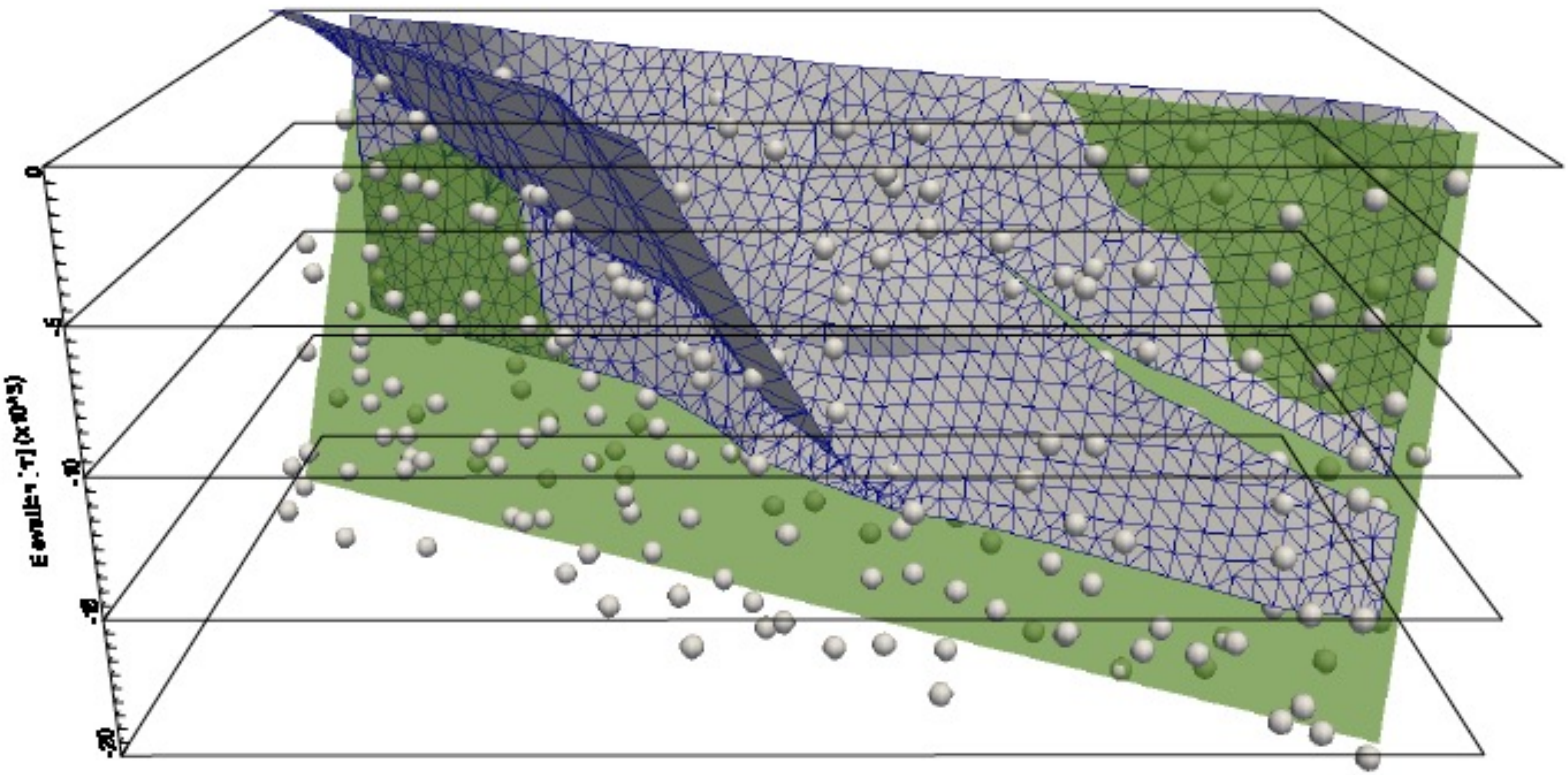


Figure 5.4: Whittier and Chino fault geometry taken from the Community Fault Model (CFM) for Southern California (Plesch et al., 2007) (blue and grey mesh), a planar representation of the Whittier fault (green) and the locations of 150 pseudo-randomly distributed source locations (grey balls).

Good structural model for training data

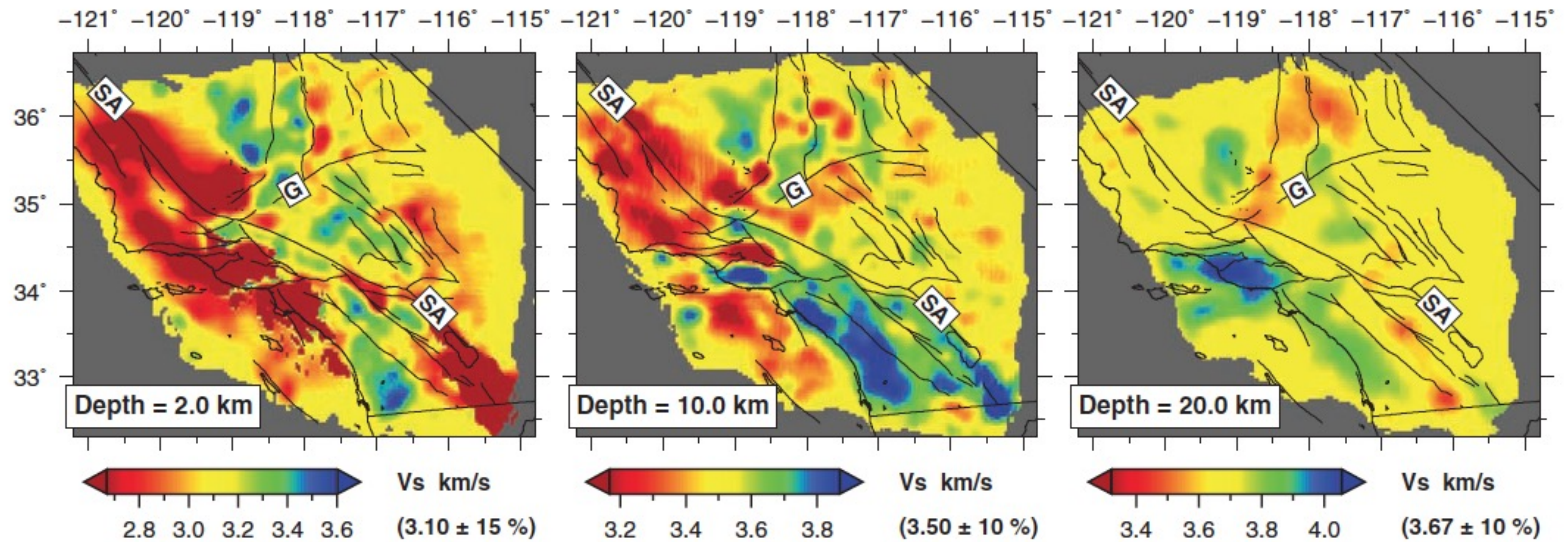


Fig. 4. Horizontal cross sections of V_S tomographic model $\mathbf{m}_{16}$ at depths of 2, 10, and 20 km. See fig. S1 for locations of major features; Garlock (G) and San Andreas (SA) faults are labeled for reference.

Tape et al. 2009

Training data:

SPECFEM3D

Tape et al. 2009

150 sources

6 component Green's functions

1866 stations

Seismograms (T>2s)

100000 cpu hours

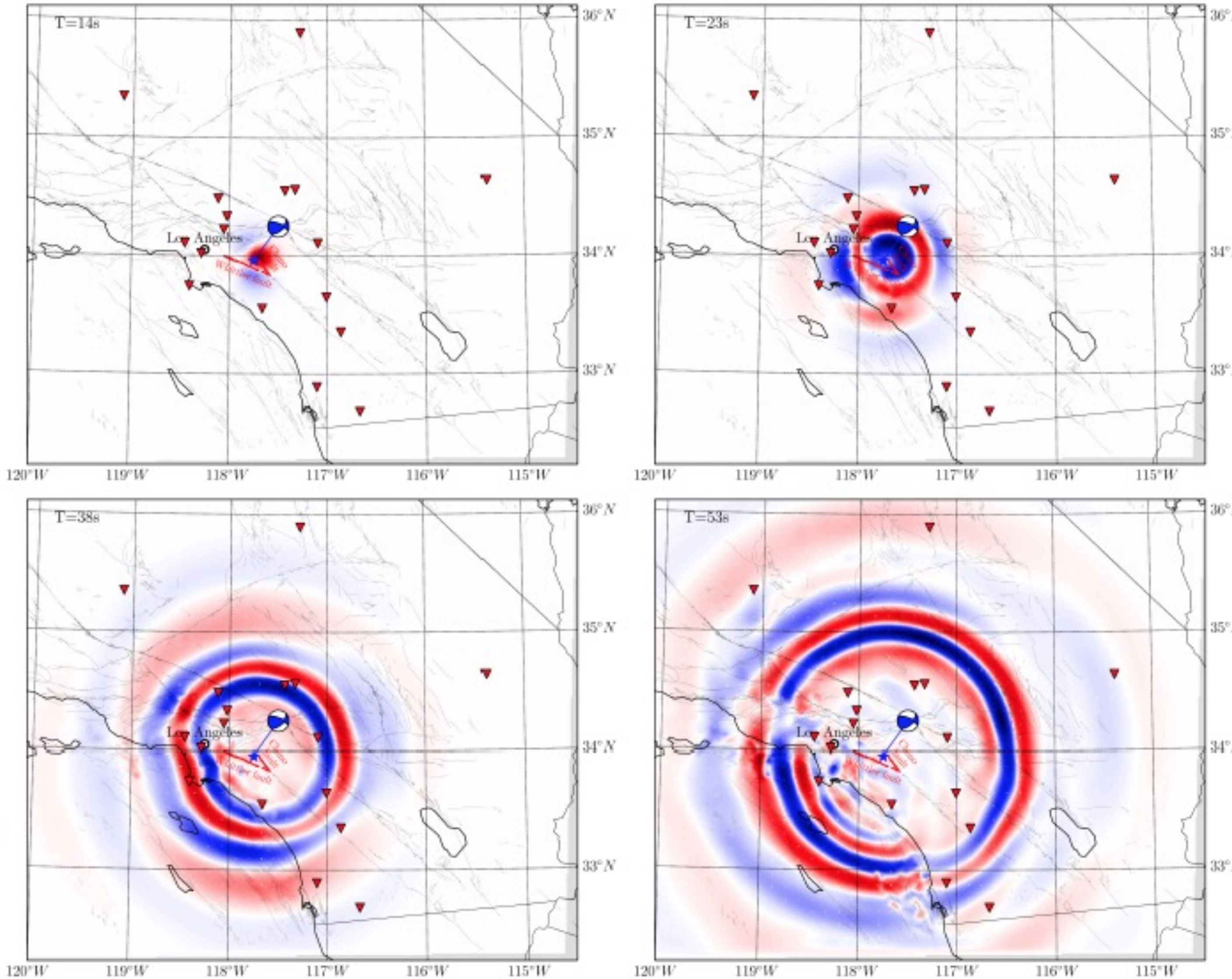
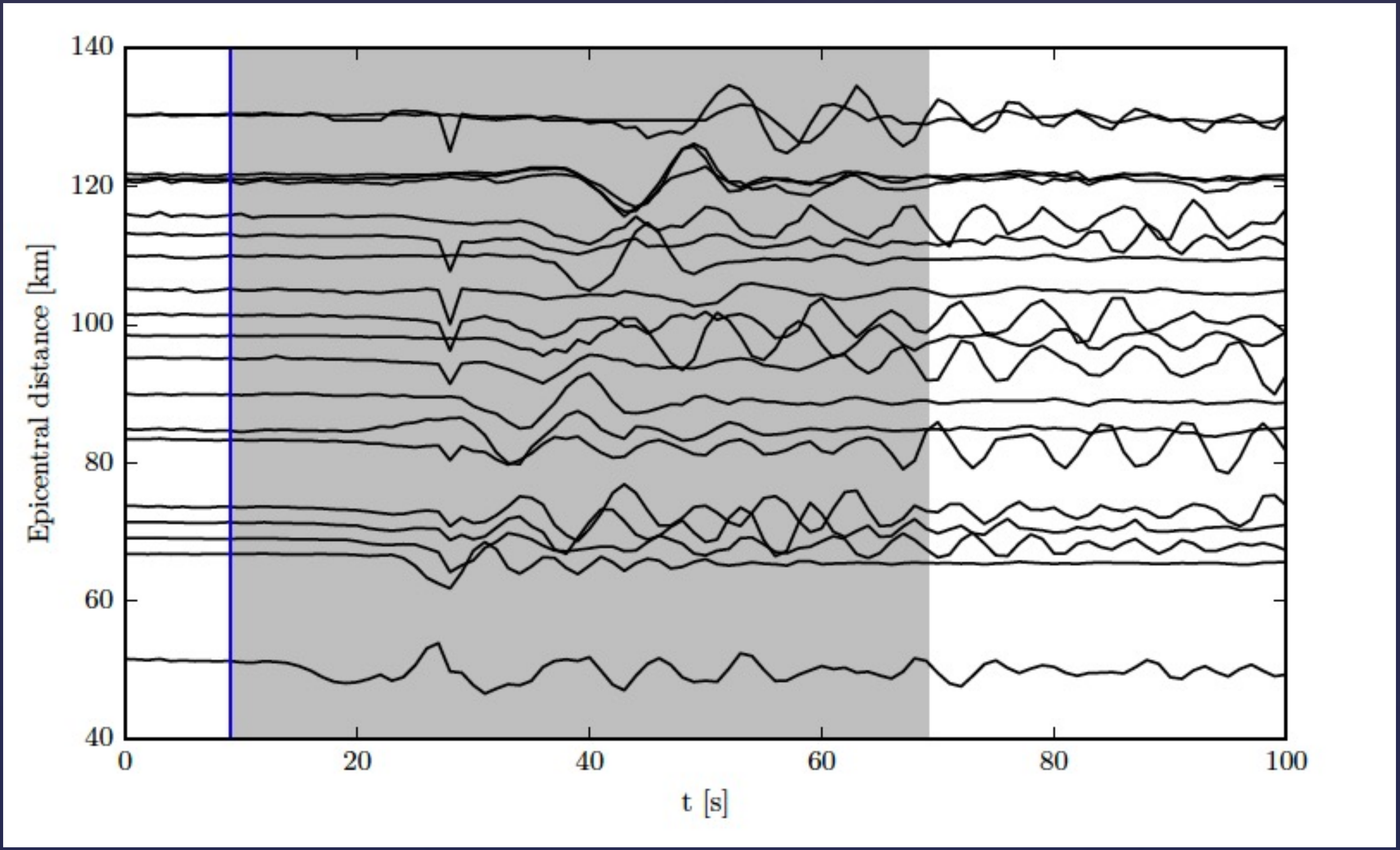


Figure C.2: Snapshots of the vertical component surface velocity field for the synthetic example of Figure 5.7 with a source half-duration of 6 s. Points in time are with respect to origin time and correspond to the endpoints of data windows of length 6 s, 15 s, 30 s and 45 s, respectively. Note that first arrivals are not clearly visible due to their small amplitudes. The first P wave arrival calculated using a 1-D average model is at 7.6 s after origin time at the closest receiver.

We use full waveforms for network training



Assessment using data not used for training

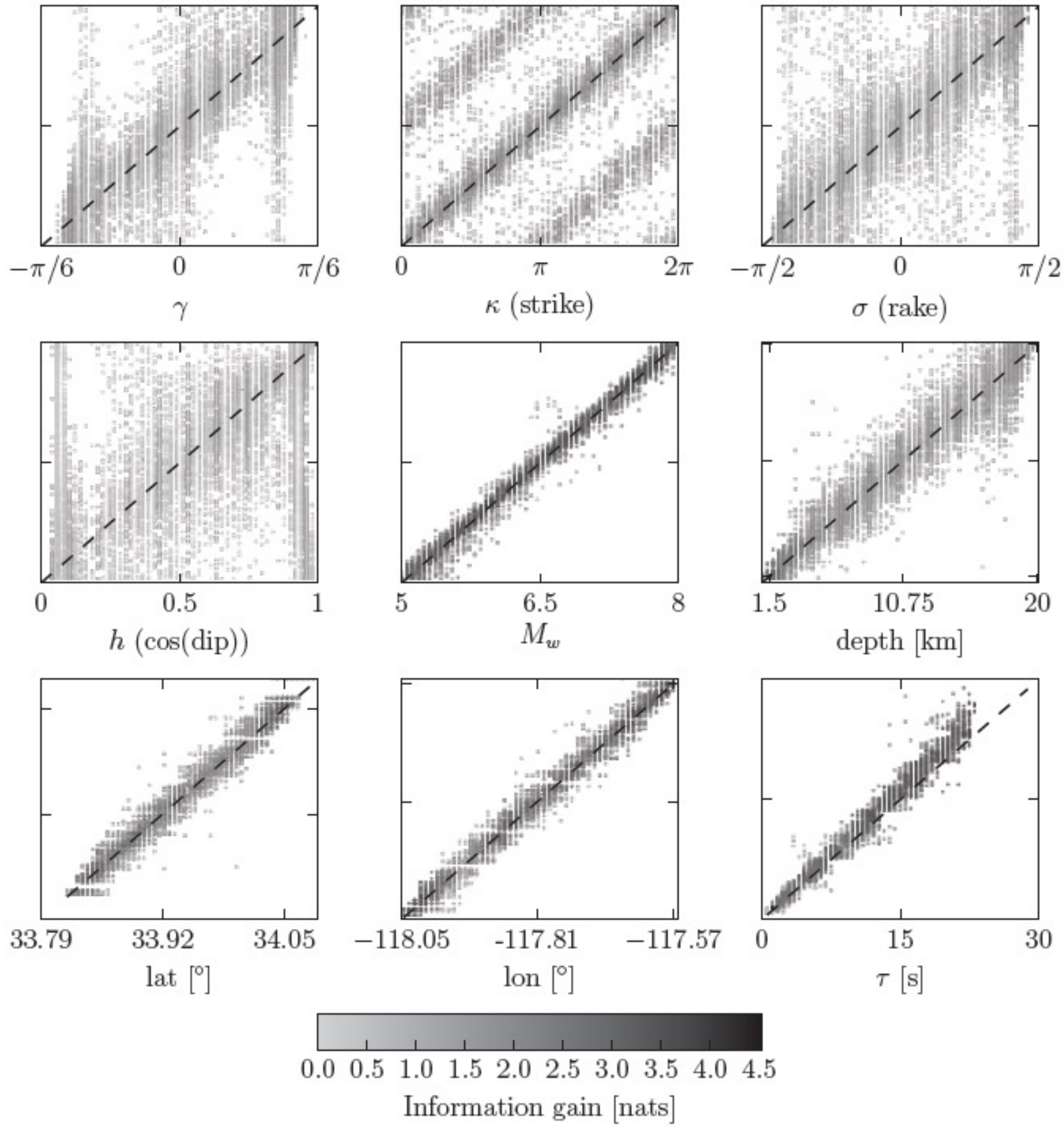
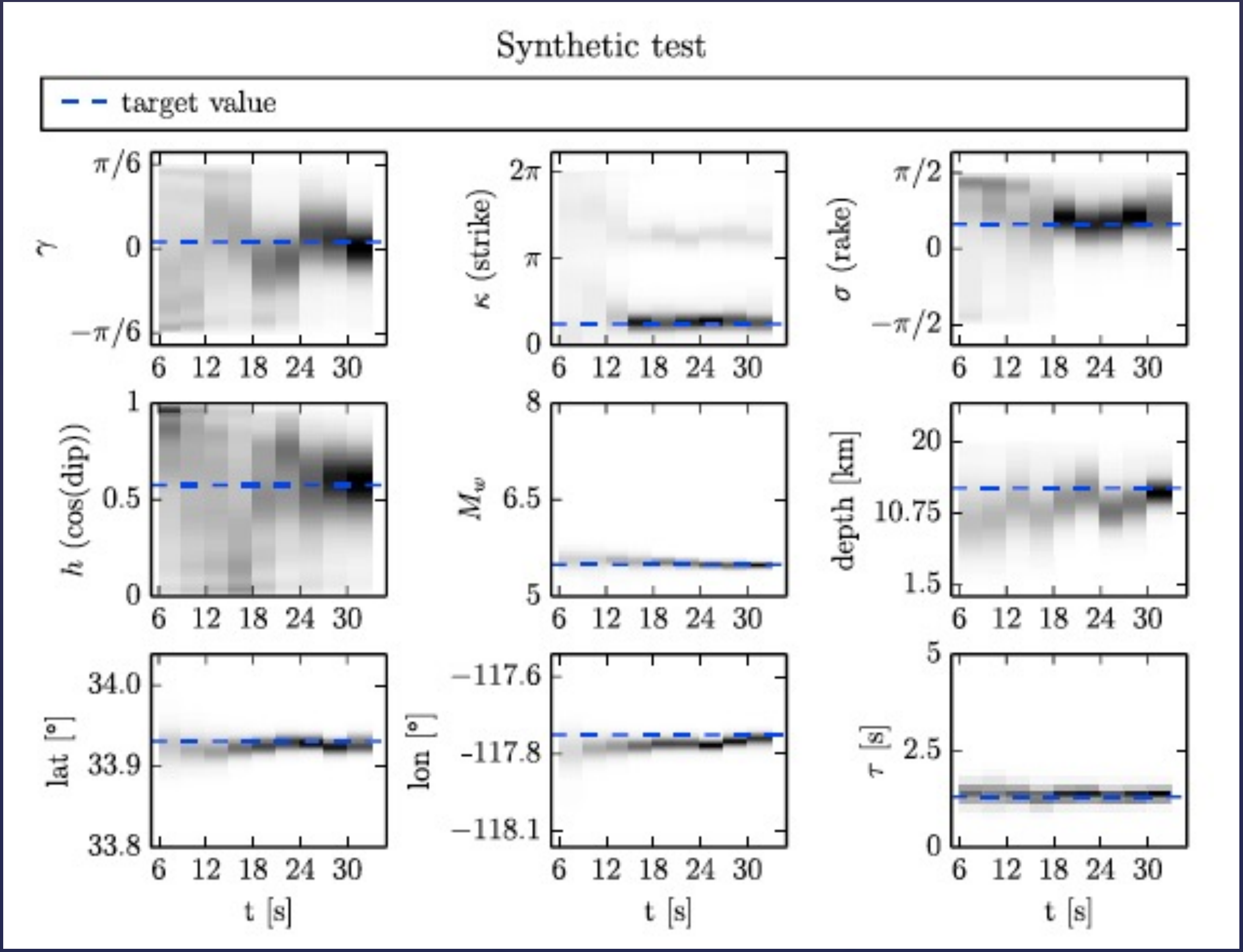


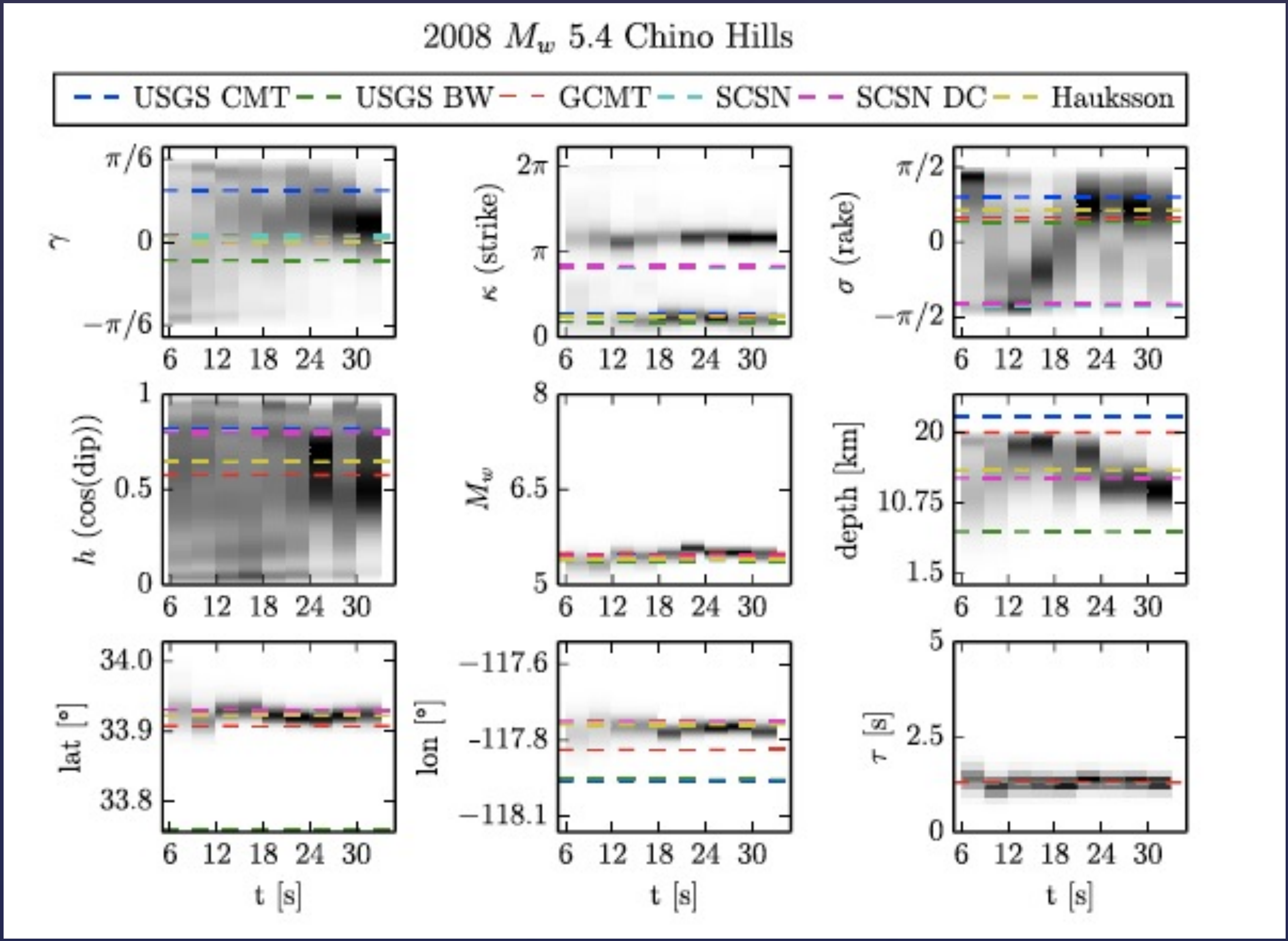
Figure 5.5: Prediction performance of network ensembles trained on the 9 source parameters using 30 s of data. Each dot corresponds to a noisy synthetic test set example, the position of the posterior mode is plotted on the horizontal, the target value on the vertical axis. The gray-scale indicates the relative information gain with respect to the prior distribution.

Results



Kaeufl et al., 2016

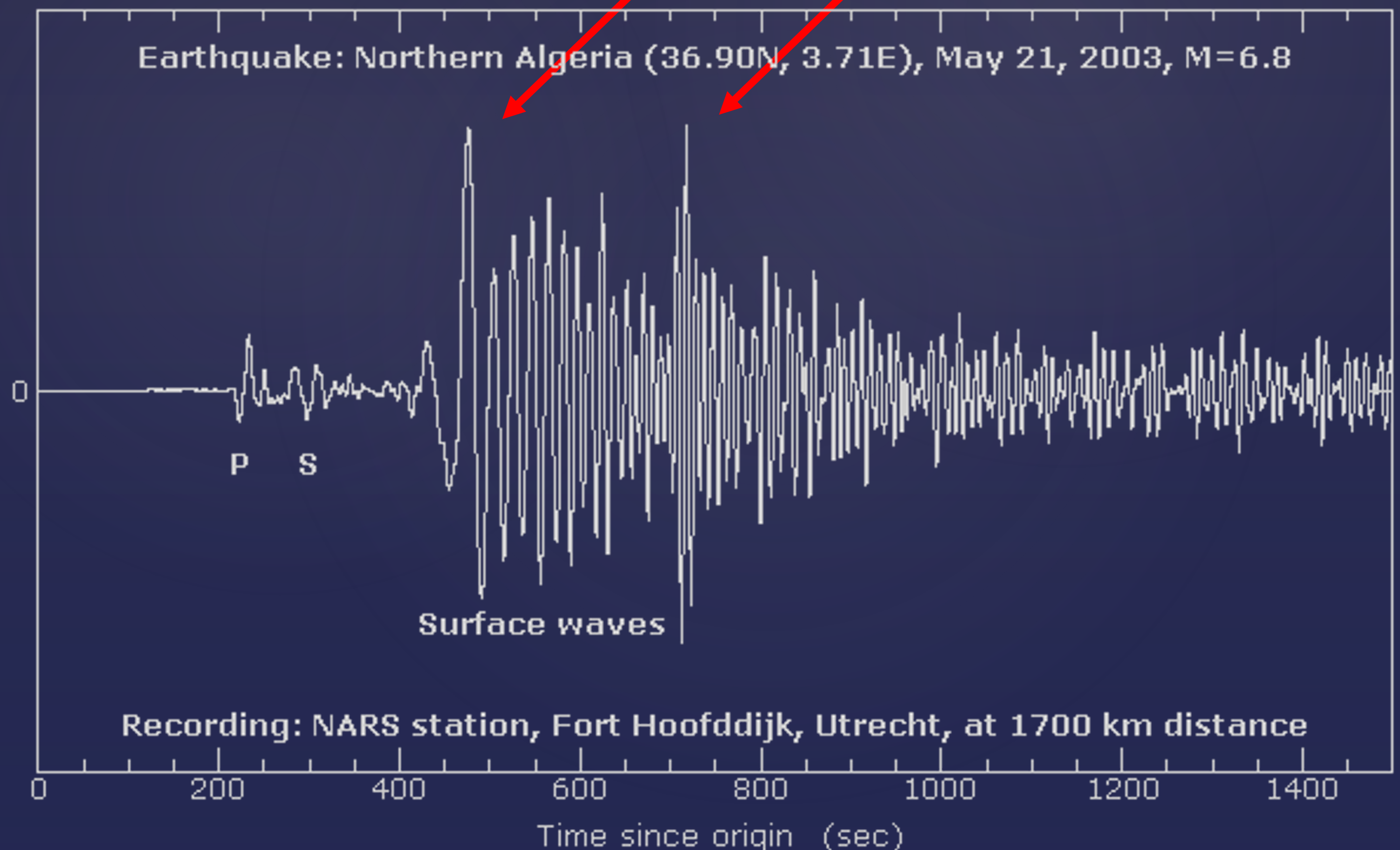
Results



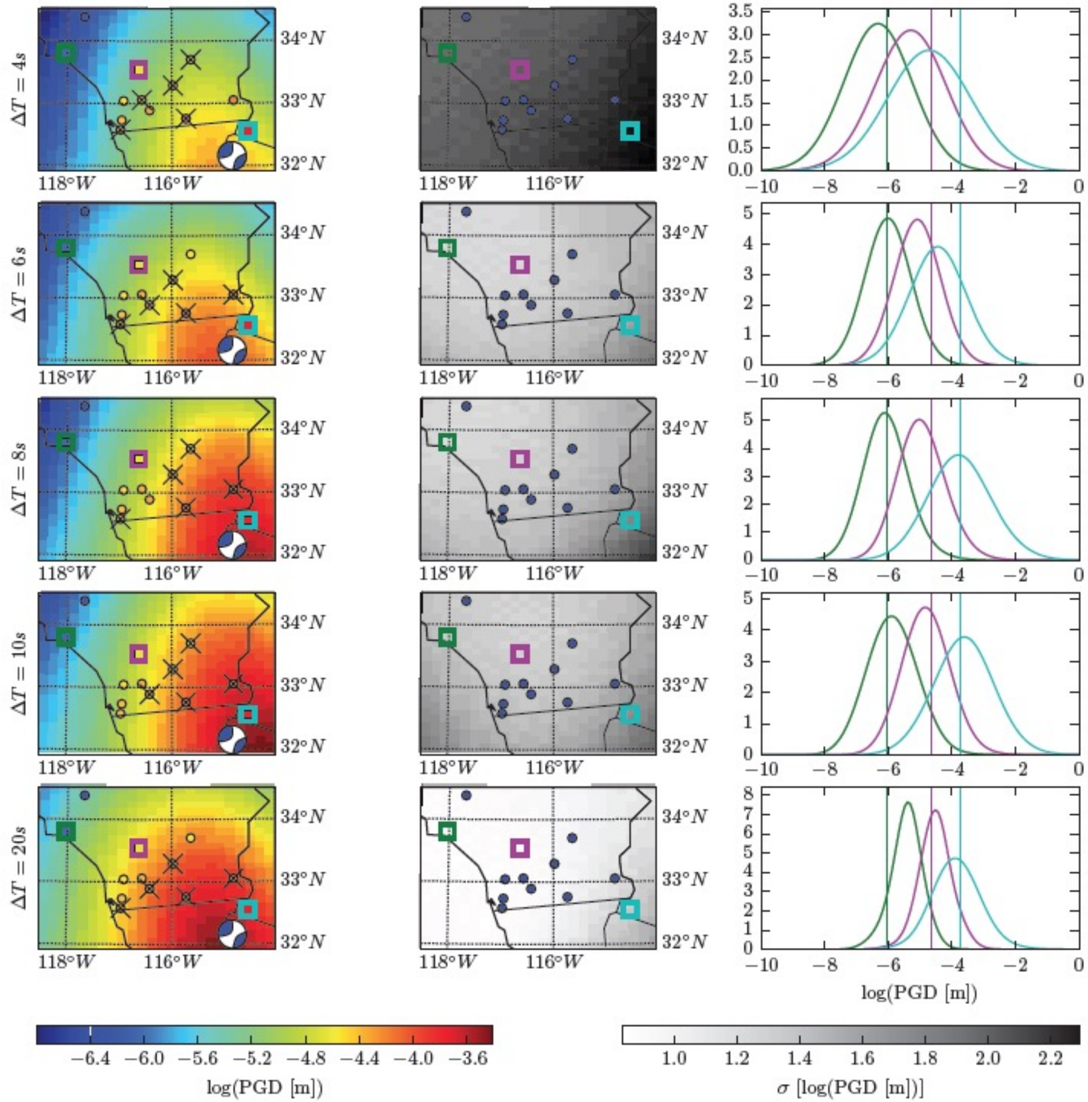
Kaeufl et al., 2016

Source physics versus peak ground displacement

Seismogram = Source * Green's function * Instrument



We can infer peak ground displacement directly



(a) Example 1

Figure 6.8: Prediction of peak ground displacement at an arbitrary location within a regional receiver network using data from five near-source receivers. **Left column**: The color scale shows the posterior mode of the PGD prediction at any point on the map given the data at the five stations marked by black crosses. The $R = 10$ receiver locations used for network training are shown by filled circles, the fill-color corresponds to the true PGD (target value) at that station. **Middle column:** Map of the uncertainty associated with the PGD predictions. **Right column:** Three posterior pdfs evaluated at the three points marked by coloured squares in the left and middle column. True PGD values (target values) are denoted by vertical lines.

Prior sampling together with neural networks are well suited for certain geophysical inference problems:

- Train on synthetic data which contain all the known physics
- All computational costs are for training
- Inference with real data is instantaneous
- Ideal for repeated inferences
- Quantitative assessment (pdfs)
- Conservative Bayesian answers
- Expert model can progressively be built-up
- Flexibility in parameter choices

A drawback is the size of the prior space, therefore this is not useful for all problems

References: 

- Kaeufl P., Valentine A.P., de Wit R.W.L., Trampert J., 2016. Solving probabilistic inverse problems rapidly with prior samples, Geophys. J. Int., 205, 1710–1728.

- Kaeufl P., Valentine A.P., Trampert J., 2016. Probabilistic point source inversion of strong-motion data in 3-D media using pattern recognition—a case study for the 2008 Mw 5.4 Chino Hills earthquake, Geophys. Res. Lett., 43, doi:10.1002/2016GL069887.