A Bayesian Approach to Earthquake Source Studies

Thesis by

Sarah Minson

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2010

(Defended June 4, 2010)

Acknowledgements

None of this would have been possible without my family: my parents Sherry and Larry Minson, my grandparents Bess and Norm Budow, and the cats Polly, Higgins, and Maynard. They are the bestest most wonderful family in the world, and I say that based purely on objective, quantifiable evidence without any bias. I really truly could not have done this work without them. So if any of this research somehow leads to a world-ending explosion, you can blame them.

Many thanks go to my thesis committee. In alphabetical order, they are as follows: Jim Beck who kindly spent an incredible amount of time teaching me about Bayesian methods. This thesis simply would not have happened without him. Tom Heaton who, in addition to being the nicest person and a great seismologist, knows all the best stories about the history of the seismology lab. (Ask him about the seismo lab in the 1970s sometime.) Don Helmberger who has more pure excitement about seismology than any three people put together and who is the king of seismology. (I have heard people say that he is the grandfather of modern seismology; but "king" seems more flattering. Of course, people also say that Hiroo Kanamori is the grandfather of modern seismology. So in that context, "grandfather" makes more sense. Most

people have two grandfathers, but having two kings tends to lead to warfare.) Hiroo Kanamori who has declined to publish more about earthquakes than the rest of us will learn in our entire lifetimes, who is the king (or grandfather) of seismology, and who always made time to answer my questions no matter how stupid or inane. Nadia Lapusta who is not only astoundingly smart (she makes everyone else at Caltech look like they are thinking with one cerebral hemisphere tied behind their back) but also has the precious ability to communicate her ideas with others. And Mark Simons who is practical. If that does not sound like much to you, then you must not have spent time with tenured faculty. Practical professors are only slightly rarer than unicorns.

Andy Kositsky is a wonderful friend who helped with typesetting the dissertation and spent many a late night strategizing with me. Nina Lin, AKA The Nicest Graduate Student In The Seismology Lab, corrected the interferometric data for tropospheric errors. Jeff Genrich and Sue Owen processed the high-rate GPS time series of the Tocopilla earthquake which were absolutely critical to this project. I also want to thank Eric Fielding for producing the wide-swath interferograms, Francisco Ortega for providing the fault geometry, and Lupei Zhu and Risheng Chu for helping with the Green's functions.

All scientific research at Caltech would undoubtedly cease to function tomorrow, and the school probably disappear as well, if it were not for the tireless work of the administrative staff. In alphabetical order, Viola Carter, Evelina Cui, Sarah Gordon, and Donna Mireles keep the seismo lab alive and well. And I must single out Rosemary Miller for taking care of me for many years now. Many thanks go to everyone in the

GPS division office, especially Dian Buchness and Marcia Hudson. And I am very grateful to Natalie Gilmore, the Assistant Dean of Graduate Studies, for shepherding me through the process of scheduling a thesis defense and submitting the dissertation.

Finally, I want to acknowledge my aunt Nancy Minson who was always rooting for me and who I miss.

Abstract

Bayesian sampling has several advantages over conventional optimization approaches to solving inverse problems. It produces the distribution of all possible models sampled proportionally to how much each model is consistent with the data and the specified prior information, and thus images the entire solution space, revealing the uncertainties and trade-offs in the model. Bayesian sampling is applicable to both linear and non-linear modeling, and the values of the model parameters being sampled can be constrained based on the physics of the process being studied and do not have to be regularized. However, these methods are computationally challenging for high-dimensional problems.

Until now the computational expense of Bayesian sampling has been too great for it to be practicable for most geophysical problems. I present a new parallel sampling algorithm called CATMIP for Cascading Adaptive Tempered Metropolis In Parallel. This technique, based on Transitional Markov chain Monte Carlo, makes it possible to sample distributions in many hundreds of dimensions, if the forward model is fast, or to sample computationally expensive forward models in smaller numbers of dimensions. The design of the algorithm is independent of the model being sampled, so CATMIP can be applied to many areas of research.

I use CATMIP to produce a finite fault source model for the 2007 Mw 7.7 Tocopilla, Chile earthquake. Surface displacements from the earthquake were recorded by six interferograms and twelve local high-rate GPS stations. Because of the wealth of near-fault data, the source process is well-constrained. I find that the near-field high-rate GPS data have significant resolving power above and beyond the slip distribution determined from static displacements. The location and magnitude of the maximum displacement are resolved. The rupture almost certainly propagated at sub-shear velocities. The full posterior distribution can be used not only to calculate source parameters but also to determine their uncertainties. So while kinematic source modeling and the estimation of source parameters is not new, with CATMIP I am able to use Bayesian sampling to determine which parts of the source process are well-constrained and which are not.

Contents

List of Figures

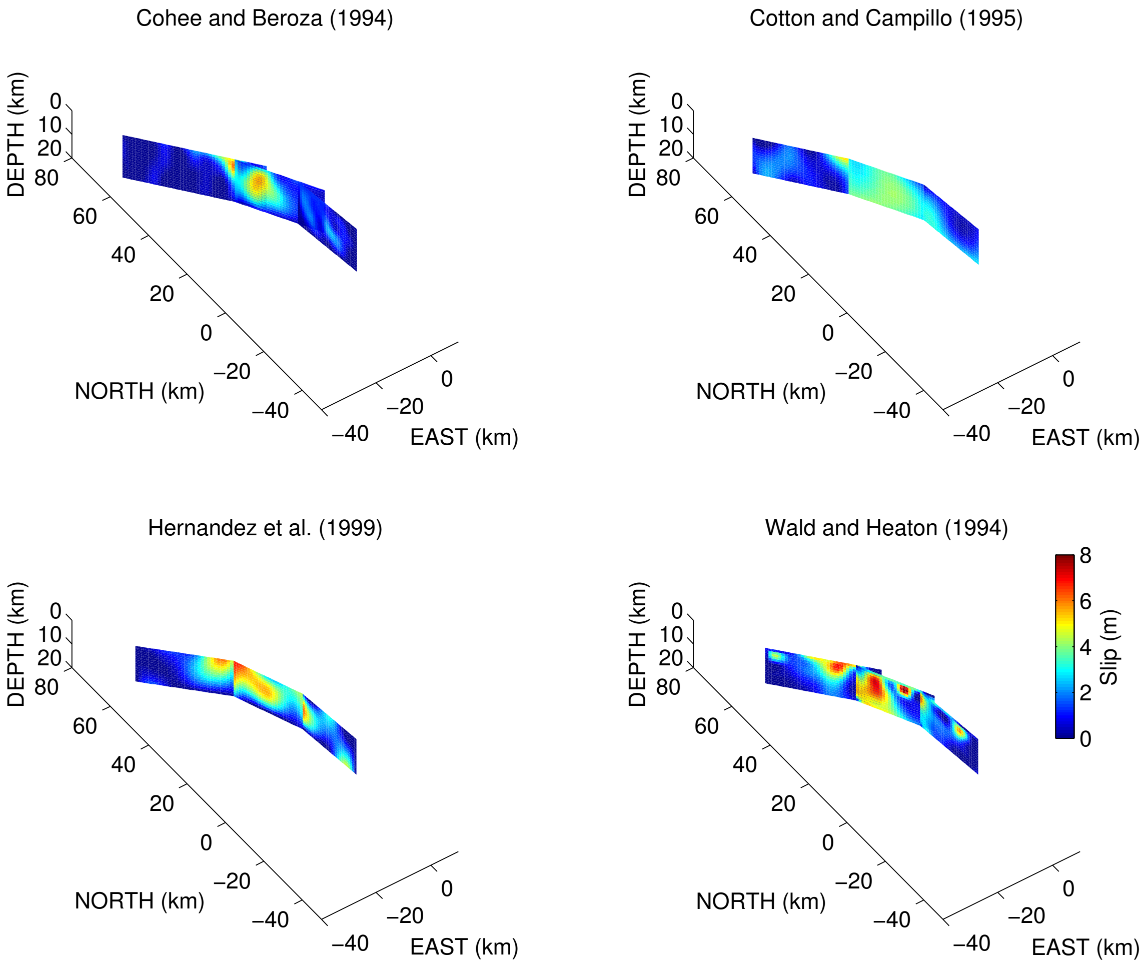
List of Tables

Chapter 1

Introduction

In order to the study the physics of earthquakes, we need observations of earthquake ruptures. But since we cannot directly observe the rupture process at depth, we must use data collected at the Earth's surface to develop models of the earthquake source which can then be used as an input for other studies. However, using limited surface observations to constrain a possibly complex and heterogeneous source process is a fundamentally under-determined inverse problem. Thus any change in the inversion methodology can lead to very different solutions, which limits our ability to infer the physics of the rupture process from a given source model (Figure 1.1).

The purpose of this work is to explain what Bayesian analysis is, to show why it is useful for solving geophysical inverse problems, to develop practical Bayesian methodologies, and to apply these results to the study of earthquake source mechanisms. In short, Bayesian analysis is important because it allows for imaging of solution spaces with multiple peaks and valleys; it works for just about any imaginable problem as it is not restricted by linearity or positivity or any other


Cohee and Beroza (1994)

Cotton and Campillo (1995)

Hernandez et al. (1999)

Wald and Heaton (1994)

Figure 1.1: *Variability of earthquake source models for the Landers earthquake*: Small differences in inversion techniques can lead to large differences in inferred earthquake slip models. As an example, consider these four published slip models for the 1992 Mw 7.3 Landers, California earthquake (*Cohee and Beroza*, 1994a; *Cotton and Campillo*, 1995; *Hernandez et al.*, 1999; *Wald and Heaton*, 1994).

common constraints; it does not require the assumption of Gaussian statistics; and it allows the inclusion of a variety of a priori information. Most importantly, it provides not one solution but an ensemble of many possible solutions along with a complete description of the model uncertainty.

Although Bayesian analysis is theoretically a panacea for most inverse modeling woes, it becomes impracticable for solving large problems with many parameters.

The number of computations required is simply too large. (In fact, throughout history, Bayesian techniques have been far more advanced than the computational power available to implement them.) Add to that the fact that the Monte Carlo samplers used in Bayesian analysis become less efficient at generating samples in high dimensions and the problem appears daunting. But I have a partial solution.

By harnessing the power of parallel computing and adding some new twists on existing sampling methods, I have built a sampling methodology capable of simulating distributions in hundreds of dimensions (for problems with fast forward models) or sampling computationally intense low-dimensional problems. Using this methodology, which I will refer to as CATMIP for Cascading Adaptive Tempered Metropolis In Parallel, full Bayesian analysis of real geophysical models with large numbers of parameters is now within our reach. The sampling methodology is independent of the data and model. So while I used this sampler to produce finite fault earthquake source models, it can be used for any kind of model.

I will begin by providing for the uninitiated some background on Bayesian techniques, including its history, mathematical underpinnings, strengths and limitations. Next the derivation of the CATMIP sampling technique is given. Finally, I will present models of the 2007 Mw 7.7 Tocopilla, Chile earthquake. This earthquake is an excellent demonstration problem for Bayesian analysis as it allows us to fuse together data from multiple sources. The Tocopilla earthquake was located under the Central Andean Tectonic Observatory (CAnTO) high-rate GPS network. These stations provide a complete near-field network of both static offsets and kinematic time series which can be used as seismograms, although these stations are located only on the hanging-wall side of the subduction interface. In addition, interferometric satellite imagery is available from different satellites and flight paths. All together, this availability of rich and diverse observations makes the Tocopilla earthquake the very model of a modern major earthquake with plentiful near-field data.

1.1 Bayesian Techniques

There is some debate as to who is the rightful founder of the probability of inference. Some argue in favor of Pierre-Simon de Laplace who independently derived many important fundamental results in 1774, and who was perhaps the first to use these techniques in a scientific context when he employed Bayesian inference for such applications as estimating the mass of the moon. As an aside, the first person to use Bayesian inference for geophysical problems is probably Sir Harold Jeffreys (e.g. *Jeffreys*, 1931, 1939). But credit for creating the theory of inverse conditional probabilities is traditionally given to the earlier work of Reverend Thomas Bayes, and his name has been attached to this field of study for posterity. Bayes never published a single paper on probability during his life, which at the very least demonstrates that members of the clergy have less pressure to publish than people in other professions. His work did not come to light until another clergyman, the preacher/philosopher/actuary Richard Price to whom Reverend Bayes left his work in his will, presented Bayes' results to the Royal Society at the end of 1763 (*Stigler*, 1990).

Bayes was interested in the relationship between inverse conditional probabilities. If you know the probability of event A if event B happens, what is the probability that event B will happen if you observe event A? Or similarly, how does the probability of a hypothesis, given some evidence, relate to the probability of that evidence, given a certain hypothesis? If that seems opaque, let us consider the specific case of data modeling. When we fit a model to data, we are essentially asking the question: Which set of model parameters is most likely to have produced the observed data? (In probability speak this is $\mathrm{P}(\theta|D)$, the probability of model θ conditional on data D.) We do not have a direct answer to that question. But we can evaluate the forward model. Although not normally looked at in this way, a forward model is equivalent to the probability of the observed data given a model. This is written as $\mathrm{P}(D|\theta)$, the conditional probability of D given θ. If a model produces a set of predicted data that are nothing like the observations, then it is highly unlikely that the proposed model accurately describes the source of those observations.

But what is the relationship between one conditional probability and its inverse? The answer is given by Bayes' theorem.

1.1.1 Bayes' Theorem

Bayes' theorem is simple to derive for discrete probabilities. The probabilities of events A and B each occurring are $\mathrm{P}(A)$ and $\mathrm{P}(B)$. We define the conditional probability that A will occur if B occurs as $\mathrm{P}(A|B)$. Let the probability of both events occurring be $\mathrm{P}(A \cap B)$. This is the joint probability of A and B. The chances of both events occurring is the product of the probability of one of the events occurring and the probability of the other event occurring given that the first event happened. In other words, $\mathrm{P}(A \cap B) = \mathrm{P}(A|B)\,\mathrm{P}(B) = \mathrm{P}(B|A)\,\mathrm{P}(A)$. Thus the relationship between the two conditional probabilities is

$$\mathrm{P}(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.1}$$

We can derive a similar relationship for the distributions of continuous variables X and Y,

$$P(x|Y = y) = \frac{P(y|X = x)P(x)}{\int P(y|X = x)P(x)dx} \tag{1.2}$$

The denominator is known as the evidence. Note that generally it is not necessary and in fact often impossible to compute the integral in the denominator. Instead, we use

$$P(x|y) \propto P(y|x)P(x) \tag{1.3}$$

For data modeling, this result looks like

$$\mathrm{P}(\theta|D) \propto \mathrm{P}(D|\theta)\,\mathrm{P}(\theta) \tag{1.4}$$

This equation says that the probability density function (PDF) describing the plausibility of all models, given our observations, is proportional to both the likelihood of these data, given those models, and the probability of the models themselves. The probability distribution of the models, P(θ), is the a priori or prior probability distribution. The left-hand side is the a posteriori or posterior distribution. P($D|\theta$) is often called the likelihood.

Notice that the Bayesian model is completely generic. The data, the model, and the form of the probability distributions are not stated. But it is this generality that allows for greater specificity. There are no simplifying assumptions. The model can be linear or nonlinear, smooth, rough, positive, negative, or anything imaginable. You can apply whatever prior information exists about the physics of the problem to create a fantastical a priori distribution. You can solve the problem as you believe the model should be expressed, and not as a simplified mathematical analog.

1.1.1.1 Likelihood

We know that P($D|\theta$) is the output of our forward modeling. One way to write this probability is using the L_2 norm

$$P(D|\theta) \propto e^{-\frac{1}{2}[d-g(\theta)]^T * C_d^{-1} * [d-g(\theta)]} \tag{1.5}$$

for forward model $g(\theta)$, and data d which has covariance C_d. This is the probability distribution for Gaussian prediction errors or the maximum information entropy

PDF[1].

Notice that the likelihood depends on the data errors. As with traditional optimization approaches, if we under-estimate our errors, we will over-fit the data (produce a posterior distribution which is too tightly peaked); and if we over-estimate our errors, we will under-fit the data (produce a posterior distribution which is too broad). In practice, the data errors may be dwarfed by the errors produced by the differences between our model and the physics of the real source process, which is sometimes termed "unmodeled dynamics." For finite fault earthquake source processes, possible error sources of this type include having the wrong source geometry, poorly located hypocenters, an incorrect elastic structure, and simply parameterizing the earthquake source evolution in a way that is not amenable to representing the "true" source process. However, it should be noted that Bayesian sampling can be used to estimate the prediction error in addition to the model being fit to the data. This is accomplished by casting the error produced by any model as a combination of a deterministic error, i.e., the error between the data and the results of the forward model, with a non-deterministic error produced by a random process. *Beck and Katafygiotis* (1998) lay out a Bayesian technique in which possible error sources are described by a probability distribution, and this prediction error is updated through sampling of the posterior distribution. This is an important but broad subject which is left to the reader to learn more about.

---

[1] See *Jaynes* (1957a) and *Jaynes* (1957b) for more on the concept of maximum information entropy.

1.1.1.2 Prior Distribution

What exactly is the prior distribution? The prior distribution represents our a priori knowledge of what the model parameters could be. Bayesian modeling requires the researcher to evaluate the model (and possibly the data) and state which possible model values are considered probable. Some feel that Bayesian inference thus introduces a bias into the problem. Many argue otherwise (e.g. *Cox*, 1961; *Jaynes*, 2003; *Loredo*, 1990; *Tarantola*, 2005). I am firmly in the latter camp.

First, it is never wrong to include information in your modeling. If you know something about what the model looks like based on some a priori information, it should be included in the modeling process. This becomes apparent if, in the interest of sticking to Gaussian statistics, you allow your optimization or sampling scheme to produce models which violate the laws of physics. Second, researchers always use their knowledge to discriminate between models. It is the rare scientist who publishes every result they ever obtained, even the ones that are clearly wrong, the ones that imply that gravity has vanished and pigs are flying.

What are we doing when we dislike a model? We are deciding that, based on our a priori knowledge, we find it highly unlikely that gravity has vanished or pigs are flying or whatever it is that seems wrong with the model. If we know that gravity exists, our model should know that too. And when we make that statement explicitly as part of the Bayesian modeling process, we let everyone know what our assumptions are.

1.1.2 Advantages of Bayesian Techniques

For clarity, when I discuss Bayesian analysis, I refer to methods which produce the complete posterior distribution $P(\theta|D)$. The literature is full of models which are described as Bayesian but which were derived via optimization. What makes them "Bayesian" in the eyes of these studies' authors is that the choice of regularization in the optimization scheme was based on some rule derived from Bayes' Theorem such as the Bayesian Information Criterion (BIC), also called the Schwarz Criterion (*Schwarz*, 1978), or the Akaike Information Criterion (AIC) (*Akaike*, 1974). These methods have the features and drawbacks common to all optimization techniques and should not be confused with the Bayesian techniques discussed here.

There are numerous advantages to Bayesian sampling over traditional optimization methods. The first, and most significant, is that it is a sampling method and not an optimization method. Unless the problem you are studying is linear, the model with smallest misfit to the data is just one model in a possibly complex solution space with multiple minima and complicated topography. So when you use an optimization scheme and produce the model with minimum misfit, assuming of course that your optimization scheme was successful in finding the global minimum and not some local minimum, what exactly have you learned? Is this the only model that best fits the data? If there are other models that fit the data equally or almost as well, are they similar to this one or are there other minima in a completely different part of the solution space? These limitations cannot generally be overcome for optimization problems. But with Bayesian sampling, we can produce models from everywhere in the solution space, with density proportional to their probability. That is, we naturally produce more models in regions which fit the data better and fewer in regions with lower probability. We can then analyze these models however we want. We can plot histograms of the distribution to image the topography of the complete solution space including the locations and sizes of its minima. We can use the distribution of samples to calculate errors on the model parameters. In Bayesian analysis, data-fitting is just the beginning and not the end.

There are advantages to Bayesian analysis even in simple cases. Linear problems which obey Gaussian statistics lend themselves to least squares analysis only if the constraints on the model parameters are no more complicated than things such as Laplacian smoothing and non-negativity constraints. Many geophysical model designs yield problems that are only invertible with appropriate regularization, and our choices of regularization are limited by what can be implemented via linear algebra. For an under-determined problem, we can follow Occam's razor and decide that we want to produce the simplest solution. But "simple" can only be defined linearly in least squares problems. Often "simple" solutions are generated by using Laplacian smoothing to produce smooth models. But a smoothed least squares inversion will never produce a point source. Is a point source solution not simple? It is conceptionally simple; it is spatially simple; but it is not simple according to our formulation of the problem. Least squares techniques do not lend themselves to producing sparse solutions. But Bayesian methods can implement the prior distribution of your choice.

One limitation shared by both Bayesian and traditional optimization methods is the model itself. To again use earthquake modeling as an example, any finite fault source assumes a fault plane geometry and earth structure. However we design our problem, there will be differences between our assumptions and the true physics of the problem, and these errors will be mapped into the models that we produce. This mapping is inescapable. In some ways this limitation underscores just how uninformative it is to calculate one optimal model from one error-ridden model design. The Bayesian approach is just as sensitive to design errors; but it yields the complete solution for a given "model class," as the model design is sometimes termed in Bayesian methods (see *Beck and Yuen*, 2004). Plus, at least in theory, techniques such as model class selection can be used to evaluate the model design itself, although these methods are often computationally intractable for geophysical problems.

Unlike traditional optimization methods, trade-offs between model parameters are not a bad thing in Bayesian sampling. Consider a case in which you are trying to fit two parameters to some data. Assume the data perfectly constrains the sum of the parameters, but the parameters are also perfectly anti-correlated and any values for the parameters are equally acceptable to the data so long as they have the required sum. Optimization methods cannot handle cases like this, generally speaking. (A singular value decomposition would be useful in this example for a linear model.) What is the optimal model when there are infinite models which fit the data? In fact, there is no optimal model. But when people model these kinds of problems they produce solutions through optimization by using some constraint to produce the smoothest model or the smallest model or something of that ilk. Bayesian sampling naturally produces the complete solution in the form of an anti-correlated distribution of samples. You immediately learn that your parameters have large covariance, a well-defined solution in the form of the sum of the model parameters, and zero resolution on the value of each independent parameter. That is the full answer. It is the correct answer.

1.2 A Brief History of Bayesian Samplers

Now that we have derived Bayes' Theorem and demonstrated why producing the full posterior probability for a problem is superior to straight optimization, we are left with one question. How do we produce the posterior distribution? If our problem had an analytical solution, which is extremely rare, our work would be done. We could simply write down the answer. Instead we will have to draw samples from the posterior probability density function (PDF). Sampling from a PDF is different from a grid search. In a grid search, models are produced from all regions of the model space. But here we need samples distributed such that their density is proportional to the posterior PDF. This means we want more samples in regions with higher probability and fewer samples in regions with low probability.

There are algorithms for directly producing random samples from certain simple PDFs. For example, the Box-Muller transformation can be used to produce random samples from a standard normal distribution. But our goal is to produce samples from an unknown target PDF. This may sound daunting, but it is in fact quite doable.

Most sampling techniques boil down to using Monte Carlo simulation to produce random samples from a known distribution (our proposal PDF), re-weighting those samples according to their probability in the target distribution, and then probabilistically choosing whether to accept the proposed sample. Eventually, and possibly with great inefficiency, you will produce samples which are distributed not according to your proposal PDF but instead distributed accordingly to your target PDF. But this process requires a lot of computational effort.

The computational expense is why the adoption of Bayesian methods has been so slow. While the theory has existed at least since Bayes' 1763 posthumous essay on probability (*Bayes*, 1763), it was not of practical use to most eighteenth century researchers. The first major applications of Monte Carlo simulation is probably the work of the Metropolis group in the twentieth century. They were working on the Manhattan project and had access to the best computers of the day. Even today, the limits of computational power have made most geophysical problems intractable using conventional sampling techniques.

Before I explain how CATMIP makes large problems tractable, I first present a primer on some existing sampling techniques. This guide is in no way intended to be a complete review of the field. It exists merely to introduce the reader to some samplers they may encounter in their own work and how the principles behind these algorithms

Table 1.1: *Rejection method.*

1. Draw a candidate sample y from known distribution g.
2. Generate a sample u from $U(0,1)$, the uniform distribution on (0,1).
3. If $u \leq \frac{f(y)}{k \cdot g(y)}$, accept y as a sample of f.
4. Return to step 1.

contribute to CATMIP.

1.2.1 Rejection Method

This technique, also known as the acceptance-rejection method, is perhaps the simplest Monte Carlo sampling technique and dates back to *von Neumann* (1951). It is also very inefficient for most practical applications. But its simplicity makes it a good introduction to the world of sampling.

As with all sampling methods, our goal is to draw samples from a target PDF $f(x)$. In the rejection method, we begin by producing samples from an easily-simulated, known distribution $g(x)$. Assume that there is a known constant k such that $f(x) \leq k \cdot g(x)$ for all x. Then we can draw samples from $f(x)$ by following the procedure in Table 1.1.

You can see intuitively how this works. We produce random samples from a known distribution and compare a normalized version of the sample's probability in the target distribution to a random variate. If the sample has a large probability in the target distribution, we are more likely to accept it. If the target PDF assigns a low probability to the sample, we are less likely to accept it. And thus we end up with a set of samples distributed proportionally to our target PDF.

There are several good features to the rejection method. From a theoretical standpoint, it produces completely independent samples, unlike the Markov Chain Monte Carlo (MCMC) methods I will present next. From a practical standpoint, it is completely parallel: the candidate samples are generated completely independently of each other making it easy to distribute sampling over a large number of computer processors. However, it is also extremely inefficient. Unless $k \cdot g(x)$ is very close to $f(x)$, the rejection rate will be large. In practice, we often have no a priori knowledge of $f(x)$ and thus cannot tune the algorithm appropriately. Also, the efficiency of the rejection method does not improve as the algorithm runs. After producing a million samples, you may have a very good idea as to what the posterior PDF looks like. But the rejection method is no more likely to produce a usable sample after the millionth iteration than it was after the first.

1.2.2 Metropolis Algorithm

The Metropolis algorithm (*Metropolis et al.*, 1953) is perhaps the most fundamental MCMC sampler; countless other samplers are simply variations on it. While its method and implementation are straightforward, the mathematical proofs behind the Metropolis algorithm in particular, and MCMC sampling in general, are surprisingly obscure. For an accessible introduction to the Metropolis algorithm, I recommend

Chib and Greenberg (1995). For more information about MCMC sampling, see *Robert and Casella* (2004) and *Liu* (2008).

The Metropolis algorithm takes the rejection method and places it in the context of a Markov process. As in the rejection method, we generate samples from a proposal distribution and then probabilistically decide whether to accept each candidate sample. Our proposal density (the PDF we use to generate candidate samples) depends on the current state of the process. Where, in the rejection method, we generated samples from the state-free distribution $g(x)$, our proposal PDF will now be $q(x, y)$ where x is the current sample and y is the candidate sample.

There are few constraints on the form of the proposal PDF used in the Metropolis algorithm other than that $q(x, y) = q(y, x)$, i.e., the probability of moving from x to y is the same as transitioning from y to x. But even the symmetry requirement has been relaxed in later samplers such as the Metropolis-Hasting algorithm (*Hastings*, 1970). *Metropolis et al.* (1953) used $q(x, y) = q'(y - x)$. Equivalently, we could say that $y = x + z$ where $z \sim q'(y - x)$. This produces a random walk through model space. While any distribution of the form $q'(y-x)$ can be used to generate candidates, the most popular choice is the multivariate normal distribution.

Using our proposal PDF and an arbitrary initial sample x_0, we can then generate samples following the procedure in Table 1.2. If the candidate sample has higher target probability than the current sample, we accept it. If it has lower probability, we still might accept it depending on the value of our random variate u. So while

Table 1.2: *Metropolis algorithm.*

- For $i = 1, 2, ..., N$

1. Draw $z \sim q'$ and compute a candidate sample $y = x_{i-1} + z$.
2. Generate a sample u from $U(0, 1)$, the uniform distribution on (0,1).
3. Compute $\alpha(x, y) = \min\{\frac{f(y)}{f(x_{i-1})}, 1\}$.
4. If $u \leq \alpha$, $x_i = y$. Otherwise $x_i = x_{i-1}$.

better models are always accepted, worse models may be accepted too, allowing the random walk to move between the peaks of a multimodal posterior PDF.

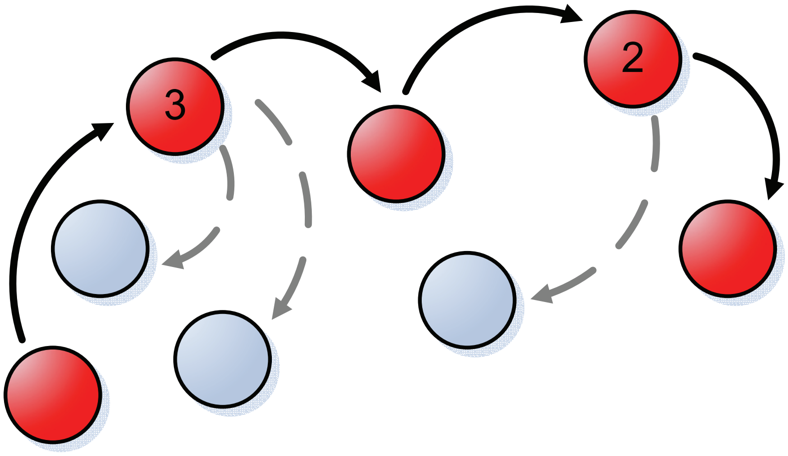


Figure 1.2: *Metropolis algorithm schematic*: This cartoon illustrates how the Metropolis algorithm produces samples through a random walk. Rejected samples are shown in gray. When a candidate sample is rejected, the current sample in the Markov chain is duplicated, and numbers indicate the frequency of duplicated models. Compare to Figure 1.3 and Figure 2.1.

This behavior illustrates the difference between probabilistic sampling, grid searches, and traditional optimization methods. On one extreme, we have gradient methods which will always transition to a better model and never transition to a worse model. They will converge to the nearest minimum even if it is not the global minimum. MCMC samplers transition to worse models with probability proportional to the candidate sample's likelihood. Even if the Markov chain is engulfed in a local maximum, the sampler will eventually accept a poor model outside of that maximum, and from there travel in a random walk to any maximum. (Note that while optimization attempts to minimize the misfit to the data given some set of constraints, this translates to trying the maximize the posterior probability in Bayesian analysis.)

On the other extreme, we have grid search methods which sample equally from all parts of the model space and waste much computational effort in areas which do not produce usable models. But MCMC samplers produce samples with density proportional to their probability. If a region of model space produces models that are twice as probable as those of another region, then the sampler will spend twice as much time in that region.

In comparison to the rejection method, there are several advantages to the Metropolis algorithm. We have eliminated any need for the typically unknown normalization factor in Bayes' Theorem. More importantly, we now have the freedom to choose a tailor-made proposal PDF for each candidate sample. However, we are not guaranteed efficient sampling. If $q(x, y)$ is not properly chosen, the rejection rate could be large. Conversely, if the acceptance rate is too high, successive samples will be highly correlated with each other. Inter-sample correlation is one of two main theoretical disadvantages to MCMC methods. The other is that MCMC methods only converge to the target distribution after some number of initial samples, but there is no way to know how many samples are needed to satisfy the "burn-in period."

As with the rejection method, the Metropolis algorithm does not gain in efficiency as it runs. Plus, we have exchanged the perfect parallelism of the rejection method for an entirely serial Markov chain. Several parallel adaptations of the Metropolis algorithm have been devised which involve mixing multiple Markov chains together, not to mention the many algorithms that use the Metropolis algorithm in conjunction with other bells and whistles such as tempering or annealing. It is one of these that we explore next.

1.2.3 Transitional Markov Chain Monte Carlo

The transitional Markov chain Monte Carlo (TMCMC) method of *Ching and Chen* (2007) offers several improvements over traditional MCMC techniques. To motivate the algorithm, let us consider a series of intermediate PDFs, $f(\theta|D)_m$, which start from our a priori distribution $\mathrm{P}(\theta)$ and converge to our target distribution $\mathrm{P}(\theta|D)$:

$$\begin{aligned} f(\theta|D)_m &\propto \mathrm{P}(\theta)\,\mathrm{P}(D|\theta)^{\beta_m} \\ & m = 0, 1, ..., M \\ & 0 = \beta_0 < \beta_1 < \beta_2 < ... < \beta_M = 1 \end{aligned} \tag{1.6}$$

If $\beta_m = 0$, then $f(\theta|D)_m \propto \mathrm{P}(\theta)$, and we can directly simulate $f(\theta|D)_m$ by drawing samples of our prior. If $\beta_m = 1$, then $f(\theta|D)_m$ is our Bayesian posterior.

For each stage m, we can generate samples of f_{m+1} by resampling f_m. The probability of choosing a specific sample of f_m, $\theta_{m,k}$, is proportional to

$$w(\theta_{m,k}) = \frac{\mathrm{P}(\theta_{m,k})\,\mathrm{P}(D|\theta_{m,k})^{\beta_{m+1}}}{\mathrm{P}(\theta_{m,k})\,\mathrm{P}(D|\theta_{m,k})^{\beta_m}} = \mathrm{P}(D|\theta_{m,k})^{\beta_{m+1}-\beta_m} \tag{1.7}$$

After resampling f_m, we have fewer unique samples of f_{m+1} than we had of f_m. But we can use MCMC sampling to replace our duplicate models with new samples. Perhaps the single most important innovation in the TMCMC algorithm is the use of an adaptive proposal density in the Metropolis algorithm. Specifically, *Ching and Chen* (2007) used a Gaussian proposal density with covariance matrix

$$\begin{aligned} \Sigma_m &= c^2 \sum_{i=1}^{N} \frac{w(\theta_i)}{\sum_{i=1}^{N} w(\theta_i)} (\theta_i - \bar{\theta})(\theta_i - \bar{\theta})^T \\ \bar{\theta} &= \sum_{i=1}^{N} w(\theta_i)\theta_i / \sum_{i=1}^{N} w(\theta_i) \end{aligned} \tag{1.8}$$

A little discussion of the preceding is required. I have normalized w in Equation 1.8. Due to a typographical error, it appears unnormalized in both *Ching and Chen* (2007) and *Muto and Beck* (2008). c is an arbitrary factor which *Ching and Chen* (2007) set to 0.2. Also somewhat arbitrary is the cooling schedule. *Ching and Chen* (2007) picked β_{m+1} such that $COV(w) = 1$ where COV denotes the coefficient of variation and is the ratio of the standard deviation of w to the mean of w. Equivalently, *Cheung* (2009) suggested that β_{m+1} should be chosen so that the number of unique samples after resampling, the effective sample size (ESS), was $\frac{1}{2}N$. The TMCMC algorithm

Table 1.3: *TMCMC algorithm.*

1. Set $m = 0$. Generate N samples $\theta_{0,1} ... \theta_{0,N}$ of the prior PDF $f_0 = \mathrm{P}(\theta)$.
2. Set $m = m + 1$. Choose β_m such that the $COV[w] = 1$.
3. Calculate Σ_m.
4. For $k = 1...N$, draw a sample $\bar{\theta}^{(m-1)}$ from $\theta_{m-1,1} ... \theta_{m-1,N}$ with probability $w(\theta_{m-1,k})$. Use the Metropolis algorithm with $N(0, \Sigma_m)$ as the proposal PDF to append a new sample to the chain of models which has $\theta_{m-1,k}$ as its starting model.
5. Repeat steps 2 to 4 until $\beta_M = 1$ is reached.

is summarized in Table 1.3.

There are two significant advantages to this algorithm. First, we are always sampling from a distribution which is approximately the equilibrium distribution, keeping our acceptance rate tolerable even in high dimensions. Second, we calculate the proposal PDF from our current best estimate of the posterior distribution. From an algorithmic standpoint, this also helps to keep our acceptance rate reasonably high. But from a modeling standpoint, we are now freed from many constraints on model parameterization. Volumes have been written on choosing the correct parameterization and coordinate systems to avoid trade-offs and covariance. However, the TMCMC algorithm uses a proposal PDF based on the current best estimate of the model covariance. Thus if two parameters trade off with each other, the sampler automatically takes large steps in the direction in which they co-vary and small steps in directions with low covariance. Of course, this technique is most efficient when the model covariance is approximately Gaussian, or at least linear. In this way, we also finally have achieved a sampler which "learns" from previous samples. At each cooling step, the algorithm uses information from the previous stage to optimize sampling.

There is still one severe limitation to the TMCMC algorithm. It is not parallel. For high-dimensional problems, which require enormous numbers of samples, stochastic simulation is only tractable if done in parallel.

Finally, there is a theoretical problem underlying the TMCMC algorithm. *Ching and Chen* (2007) assume stationarity for all $f(\theta|D)_m$, and thus ignore the "burn-in" period of MCMC sampling. But this is a false assumption. At any cooling step, we have at best reached stationarity with the target distribution, $f(\theta|D)_m$. We then resample with probability w, yielding a set of samples which is approximately, but not exactly, distributed according to $f(\theta|D)_{m+1}$, and thus we are not at stationarity with our new target distribution. Therefore, there will be a burn-in period which is ignored in the TMCMC algorithm.

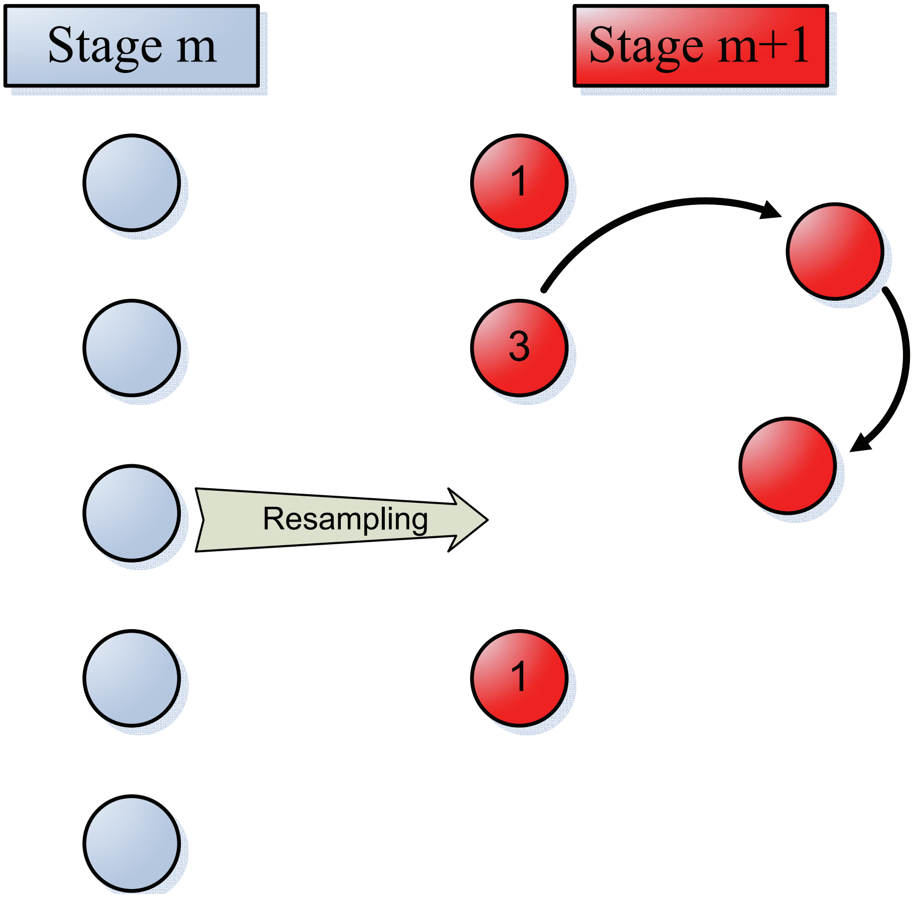


Figure 1.3: *TMCMC algorithm schematic*: This cartoon illustrates one complete cooling stage of the TMCMC algorithm. The five samples from β_m are resampled and then the Metropolis algorithm is run to replace the unique samples lost through resampling. Numbers indicate the frequency of each model after resampling. The five red samples comprise the posterior distribution for β_{m+1}. The algorithm is plotted with a 100% acceptance rate for simplicity. Compare to Figure 1.2 and Figure 2.1.

Chapter 2

Cascading Adaptive Tempered Metropolis In Parallel: CATMIP

I now introduce a new solution for efficient sampling of high-dimensional models, which I call CATMIP for Cascading Adaptive Tempered Metropolis In Parallel. It is not intended to be the last word in MCMC simulation. My goal is simply to produce a useful methodology for Bayesian analysis of finite fault earthquake source models. But as a byproduct, this technique can be used for many geophysical problems.

I begin with the combination of tempering and resampling used in TMCMC. I then introduce a different Metropolis sampling scheme and parallelize the algorithm. The additional sampling increases CATMIP's exploration of the model space. With sufficiently long Markov chains, our final distribution will come from samples taken after the burn-in period, unlike TMCMC. Finally, I develop a method for integrating different data sets and parameter spaces. I refer to this process as cascading. Cascading increases the efficiency of sampling when the model fuses together multiple data sets that are dependent on different model parameters, especially when combining multiple forward models of different computational costs. For kinematic earthquake source models, we will encounter both of these situations. The static data depend on the final slip distribution only, while kinematic data contain information about the time history of the source process. Furthermore, the static forward model is much faster and simpler than the full kinematic forward model. This will be discussed in more detail in Section 2.2 and Section 3.2.

2.1 Parallel Adaptive Tempered Metropolis Sampling

Following *Ching and Chen* (2007), I sample from a series of intermediate distributions:

$$\begin{aligned} f(\theta|D)_m \quad &\propto \quad \mathrm{P}(\theta)\,\mathrm{P}(D|\theta)^{\beta_m} \\ &m = 0, 1, ..., M \\ &0 = \beta_0 < \beta_1 < \beta_2 < ... < \beta_M = 1 \end{aligned} \tag{2.1}$$

For each stage m, I calculate a new "temperature" β_{m+1}. Any annealing schedule can be employed. I use the criterion of *Ching and Chen* (2007) that $COV(w_m) = 1$ where w_m is a vector of weights given by

$$w_m = \frac{\mathrm{P}(\theta_m)\,\mathrm{P}(D|\theta_m)^{\beta_{m+1}}}{\mathrm{P}(\theta_m)\,\mathrm{P}(D|\theta_m)^{\beta_m}} = \mathrm{P}(D|\theta_m)^{\beta_{m+1}-\beta_m} \tag{2.2}$$

I then resample the current set of models $\{\theta_m\}$ according to the probability distribution w_m to produce $\{\hat{\theta}_m\}$. Note that $\{\hat{\theta}_m\}$ will likely contain duplicate samples. Each sample $\hat{\theta}_m$ of $\{\hat{\theta}_m\}$ is then used as the seed for one of N instances of the Metropolis algorithm each of which has a chain length of N_{steps}. These samplers are run in parallel, and the final model from each Markov chain is collected to make $\{\theta_{m+1}\}$. The proposal density for the Metropolis sampler is $q(x,y) = N(x, \Sigma_m)$ where

$$\begin{aligned} \Sigma_m &= c_m^2 \sum_{i=1}^{N} \frac{w(\theta_i)}{\sum_{i=1}^{N} w(\theta_i)} (\theta_i - \bar{\theta})(\theta_i - \bar{\theta})^T \\ \bar{\theta} &= \sum_{i=1}^{N} w(\theta_i)\theta_i / \sum_{i=1}^{N} w(\theta_i) \end{aligned} \tag{2.3}$$

$c_m = a + bR$ where R is the acceptance rate of the Metropolis sampling and a and b are arbitrary constants (Matthew Muto, personal communication). Thus we rescale our proposal density by the acceptance rate of our sampler. When the acceptance rate is higher, we increase the size of our random walk steps, allowing greater exploration of the model space. When our acceptance rate decreases, we take smaller steps to increase the chances that a candidate model will be accepted. For the performance tests and earthquake modeling presented here, I use $a = \frac{1}{9}$ and $b = \frac{8}{9}$.

In the original TMCMC algorithm, the purpose of the Monte Carlo sampling was simply to “jitter” the samples so that we did not have repeated models, and the length of each random walk chain was proportional to the probability of its seed model. This results in very little exploration of the model space. TMCMC also keeps all of the samples including the ones from the burn-in period. I have replaced

Table 2.1: *CATMIP algorithm.*

1. Set $m = 0$. Generate N samples $\{\theta_0\} = \{\theta_1 ... \theta_N\}$ of the prior PDF $f_0 = \mathrm{P}(\theta)$.
2. Set $m = m + 1$. Choose β_m such that the $COV[w]$ equals some target value.
3. Calculate Σ_m and c_m.
4. Draw N samples from $\{\theta_{m-1}\}$ with probability w. The set of resampled models is $\{\hat{\theta}_{m-1}\}$.
5. Use each resampled model $\hat{\theta}_{m-1}$ in $\{\hat{\theta}_{m-1}\}$ as the seed for generating N_{steps} models from the Metropolis algorithm with proposal density Σ_m.
6. $\{\theta_m\}$ is comprised of the final model from each Markov chain. Thus the total number of samples is unchanged.
7. Repeat steps 2 to 6 until $\beta_M = 1$ is reached.

this part of the algorithm with independent runs of the Metropolis algorithm using Markov chains of length N_{steps}. In TMCMC, if after resampling a particular model has been chosen five times, those five copies of the model are replaced with the five models generated from running the Metropolis algorithm for five steps. In my version, the five models are used as the seeds for five independent instances of the Metropolis algorithm where each Markov chain is N_{steps} long. The five seed models are then replaced with the final model from each Markov chain. The expectation is that, with these changes, we have improved CATMIP's chance of visiting new parts of the model space and decreased our inter-sample correlation while eliminating the effects of the burn-in period. CATMIP's more robust sampling comes at the cost of increasing the number of model evaluations made over the lifetime of the algorithm by approximately a factor of N_{steps}. But since each Markov chain is run in parallel, this algorithm is extremely efficient when distributed over many computer processors.

The basic CATMIP algorithm is given in Table 2.1 and illustrated in Figure 2.1. The behavior of the CATMIP algorithm while sampling a biased mixture of two-dimensional Gaussians is shown in Figure 2.2.

2.2 Cascading

To handle even larger sample spaces, I use an approach I call “cascading”, which allows us to model a subset of the data and model parameters and then apply that posterior PDF to the full problem. (This technique should not be confused with the completely unrelated but similarly named approach of *Tarantola* (2005).) Consider a case in which we have two data sets D_1 and D_2. D_1 informs only model parameters θ_1 while D_2 informs both θ_1 and additional model parameters θ_2. We can write our a posteriori distribution as

$$\begin{aligned} \mathrm{P}(\theta|D) &\propto \mathrm{P}(\theta)\,\mathrm{P}(D|\theta) \\ &\propto \mathrm{P}(\theta_1)\,\mathrm{P}(\theta_2)\,\mathrm{P}(D_1|\theta_1)\,\mathrm{P}(D_2|\theta_1,\theta_2) \\ &\propto [\mathrm{P}(\theta_1)\,\mathrm{P}(D_1|\theta_1)]\,\mathrm{P}(\theta_2)\,\mathrm{P}(D_2|\theta_1,\theta_2) \\ &\propto \mathrm{P}(\theta_1|D_1)\,\mathrm{P}(\theta_2)\,\mathrm{P}(D_2|\theta_1,\theta_2) \qquad (2.4) \end{aligned}$$

Thus the full a posteriori distribution is proportional to the posterior distribution of

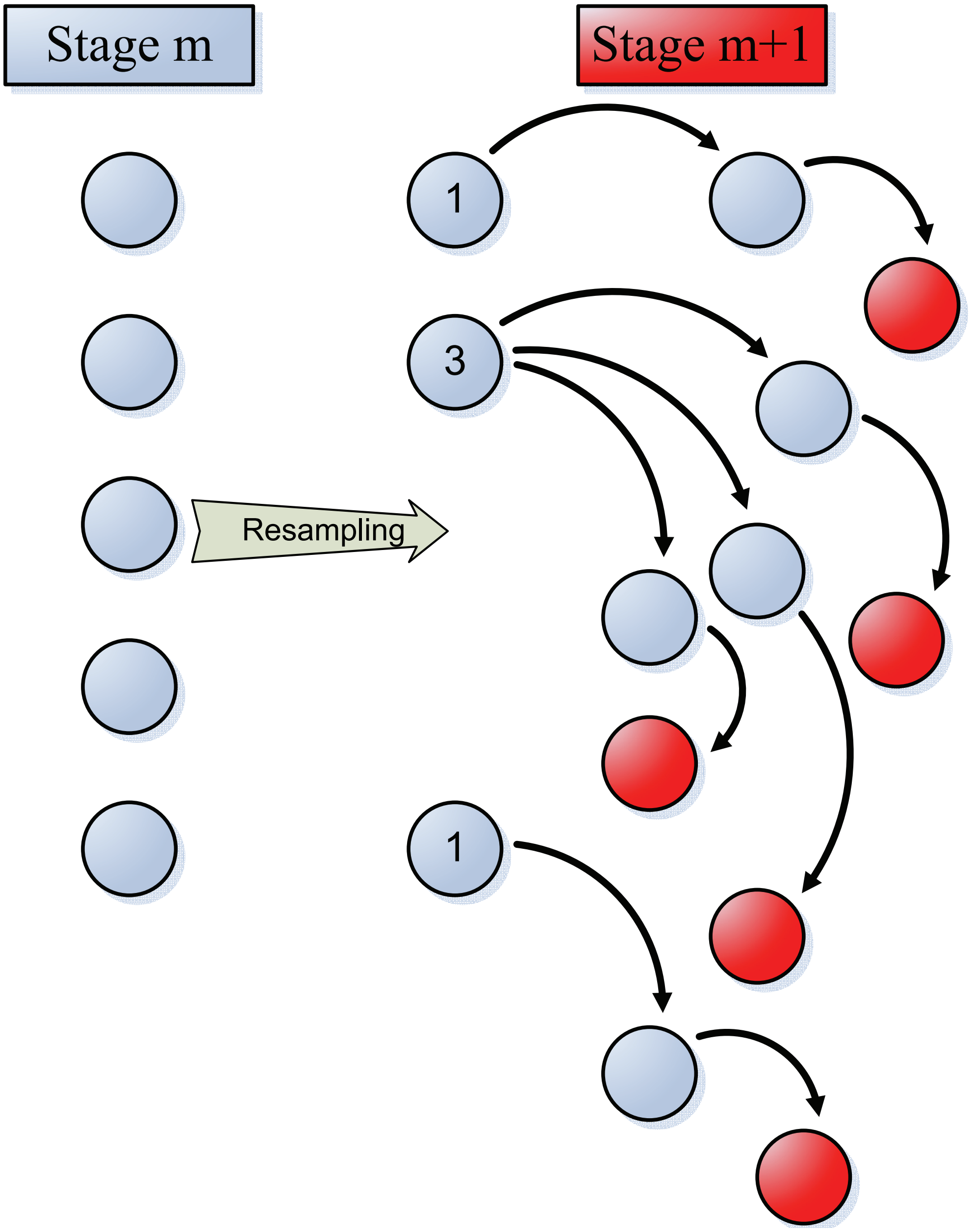


Figure 2.1: *CATMIP algorithm schematic*: This cartoon illustrates one complete cooling stage of the CATMIP algorithm. The five samples from β_m are resampled and then an instance of the Metropolis algorithm is run for each of the resulting samples. Numbers indicate the frequency of each model after resampling. The five red samples comprise the posterior distribution for β_{m+1}. The algorithm is plotted with a 100% acceptance rate for simplicity. Compare to Figure 1.2 and Figure 1.3.

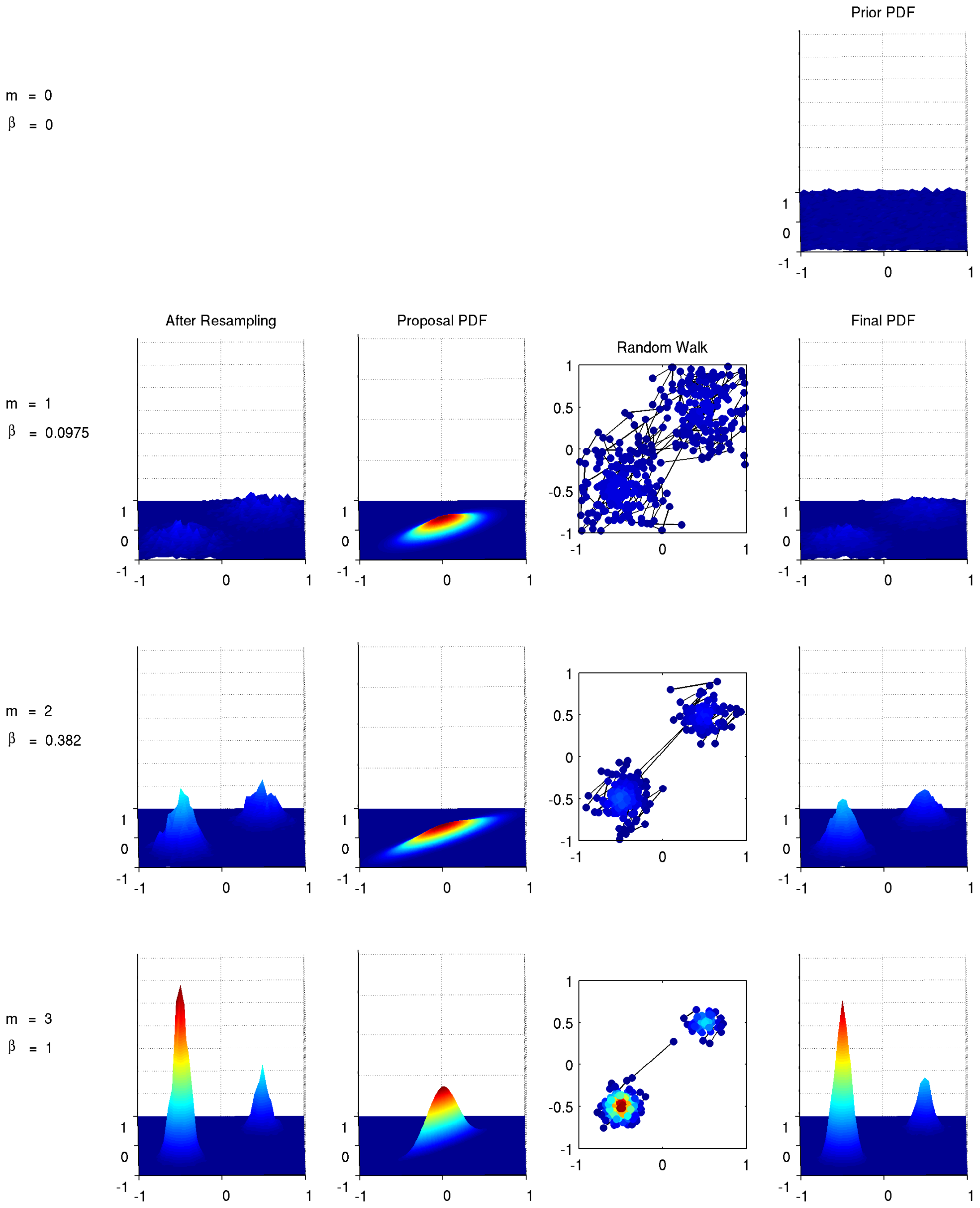


Figure 2.2: *CATMIP algorithm example*: As outlined in Box 2.1, the CATMIP algorithm begins by directly sampling the prior distribution (row 1). A new value for β is calculated and the distribution is resampled (column 1). The covariance of samples and acceptance rate is used to design a proposal PDF (column 2) for use in the Metropolis algorithm (column 3). The final samples from each Markov chain comprise the new PDF (column 4). In this example, the target distribution is the sum of two Gaussians, one of which has a factor of three greater amplitude than the other, with a uniform prior distribution.

Table 2.2: *Cascading.*

1. Sample $\mathrm{P}(\theta|D)_m \propto \mathrm{P}(\theta_1)\,\mathrm{P}(\theta_2)\,\mathrm{P}(D_1|\theta_1)^{\beta_m}$, $0 \leq \beta_m \leq 1$.
2. Sample $\mathrm{P}(\theta|D)_n \propto \mathrm{P}(\theta_1)\,\mathrm{P}(\theta_2)\,\mathrm{P}(D_1|\theta_1)\,\mathrm{P}(D_2|\theta_1,\theta_2)^{\alpha_n}$, $0 \leq \alpha_n \leq 1$.

the first set of data and parameters. Note that no approximations or simplifications have been made in arriving at this result.

We can incorporate Equation 2.4 into our tempering scheme by rewriting our transitional distributions as

$$\mathrm{P}(\theta|D)_{m,n} \propto \mathrm{P}(\theta)\,\mathrm{P}(D|\theta)_{m,n} = \mathrm{P}(\theta_1)\,\mathrm{P}(\theta_2)\,\mathrm{P}(D_1|\theta_1)^{\beta_m}\,\mathrm{P}(D_2|\theta_1,\theta_2)^{\alpha_n} \tag{2.5}$$

To sample this distribution, simply follow the steps in Box 2.2.

I will return to the cascading technique and how it relates to finite fault earthquake source models in Section 3.2.

2.3 Performance Tests

In this section, I will present the results of synthetic tests designed to determine the computational cost and expected results of CATMIP. Before applying the sampler to real data where the earthquake source process is unknown, or even synthetic analogs of the finite fault problem where model resolution is limited, I first explore how many samples it takes to reproduce a known high-dimensional probability distribution. Then I test the ability of CATMIP to sample a static finite fault model using synthetic data. Finally, I compare CATMIP to TMCMC and the Metropolis algorithm.

2.3.1 Multivariate Normal Distributions

We start with a simple problem: an n-dimensional Gaussian target distribution with a uniform a priori distribution. The results of this test for increasing numbers of free parameters is shown in Figure 2.3. For each dimensional space, I plot the quality of the posterior distribution for different numbers of Markov chains and varying lengths of those chains. The misfit between the mean of the target distribution and the mean of the posterior distribution is quantified by the model variance reduction (VR),

$$VR = [1 - \frac{\|m^{target} - m^{est}\|_2^2}{\|m^{target}\|_2^2}] \cdot 100\% \tag{2.6}$$

where m^{target} is the mean of the target distribution, m^{est} is the mean of the samples, and $\|\|_2^2$ denotes the square of the L_2 norm.

There are several conclusions we can draw from Figure 2.3. The variation in the total number of model evaluations made over the lifetime of the algorithm, which includes the number of cooling stages as well as the number of Markov chains, N, and their lengths, N_{steps}, is small compared the range of N and N_{steps} under consideration. Thus the density of sampling does not much affect the cooling schedule. Although

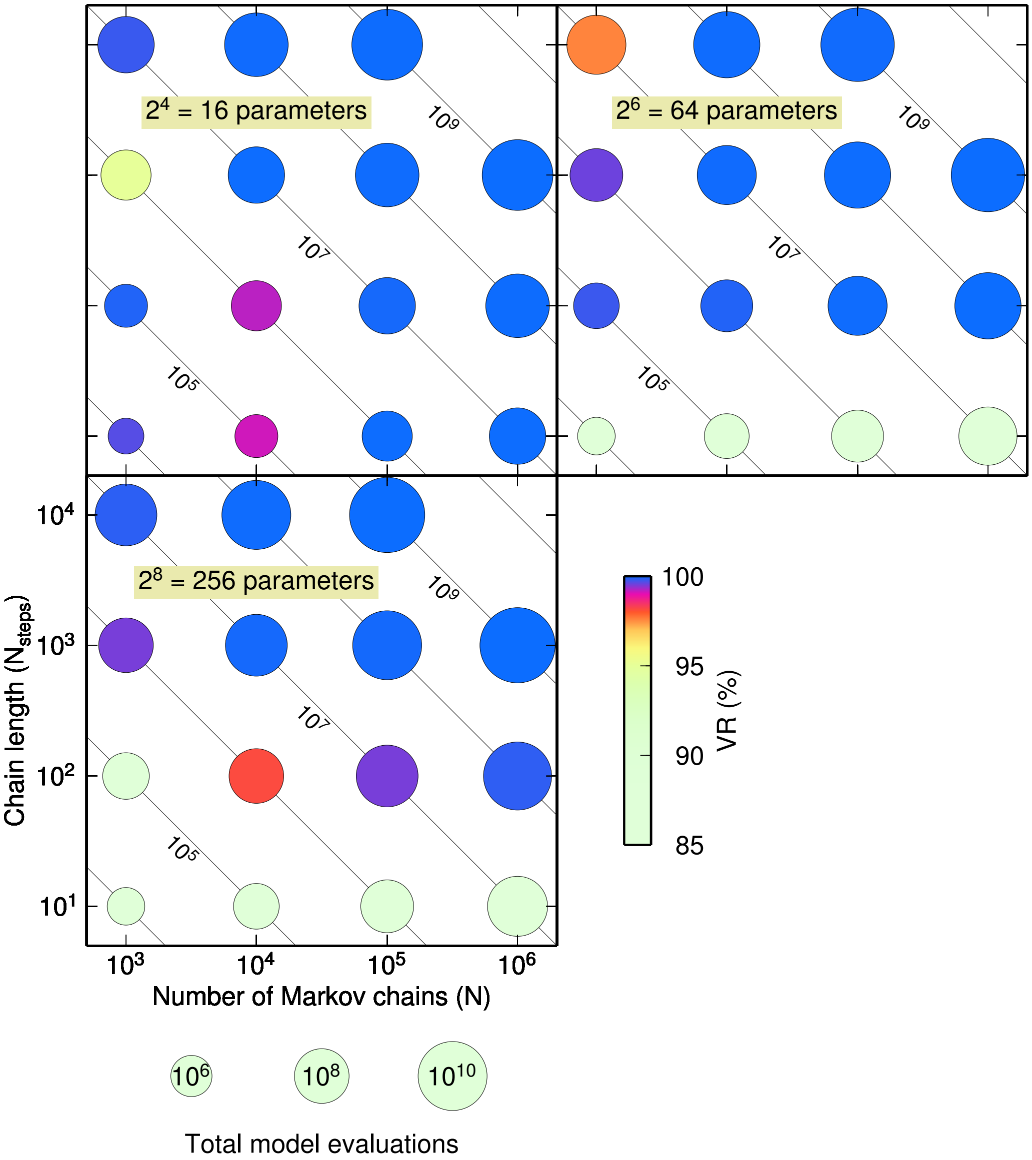


Figure 2.3: *Sampling a Gaussian distribution*: The number of samples required to reproduce a given distribution increases greatly with increasing numbers of model parameters. This is called the Curse of Dimensionality. Symbol colors denote the difference, quantified by the model variance reduction, between the mean of the target distribution and the mean of samples produced by CATMIP for a multivariate normal distribution as a function of the number of samples N and the length of Markov chains N_{steps}, with contour lines representing equal numbers of model evaluations per cooling step. The size of the symbols represents the total number of model evaluations over the lifetime of the algorithm including all cooling stages from the prior to the posterior.

the number of model evaluations needed to reproduce the target distribution for the higher-dimensional cases considered is large, it is not computationally intractable. And we can see that when the posterior distribution is on the cusp of being under-sampled, we may or may not reproduce the target distribution. Also, longer chain lengths tend to improve the quality of the solution. This may explain why CATMIP outperforms TMCMC when we compare them head-to-head (Section 2.3.3).

2.3.2 Synthetic Finite Fault Models

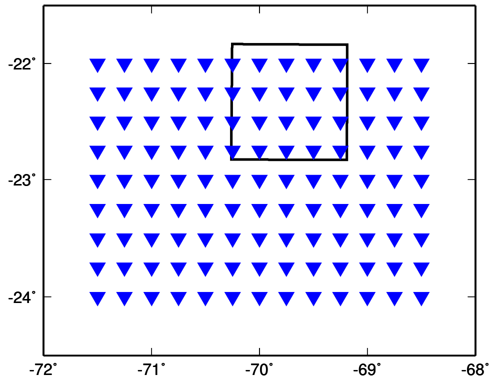


Figure 2.4: *Source-receiver geometry for synthetic static finite fault model*: Synthetic three-component GPS data were calculated for each of the stations denoted by triangles. The surface projection of the fault plane is shown with a thick black line. The fault has a strike of 0° and a dip of 18°. The depth to the top of the fault is 40 km. The assumed data uncertainty is 1 cm. Sampling results are shown in Figure 2.5.

Next, I look at the results of using CATMIP to produce a slip model using a dense network of synthetic GPS observations. This a synthetic version of the static finite fault modeling that I will introduce in Section 3.2 and then apply to the 2007 Mw 7.7 Tocopilla, Chile earthquake in Chapter 4. Since the slip distribution consists of two

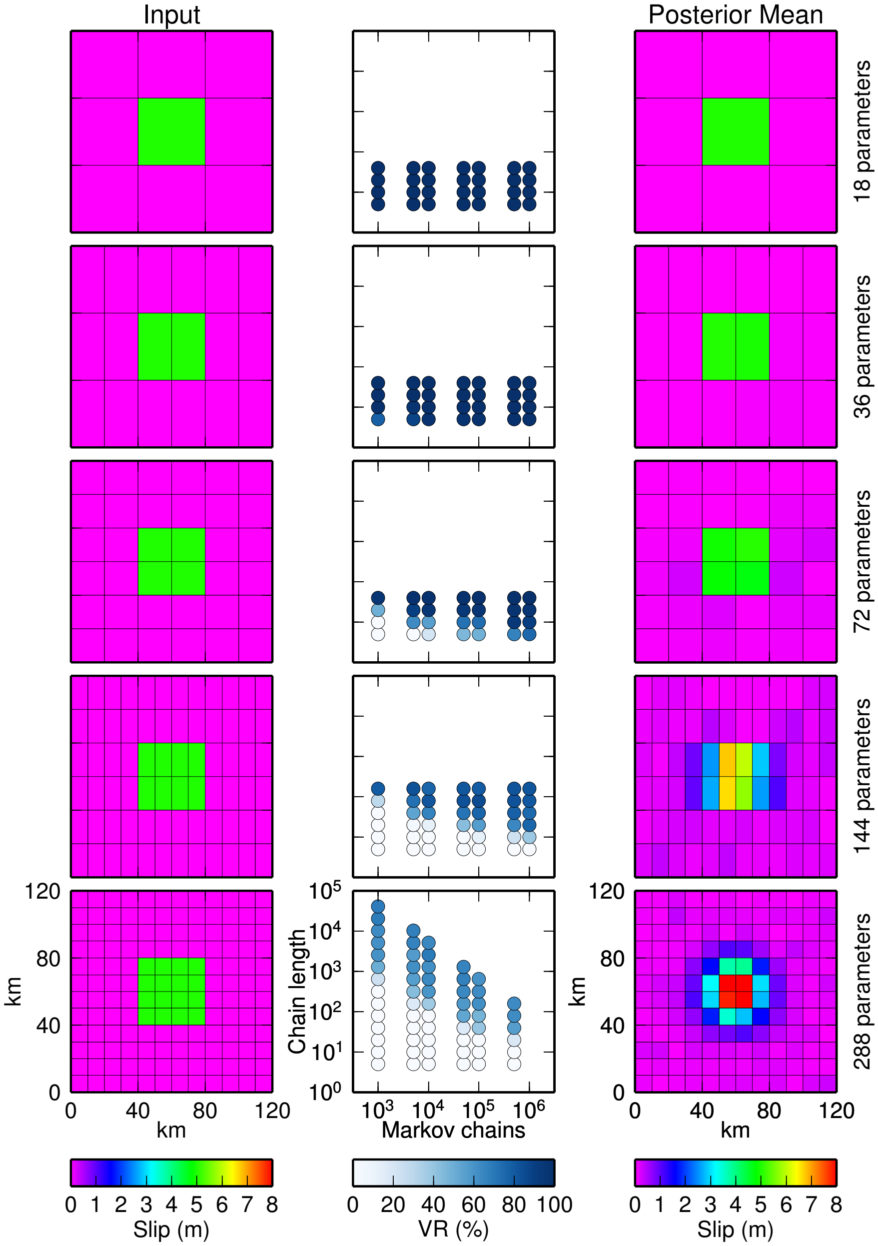


Figure 2.5: *Sampling a synthetic static finite fault model*: The left column shows the slip distributions used to create the synthetic surface displacements for each performance test. The quality of the output of CATMIP sampling is shown in the middle column. The mean of the posterior distribution for the CATMIP run with the largest number of samples is shown in the right column. The source-receiver geometry for this test is mapped in Figure 2.4.

components of motion on each fault patch, the total number of free parameters in each synthetic model is twice the number of fault patches. The source-receiver geometry for this performance test is shown in Figure 2.4.

Figure 2.5 shows that as the number of patches in the fault model increases, the number of samples required to reproduce the input slip distribution increases. At some point, the quality of the solution begins to decline, not because of undersampling, but because the patch size has become so small that the data can no longer resolve the model given this source-receiver geometry. When model resolution is lost, the displacements on neighboring patches begin to trade off with each other. The mean of all of these possible models results in a smoother slip distribution than the input model. This is not an incorrect result. Our data do not resolve the slip on each patch, so the posterior PDF for each patch is no longer highly peaked. There are many possible solutions which satisfy our data and, in real applications, we should consider all of these models.

2.3.3 CATMIP vs. TMCMC vs. Metropolis

Finally, I compare CATMIP to TMCMC and the Metropolis algorithm. Loosely based on Example 2(VIII) in *Ching and Chen* (2007), I run all three samplers with a prior distribution of $U(-2,2)$ and a target distribution which is a biased mixture of

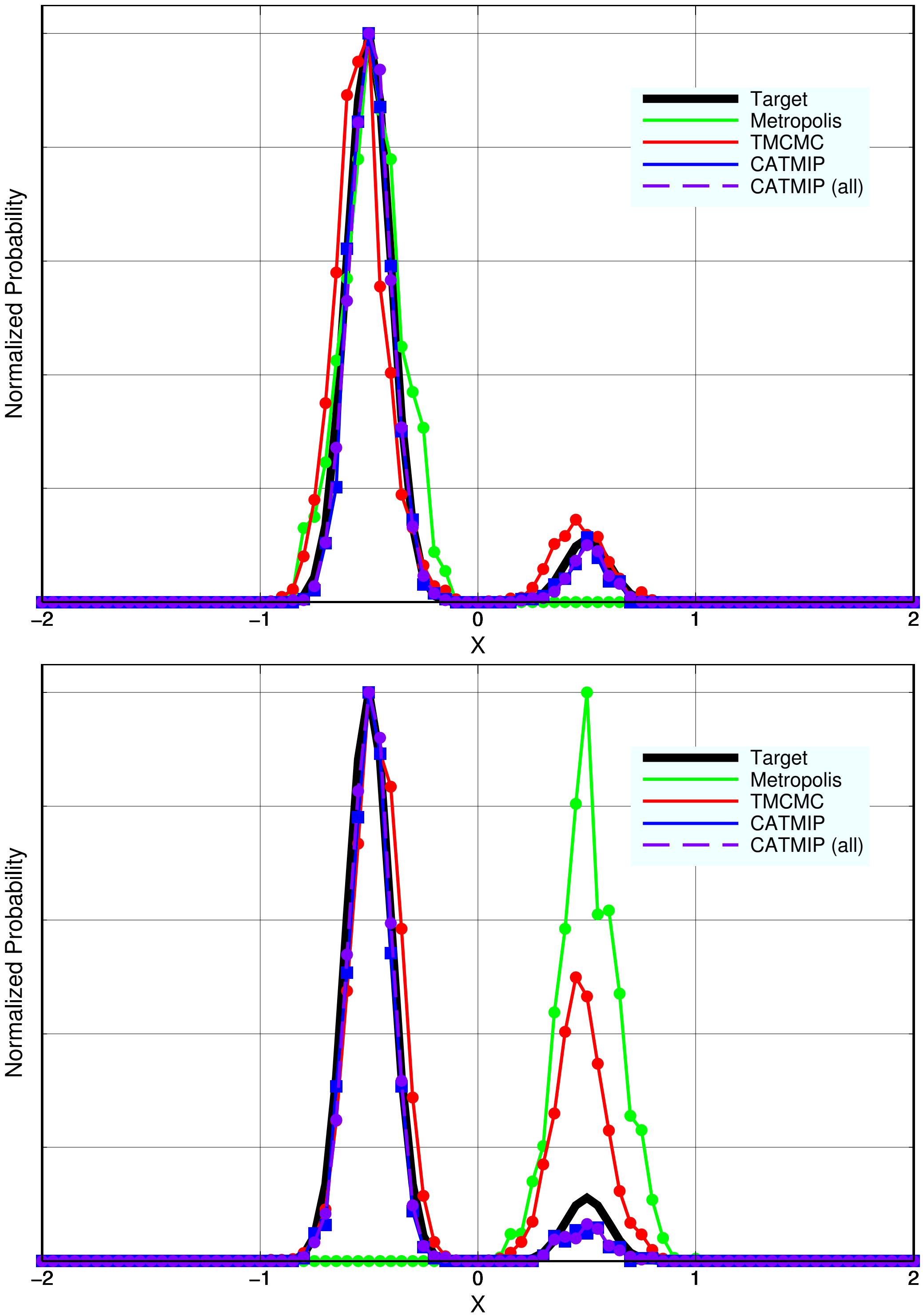


Figure 2.6: *Comparison of CATMIP, TMCMC, and Metropolis algorithms*: (Top) Marginal distribution for mixture of six-dimensional Gaussians. (Bottom) Marginal distribution for mixture of ten-dimensional Gaussians. See text for details. Distributions are normalized to have the same peak amplitude.

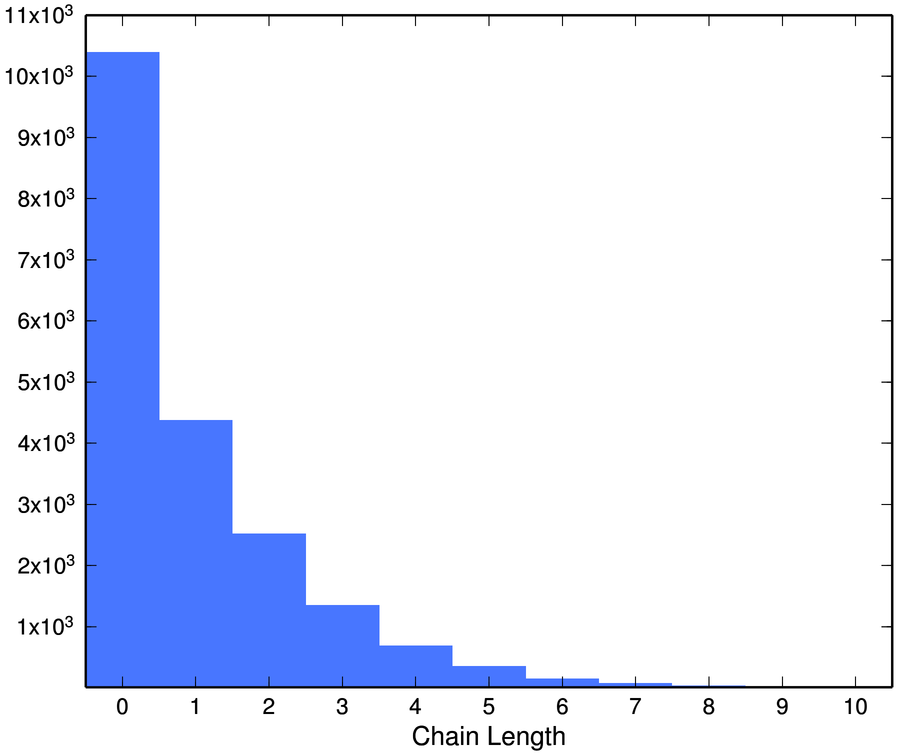


Figure 2.7: *Distribution of Markov chain lengths in TMCMC algorithm*: This is the distribution of Markov chain lengths from the final resampling in the TMCMC run with 20,000 samples for a mixture of ten-dimensional Gaussians (Figure 2.6). Over 20% of the models are in Markov chains of length one.

six-dimensional Gaussians:

$$\begin{aligned} 0.1N(\mu_1,\sigma) + 0.9N(\mu_2,\sigma) \\ \mu_1 = \left[\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2}\right] \\ \mu_2 = -\mu_1 \\ \sigma = 0.1\underline{I} \end{aligned} \tag{2.7}$$

Both CATMIP and TMCMC are run with a target COV of 1. All three samplers were set up so that they would produce approximately the same number of samples of the posterior distribution. But this is not precise because both CATMIP and TMCMC choose the cooling schedule dynamically, and TMCMC chooses the Markov chain lengths as well. The Metropolis algorithm was run for 55,000 samples. TMCMC was run with 5,000 samples; it took 11 cooling steps to complete, for a total of 55,000 samples over the lifetime of the algorithm, although only the 5,000 samples from the last cooling step are usable. CATMIP was run with 500 Markov chains, each with a length of 10 steps, or 5,000 samples per cooling step; it completed in 10 cooling steps, for a total of 50,000 samples over the lifetime of the algorithm.

Because in TMCMC not all of the samples are updated at each cooling step, the total number of model evaluations per cooling step is smaller than the total number of samples. So to compare the performance of the three samplers for a given computational expense, I ran TMCMC, tallied the total number of forward model evaluations over its lifetime, and then ran CATMIP and the Metropolis algorithm with sampling parameters such that the total number of forward model evaluations in each run would be approximately the same as for the TMCMC algorithm.

For the next experiment, I increased to ten the number of dimensions in the target distribution. TMCMC was run with 20,000 samples. The algorithm completed after 16 cooling steps with a total of 184,271 forward model evaluations. I ran CATMIP with 1,000 Markov chains, each with a length of 16 steps. The run finished after 12 cooling steps during which the forward model was evaluated 181,000 times. The Metropolis algorithm was run for 181,000 steps, the same number of model evaluations as CATMIP. For both this test and the last, the proposal distribution for the Metropolis algorithm is the same as the proposal distribution for the final cooling step ($\beta = 1$) of CATMIP and the Markov chain seed sample is $\underline{0}$.

The marginal distributions for one dimension of the target distribution are shown in Figure 2.6. For CATMIP, I plot the results if all samples from the last cooling step are used as well as if only the last sample from each Markov chain is kept. The Metropolis algorithm has difficulty with the multimodal distribution, and becomes trapped in the main peak of the target PDF for the six-dimensional case, and becomes trapped in the less probable peak in the ten-dimensional test. (Of course, given enough samples, the Metropolis random walk would eventually find the other peak in both cases.) TMCMC and CATMIP image both peaks of the distribution. Both do a good job of recovering the relative bias between the two peaks in the six-dimensional example, but TMCMC fails to reproduce the relative probability of the two peaks in the ten-dimensional problem. Overall, CATMIP can better reproduce the target distribution with less computational expense.

Since the CATMIP and TMCMC runs in Figure 2.6 were done with the same cooling scheme, resampling, prior distributions, similar numbers of samples, the better performance of CATMIP must be due to the different implementation of the Metropolis algorithm. In TMCMC, the length of each Markov chain is proportional to the likelihood of its seed model. Notice in Figure 2.7 that more than 20% of all Markov chains have a length of one, and are not updated after resampling. No Markov chain has a length greater than ten steps (and only four chains are that long) while, in the corresponding CATMIP run, all instances of the Metropolis algorithm are run for 16 steps. TMCMC does very little exploration of the solution space and relies on tempering and resampling to reach the posterior distribution; the Metropolis algorithm contributes very little to the solution. But CATMIP uses the Metropolis algorithm to more extensively random walk through the solution space in addition to utilizing tempering and resampling. Even though this additional MCMC sampling makes CATMIP more computationally expensive than TMCMC for a given number of samples, CATMIP can produce better results than TMCMC with fewer samples, making it overall computationally cheaper than TMCMC.

2.4 CATMIP: A Qualitative Discussion

Let us put aside the math and performance tests for a little while, and discuss qualitatively what CATMIP is and how it works. It has embedded in it the Metropolis algorithm (*Metropolis et al.*, 1953). The Metropolis algorithm is the figurative grandfather of just about every Markov Chain Monte Carlo (MCMC) method in existence. It uses a random walk to explore the model space and probabilistically chooses whether or not to take a proposed step based on the probability associated with the candidate sample, i.e., the probability of the proposed new location in model space.

CATMIP and TMCMC belong to a class of samplers which use tempering. This approach is similar to simulated annealing optimization (*Kirkpatrick et al.*, 1983;

Cerny, 1985) except that simulated annealing does not produce probabilistically distributed samples. Because we start with a broad posterior distribution and then slowly “cool” it to the true distribution, it is easier for the sampler to find all of the peaks of the posterior distribution. This is why the CATMIP and TMCMC algorithms out-performed the Metropolis algorithm at sampling a mixture of Gaussians. Although its origins are unclear, the use of tempering in Bayesian sampling dates back at least to *Marinari and Parisi* (1992).

Both CATMIP and TMCMC employ resampling. This produces a behavior similar to genetic algorithms (e.g. *Holland*, 1992) in that “less fit” Markov chains (those which have lower posterior probability) are killed off in favor of more probable models. Looked at in this manner, the samples in CATMIP’s genetic pool reproduce through replication (resampling) and mutation (random walking) but do not cross-breed. While the Markov chains do not mix with each other, information from all Markov chains is combined to calculate the model covariance and proposal density.

The main difference between CATMIP and TMCMC is in how the Metropolis algorithm is employed. In TMCMC, the more probable models are assembled into longer Markov chains. In CATMIP, more probable models spawn more Markov chains, leading to a concentration of multiple chains in regions of high probability. In some ways the TMCMC approach is more firmly in the realm of evolutionary computation because better models have more offspring than worse models. CATMIP is more similar to the Neighbourhood algorithm (*Sambridge*, 1999a,b) which is not a Bayesian sampler but which explores the solution space by concentrating sampling in regions which produce better models.

Chapter 3

Finite Fault Earthquake Source Models

We now turn our attention to the geophysical problem I want to solve, namely producing finite fault earthquake source models using Bayesian sampling. Over the last decade or so, there has been a tremendous increase in the data available to constrain the source processes of earthquakes. High-rate GPS networks have become more common and, if an earthquake is located near stations of this type, we can use their time series as high-quality displacement seismograms (e.g. *Miyazaki et al.*, 2004; *Ji et al.*, 2004; *Hartzell et al.*, 2007; *Bilich et al.*, 2008). Satellite radar interferometry has blossomed, yielding spatially dense sampling of the entire surface displacement field. Putting all these data together allows us to study the details of the slip evolution of earthquakes as never before. And by determining the kinematics of the rupture process, we can unravel the underlying physics. The physics of earthquakes is mostly beyond the scope of what I will be presenting, although I will discuss some applications of Bayesian finite fault modeling in Chapter 5. But it is important to realize finite fault earthquake source models are more than an end in themselves.

3.1 A Brief History of Finite Fault Modeling

Researchers have been producing finite fault models for many years now. An exhaustive review of the existing literature cannot be included here, and review papers have been written on the subject (e.g. *Ide*, 2007). But I want to discuss a few notable existing modeling techniques.

The finite fault source model is at its root linear. An observed slip distribution can be represented by the linear combination of many dislocations with different locations, orientations, strengths, and origin times. But the physics of the problem dictate that the constraints on these source parameters are not linear. Some of these constraints can be written to fit into a least squares framework; for example, backslip can be avoided by using non-negative least squares. And thus the combination of non-negative least squares and Laplacian smoothing is often sufficient to produce static slip models. But the main difficulty with a linear approach to kinematic finite fault modeling lies in imposing causality on the temporal evolution of slip.

We expect rupture to propagate across a fault plane, triggering slip as it goes. The least restrictive interpretation of this statement is that slip at a particular point cannot occur before the stress change has propagated to that location. The more restrictive interpretation is that slip begins when the propagation front arrives at said point.

But there is no simple linear constraint that expresses either of these statements.

This does not mean that there are not linear causal finite fault source models. *Hartzell and Heaton* (1983) produced a finite fault model of the 1979 Imperial Valley earthquake via least squares inversion by fixing the rupture velocity in their model and allowing slip to occur in prescribed time windows. Similarly, *Cohee and Beroza* (1994a) used a fixed rupture velocity to model the 1992 Landers earthquake, with either one dislocation per fault patch or slip in three time windows, by using a combination of linear and linearized inversions. *Kikuchi and Kanamori* (1982) used a fixed rupture velocity in their modeling methodology but allowed for complex source-time functions. These are reasonable modeling choices for an optimization approach because source dimension is known to trade-off with rupture velocity. In the Bayesian approach, we do not worry about these trade-offs.

The non-linear optimization approach of *Ji et al.* (2002) allows for simultaneous solution of variable rupture velocity and source-time functions. There are several large caveats that go along with this technique. One is that the optimization employs simulated annealing, a process which is almost as computationally intensive as Bayesian analysis, yet has all of the disadvantages of optimization. Another is that while *Ji et al.* (2002) do solve for the source-time function and rupture velocity independently, the source-time parameterization may be over-simplified, and it is very difficult to enforce causality in this optimization scheme.

3.2 A Bayesian Finite Fault Parameterization

Here I will discuss the particular finite fault parameterization I use in my modeling. For Bayesian sampling we need a model with a tractable number of free parameters, a fast forward model, and a prior distribution. Let us take these issues in order.

As the name suggests, a finite fault earthquake source model consists of the history of an earthquake source (its spatial and temporal evolution) over a fault surface of finite extent. Thus we must determine all faulting parameters at many points in space. There is not much flexibility when choosing the number of points or patches in your model. It is mostly determined by the spatial resolution of your data and the frequency content of the kinematic data being used. So our only option is to describe the rupture source at each patch with as few parameters as possible. In an attempt to balance computational and sampling cost with reasonable flexibility in the source model, we use four faulting parameters per patch: slip in two directions, rupture velocity, and source duration.

3.2.1 Static Source Model

Because Bayesian sampling will require that we compute the forward model millions or billions of times, it is necessary to design the forward problem so that it can be evaluated as quickly as possible. The static model is entirely linear. We simply evaluate $(d - G * m)^T * C_d^{-1} * (d - G * m)$ where m is a vector containing the two components of slip on each patch, G is a matrix of Green's functions, d is a vector of observed surface displacements, and C_d is the covariance of d. Except in the case of very large amounts of data and model parameters, the forward model is quite fast if you pre-compute C_d^{-1} from C_d.

In addition to the parameters describing the slip model itself, we must also introduce several nuisance parameters. InSAR data measure the change in line-of-sight distance between the satellite and the ground. They are interferograms measuring changes in phase, and by necessity they contain an arbitrary constant offset in derived surface displacements. Furthermore, slight ambiguities in the satellite's orbit commonly manifest themselves as spurious ramp-like displacement fields. Thus, for each interferogram used, I also solve for three parameters (a, b, and c) which fit a ramp to the InSAR data in the form $a + b\underline{x} + c\underline{y}$ where $\underline{x}$ and $\underline{y}$ are the locations of the data in local Cartesian coordinates.

3.2.2 Kinematic Source Model

The kinematic model similarly includes a linear combination of fundamental Green's functions, although these Green's functions are vectors of time series instead of scalar values. However, we must also convolve our basis functions with the source-time function of each patch (which we parameterize as a triangle with rise τ_r) and time-shift them based on the spatial distribution of rupture velocities, V_r. There is no method of convolution that is sufficiently fast for our purposes. So we pre-compute a set of Green's functions for each source-receiver pair and a wide variety of possible source-time functions. For each forward model evaluation, we simply make a linear combination of the appropriate pre-computed synthetic sources.

To determine how much to time-shift each source in our model, we must map our hypocenter location and rupture velocity field to initial rupture times at each patch. This mapping can be done quickly and efficiently using the Fast Sweeping Algorithm (*Zhao*, 2005), a level-set method which uses a Godunov upwind difference scheme (*Rouy and Tourin*, 1992) for solving the Eikonal equation,

$$\begin{aligned} |\nabla u(\underline{x})| &= f(\underline{x}), \qquad && \underline{x} \in R^n,\ f(\underline{x}) > 0 \\ u(\underline{x}) &= 0, && \underline{x} \in \Gamma \subset R^n \end{aligned} \tag{3.1}$$

where $u(\underline{x})$ is the first arrival time at $\underline{x}$ of a wavefront which initiates at Γ and propagates with normal velocity $\frac{1}{f(\underline{x})}$. $f(\underline{x}) = \frac{1}{V_r(\underline{x})}$ for our earthquake source model. (We can also use a variable hypocenter and solve for the distribution of possible hypocenter locations as part of our modeling, although the models presented in the following chapter use a fixed hypocenter.)

The kinematic model has one triangular source-time per patch while rupture velocity is allowed to freely vary. This is almost the opposite approach to *Kikuchi and Kanamori* (1982) who used a complex source-time function with a fixed rupture velocity. *Cohee and Beroza* (1994b) concluded, somewhat unsurprisingly, that the former approach does a better job of recovering the rupture velocity of the source at the expense of doing a worse job at estimating rise time; but they also found that source models which can only rupture once do a better job at estimating the seismic moment. And while the results of single-window and multiple-window models, in the terminology of *Cohee and Beroza* (1994b), are fairly similar in quality, the quality of the solution will be poor unless the slip distribution is constrained by independent data. This should not be a problem for the models in Chapter 4 because there is an abundance of static surface displacements which can be used to determine the slip distribution.

3.2.3 Choice of Prior Distribution

Finally, we must design a prior distribution to use with our model. The hypocenter location and ramp parameters may be assumed to be normally distributed. But we know very little a priori about the distribution of source-time functions or rupture velocities. So for lack of a better model, we use uniform priors on those parameters. In fact, the one source parameter we truly know well is the seismic moment tensor. This does not tell us anything about the distribution of slip; but it tells us something about the total amount of slip. We also have a good idea of the overall slip direction, or rake, based on teleseismic observations. To turn these observations into a prior distribution of slip, we use a rotated coordinate system with one axis, $U_{\parallel}$, aligned with the teleseismic rake direction, and the other axis, $U_{\perp}$, perpendicular to the rake

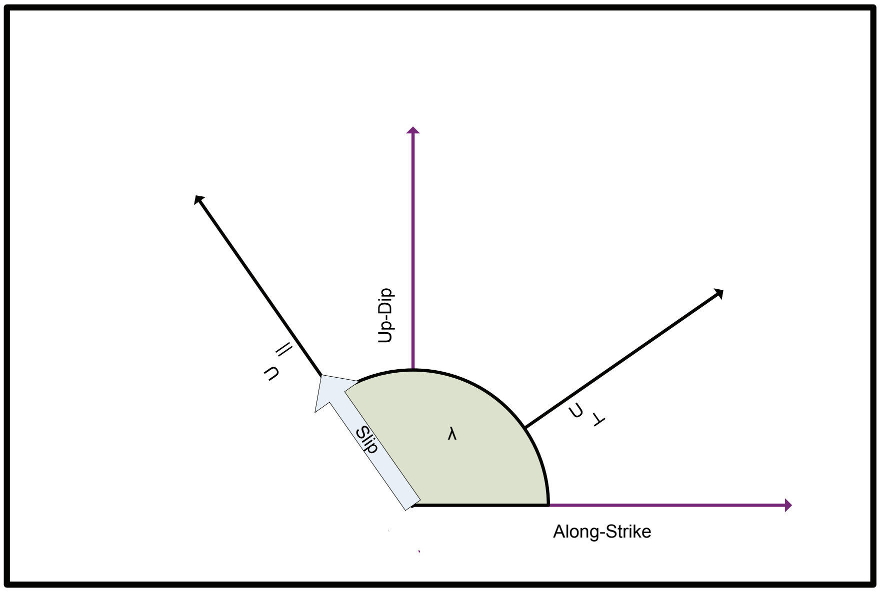


Figure 3.1: *Slip coordinate system*: The Bayesian finite fault parameterization uses components of slip, $U_{\perp}$ and $U_{\|}$, which lie in the fault plane but are orthogonal to each other. $U_{\perp}$ is aligned with the direction of hanging-wall motion given by the rake angle, λ

direction (Figure 3.1). The prior on $U_{\perp}$ is a Gaussian with zero mean; we allow some variation in the rake but assume it should be minimal. Our prior on slip in the $U_{\|}$ direction is the Dirichlet distribution which is applicable to problems where a finite number of states must sum to a specified value (see e.g. *Gelman et al.*, 2004). Thus the total moment for each sample of the prior is pre-determined. However, we do not use a constant value of moment for all samples. Instead, we assume that there is a Gaussian uncertainty on the catalog magnitude for the earthquake. For each a priori sample we draw a magnitude from a Gaussian distribution and then generate random slips on each patch using the Dirichlet distribution so that the total slip equals the proposed magnitude (Figure 3.2). After drawing samples of the prior distribution,

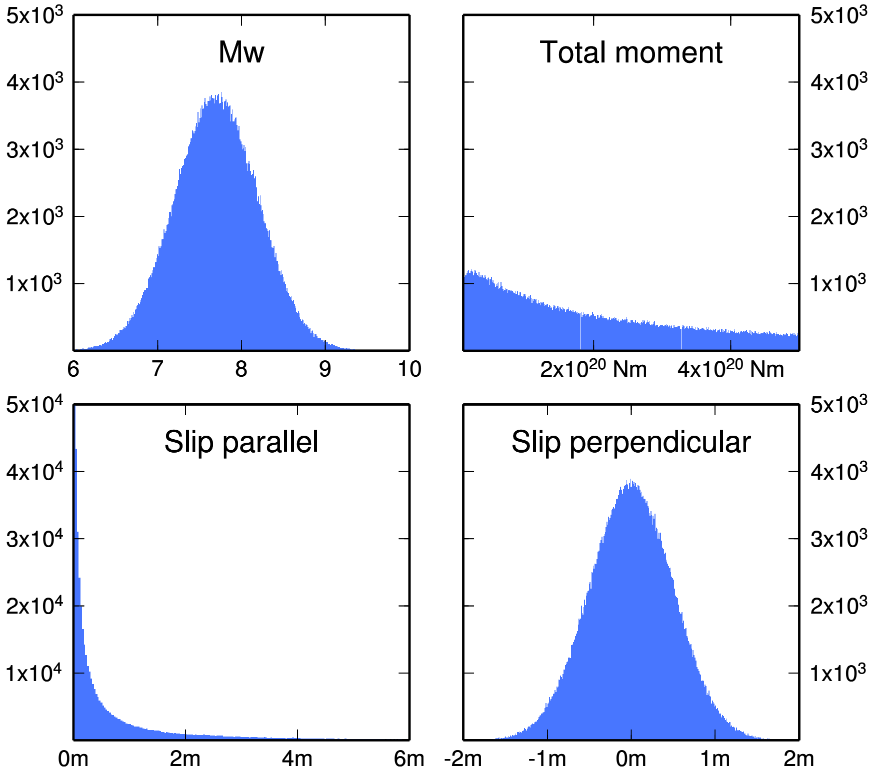


Figure 3.2: *Prior distribution on slip*: We describe the magnitude of the earthquake by a normal distribution centered on the catalog magnitude for the event (top left). This yields a log-normal distribution for scalar seismic moment (top right). We then use the Dirichlet distribution to translate the distribution of total moment for the earthquake into a distribution of slip on each patch, and we assume that this slip is aligned with the rake direction of the earthquake (bottom left). The earthquake is allowed to slip in two directions, and the distribution on slip in the direction perpendicular to the rake direction is assumed to be Gaussian (bottom right). The distributions plotted here are the same as for the Tocopilla earthquake study (Table 4.2).

we do not impose the Dirichlet prior on succeeding cooling steps. Instead, we use a one-sided semi-infinite uniform prior probability: we allow any positive value for $U_\parallel$, but forbid large amounts of back-slip. The minimum allowed slip value is less than zero to avoid under-sampling small slip values due to the hard bound on minimum $U_\parallel$.

Note that for efficiency we rotate our Green's functions into our ($U_\perp$, $U_\parallel$) coordinates thus obviating the need to transform the forward model for each likelihood evaluation. It should also be noted that the Dirichlet distribution is not the best or only prior distribution for slip. But it is a probability distribution that satisfies our requirements that the slip on each patch be in the forward rake direction ($U_\parallel \geq 0$), and that the total slip satisfies the moment the of the earthquake.

3.2.4 Implementation of Cascading

Because I have both kinematic and static data, I can take advantage of the cascading technique. As the static data depend on the slip parameters and the InSAR nuisance parameters only, we can use a Bayesian static model as a prior distribution for a full kinematic model which also includes rupture velocity, slip duration, and possibly hypocenter location. Let us call our static parameters (slip and InSAR ramps) θ_s and our kinematic-only parameters (rupture velocity, rise time, and hypocenter location)

θ_k. The full model $\theta = [\theta_s, \theta_m]$ can be written as

$$\mathrm{P}(\theta|D) \propto \mathrm{P}(\theta_s|D_s)\,\mathrm{P}(\theta_k|D_k, D_s) = \mathrm{P}(\theta_s)\,\mathrm{P}(\theta_k)\,\mathrm{P}(D_s|\theta_s)^{\beta}\,\mathrm{P}(D_k|\theta_k, \theta_s)^{\alpha} \tag{3.2}$$

When fusing static and kinematic earthquake data into a single source model, I solve the static problem first. Then, through cascading, I use the posterior distribution from modeling the static data to make the prior distribution for the full kinematic problem. This significantly decreases the computational cost in two ways. First, I can treat the model as two problems: one which has half as many free parameters and one which has a small solution space when compared to the full problem without cascading. Second, I can explore the solution space of the slip distribution using only the static forward model, which is significantly faster than the full kinematic forward model. For the models presented in the following chapter, it takes about 0.005 seconds to evaluate the static forward model, while the kinematic forward model requires approximately 0.03 seconds more, for a total of around 0.035 seconds to evaluate the joint probability $\mathrm{P}(\theta_s|D_s)\,\mathrm{P}(\theta_k|D_k, D_s)$.

Chapter 4

The 2007 Mw 7.7 Tocopilla, Chile Earthquake

4.1 Tectonic Setting

My demonstration earthquake for Bayesian finite fault modeling is the 2007 Mw 7.7 Tocopilla, Chile earthquake. The Peru-Chile subduction zone is a complex, seismically active region which seems to exhibit different behaviors in different regions. Its great seismic productivity is in large part due to the high Nazca-South America convergence rate, approximately 68 mm/yr in northern Chile (*Norabuena et al.*, 1998). The largest earthquake ever recorded is the Mw 9.5 great 1960 Chile earthquake. North of that rupture, the subduction interface has historically failed in magnitude 7 and 8 events (Figure 4.1). *Barazangi and Isacks* (1976) suggested that the subduction zone is split into five segments, with the moderately dipping southern Ecuador, southern Peru-northern Chile, and southern Chile segments separated by shallowly dipping segments in north-central Peru and central Chile. The shallowly dipping segments are marked by a notable lack of Quaternary volcanism.

Even finer regional segmentation of the subduction interface appears to exist. *Lomnitz* (2004) points out that the 1906 M 8.6 Valparaiso earthquake only ruptured part of what was then a seismic gap, with the rest of the seismic gap failing in the 1928 M 8.4 Talca earthquake. Similarly, the 2007 Mw 7.7 Tocopilla, Chile earthquake extends north from the 1995 Mw 8.1 Antofagasta earthquake, with the Mejillones Peninsula marking the termination of both ruptures (*Loveless et al.*, 2009). *Von Huene and Ranero* (2003) theorize that the Mejillones Peninsula is just one of several seaward-dipping blocks in the region which may form barriers to seismic rupture, although it has been observed to slip aseismically (*Pritchard and Simons*, 2006).

4.2 Data and Green's Functions

There are numerous data available for the Tocopilla earthquake because it ruptured beneath the Central Andean Tectonic Observatory (CAnTO) GPS network from which we obtained both static offsets and high-rate time series that can be used in kinematic modeling (Figures 4.2, 4.3, 4.4, and 4.5). Two tracks from each of three different kinds of InSAR imagery are available: ascending ALOS data, descending normal-swath and ascending wide-swath ENVISAT data (Figure 4.6). GPS data from the CAnTO GPS network were processed into daily positions and 1 Hz time series by Jeff Genrich. Both data sets were made with IGS refined orbits (*Dow et al.*, 2009), GAMIT and GLOBK software, and were not sidereally filtered. Independent daily

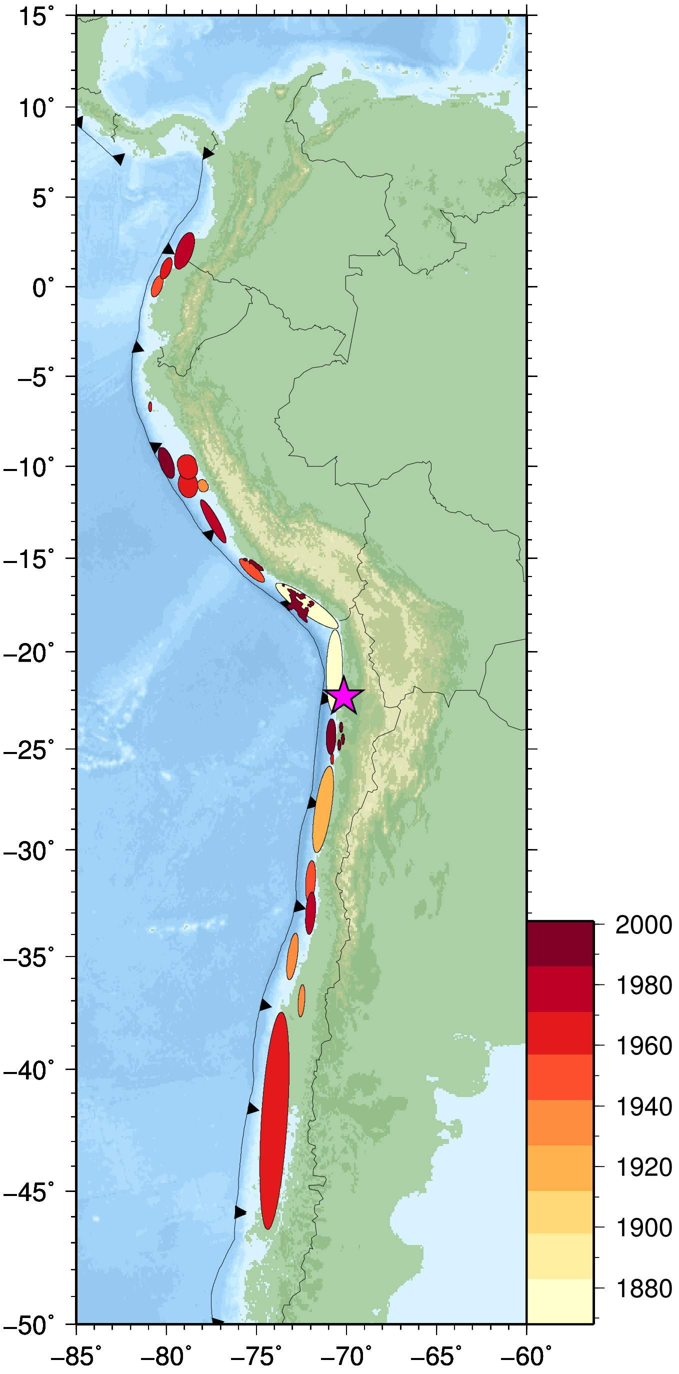


Figure 4.1: *Historical seismicity of the Peru-Chile subduction zone*: The location of the 2007 Tocopilla earthquake is shown with a star.

positions were made by averaging 30 second epoch regional solutions. The 1 Hz time series were processed with TRACK software using IGS station AREQ in Arequipa, Peru as a reference.

The ALOS and normal-swath Envisat interferograms were created with ROIPAC processing software. Initial processing of the wide-swath interferograms was done by Eric Fielding. The orientation and depth of the fault plane are derived from the work of Francisco Ortega (Table 4.1). Nina Lin provided corrections for tropospheric errors using the method of *Lin et al.* (2010). The corrected interferograms were then resampled using the technique of *Lohman and Simons* (2005) (Figure 4.7). Static and kinematic Green's functions were calculated using frequency-wavenumber integration (*Zhu and Rivera*, 2002) for a local 1-D tomographic velocity model (*Husen et al.*, 1999) (Figure 4.8).

For static modeling, I used all six resampled interferograms and GPS offsets derived from the daily positions. The resampling process calculates a data covariance matrix for the output data points. To estimate the uncertainties in the GPS data, I calculated the variance (relative to a linear tectonic displacement model) of the daily time series during a seismically inactive period. I first made a preliminary GPS-only model which I used to forward-predict the radar data and calculate a residual displacement field. I then fit a ramp to the residual displacements, and used the resulting model values and uncertainties to estimate standard deviations for the Gaussians that I then used as prior distributions in my final static modeling. For the kinematic model, I used the same prior distributions and, through cascading, the posterior slip distributions

Table 4.1: *Fault geometry.*

Parameter	**Prior distribution**
Strike	5°
Dip	20°
Rake	99°
Epicenter	$(-22.33°, -70.16°)$
Source Depth	41 km

from the static model. The kinematic data are the horizontal components of the 1 Hz GPS time series, as the vertical components are too noisy to be usable. I filtered these time series using a causal butterworth low pass filter with a corner period of 3 seconds and arbitrarily assume a data variance of 1 cm^2. The earthquake epicenter comes from *Delouis et al.* (2009). The different data sets are not weighted relative to each other in the joint kinematic modeling.

4.3 Static Model

As discussed in Section 3.2, the static model consists of two components of slip for each fault patch as well as 18 InSAR nuisance parameters, for a total of 498 free parameters. The prior distributions used for this model and the following kinematic model are given in Table 4.2.

After completing the CATMIP algorithm, our solution is not a single model but an enormous ensemble of samples of the posterior probability distribution. This is a completely different result from traditional optimization methods. Our main goal now is to analyze the posterior distribution to learn about the family of models which are

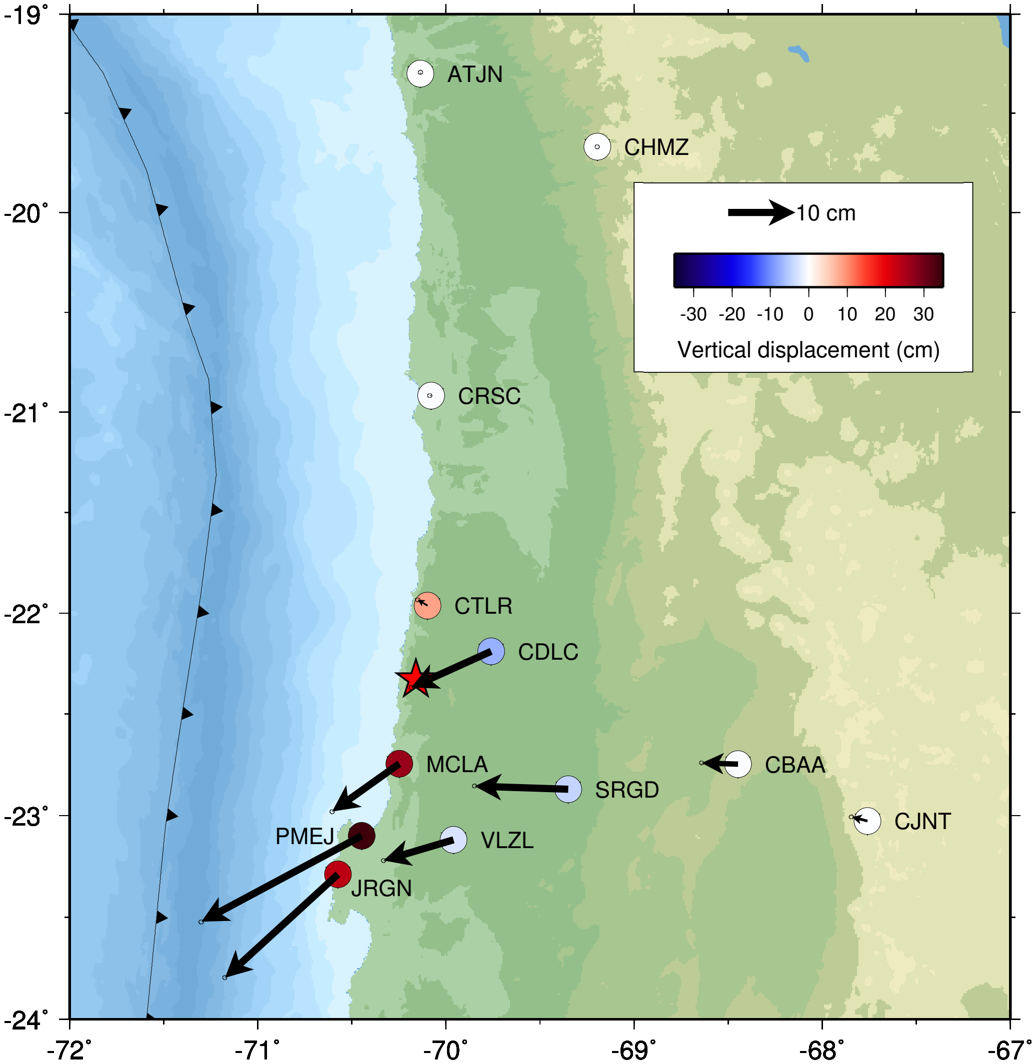


Figure 4.2: *GPS offsets from daily positions*: The star marks the location of the earthquake hypocenter.

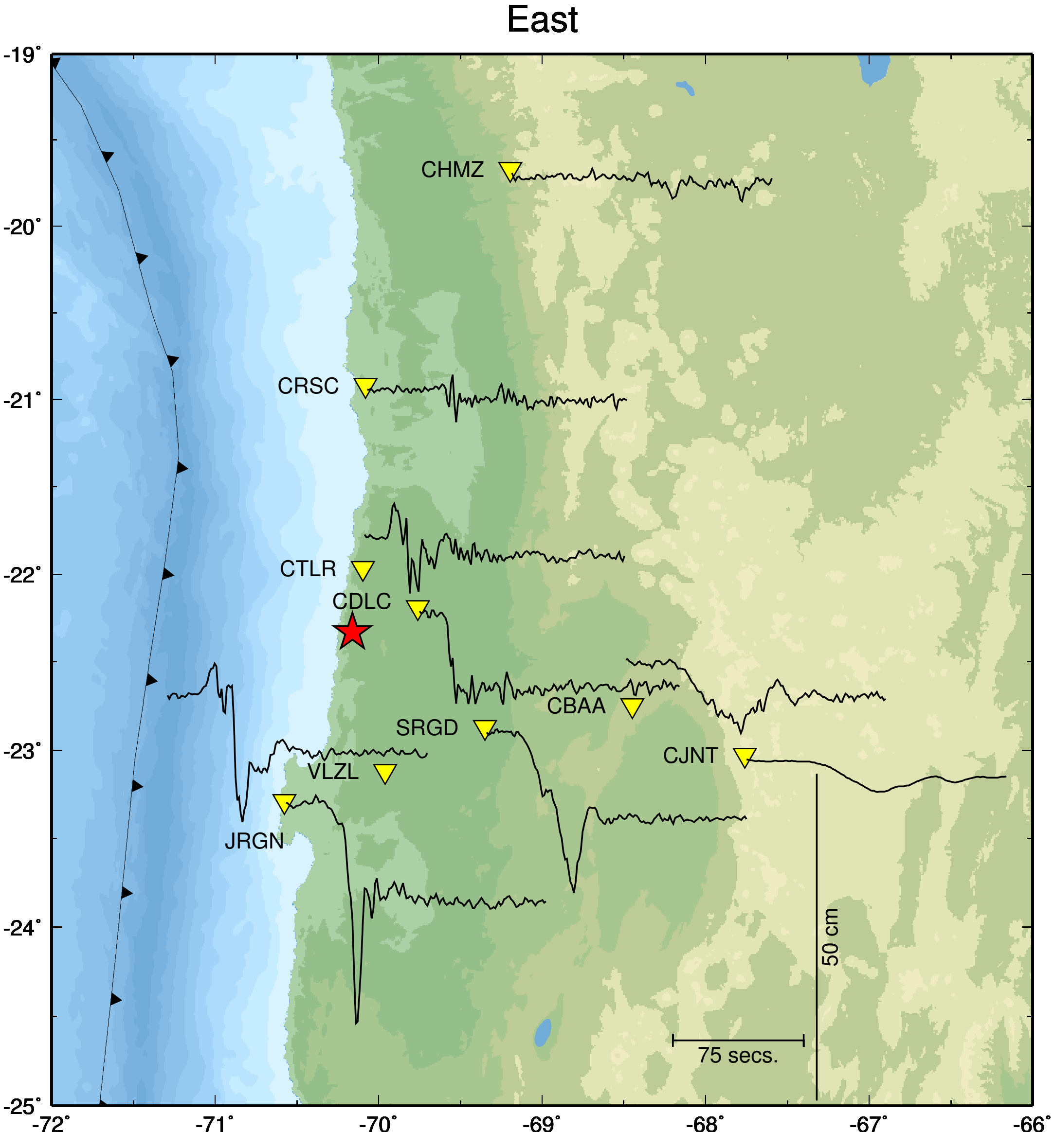


Figure 4.3: *East component of 1 Hz GPS time series.*

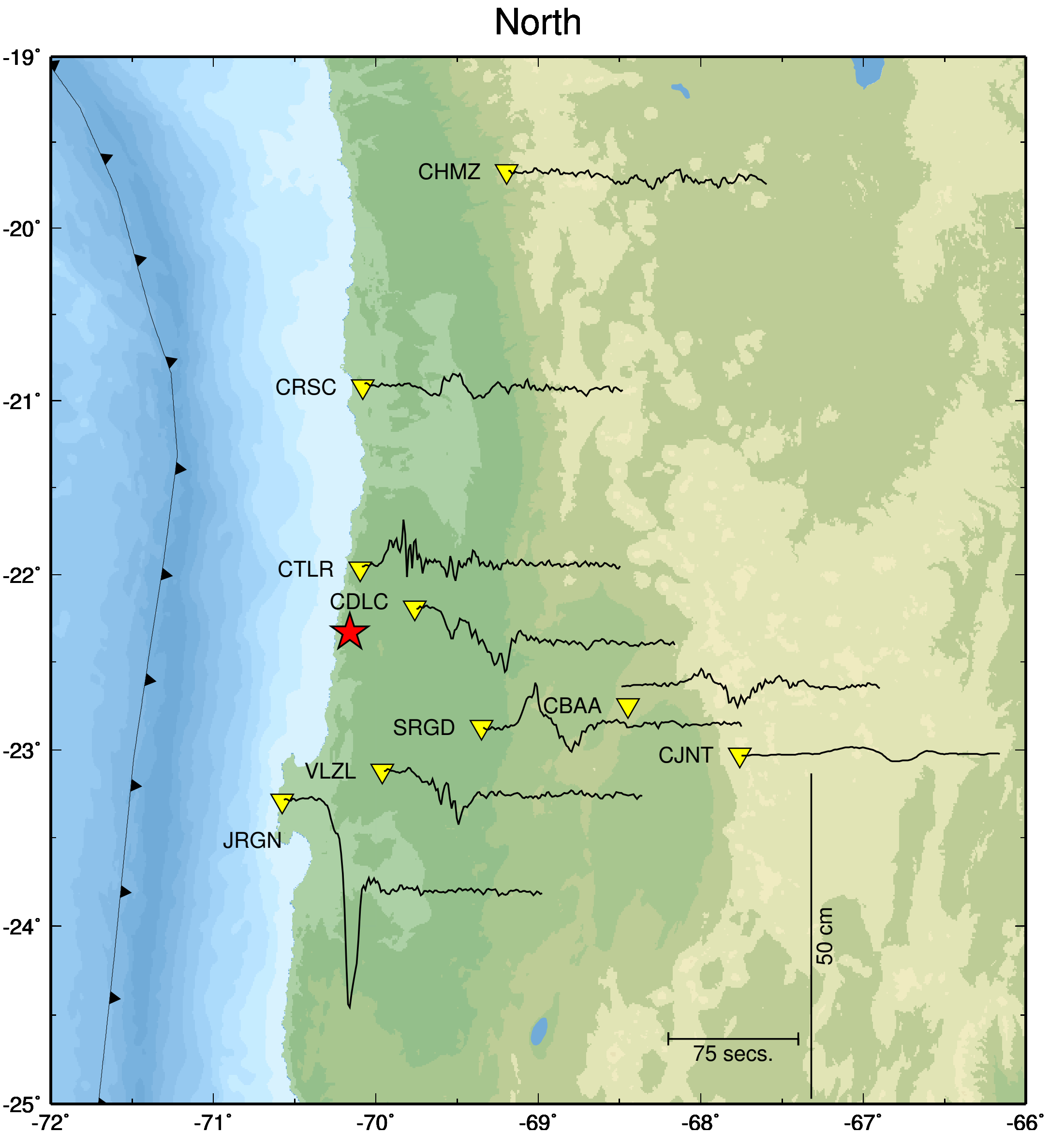


Figure 4.4: *North component of 1 Hz GPS time series.*

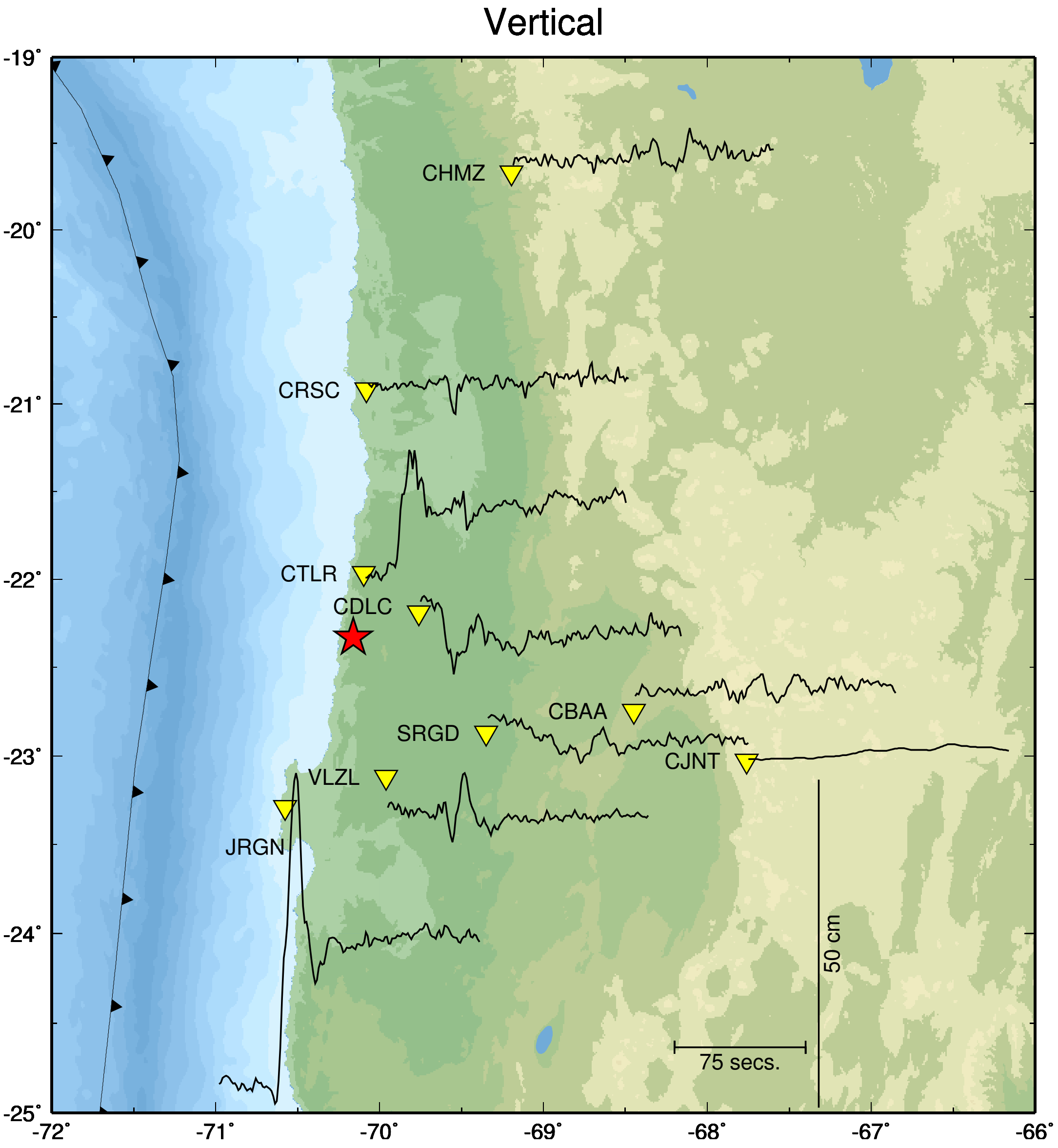


Figure 4.5: *Vertical component of 1 Hz GPS time series.*

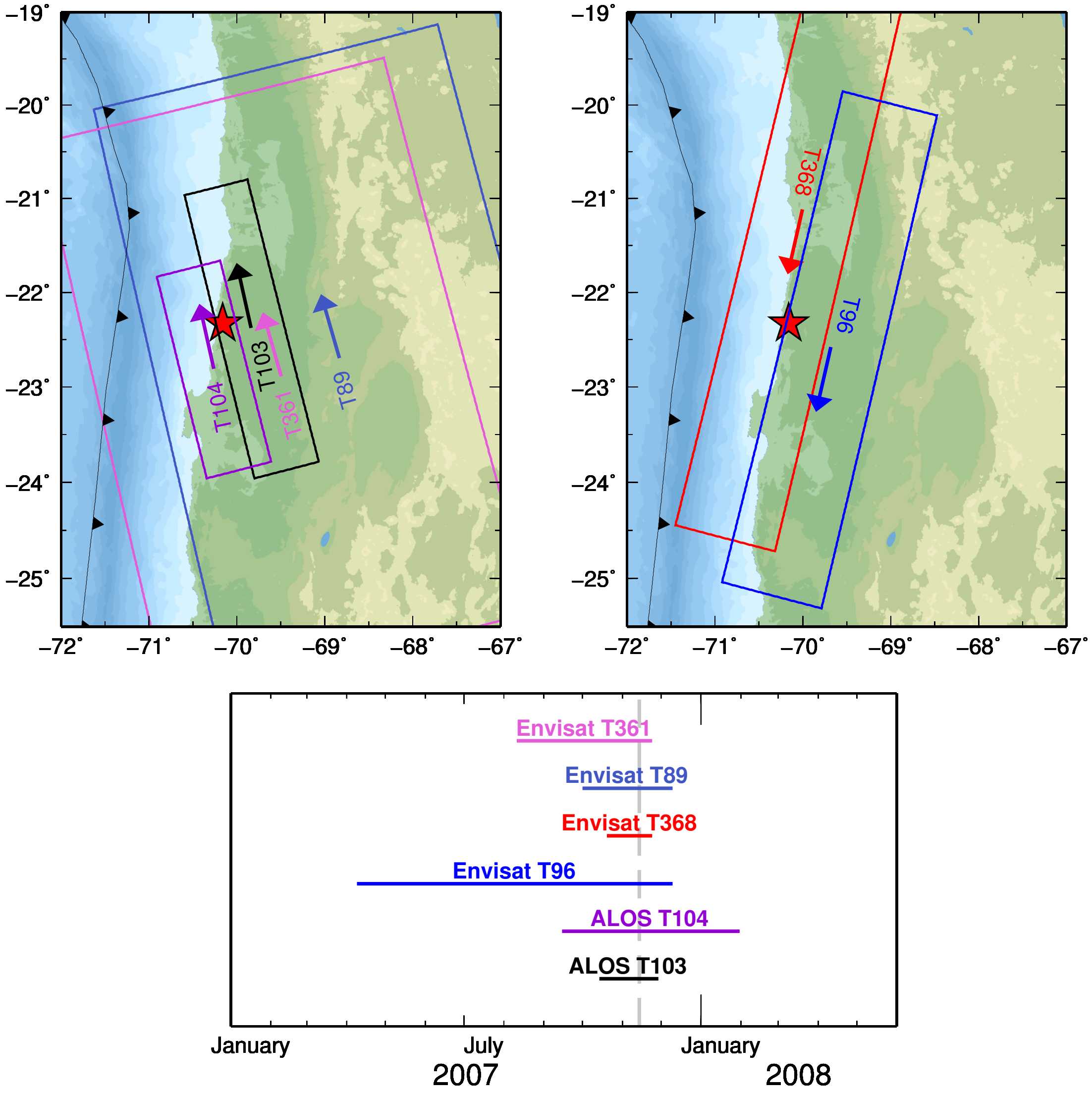


Figure 4.6: *InSAR data*: (Top) InSAR satellite flight paths. The footprint of each scene is shown by the colored rectangles, with arrows showing the corresponding satellite flight path for ascending (left) and descending (right) tracks. (Bottom) Dates that radar imagery was captured relative to the date of the Tocopilla earthquake. Horizontal lines, labeled with the satellite name and track number, show the time elapsed between the first and second scene in each interferometric pair. The dashed vertical line marks the day of the Tocopilla earthquake.

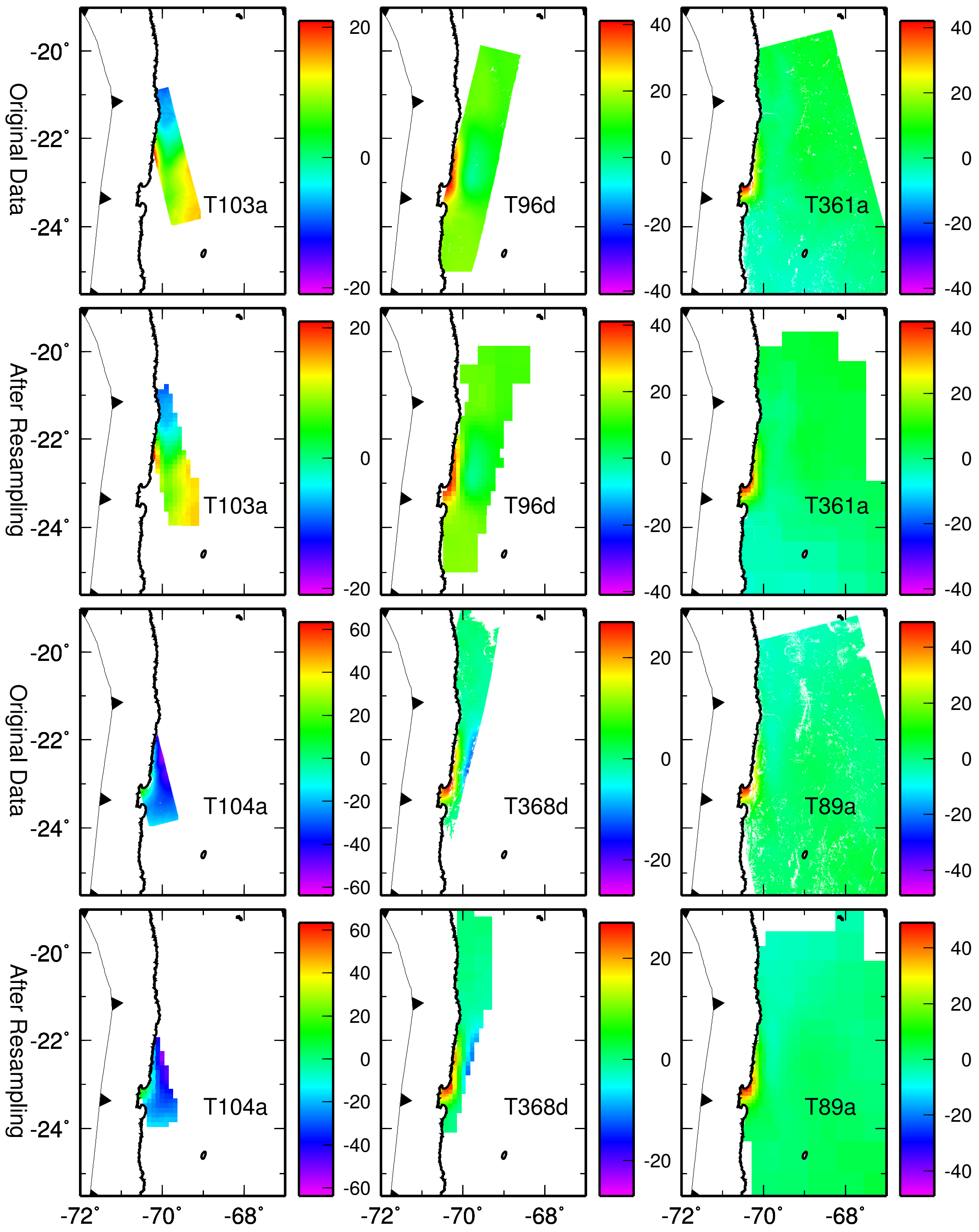


Figure 4.7: *Resampled InSAR data*: Six InSAR interferograms were used in this model. The original data were resampled using the method of *Lohman and Simons* (2005), and it is the resampled data points that were used in the modeling.

Table 4.2: *Prior distributions*: The prior distributions used in the static and kinematic models.

Parameter	**Prior distribution**
Magnitude	Normal distribution: $\mu = 7.7, \sigma = 0.5$
Slip Parallel	Dirichlet distribution: $\alpha = 1$, slip $\geq -1\,\mathrm{m}$
Slip Perpendicular	Normal distribution: $\mu = 0\,\mathrm{m}, \sigma = 0.5\,\mathrm{m}$
Rise Time	Uniform distribution: $0.25\,\mathrm{s} \leq \tau_r \leq 7.75\,\mathrm{s}$
Rupture Velocity	Uniform distribution: $0\,\mathrm{km/s} < V_r \leq 8\,\mathrm{km/s}$
InSAR Ramp 1: T103a	3 Normal distributions:
	$a + bx + cy$ (μ, σ)
	a $(0, 190.292)$
	b $(0, 39.6504)$
	c $(0, 154.065)$
InSAR Ramp 2: T104a	3 Normal distributions:
	$a + bx + cy$ (μ, σ)
	a $(0, 504.560)$
	b $(0, 72.9119)$
	c $(0, 523.016)$
InSAR Ramp 3: T96d	3 Normal distributions:
	$a + bx + cy$ (μ, σ)
	a $(0, 18.0386)$
	b $(0, 9.51263)$
	c $(0, 14.4670)$
InSAR Ramp 4: T368d	3 Normal distributions:
	$a + bx + cy$ (μ, σ)
	a $(0, 43.9352)$
	b $(0, 20.8235)$
	c $(0, 24.8122)$
InSAR Ramp 5: T361a	3 Normal distributions:
	$a + bx + cy$ (μ, σ)
	a $(0, 28.9768)$
	b $(0, 7.60256)$
	c $(0, 36.4127)$
InSAR Ramp 6: T89a	3 Normal distributions:
	$a + bx + cy$ (μ, σ)
	a $(0, 32.9630)$
	b $(0, 2.44874)$
	c $(0, 30.6867)$
Hypocenter (along-strike, down-dip)	Normal distribution: $\mu = (130.5\,\mathrm{km}, 49.5\,\mathrm{km})$ $\sigma = (0\,\mathrm{km}, 0\,\mathrm{km})$

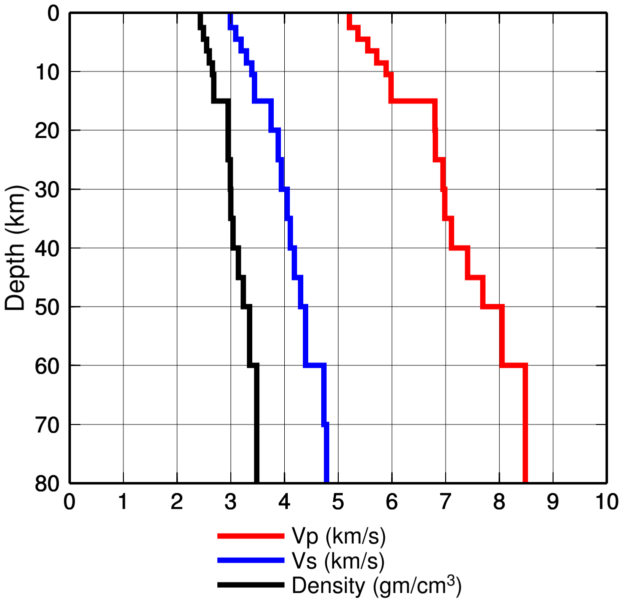


Figure 4.8: *Velocity model*: The P-wave and S-wave velocities, v_p and v_s, are from *Husen et al.* (1999). The density in $\frac{\text{gm}}{\text{cm}^3}$ is given by $\rho = 0.77 + 0.32 \cdot v_p$. The P-wave and S-wave quality factors are $Q_\alpha = 1500$ and $Q_\beta = 600$, respectively.

consistent with the data and the specified prior information. Let us begin by looking at some statistics of the posterior distribution, in particular three models: the model with the greatest posterior probability (this is called the maximum a posteriori or MAP), the mean of all models, and the median of all models. (Note that the MAP is a particular sample from the posterior distribution while the mean and the median are models statistically derived from the ensemble of samples.) Each of these three models has a reasonable slip distribution (Figure 4.9). If any of these models were the solution produced by an optimization algorithm, we would probably consider it a successful run of the numerical solver.

The MAP model has several separate asperities in a narrow band while the mean and median of the distribution are slightly smoother. All three models are very similar. All three fit the data well (Figures 4.10, 4.11, 4.12, and 4.13). In the MAP model, there are several patches near the edge of the fault with significant amounts of slip. (This pattern is commonly found in finite fault models made with traditional optimization methods.) It is possible that this slip is a real feature of the earthquake source process; but more likely it is simply extra model complexity which helps fit noise in the data, or real data signal which is inconsistent with our source model design, or simply a random fluctuation in the model parameters. If it is the latter, we would expect these extra slip patches to have different amounts of slip or be located in different parts of the fault with each new model. This is what we see. While the MAP model has several of these random patches, they average out and disappear in the mean and median models. So we can infer that these slips are not persistent features of the earthquake source process.

We must be careful when evaluating these individual models (the MAP, the mean, and the median) because they are just that: individual models. If the posterior is broad or has multiple peaks, the MAP may be misleading. Similarly, if the posterior distribution is not Gaussian, the mean and median of the distribution may be meaningless values. To truly understand the shape of the solution space, we have to look at the posterior distribution itself.

Given the full distribution, we can answer many questions about the source model. For example, how well resolved is the slip on a given patch? Figure 4.14 shows that the posterior distribution is very different from our prior distribution. So the data are informative. The distribution on the $U_\perp$ component of slip is highly peaked and

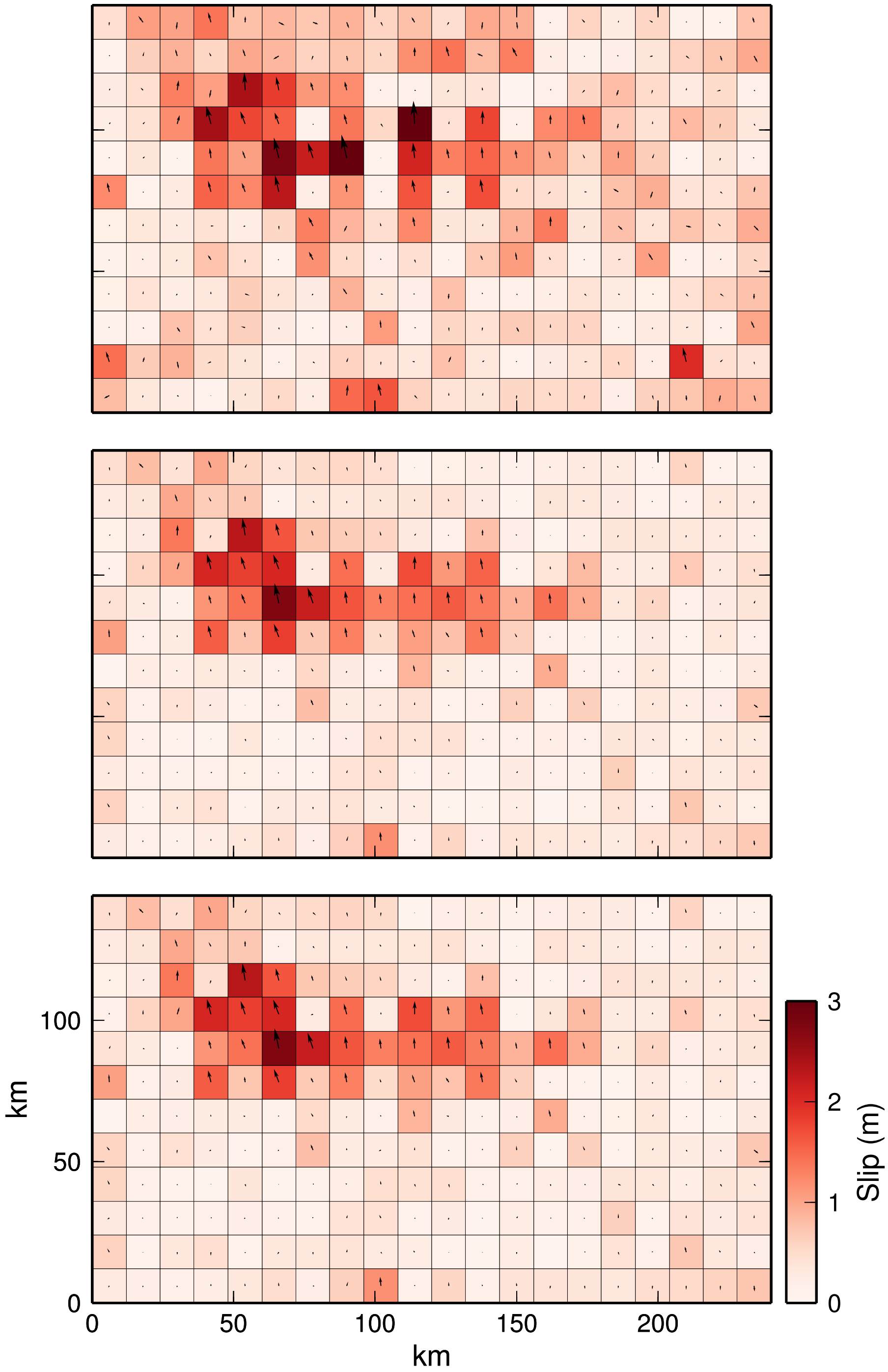


Figure 4.9: *Three models from the posterior distribution*: These are the MAP (top), mean (middle) and median (bottom) models of the posterior distribution. The magnitude of slip is shown with color and the arrows indicate the direction of slip.

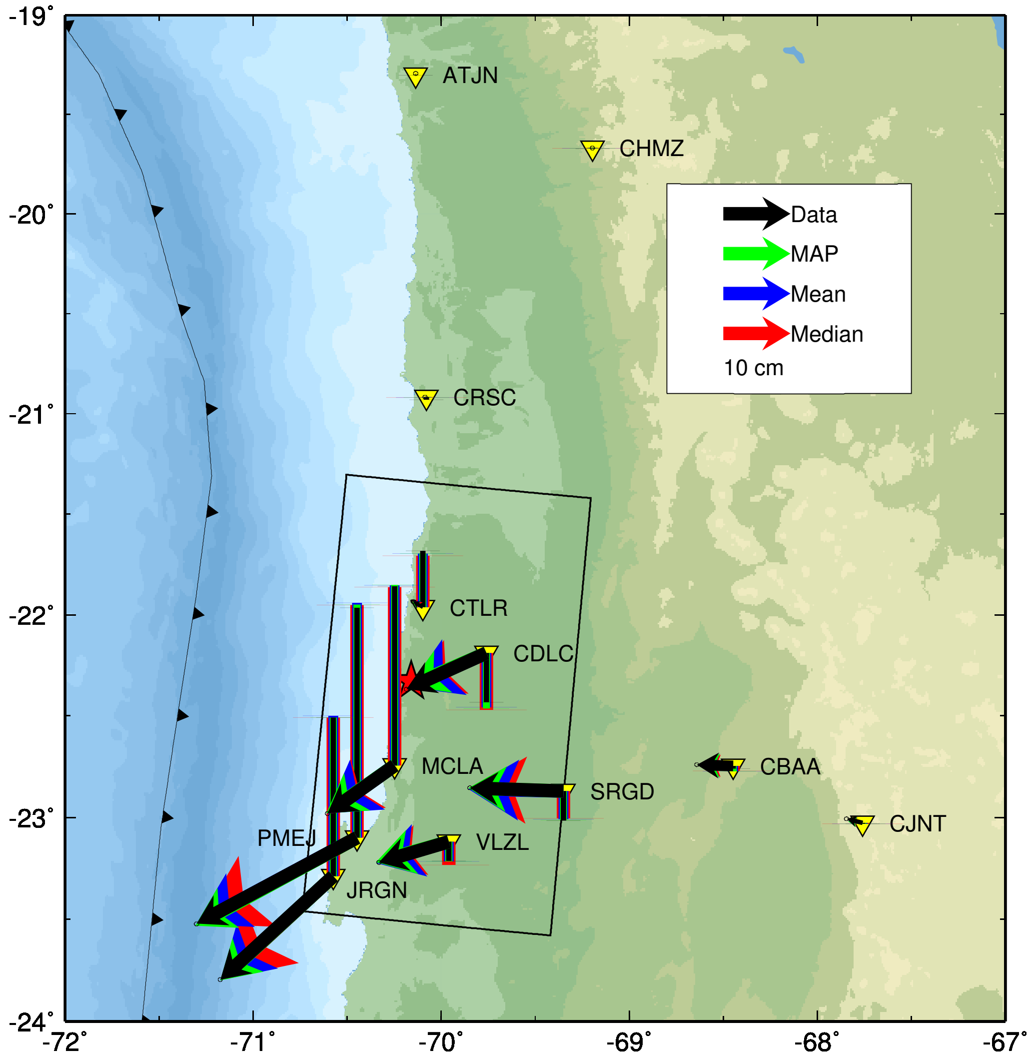


Figure 4.10: *GPS predictions from three models*: The predicted static displacements from the MAP, mean, and median of the posterior distribution are compared to the GPS offsets. The surface projection of the fault plane is shown with a thick black line.

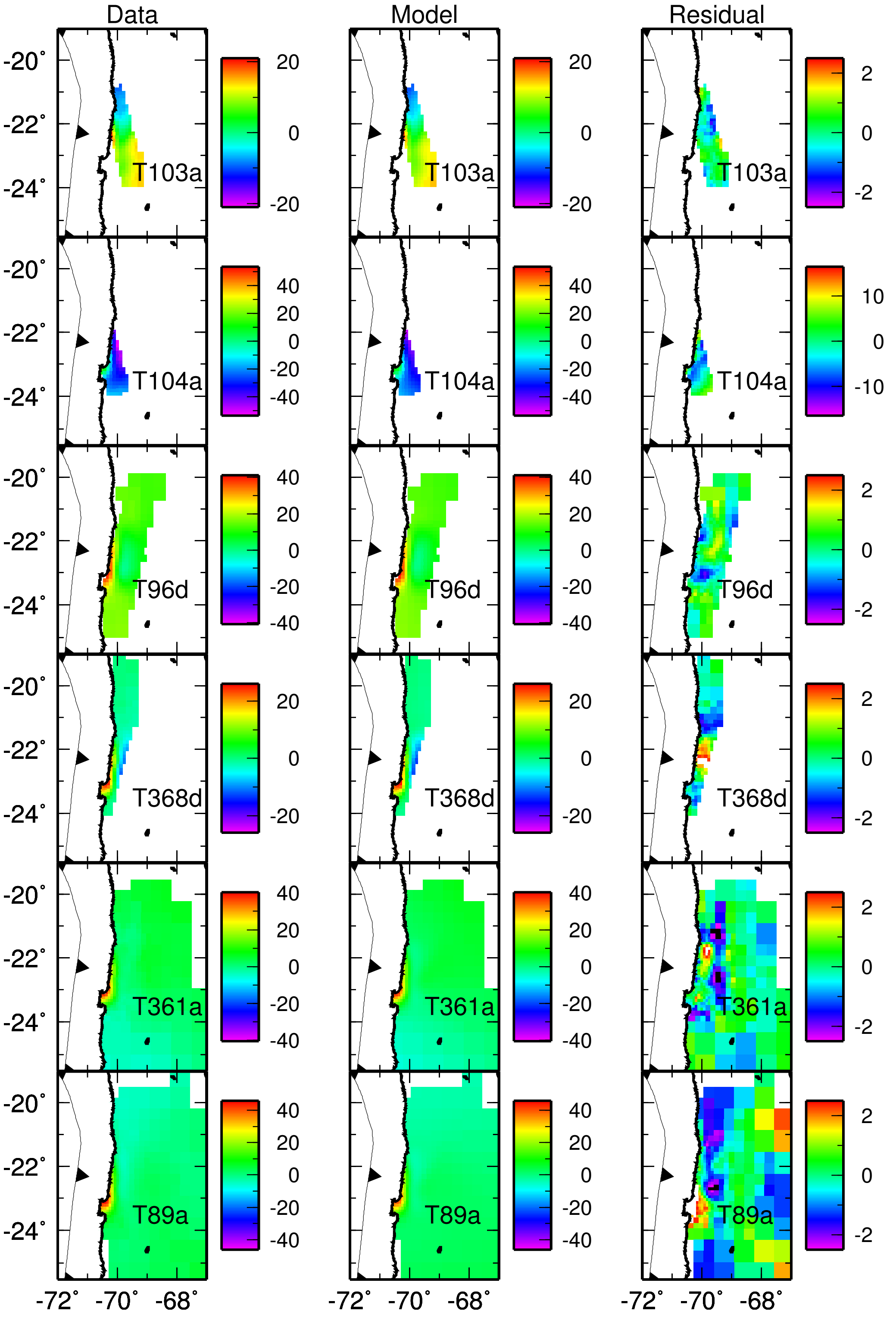


Figure 4.11: *InSAR predictions from the posterior distribution: Part I*: The predicted static displacements in cm from the MAP of the posterior distribution are compared to the InSAR data.

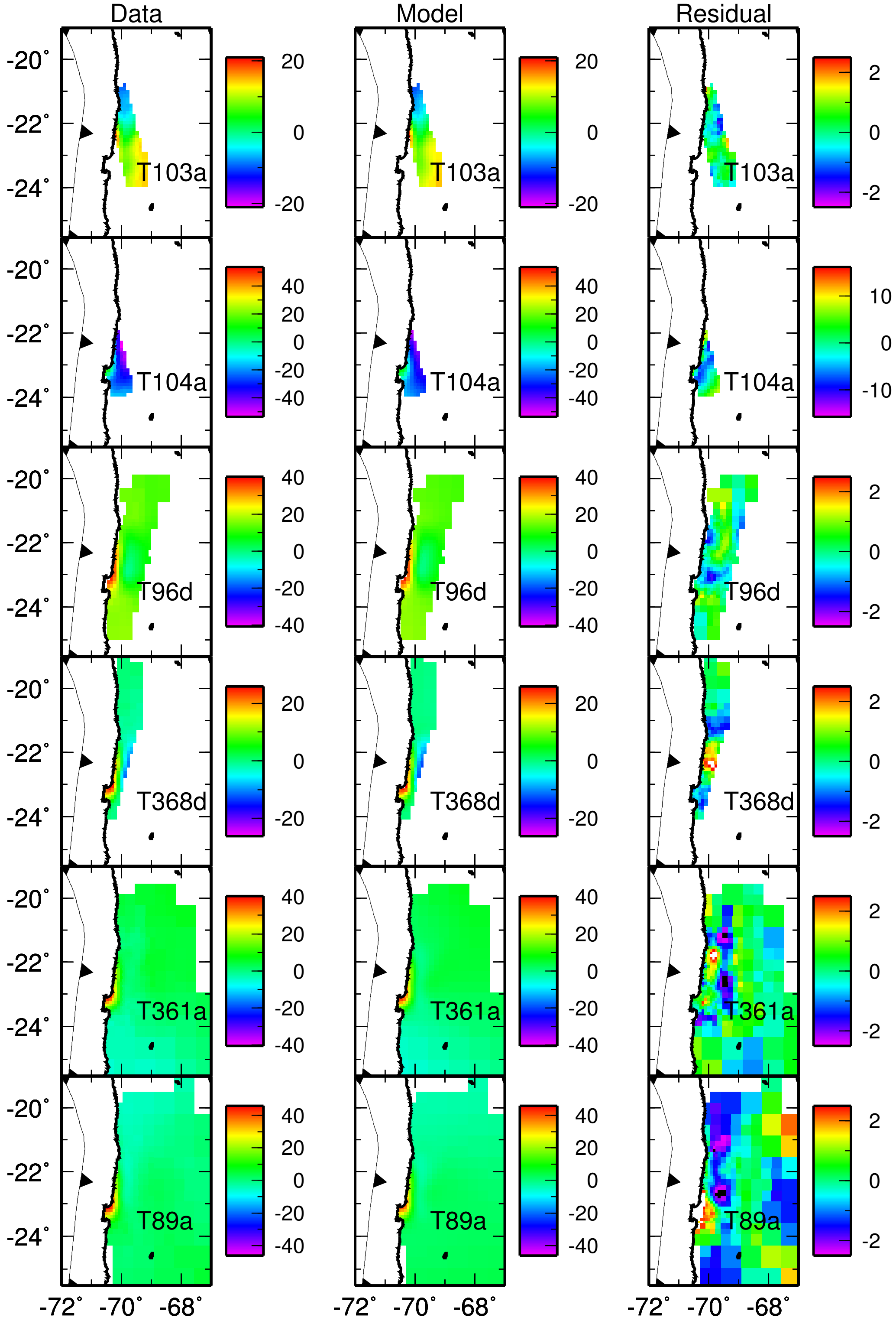


Figure 4.12: *InSAR predictions from the posterior distribution: Part II*: Same as Figure 4.11 for the mean of the posterior distribution.

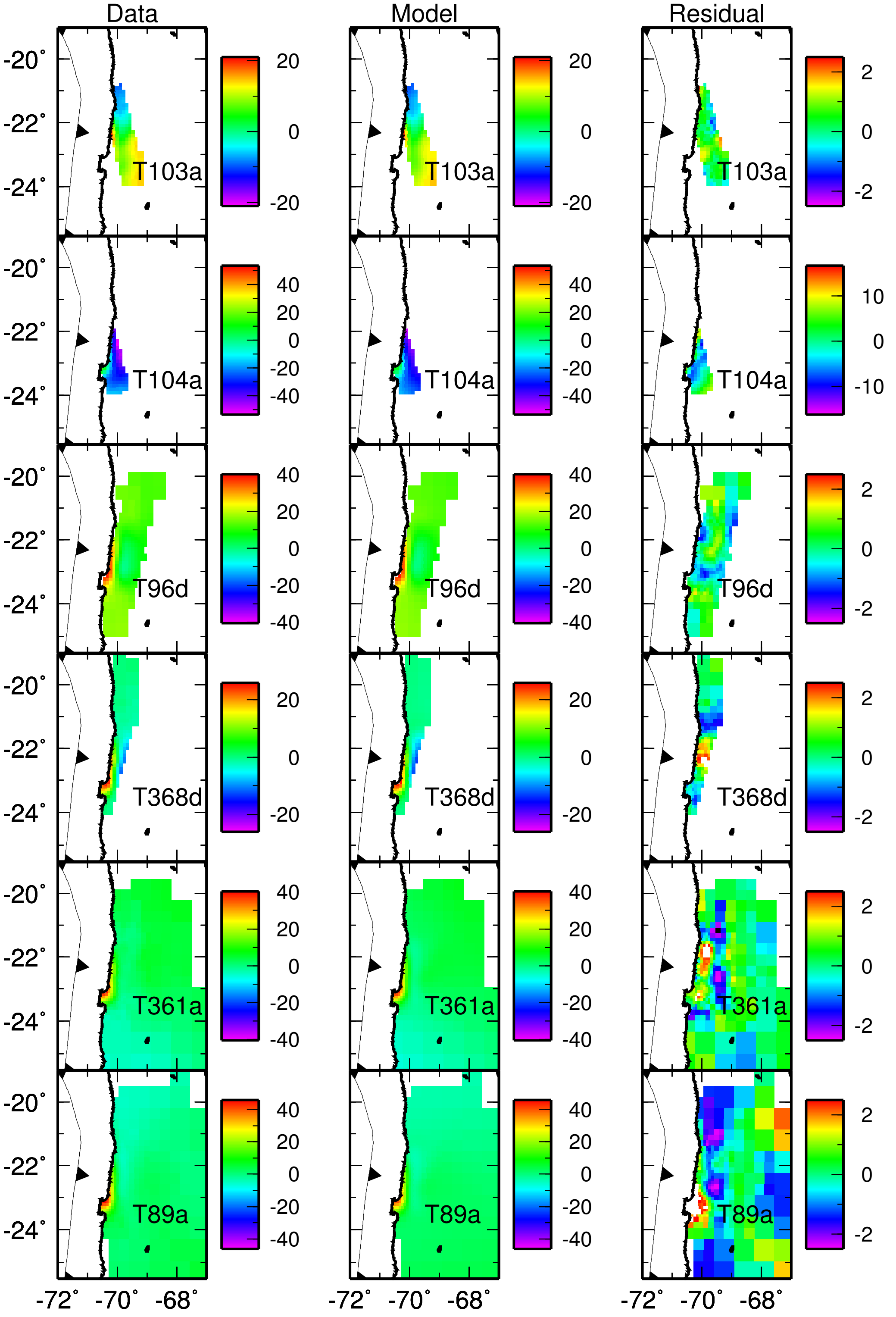


Figure 4.13: *InSAR predictions from the posterior distribution: Part III*: Same as Figure 4.11 for the median of the posterior distribution.

near zero, so the slip direction is clearly close to the a priori rake direction. Both components of slip have Gaussian-looking posteriors, explaining the similarity in the MAP, mean, and median models. However, the distribution on the $U_{\|}$ component of slip is broad, indicating that its value is not as well resolved, most likely due to trade-offs with neighboring patches. This brings us to our next logical question: What is the model covariance of our posterior distribution? I present the correlation between one patch and all other patches in Figures 4.16, 4.17, 4.18, and 4.19. We see that, for the $U_{\perp}$ component of slip, there is very little correlation between the given patch and its neighbors, while there is significant anti-correlation for the $U_{\|}$ component of slip. This may be why the posterior distribution on $U_{\perp}$ is tighter than the distribution on $U_{\|}$. Together, these observations show that the direction of slip is well-resolved, but the data have less power to resolve the magnitude of slip on each patch.

Finally, let us consider the complete distribution of 480 slip values as a whole. Figure 4.20 shows two contour lines. One contour circumscribes the region of the model space which contains 95% of all samples in the posterior distribution, the other contains 67% of all models. We can see that the rake direction is well-constrained. And we can also see that neighboring patches trade-off with each other leading to a range of possible slip amplitudes on each patch. But overall, the solution is well-constrained. The rupture area is definitely confined to a narrow band elongated in the north-south direction with the greatest slip at the southern end.

Figure 4.20 encapsulates almost the entire solution space. It shows the distribution of

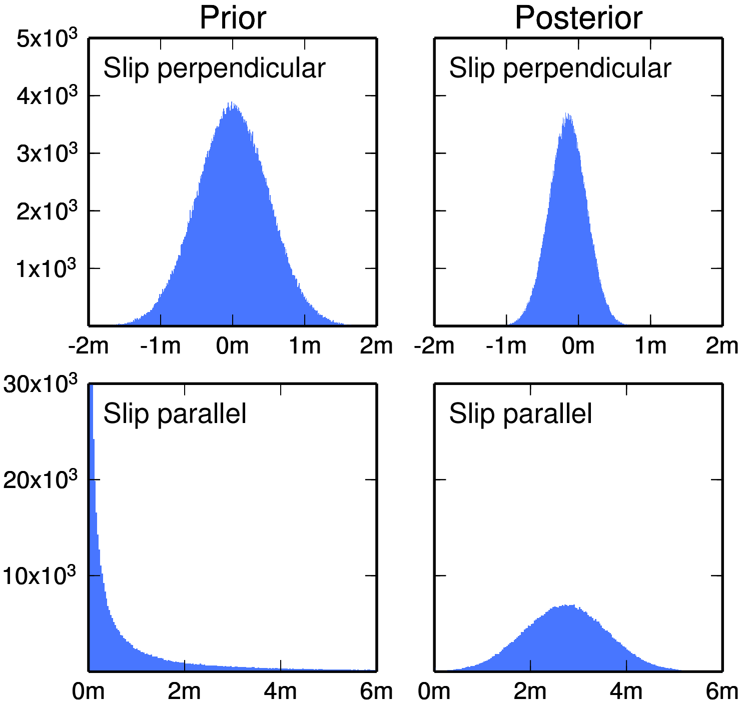


Figure 4.14: *Evolution of the posterior distribution*: Comparison of the prior slip distribution (left column) to the posterior distribution (right column) for the patch in Figure 4.17 (patch II in Figure 4.15).

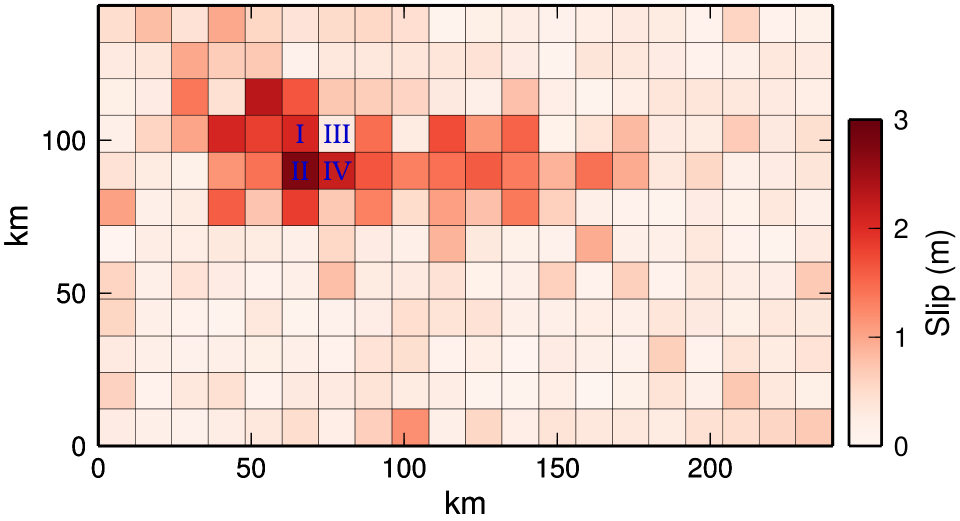


Figure 4.15: *Key to model correlation plots*: The model correlations with respect to each of the patches marked with Roman numerals are presented in Figure 4.16 through Figure 4.19.

slip on each patch, and thus the variance on each component slip and the covariance between the two components of slip on each path. (The patch-to-patch model covariance was explored in Figures 4.16, 4.17, 4.18, and 4.19). For each model parameter, it shows all possible values of that parameter which are consistent with the data.

Using surface observations to infer a slip distribution at depth is an under-determined process. Figure 4.20, with its uncertainties and trade-offs in slip, is the complete finite fault solution. A single slip model produced through optimization is just one sample from this solution space.

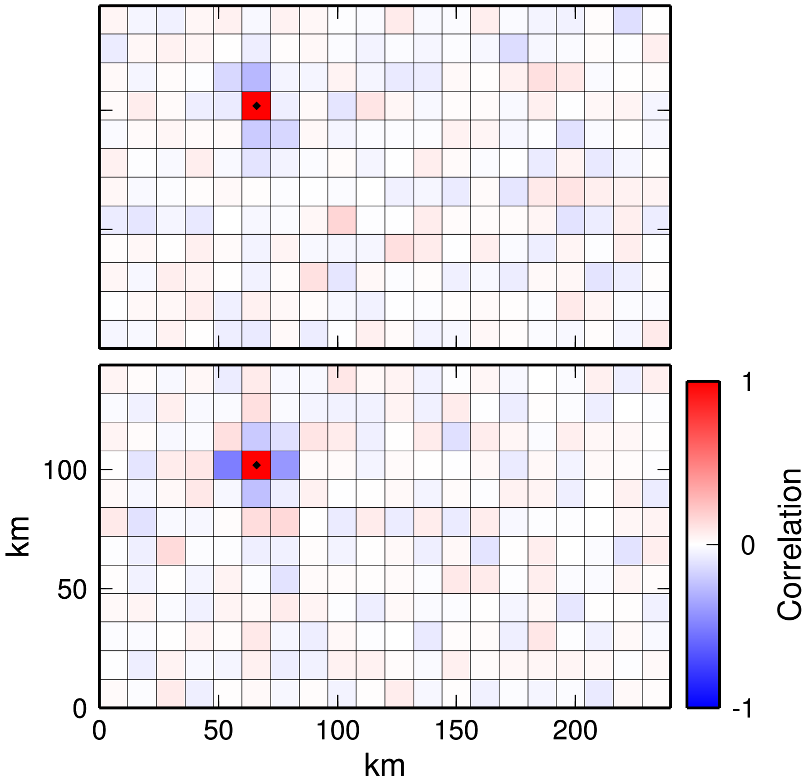


Figure 4.16: *Model correlation I*: The correlation between the patch marked with a diamond and all other patches: $U_{\perp}$ (top), $U_{\parallel}$ (bottom).

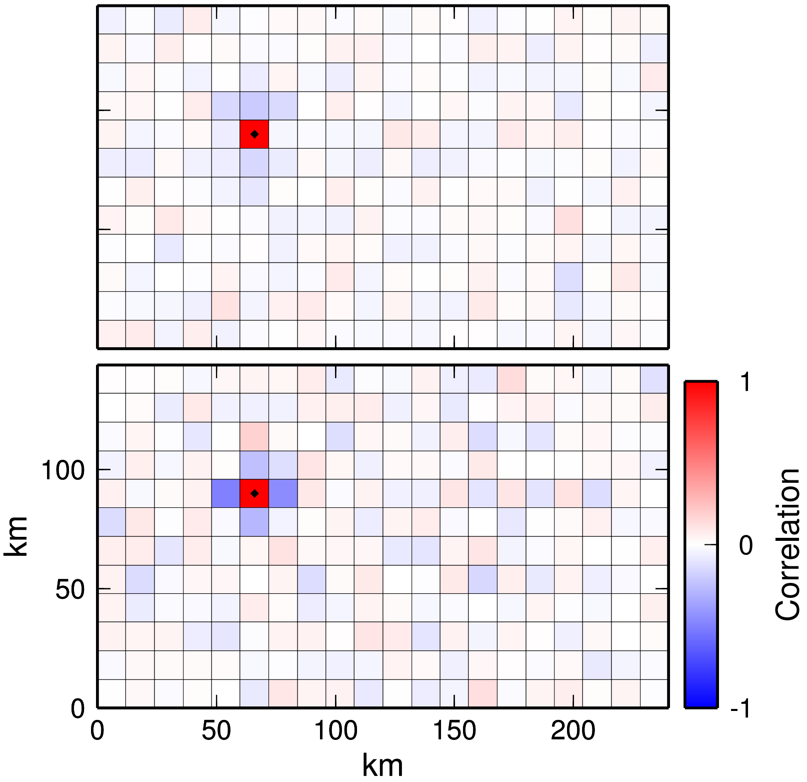


Figure 4.17: *Model correlation II*: The correlation between the patch marked with a diamond and all other patches: $U_\perp$ (top), $U_\parallel$ (bottom).

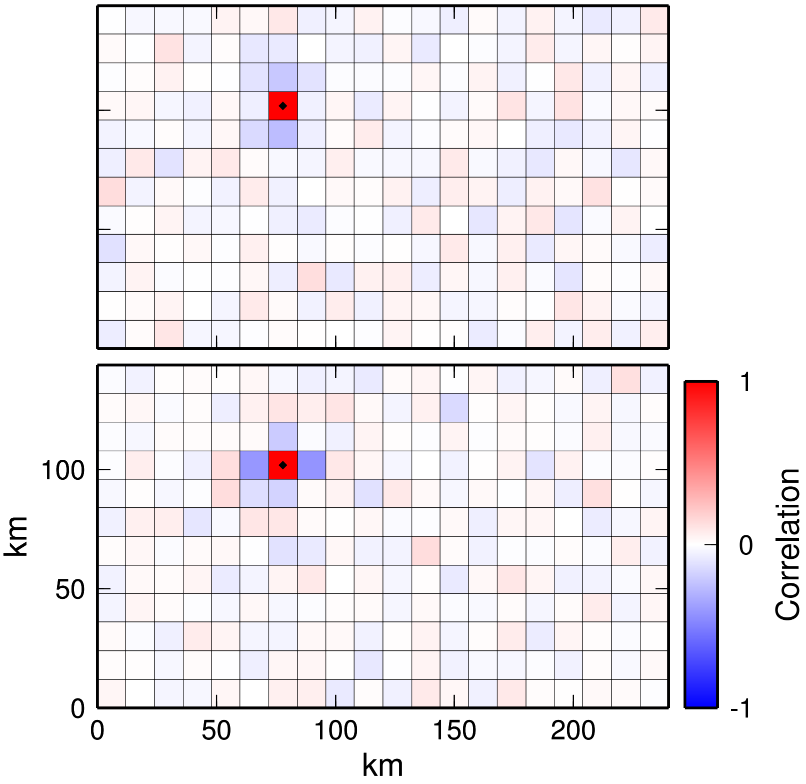


Figure 4.18: *Model correlation III*: The correlation between the patch marked with a diamond and all other patches: $U_\perp$ (top), $U_\parallel$ (bottom).

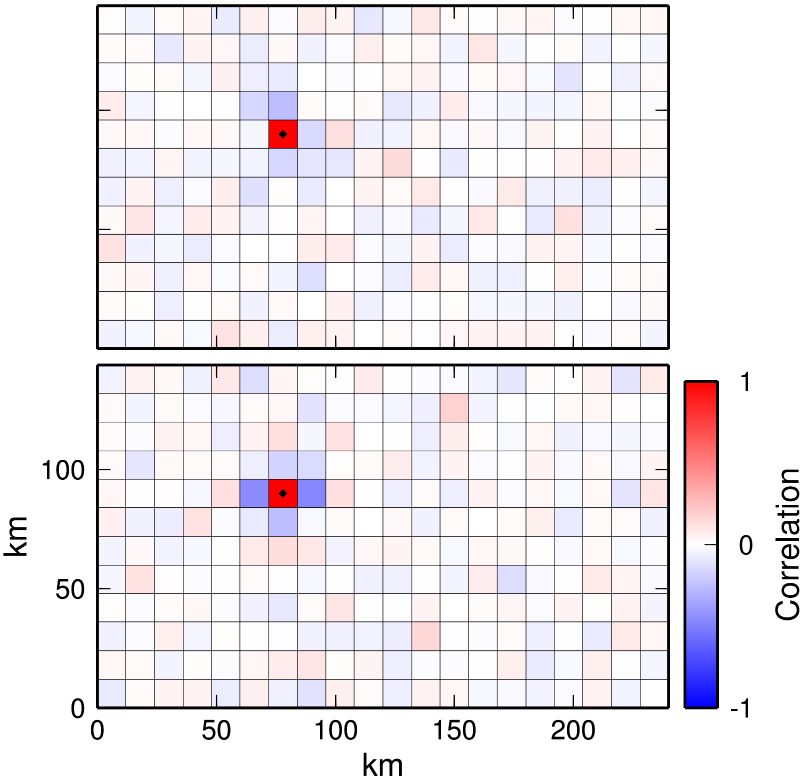


Figure 4.19: *Model correlation IV*: The correlation between the patch marked with a diamond and all other patches: $U_{\perp}$ (top), $U_{\parallel}$ (bottom).

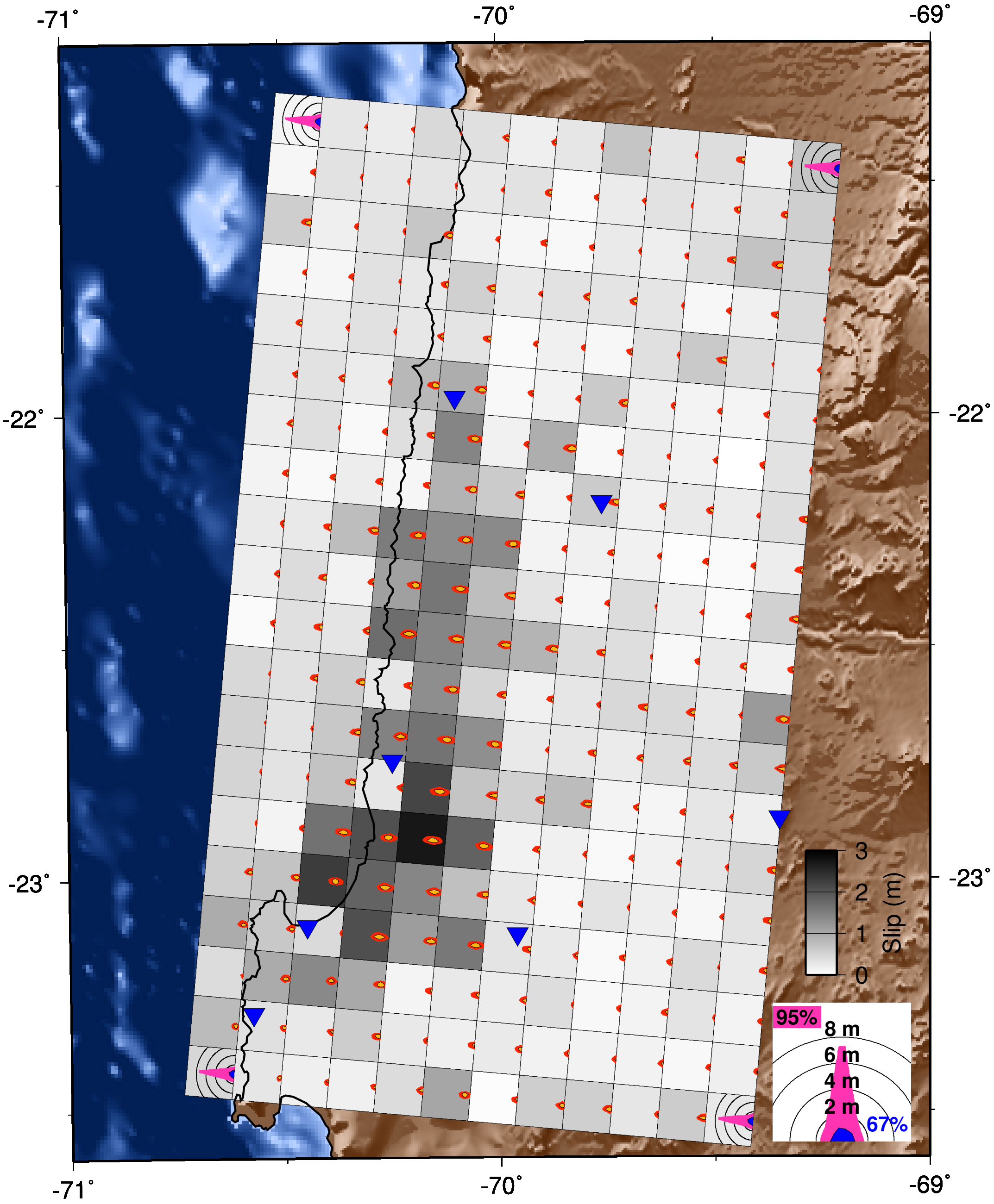


Figure 4.20: *Posterior slip distribution from static modeling*: Two contour lines are drawn for each patch. Each contour illustrates the range of slip values which contain 67% and 95% of all posterior models. The patches at the corners illustrate the range of slip values found in the prior distribution. The background grayscale intensity of each patch is the mean of the slip distribution for that patch. Locations of GPS stations are indicated by triangles.

4.4 Interlude: Conventional Kinematic Finite Fault Models

Before I delve into the Bayesian kinematic modeling of the Tocopilla earthquake, I want to show the types of source models that can be produced with existing methods. I present three models created with the simulated annealing optimization method of *Ji et al.* (2002) using teleseismic waveforms, static GPS offsets, and the two descending ENVISAT interferograms. The velocity model, data processing, and fault geometry are very similar to the Bayesian static model (Table 4.3, Figure 4.21). The solution is regularized to match a predetermined moment of $5.7 \cdot 10^{27}$ dyne*cm and the slip distribution is spatially smoothed. Spatial smoothing is necessary to eliminate solutions with oscillating slip distributions (*Harris and Segall*, 1987; *Du et al.*, 1992). The moment constraint is used because this joint inversion technique is prone to producing solutions with both reasonable rise times and rise times which are so long that the dislocations on those patches do not radiate seismically, and by this method the solution attempts to reconcile differences between the seismic and static data. But such a source process is physically unlikely assuming that both data sets are of a good quality and the static data is not contaminated by post-seismic deformation.

With optimization approaches, we spend a lot of time worrying about the amount of smoothing to use in the inversion scheme. Use too little smoothing, and the inversion will over-fit the data by producing a very rough slip model. Use too much smoothing, and the solution will under-fit the data. Our knowledge of the physics of the earthquake source process does not tell us a priori how rough the solution should be. Uniform slip is a perfectly valid source model. But so is a model with highly heterogeneous slip due to the material properties of the fault or the stress distribution on the fault, among other reasons. We may be able to discriminate between reasonable and unreasonable source models, but we cannot know how smooth the slip distribution of an earthquake should be. Thus another advantage of the Bayesian approach is that we do not need to specify a model roughness. We will sample all possible models including smooth models and rough models and we will determine the distribution of possible slip distributions that fit the data without fixing the roughness of the solution a priori.

Each of the three models presented here has different smoothing weights, and the choice of weights is arbitrary (Table 4.4). All three models produce good fits to the data (Figure 4.23 through Figure 4.34). But are these the appropriate data fits? Should we decrease the smoothing and increase the fit to the data?

All three inversions produce reasonable source models (Figures 4.35 and 4.36) and are consistent with published solutions (e.g. *Delouis et al.*, 2009; *Loveless et al.*, 2009). But each model has different amounts of slip at the edge of the fault. Are these features real or are they attempts by the numerical solver to fit complexity in the data that may be signal or may be noise? As discussed in the previous section, individual models may contain random model fluctuations that disappear when you look at the average of all possible models.

Table 4.3: *Fault geometry for simulated annealing.*

Parameter	**Prior distribution**
Strike	0°
Dip	18°
Epicenter	(-22.2018, -70.1411)
Source depth	52 km
Hypocenter (along strike, down-dip)	(139.5 km, 76.5 km)

Table 4.4: *Model regularization.*

Model	**Moment Constraint**	**Slip Smoothing**
1	0.3	0.1
2	0.3	0.02
3	0.1	0.02

This ambiguity illustrates why Bayesian analysis is so useful for finite fault modeling. With optimization methods, we have no way of evaluating the uncertainty associated with these solutions or finding which other models might fit the data. These three solutions are models which fit the data, and should be accepted as possible source-time histories for the Tocopilla earthquake. But we do not know if these solutions are representative of the whole solution space. If they are similar to each other, it may because because the data constrain the source process well, or it could be that all three solutions are from the same peak in a complex solution space with multiple peaks. If they are dissimilar, it could be because the data have little resolving power and the solution space is broad, or it could be because the solution space has many tight peaks and each model comes from a different region of the solution space. In order to image the solution space, we will have to create a fully Bayesian source model.

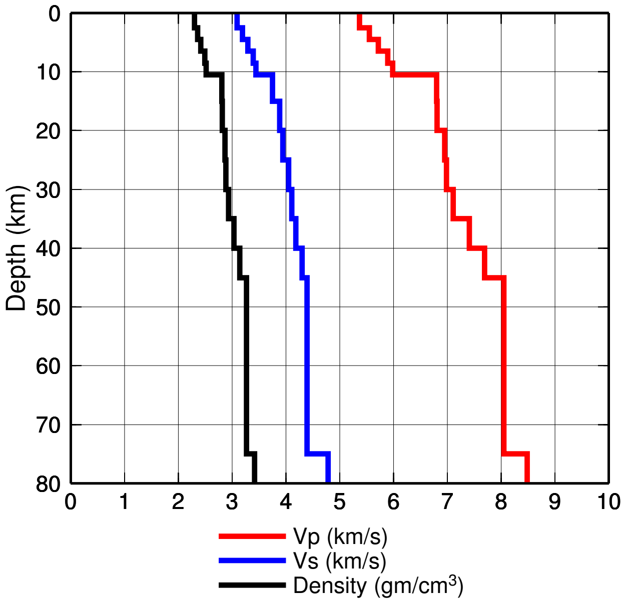


Figure 4.21: *Velocity model for simulated annealing*: v_p and v_s are a variation on the elastic structure from *Husen et al.* (1999). The P-wave and S-wave quality factors are $Q_\alpha = 1500$ and $Q_\beta = 600$, respectively.

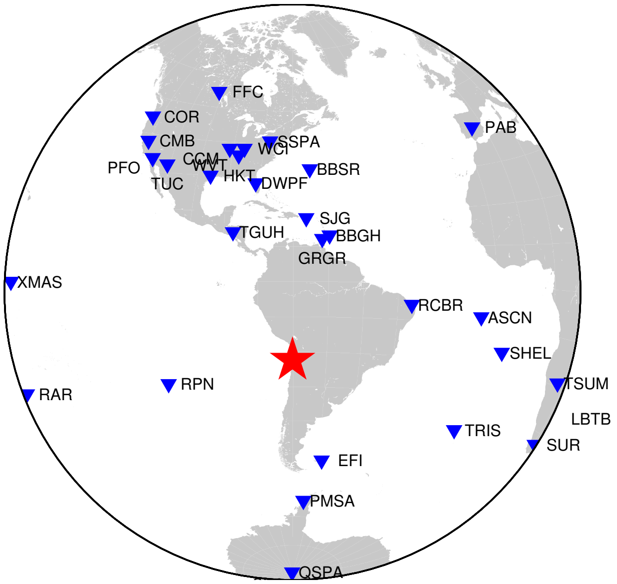


Figure 4.22: *Teleseismic station map*: These are the seismic stations used in simulated annealing modeling. Location of the Tocopilla earthquake is shown with a star.

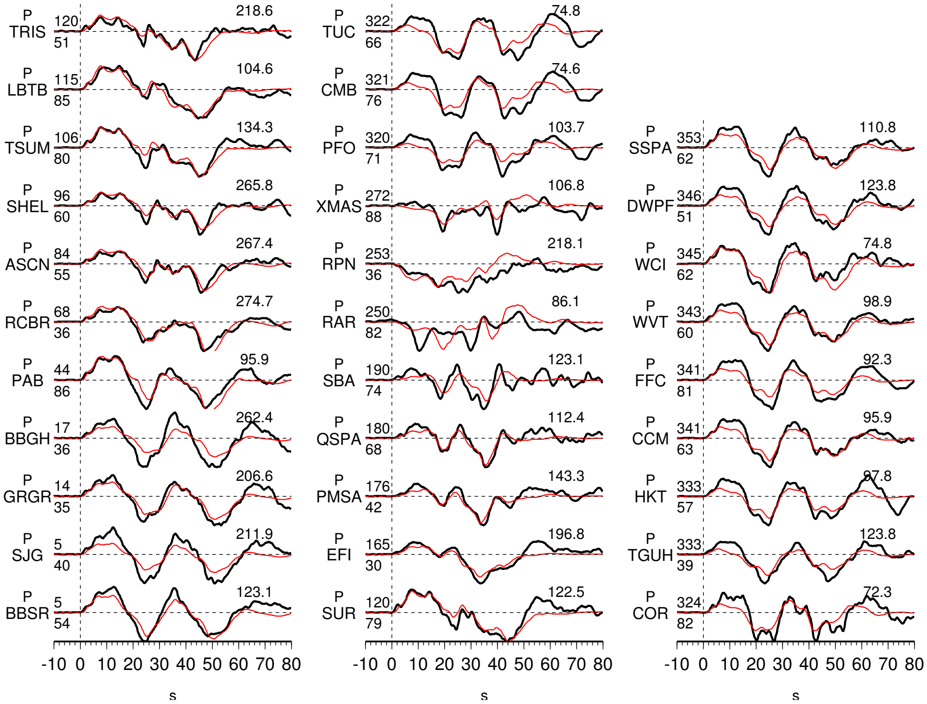


Figure 4.23: *Teleseismic P-wave fits for simulated annealing Model 1*: Data are plotted in black, synthetics in red. The bottom and top numbers on the left are the distance and azimuth of each station in degrees. To the right is the maximum displacement in microns.

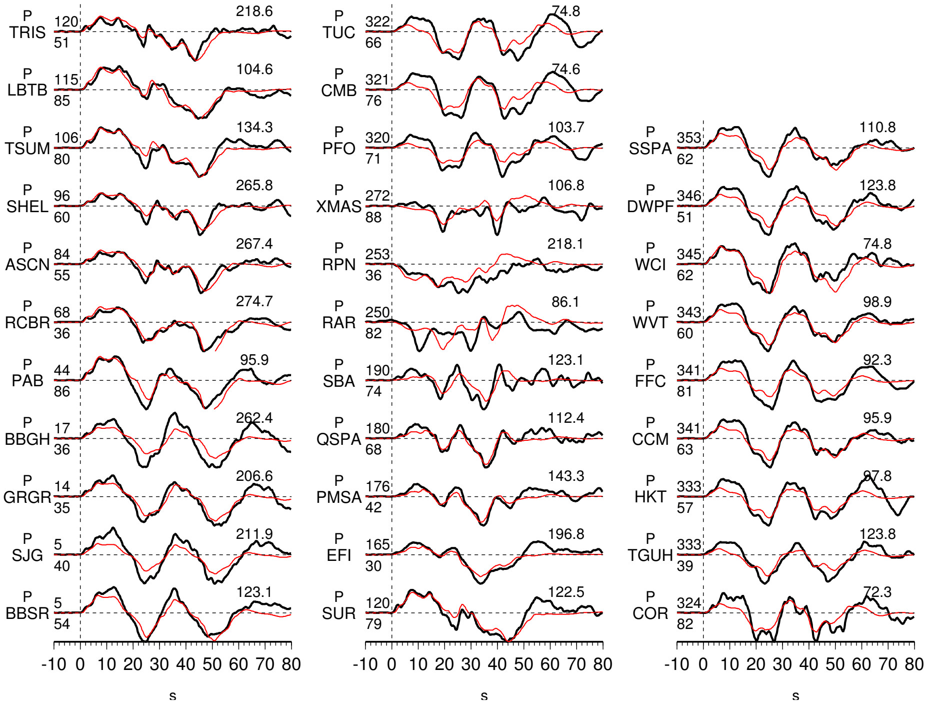


Figure 4.24: *Teleseismic P-wave fits for simulated annealing Model 2*: See Figure 4.23 for description.

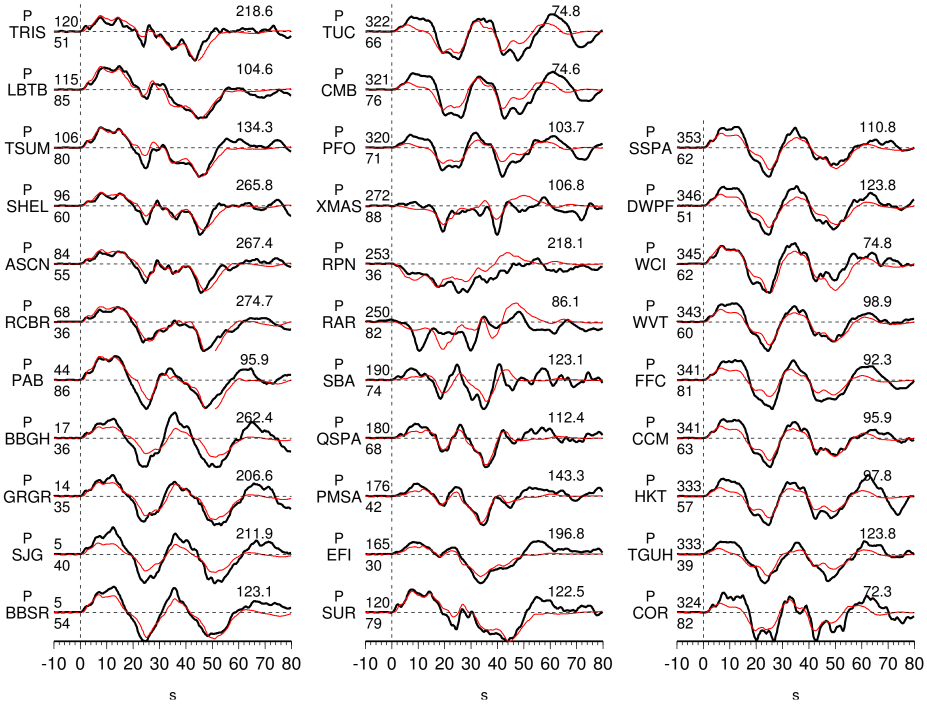


Figure 4.25: *Teleseismic P-wave fits for simulated annealing Model 3*: See Figure 4.23 for description.

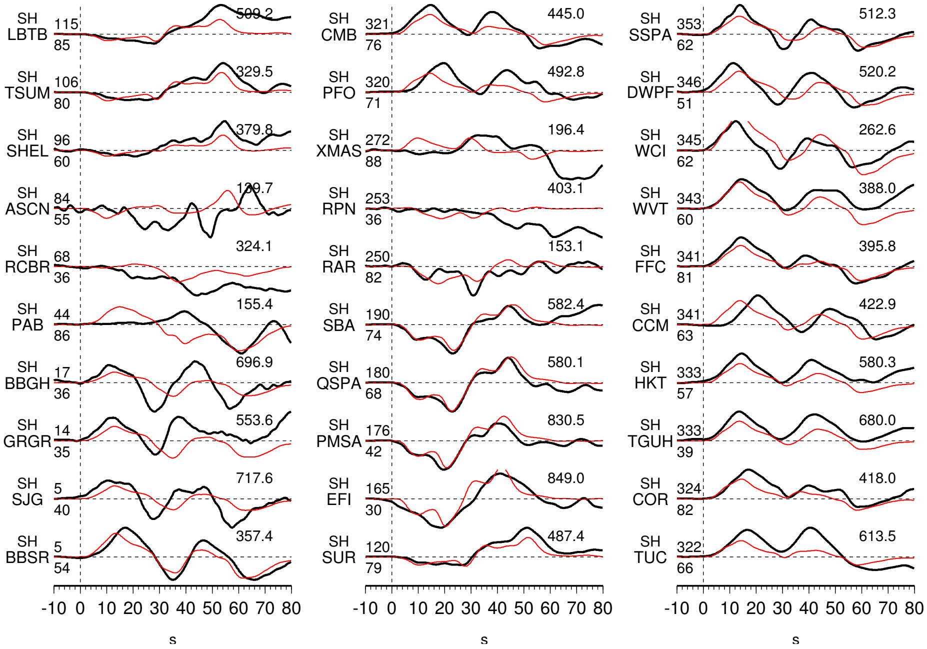


Figure 4.26: *Teleseismic SH-wave fits for simulated annealing Model 1*: See Figure 4.23 for description.

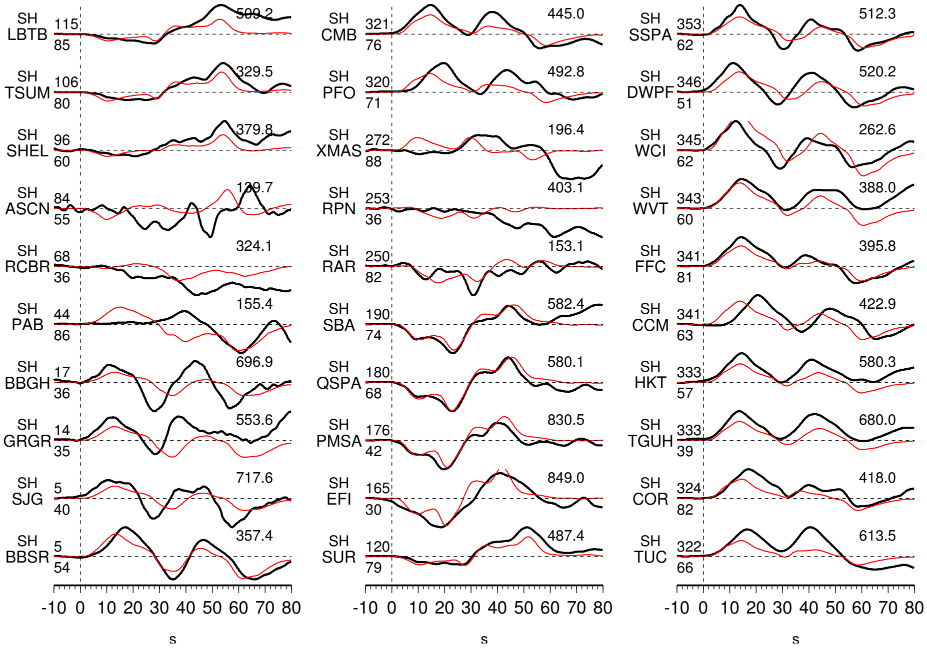


Figure 4.27: *Teleseismic SH-wave fits for simulated annealing Model 2*: See Figure 4.23 for description.

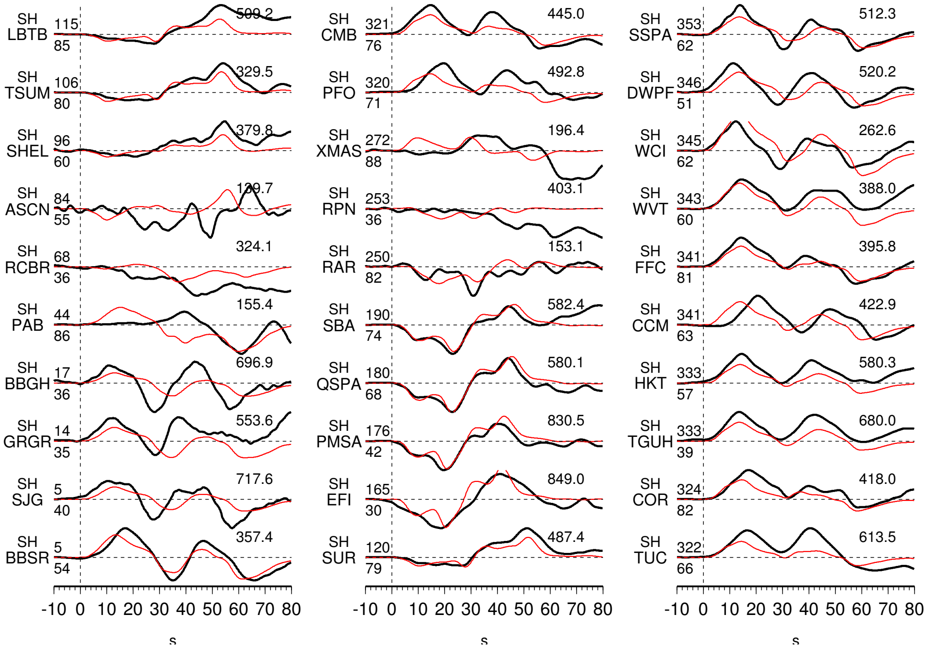


Figure 4.28: *Teleseismic SH-wave fits for simulated annealing Model 3*: See Figure 4.23 for description.

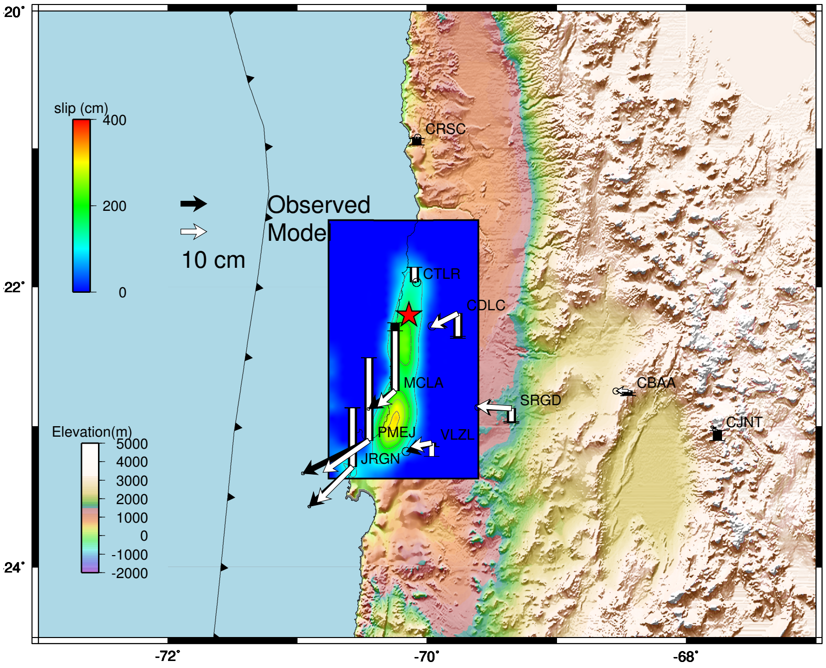


Figure 4.29: *GPS fits for simulated annealing Model 1*: Data are plotted in black. Synthetics are shown in white. Hypocenter is marked with a star.

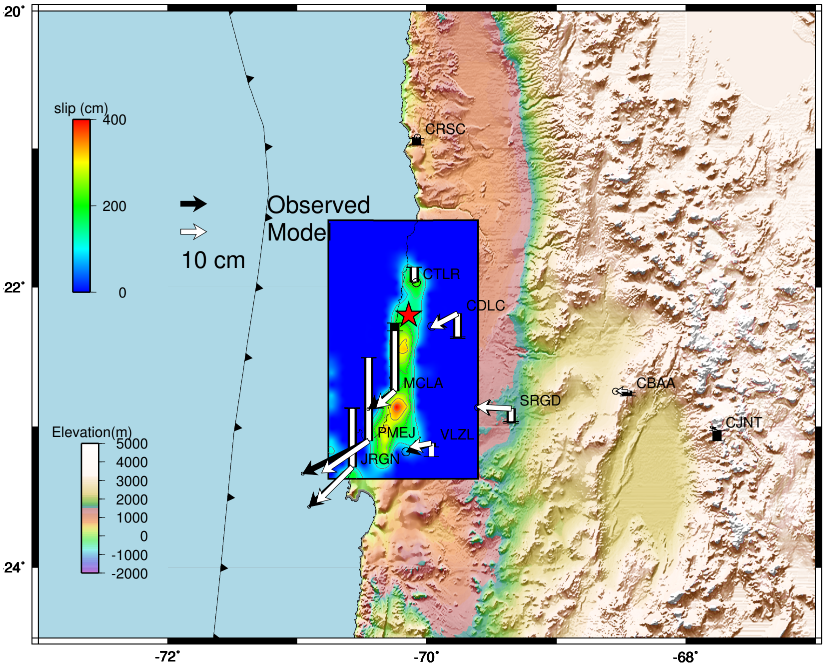


Figure 4.30: *GPS fits for simulated annealing Model 2*: Data are plotted in black. Synthetics are shown in white. Hypocenter is marked with a star.

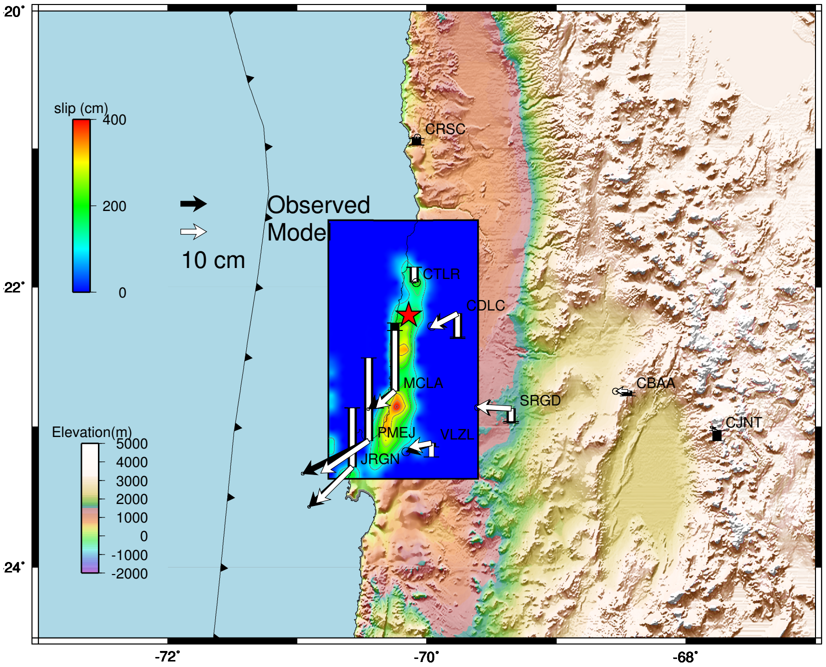


Figure 4.31: *GPS fits for simulated annealing Model 3*: Data are plotted in black. Synthetics are shown in white. Hypocenter is marked with a star.

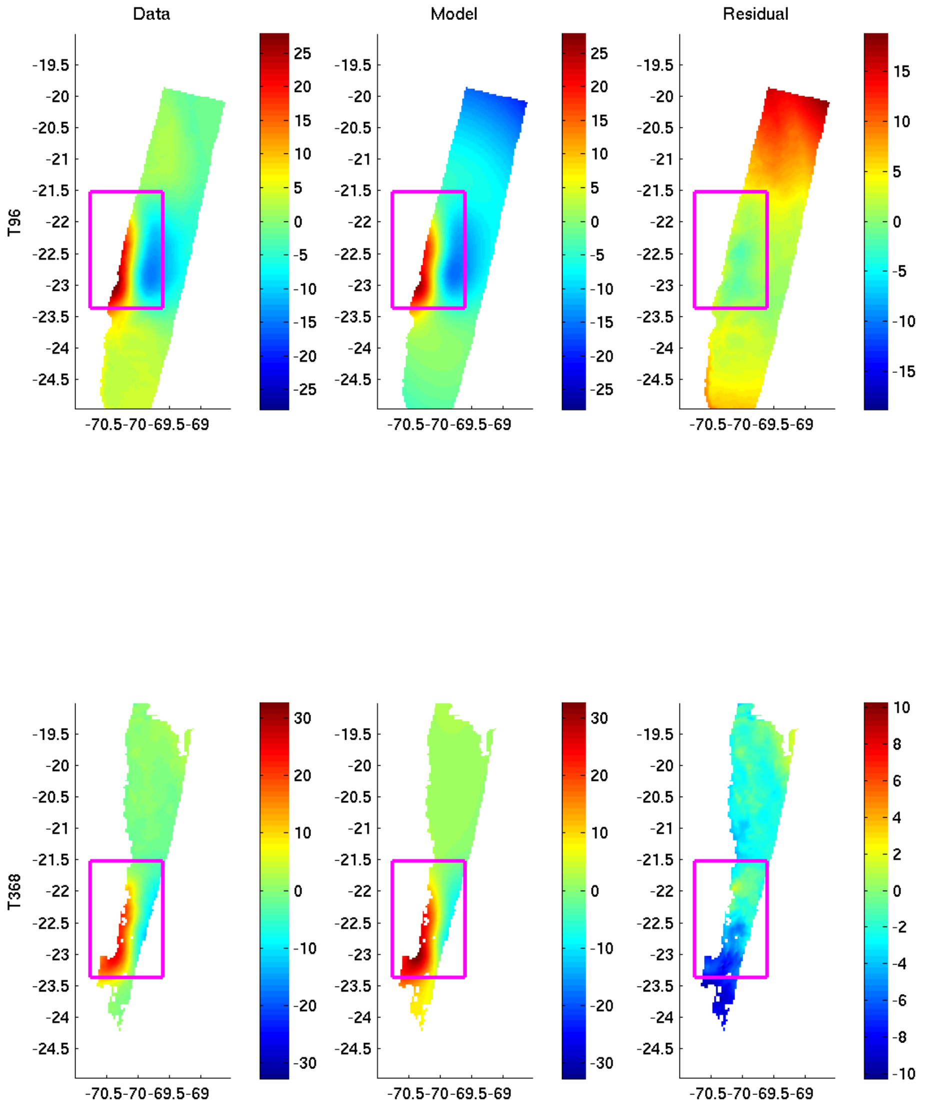


Figure 4.32: *InSAR fits for simulated annealing Model 1*: Displacements are in cm.

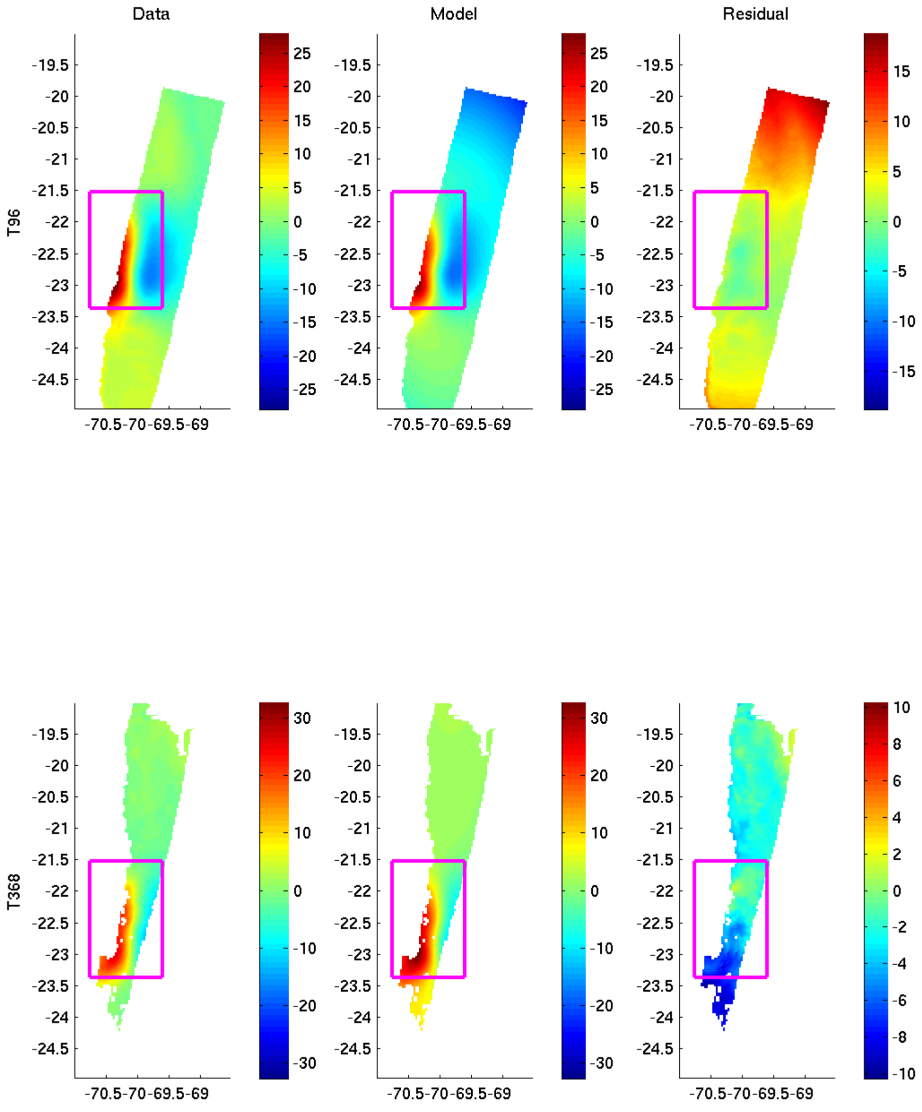


Figure 4.33: *InSAR fits for simulated annealing Model 2*: Displacements are in cm.

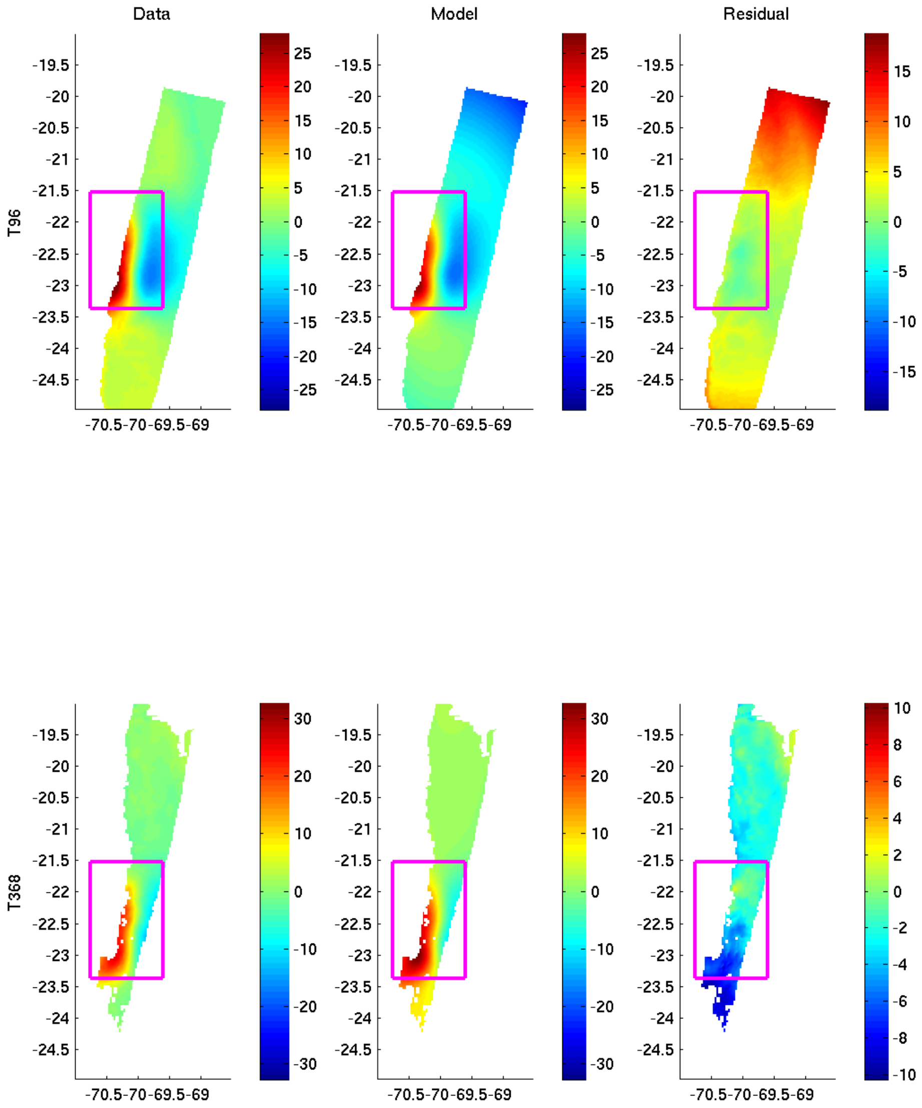


Figure 4.34: *InSAR fits for simulated annealing Model 3*: Displacements are in cm.

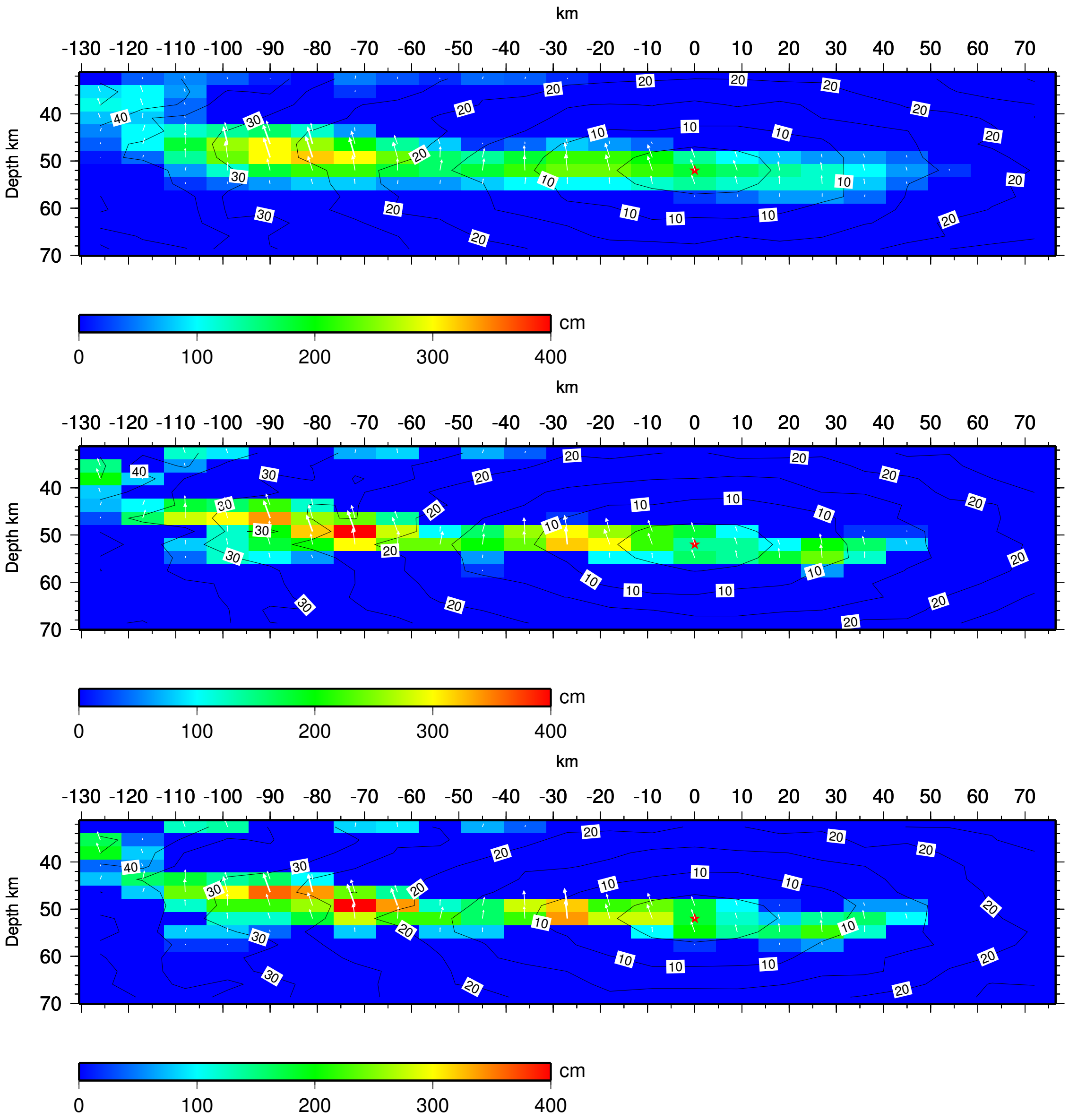


Figure 4.35: *Slip models from simulated annealing*: Slip is shown with color. Contour lines indicate location of the rupture front. The hypocenter is marked with a star. (Top) Model 1, (middle) Model 2, (bottom) Model 3.

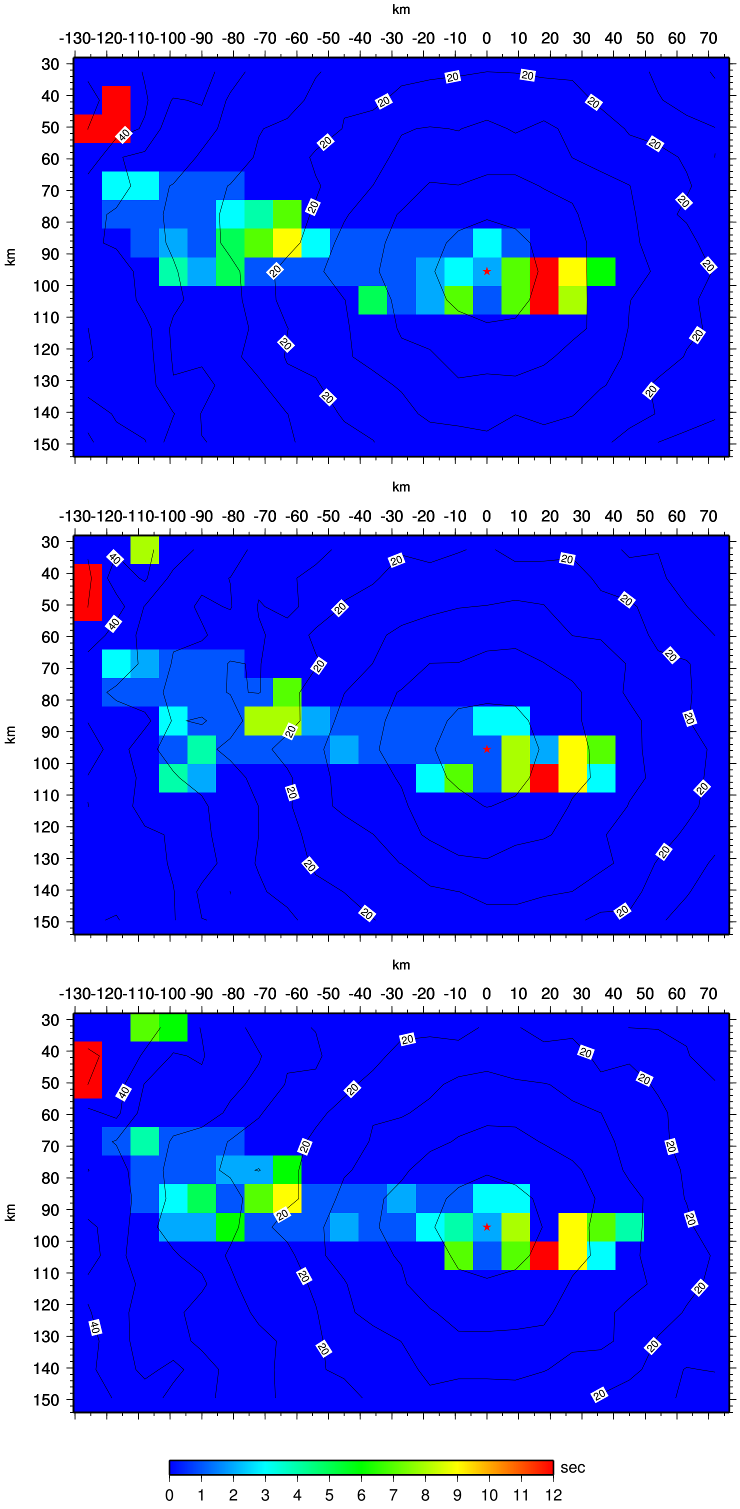


Figure 4.36: *Rupture evolution from simulated annealing*: Rise time is shown with color. Contour lines indicate location of the rupture front. The hypocenter is marked with a star. (Top) Model 1, (middle) Model 2, (bottom) Model 3.

4.5 Kinematic Model

We now come to the ultimate challenge: creating a fully Bayesian kinematic earthquake source model. I used the cascading technique described in Section 2.2 and Section 3.2 to produce a posterior distribution from kinematic GPS data, static GPS offsets, and six interferograms. The prior distributions are the same as in Table 4.2, but the samples of the prior slip distribution and InSAR nuisance parameters are given by the posterior distribution of the static model (Section 4.3).

Let us again start with individual samples from the posterior distribution. The maximum a posteriori (MAP), mean, and median of the posterior distribution are plotted in Figures 4.37 and 4.38. The temporal evolution of the rupture is shown in Figure 4.39, and the evolution of slip on each patch is shown in Figures 4.40, 4.41, and 4.42. The posterior slip distribution from joint static and kinematic modeling is very similar to that of the static model although slightly more compact. The rupture expands smoothly with any complexity occurring in regions without slip and thus little constraint on the source time history. The displacements from these three models are compared to the static data in Figures 4.43, 4.44, 4.45, and 4.46; the quality of fit is not worse than for the static-only model, which is not surprising considering how similar the joint slip distribution is to the static distribution. The synthetic time series from the MAP, mean, and median models are compared to the observed high-rate GPS displacements in Figure 4.47. Overall, the fits are excellent although the model fails to reproduce some of the finer details in the data; but it is unclear whether this is real signal or noise. This lack of high-frequency content is more pronounced for the mean and median models which are smoother than the MAP.

Let us now examine the probability distributions and answer the same questions we posed for the static model. How well is the distribution of model parameters on a given patch resolved? The answer is in Figures 4.48 and 4.49. Notice that because we have used the cascading method, the prior slip distribution is the posterior distribution from the static model while the prior distributions for rise time and rupture velocity are uniform distributions. Despite the already well-resolved static slip distribution, the addition of the kinematic data increases our model resolution because we have an abundance of near-field data. The joint posterior distribution on $U_{\|}$ is shifted relative to the prior distribution from the static model, but this change must be beyond the resolution of the static data since the joint posterior fits the static data well. For both rupture velocity and rise time, the posterior distribution is nearly Gaussian, relatively well-constrained, and significantly narrower than the prior distribution, which indicates that the model is well-resolved by the data.

The patch-to-patch correlation, or trade-off, is plotted in Figures 4.51, 4.52, 4.53, and 4.54. The only model parameters which show poor spatial resolution is the rake-parallel component of slip and, for some patches, the rupture velocity. For all other model parameters, the given patch is independent of its neighbors.

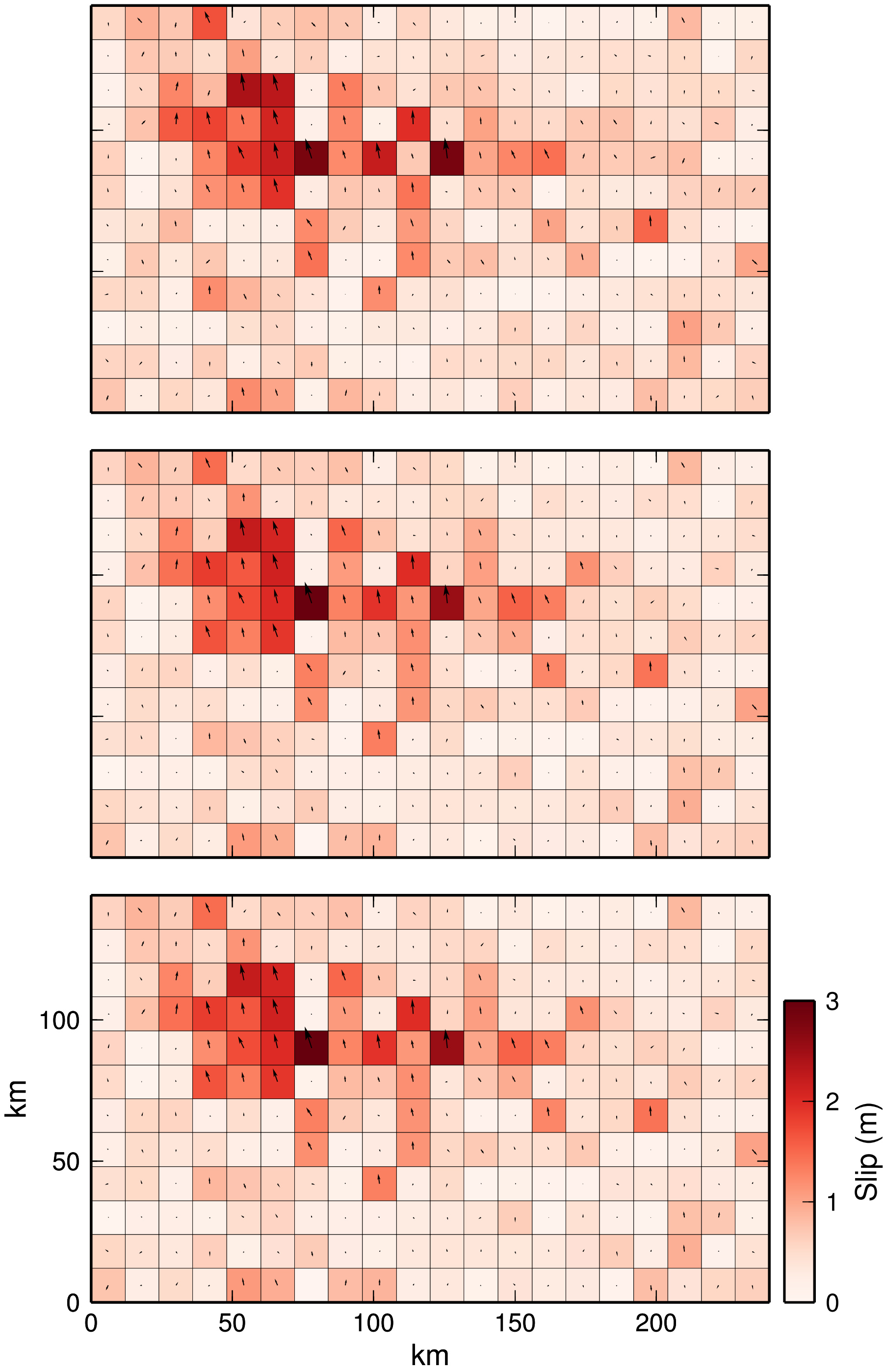


Figure 4.37: *Three models from the posterior slip distribution*: These are the MAP (top), mean (middle) and median (bottom) models of the posterior slip distribution. The magnitude of slip is shown with color and the arrows indicate the direction of slip.

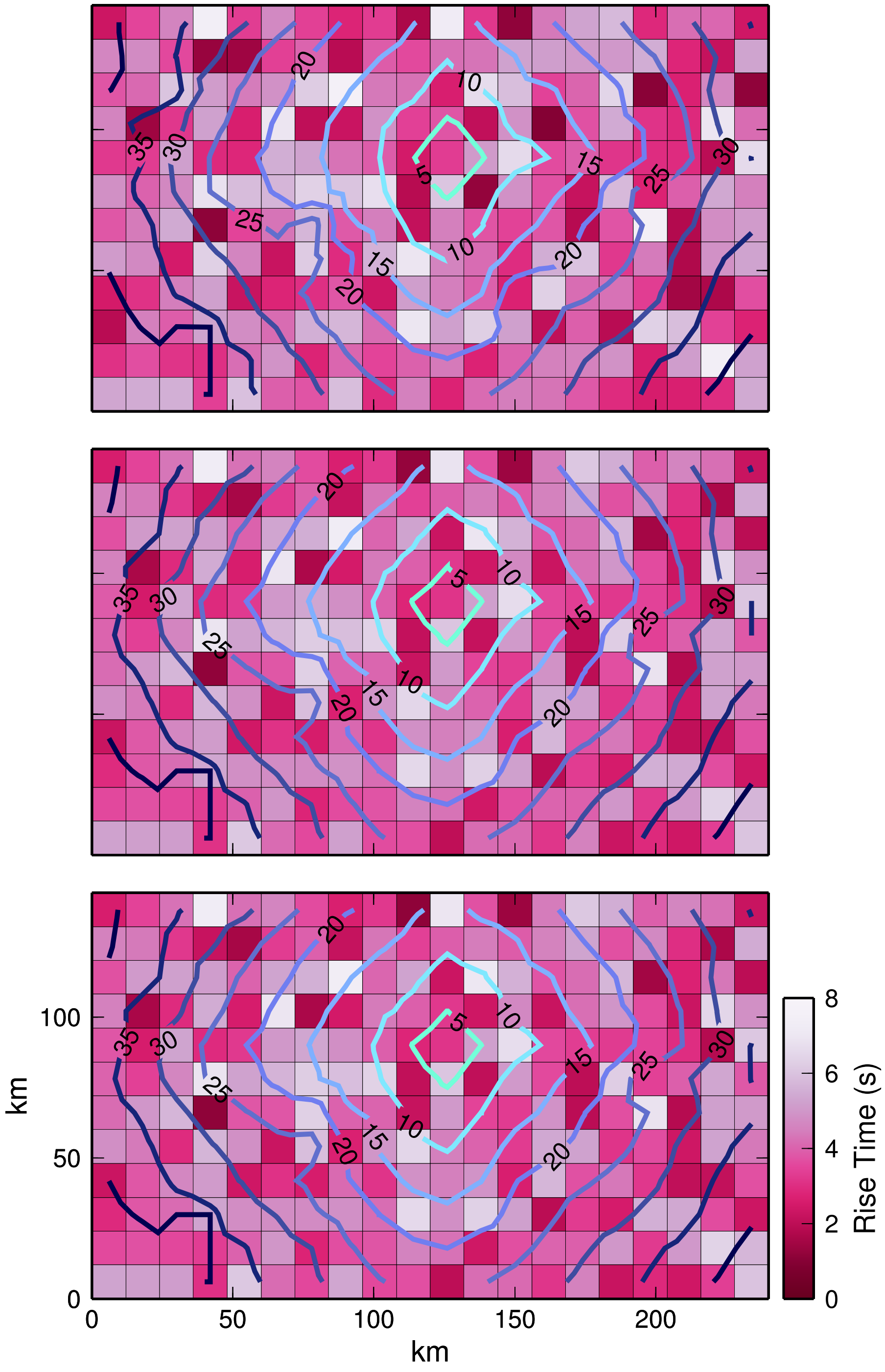


Figure 4.38: *Three models from the posterior kinematic distribution*: These are the MAP (top), mean (middle) and median (bottom) models of the posterior rupture velocity and rise time distributions. The rise time on each patch is shown in color. Contour lines illustrate the evolution of the rupture front; times are measured in seconds after the origin time.

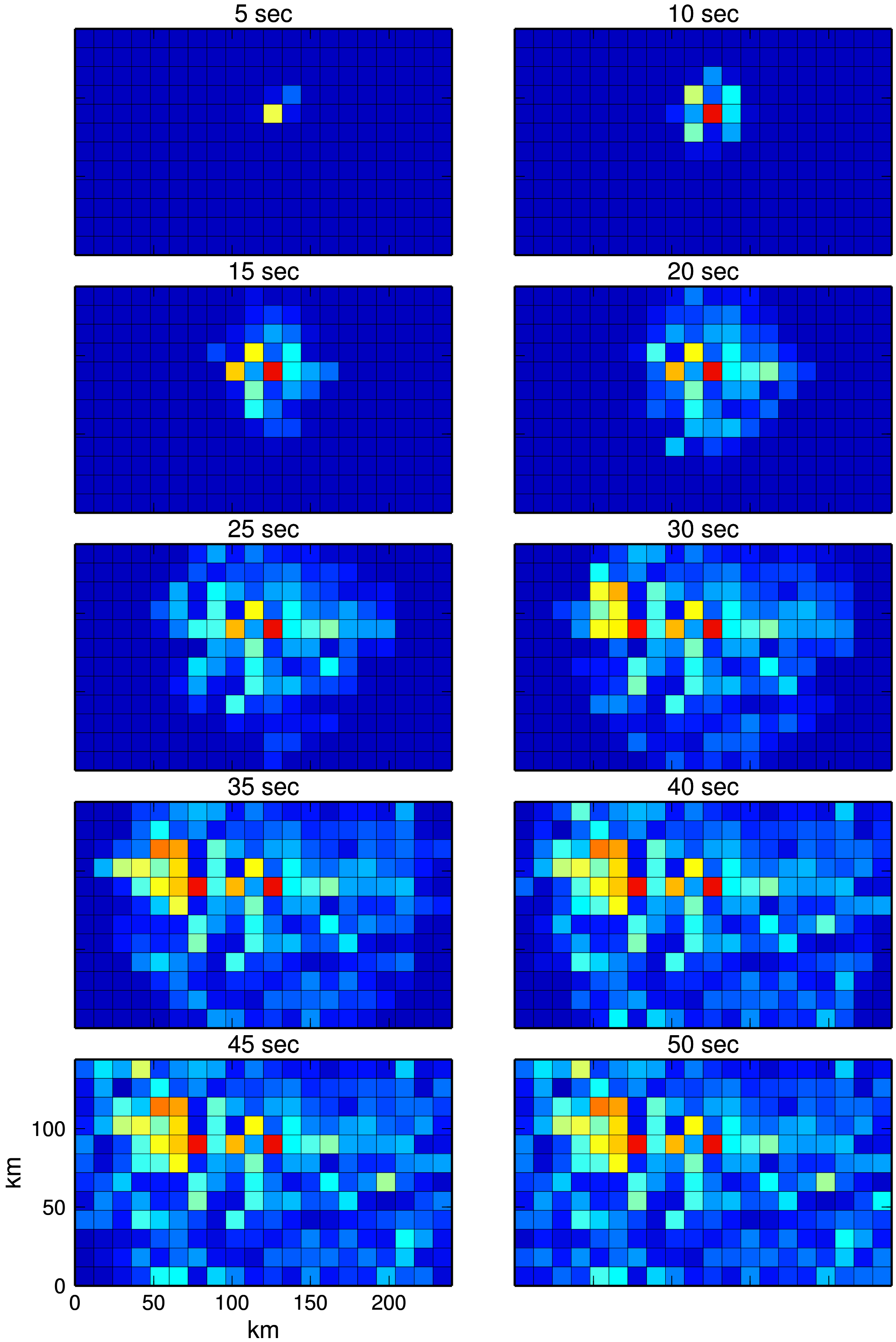


Figure 4.39: *Rupture evolution*: Snapshots of the MAP model are plotted for different points in time. All times are relative to the origin time.

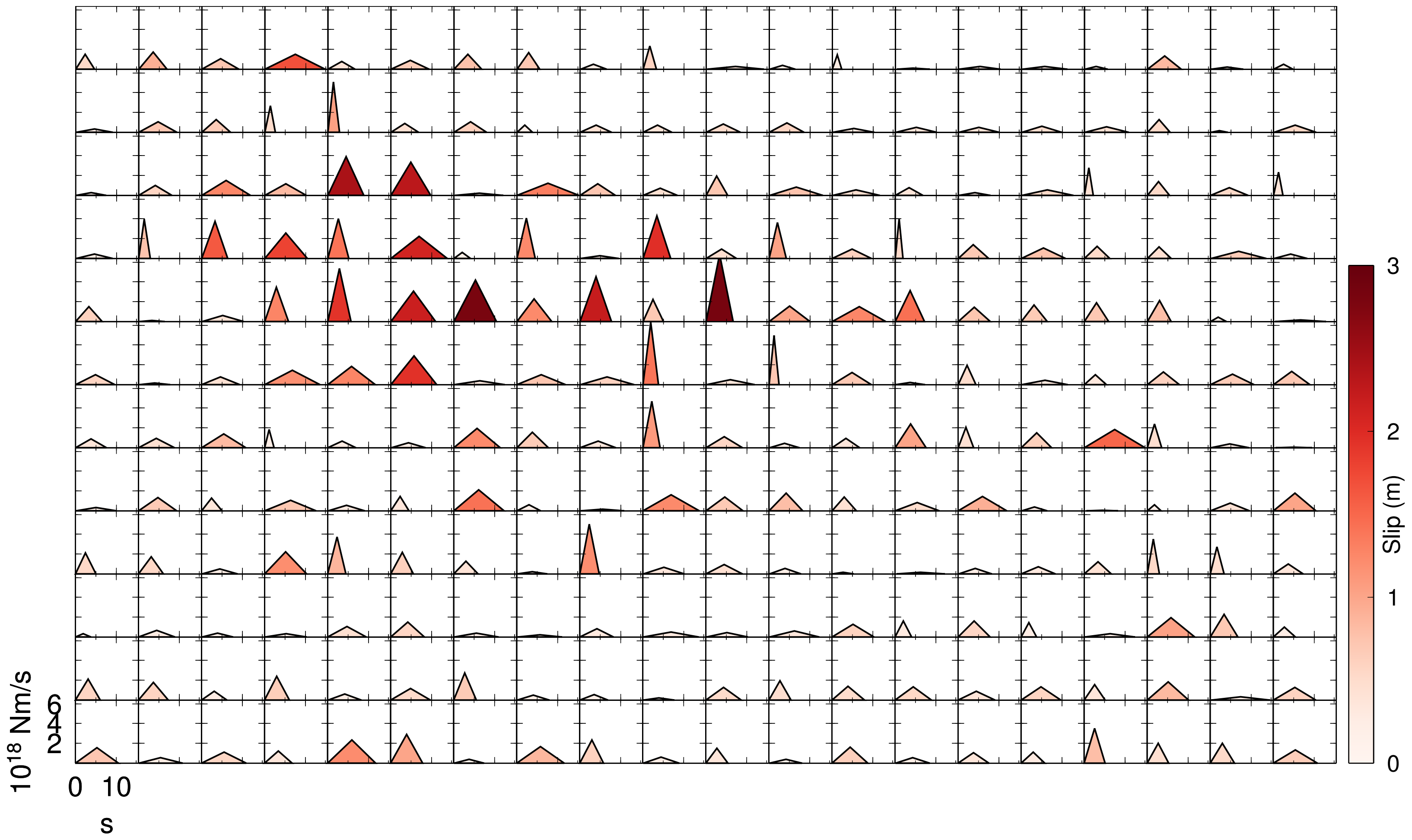


Figure 4.40: *Source-time function: Part I*: This the evolution of slip on each patch for the MAP of the posterior distribution. Time is relative to the initial rupture time on each patch.

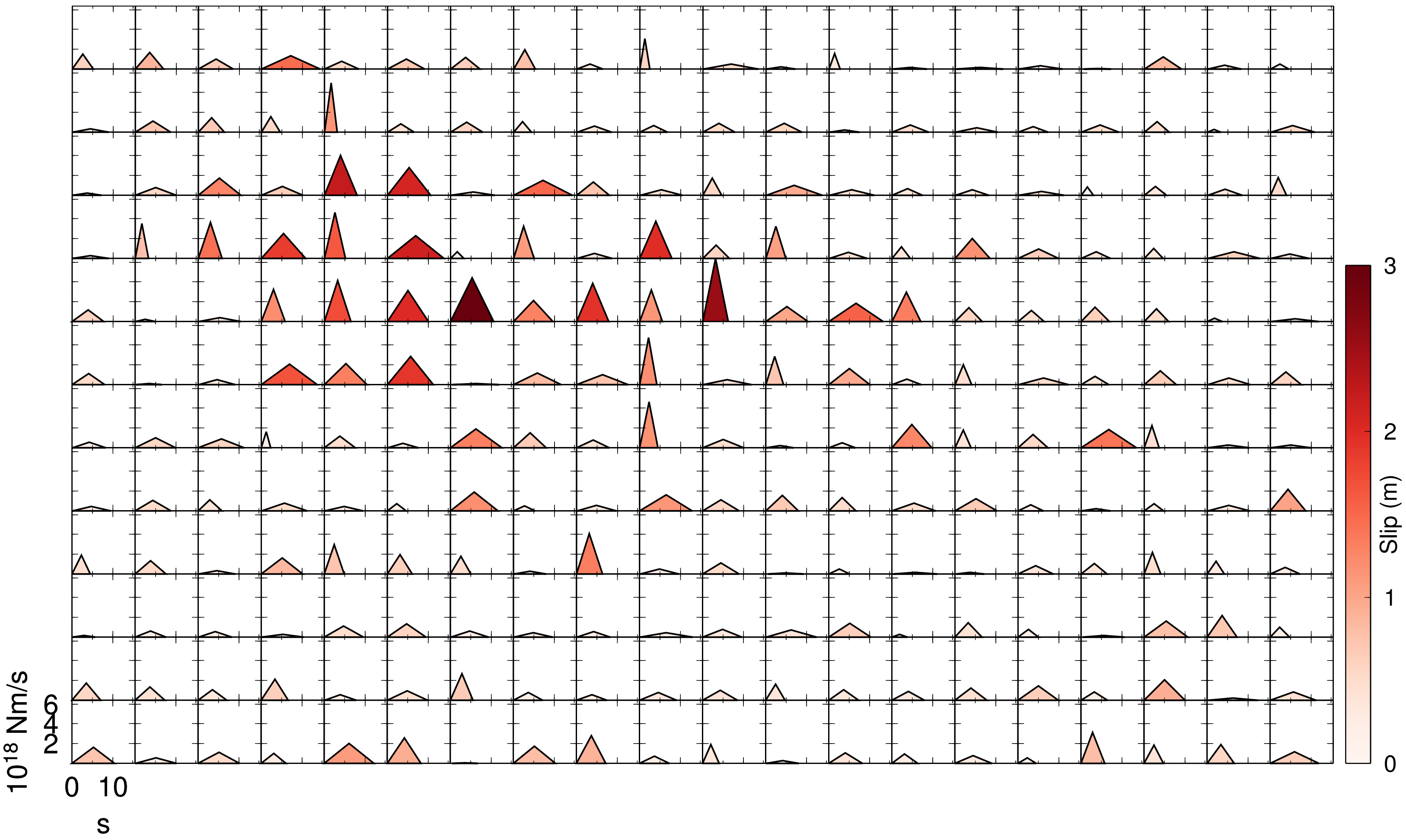


Figure 4.41: *Source-time function: Part II*: Same as Figure 4.40 for the mean of the posterior distribution.

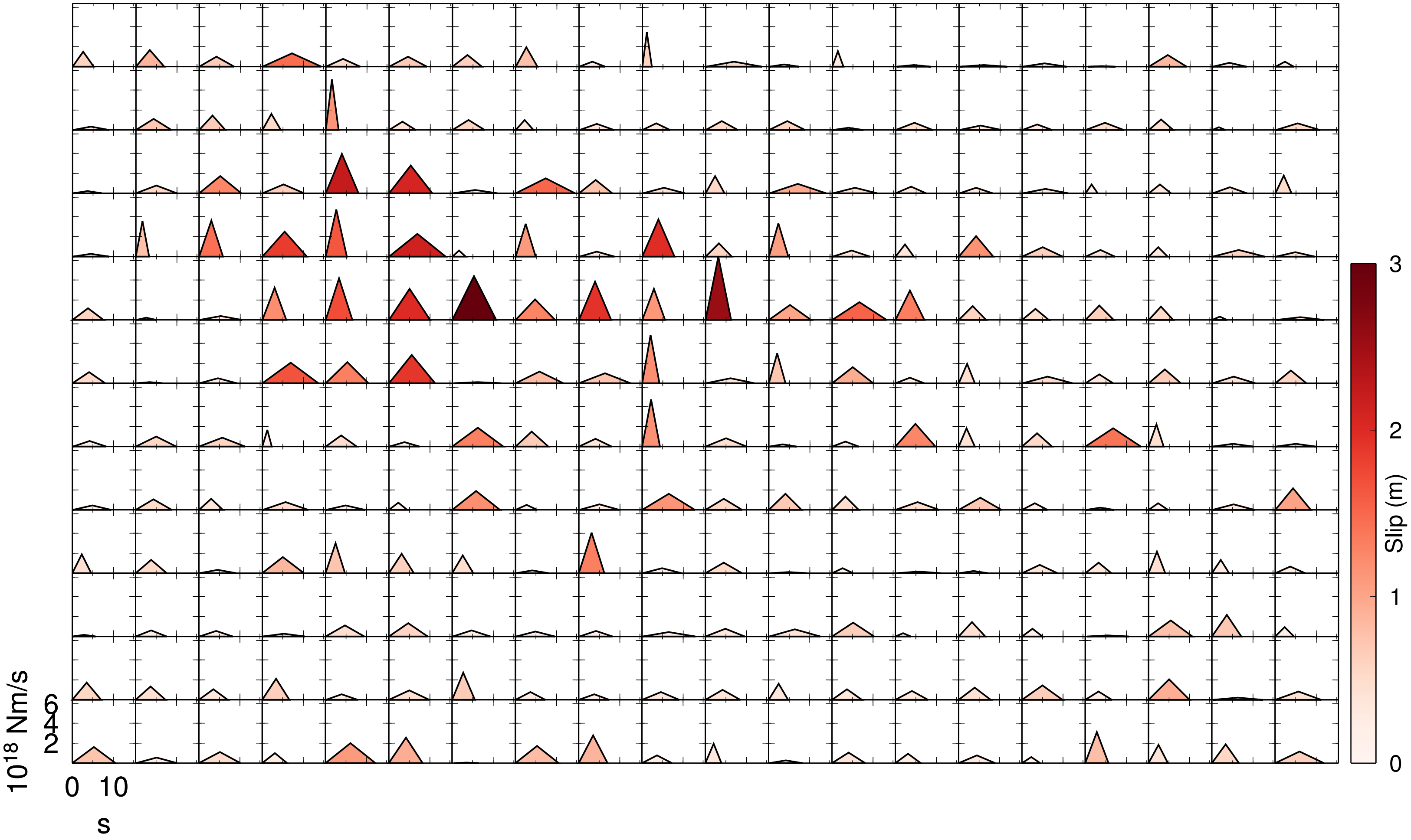


Figure 4.42: *Source-time function: Part III*: Same as Figure 4.40 for the median of the posterior distribution.

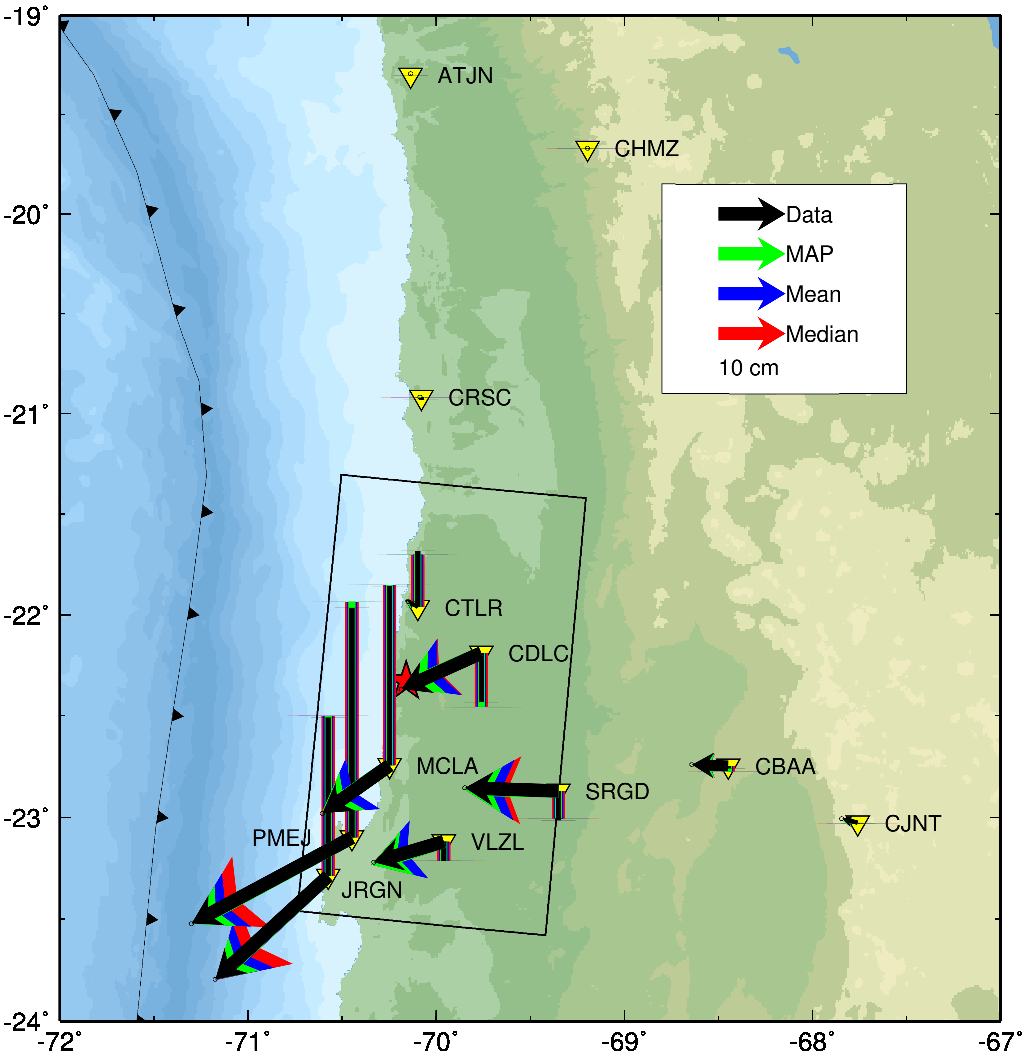


Figure 4.43: *GPS predictions from three models*: The predicted static displacements from the MAP, mean, and median of the posterior distribution are compared to the static GPS data. The surface projection of the fault plane is shown with a thick black line.

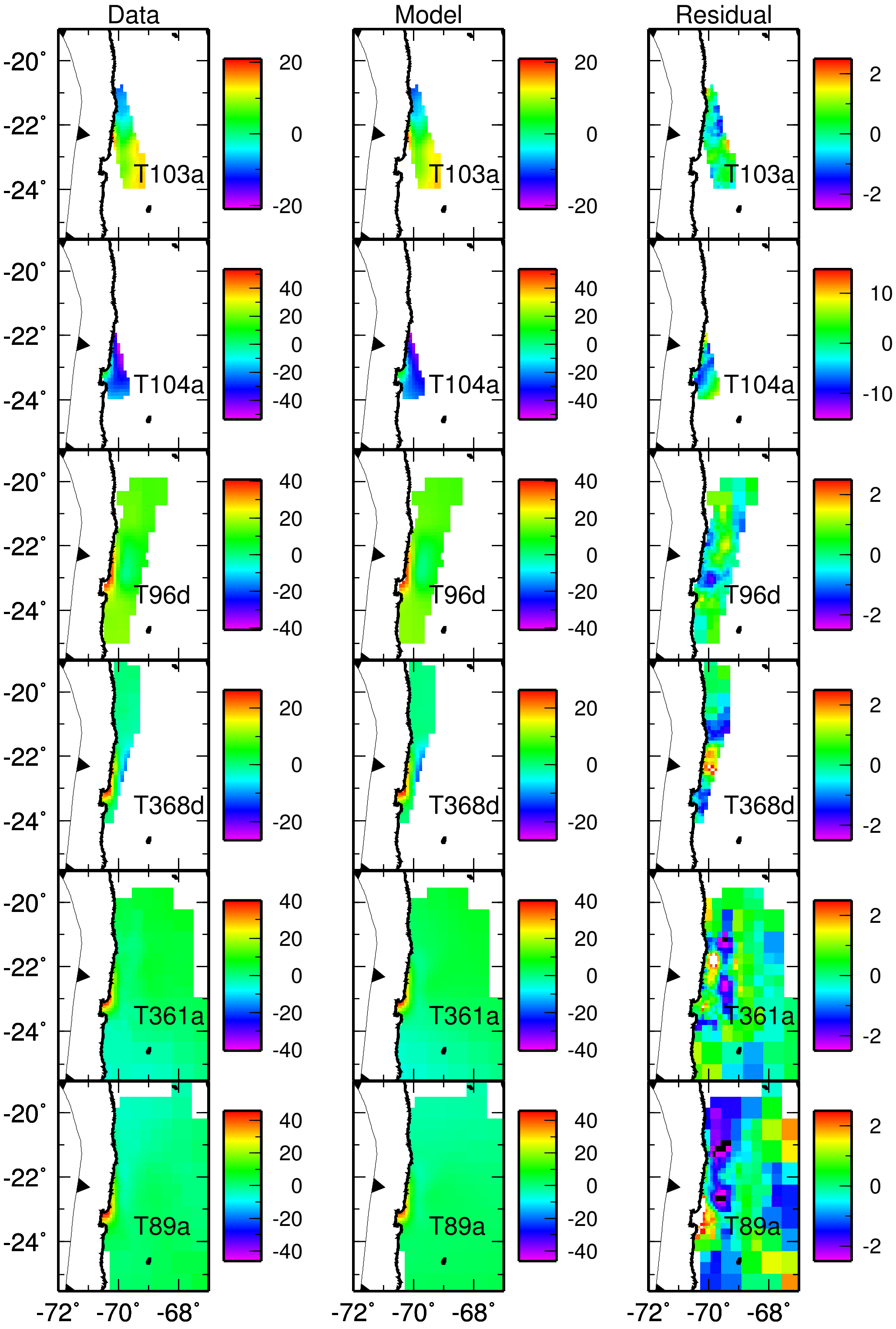


Figure 4.44: *InSAR predictions from the posterior distribution: Part I*: The predicted static displacements in cm from the MAP of the posterior distribution are compared to the InSAR data.

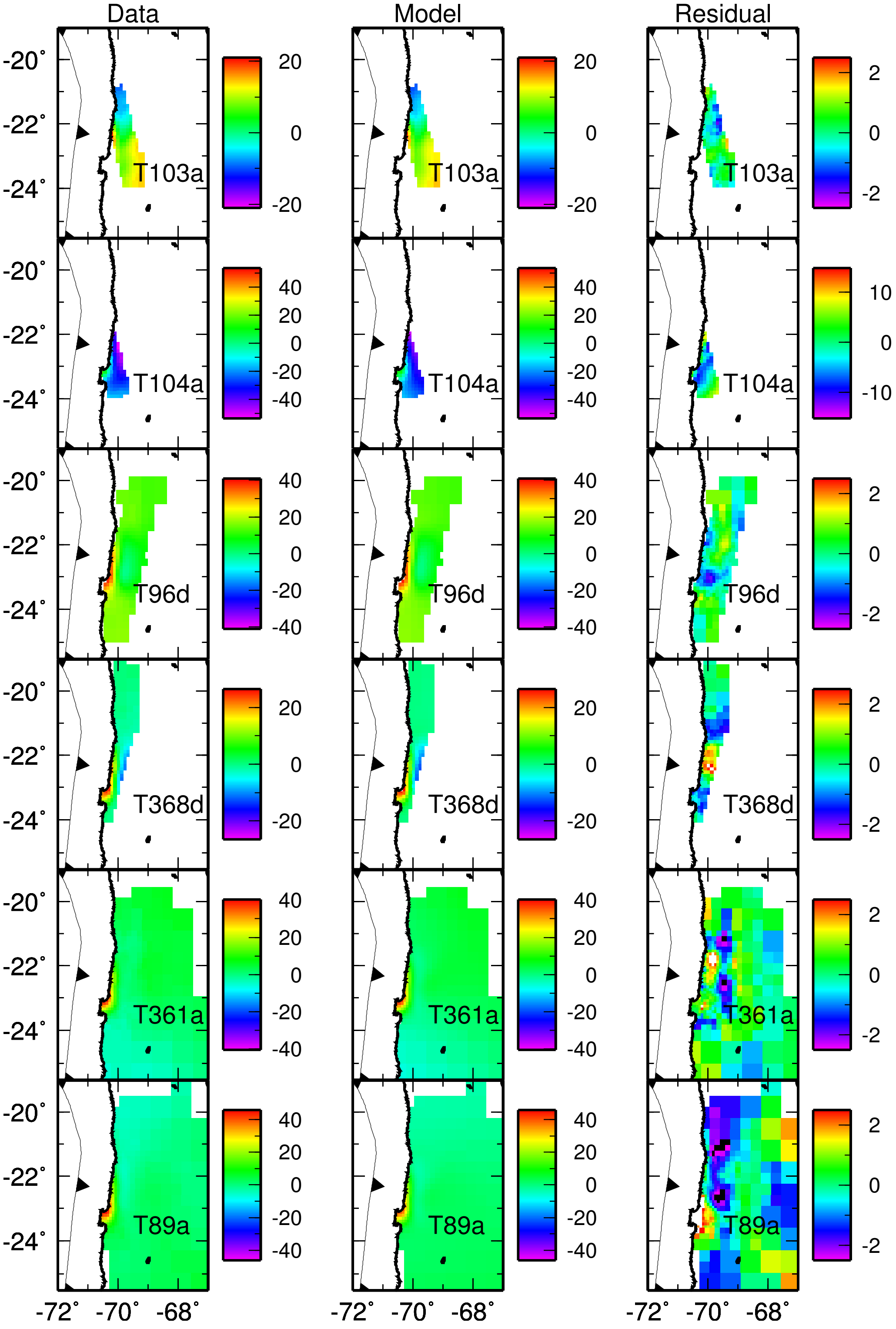


Figure 4.45: *InSAR predictions from the posterior distribution: Part II*: Same as Figure 4.44 for the mean of the posterior distribution.

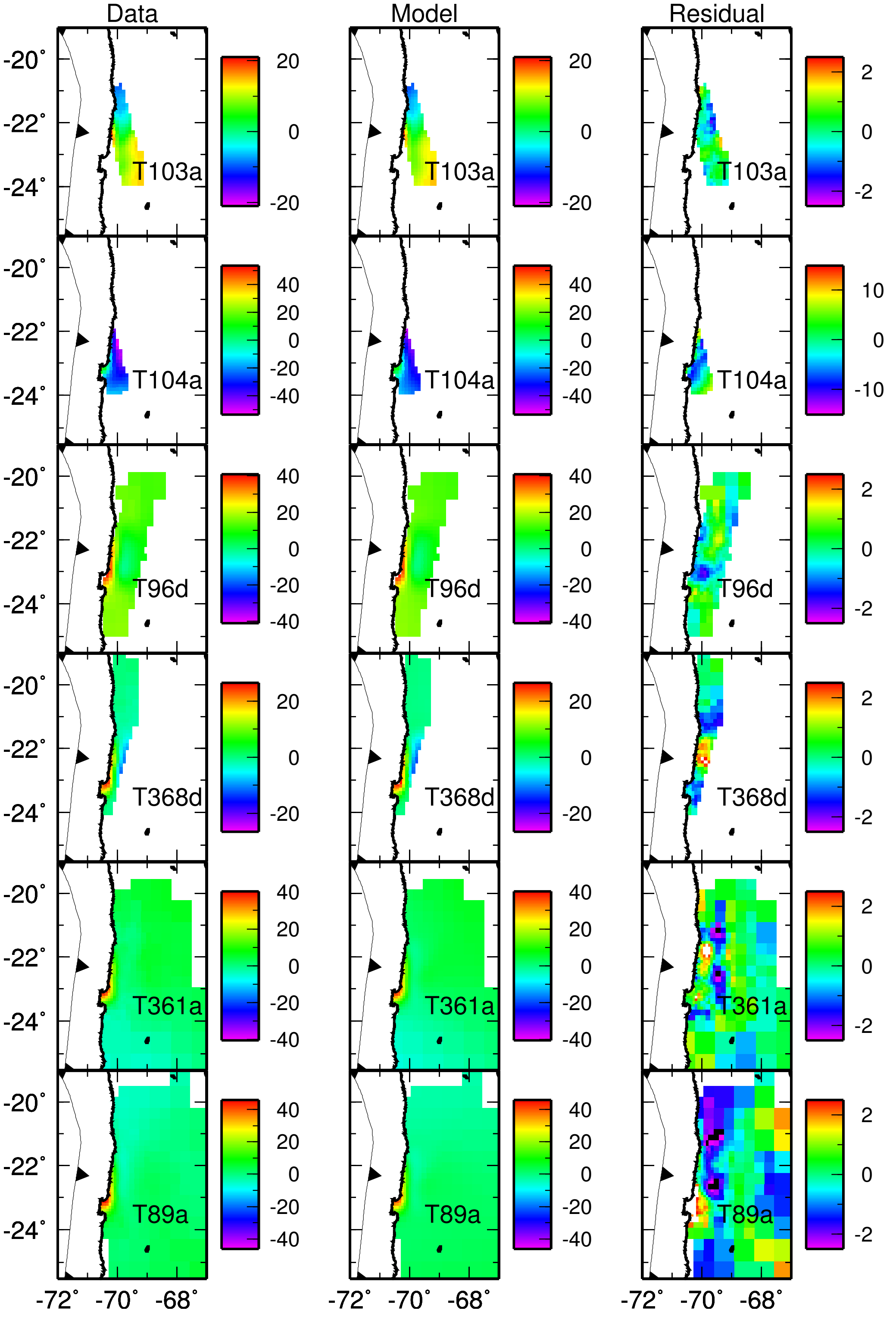


Figure 4.46: *InSAR predictions from the posterior distribution: Part III*: Same as Figure 4.44 for the median of the posterior distribution.

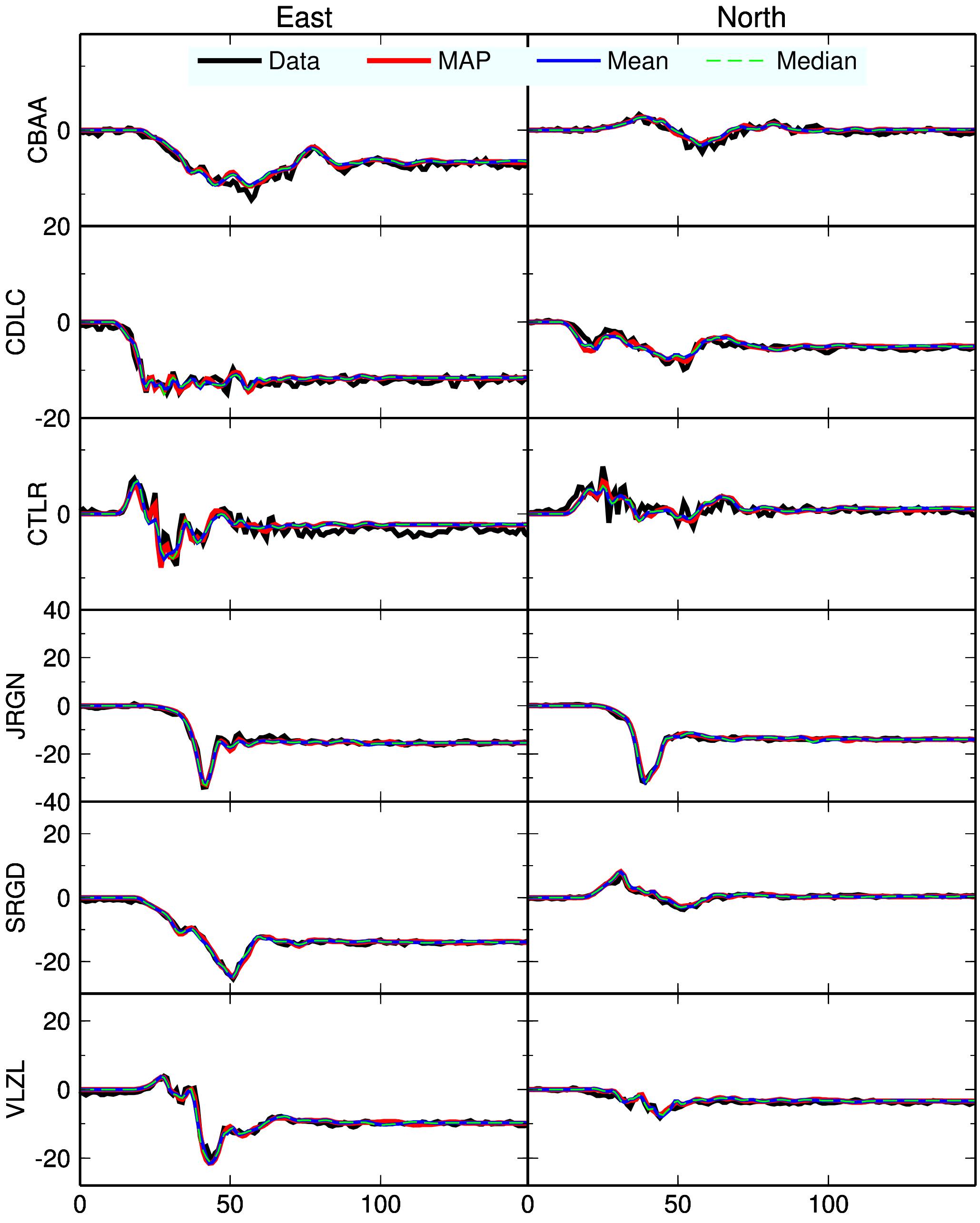


Figure 4.47: *Kinematic GPS predictions from three models*: The predicted time series from the maximum a posteriori (MAP), mean, and median models are compared to the high-rate GPS data. Displacements are in cm.

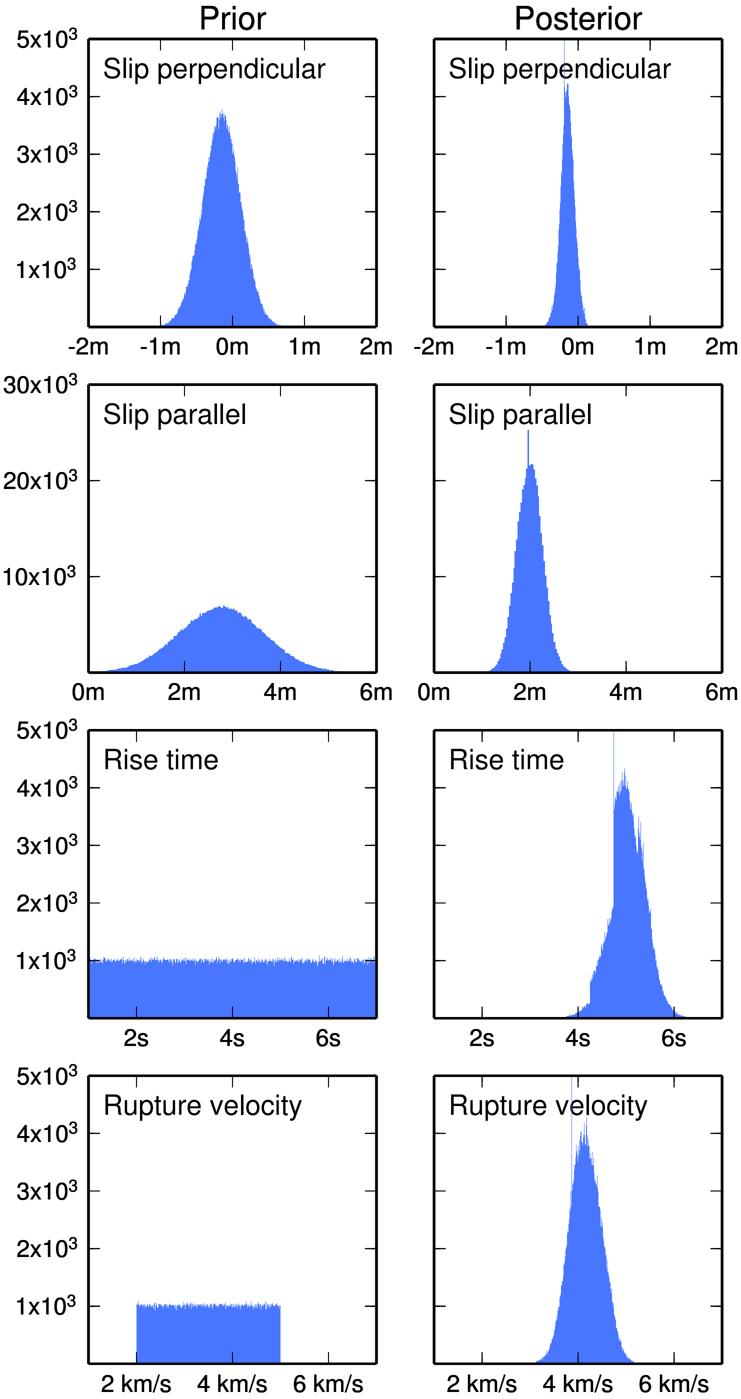


Figure 4.48: *Evolution of the posterior distribution*: Comparison of the prior slip distribution (left column) to the posterior distribution (right column) for the patch in Figure 4.52 (patch II in Figure 4.50). The prior distributions on slip are the posterior distributions from the static model. The prior distributions on rise time and rupture velocity are uniform. This is the same patch as is plotted in Figure 4.14.

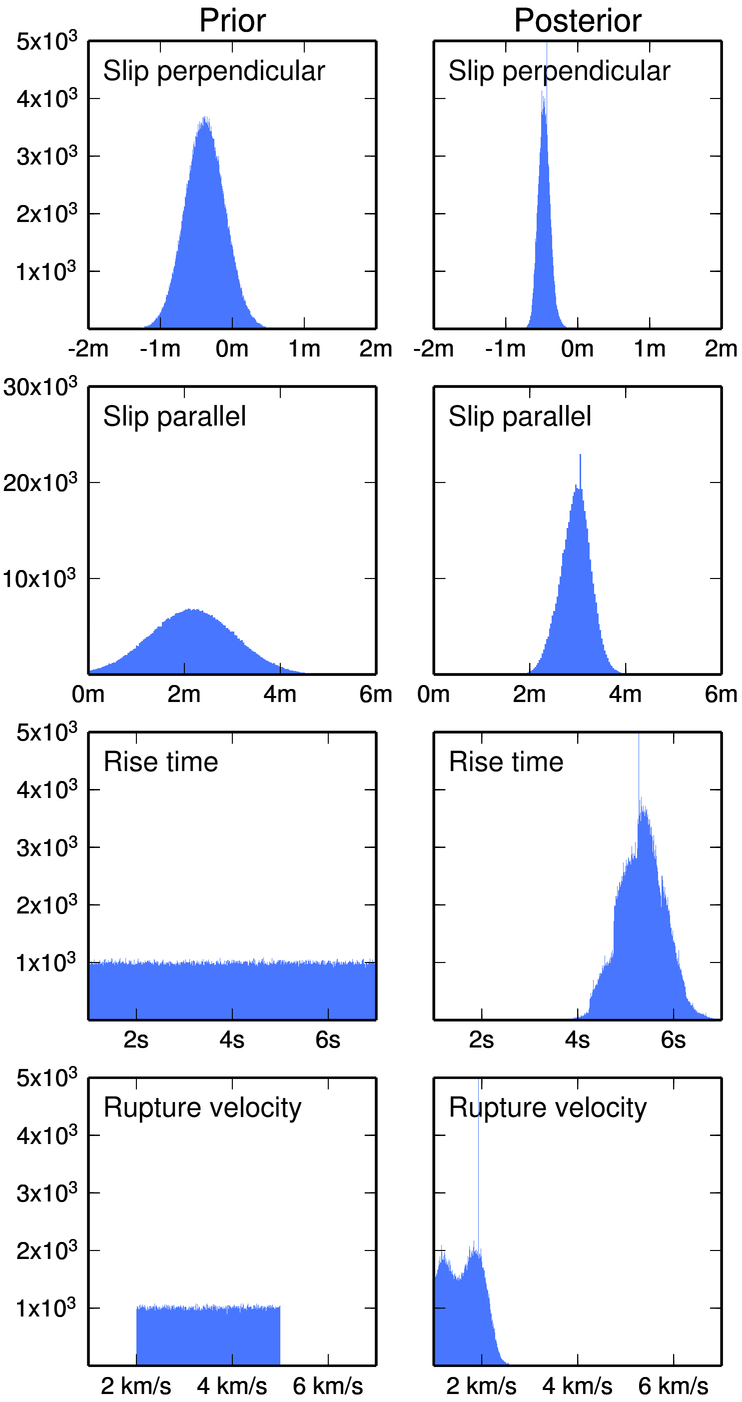


Figure 4.49: *Evolution of the posterior distribution*: Same as Figure 4.48 for the patch in Figure 4.54 (patch IV in Figure 4.50).

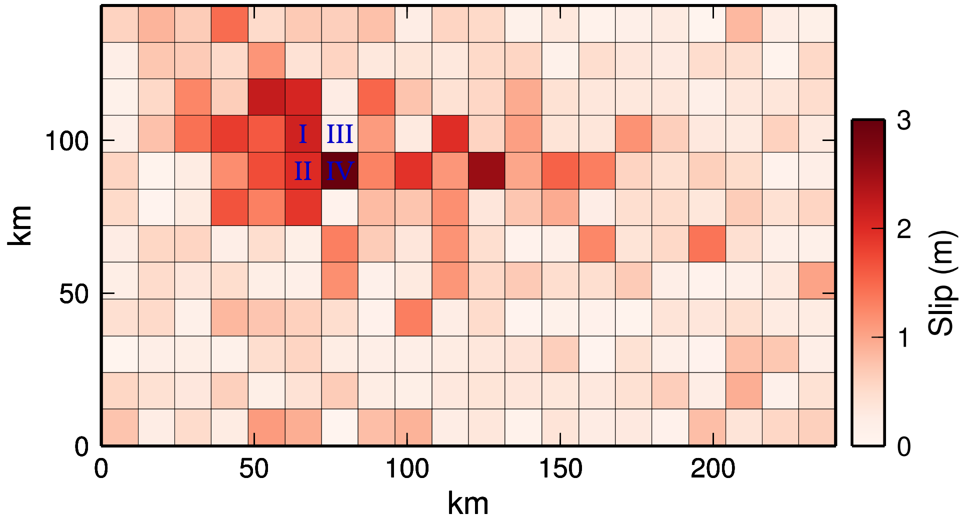


Figure 4.50: *Key to model correlation plots*: The model correlations with respect to each of the patches marked with Roman numerals are presented in Figure 4.51 through Figure 4.54.

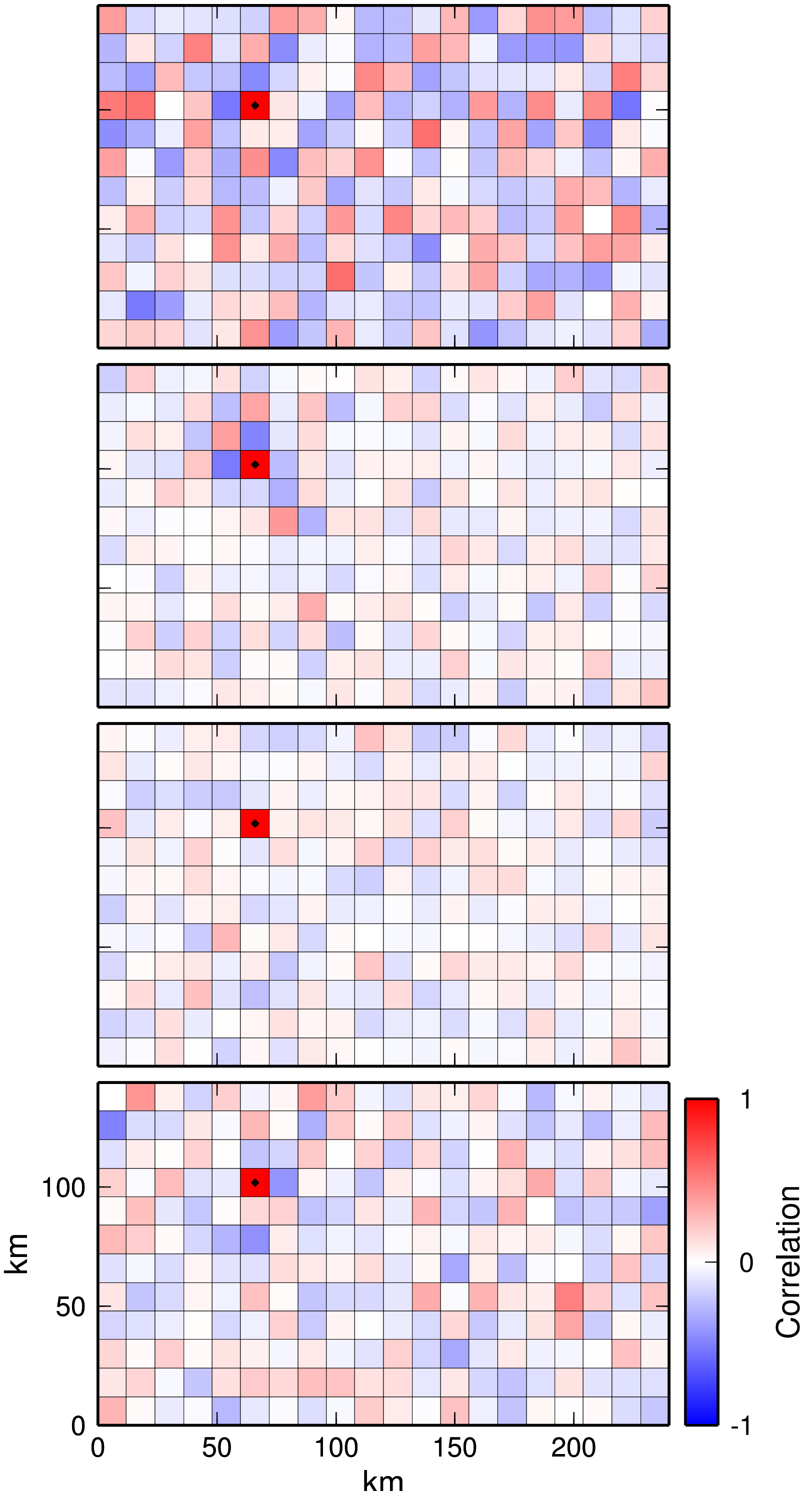


Figure 4.51: *Model correlation I*: The correlation between the patch marked with a diamond and all other patches: $U_\perp$ (top), $U_\parallel$ (row two), τ_r (row three), V_r (bottom).

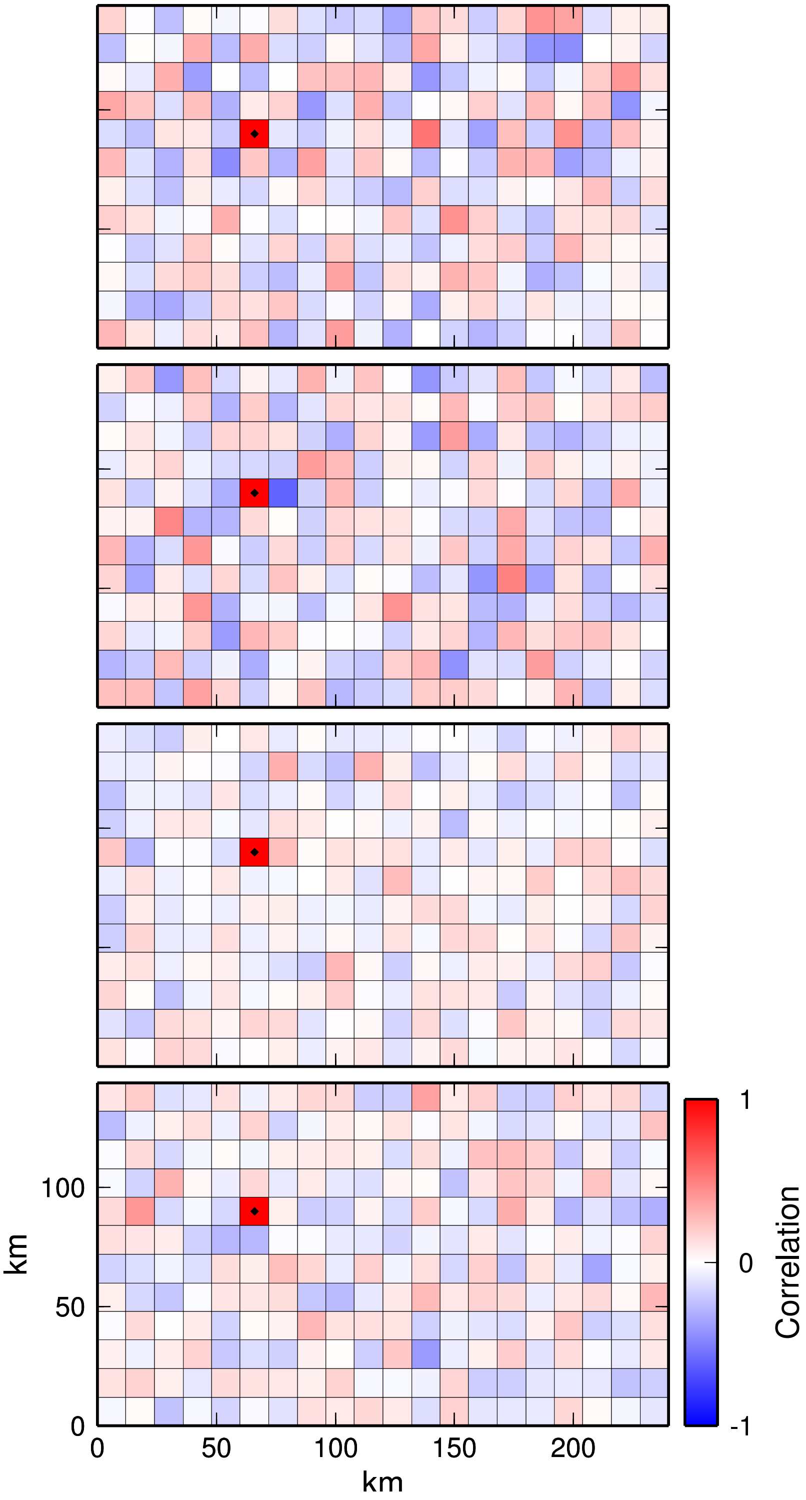


Figure 4.52: *Model correlation II*: The correlation between the patch marked with a diamond and all other patches: $U_\perp$ (top), $U_\parallel$ (row two), τ_r (row three), V_r (bottom).

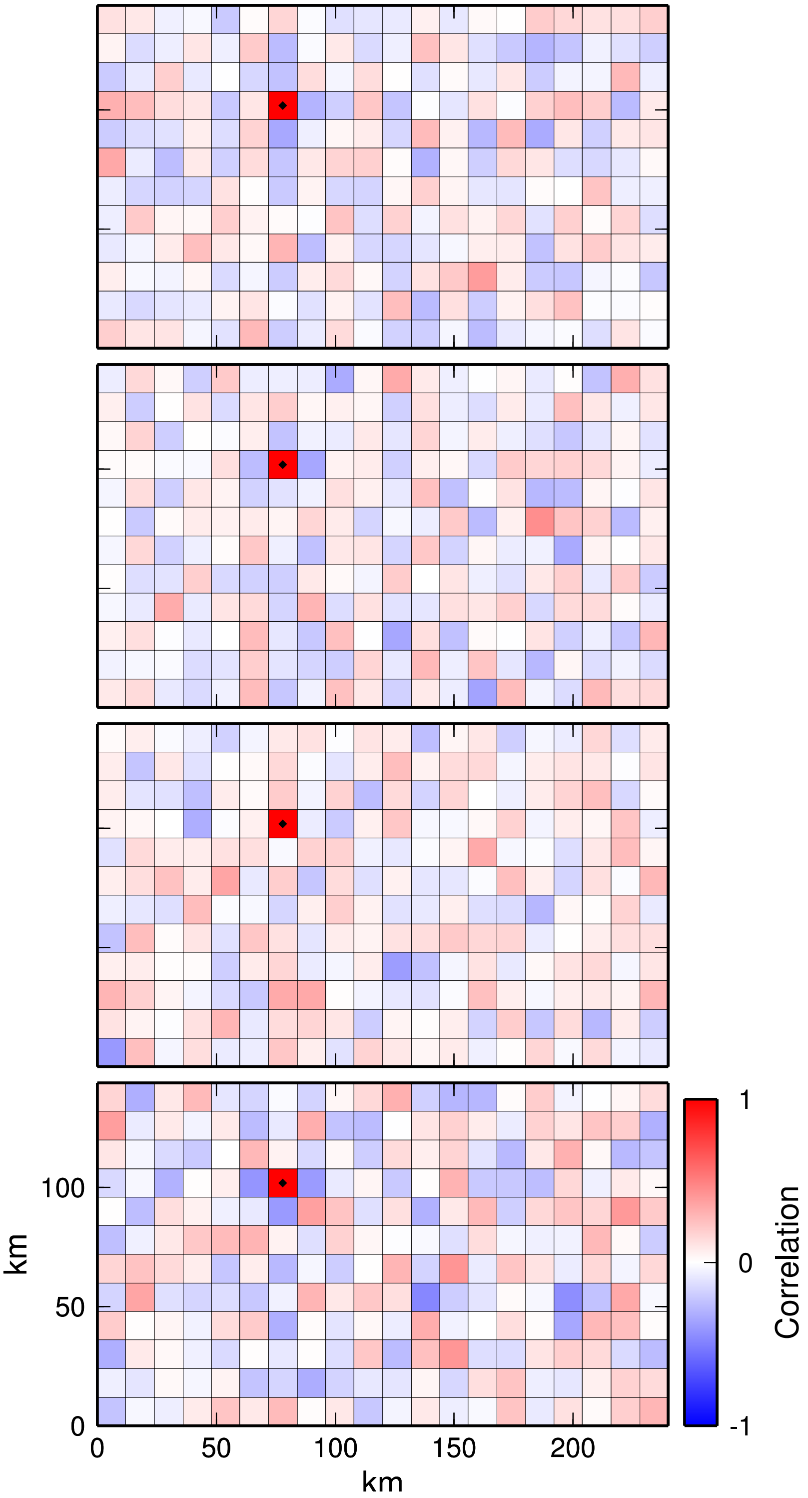


Figure 4.53: *Model correlation III*: The correlation between the patch marked with a diamond and all other patches: $U_\perp$ (top), $U_\parallel$ (row two), τ_r (row three), V_r (bottom).

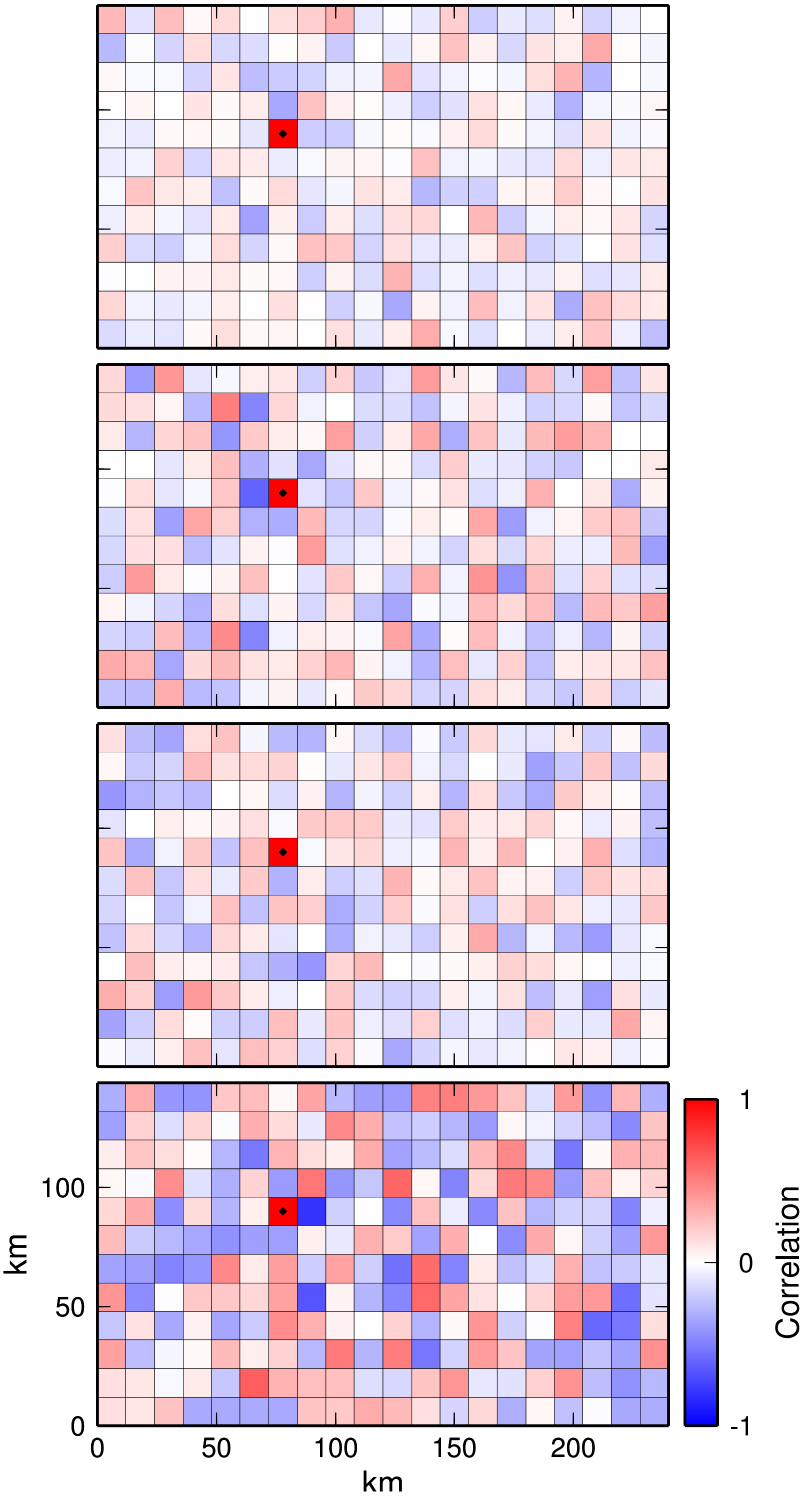


Figure 4.54: *Model correlation IV*: The correlation between the patch marked with a diamond and all other patches: $U_\perp$ (top), $U_\parallel$ (row two), τ_r (row three), V_r (bottom).

Finally, we will examine the complete slip distribution as shown in Figure 4.55 and a zoomed-in version of the same plot in Figure 4.56. What is immediately apparent is that the dense network of high-rate GPS stations does a good job of constraining the slip distribution. The kinematic posterior slip distribution is highly peaked, more so than the static-only posterior distribution (Figure 4.20). The rupture area is smaller, and slip is accumulated in well-defined asperities.

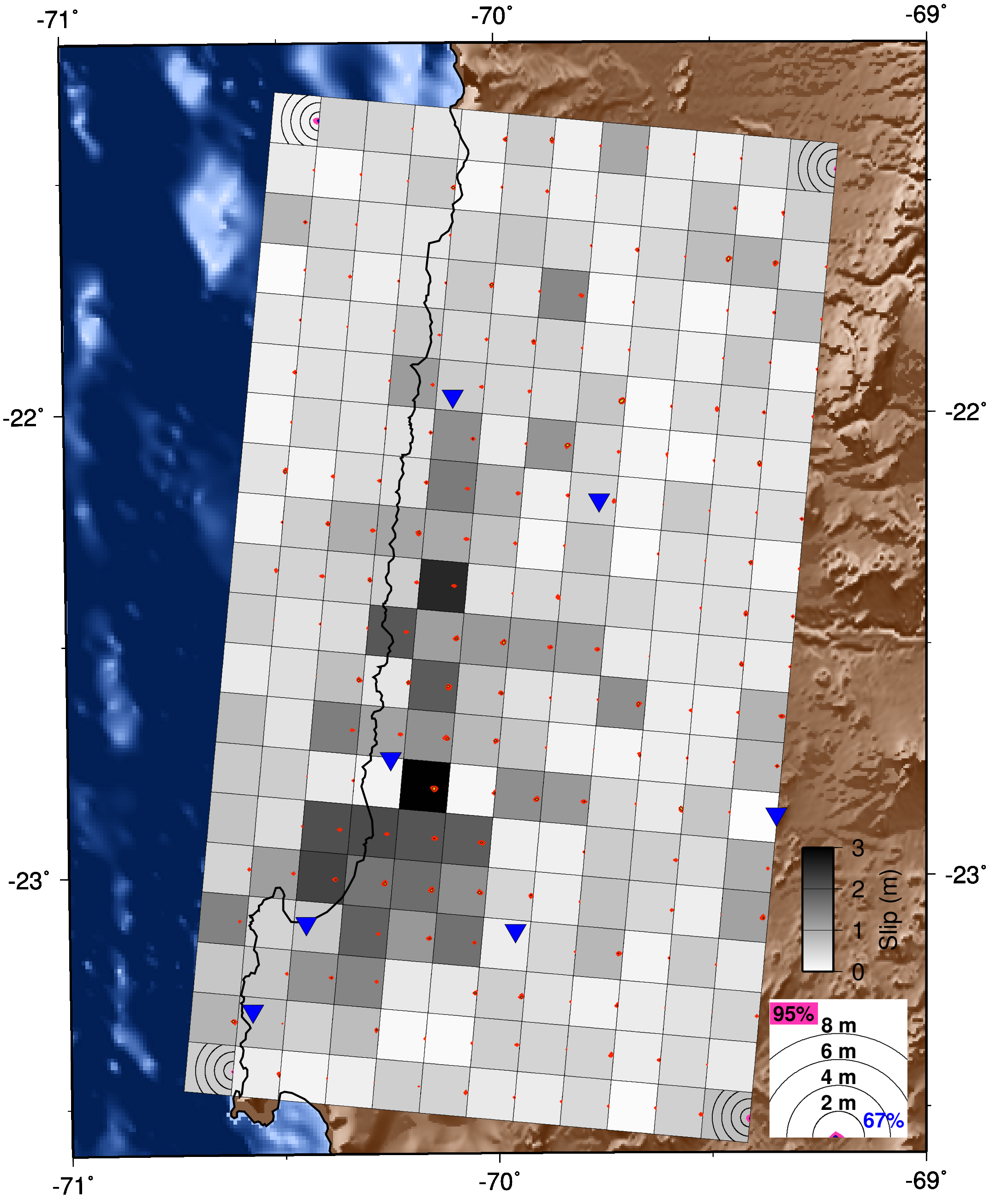


Figure 4.55: *Posterior slip distribution from kinematic modeling*: Two contour lines are drawn for each patch. Each contour illustrates the range of slip values which contain 67% and 95% of all posterior models. The patches at the corners illustrate the range of slip values found in the prior distribution. The background grayscale intensity of each patch is the mean of the slip distribution for that patch. Locations of GPS stations are indicated by triangles.

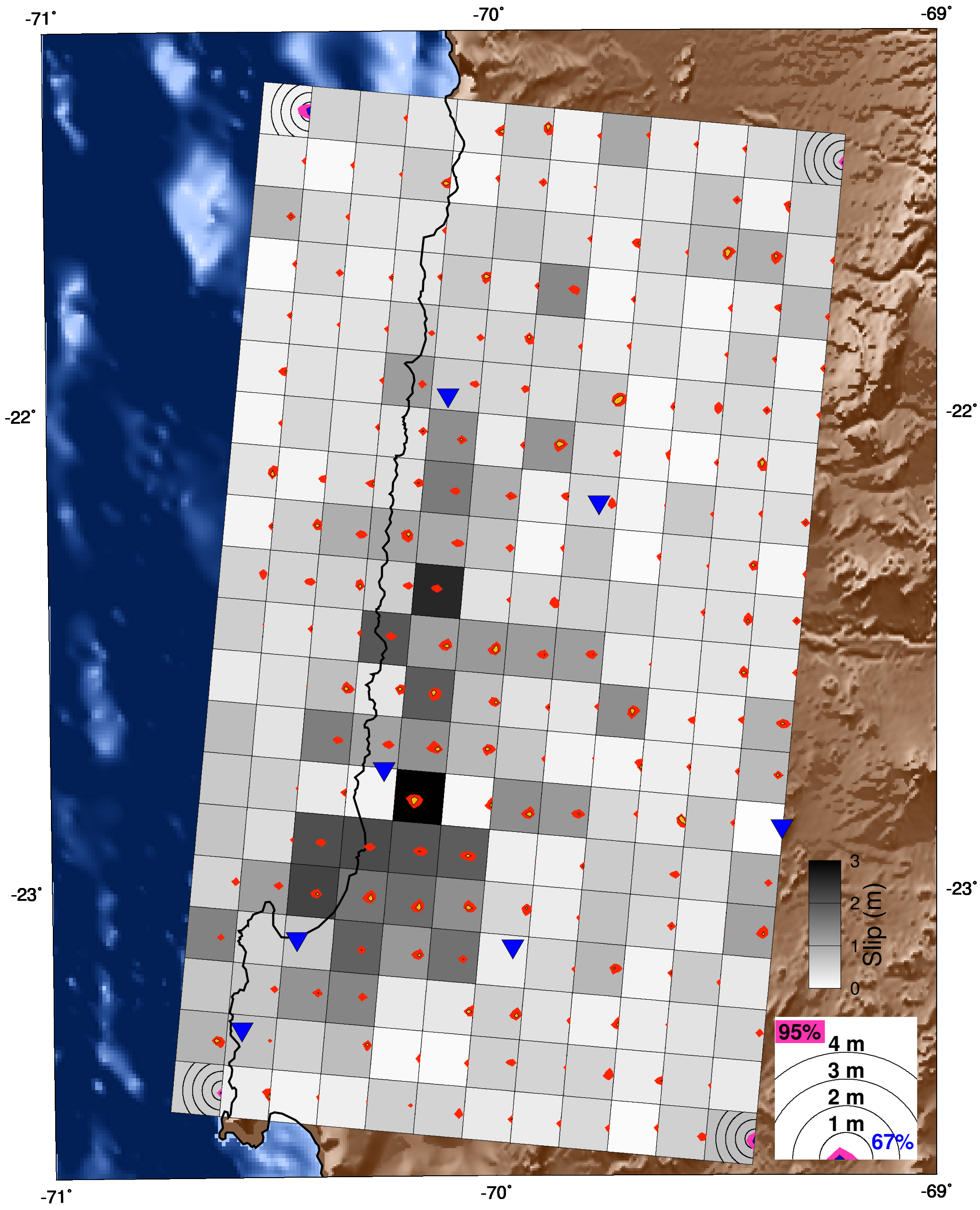


Figure 4.56: *Zoomed posterior slip distribution from kinematic modeling*: Two contour lines are drawn for each patch. Each contour illustrates the range of slip values which contain 67% and 95% of all posterior models. The patches at the corners illustrate the range of slip values found in the prior distribution. The background grayscale intensity of each patch is the mean of the slip distribution for that patch. Locations of GPS stations are indicated by triangles.

Figure 4.57 shows the complete posterior distribution for rupture velocity and rise time. Rupture velocity may be, in general, better resolved than rise time since the posterior distribution is slightly broader along the rise time axis in Figure 4.57. Both rupture velocity and rise time are better constrained on patches with significant amounts of slip, which is not surprising since, if there is little slip on a patch, then that patch does not contribute much to observed kinematic surface displacements. Neither rupture velocity nor rise time seem to be correlated with either the amount of slip or the roughness of the slip distribution.

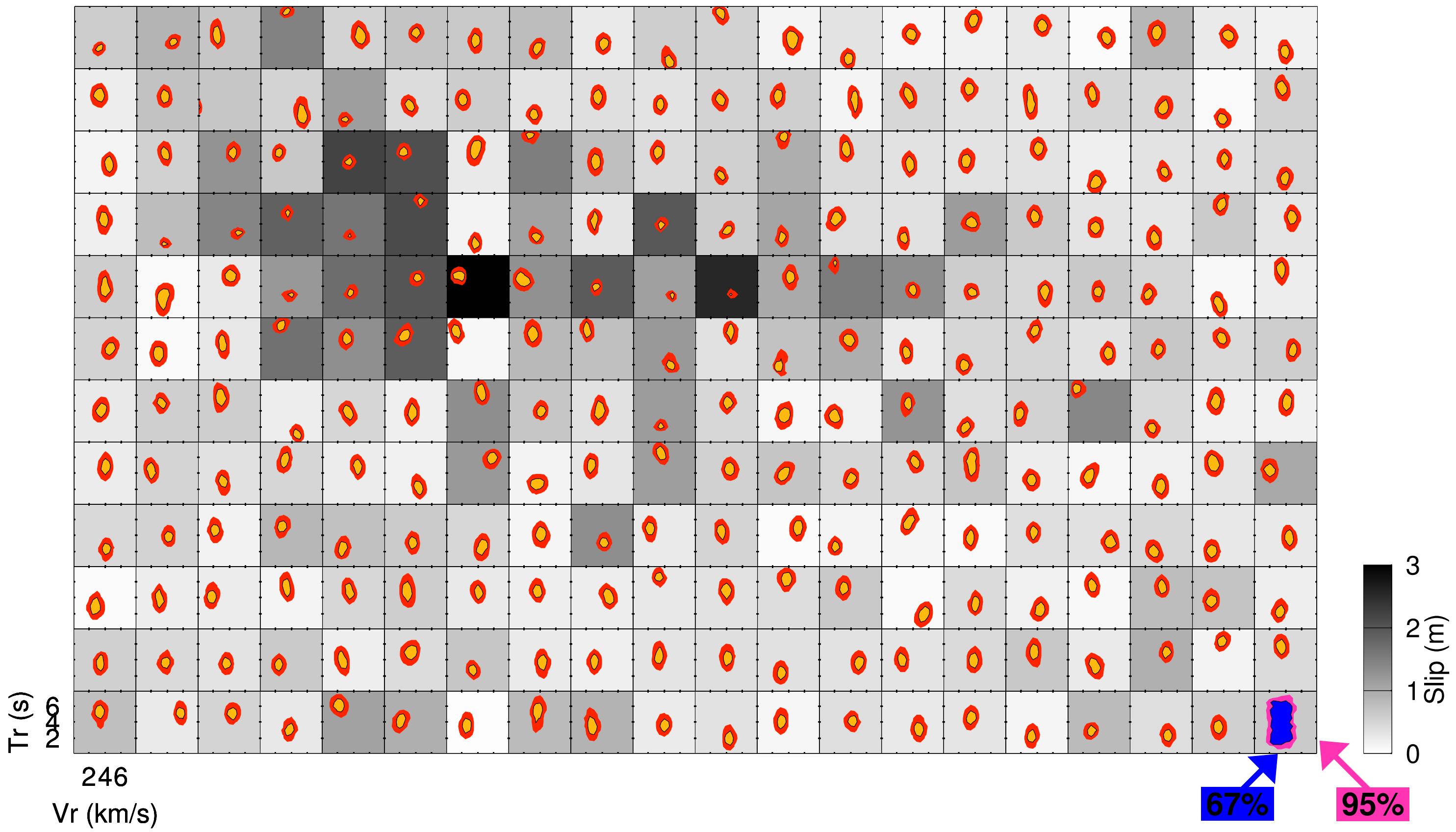


Figure 4.57: *Posterior kinematic distribution from kinematic modeling*: Two contour lines are drawn for each patch. Each contour illustrates the range of slip values which contain 67% and 95% of all posterior models. The patch at the bottom-right corner illustrates the range of slip values found in the prior distribution. The background grayscale intensity of each patch is the mean of the slip distribution for that patch.

Chapter 5

Source Properties of the Tocopilla Earthquake

So far I have derived a new sampling method for Bayesian analysis, used it to produce a source model for the 2007 Tocopilla, Chile earthquake, and analyzed the posterior probability distribution to understand the features and uncertainties in the source model. This is a new way of looking at finite fault source models. But slip distributions are not the ultimate objective. They are best used as inputs to help answer questions about the physics of the source process. And because we have not one model, but instead a complete ensemble of all possible models which are consistent with the data, we can use the entire distribution to calculate bounds on derived source parameters.

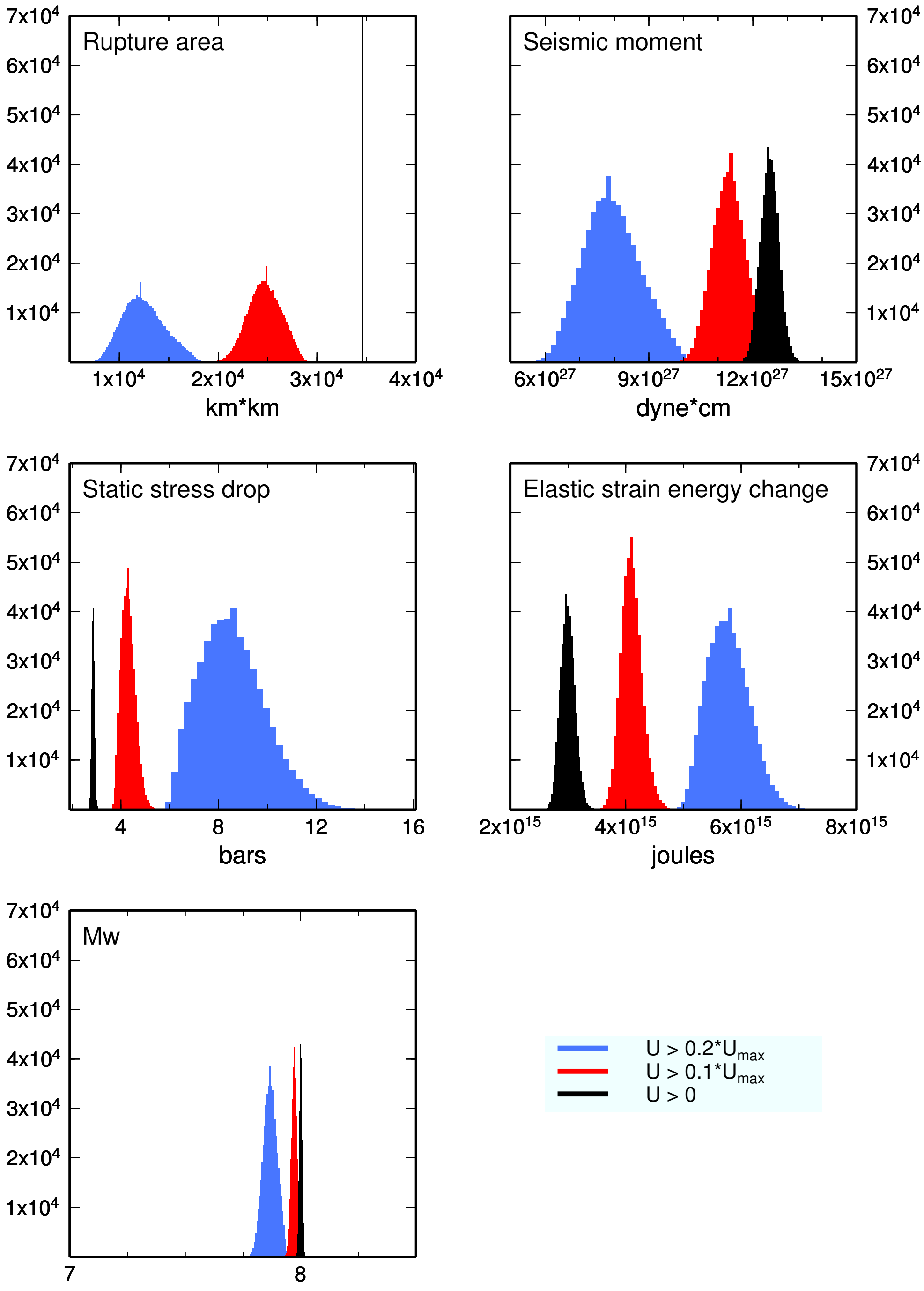


Figure 5.1: *Assorted source characteristics*: Probability distributions for rupture area, seismic moment, static stress change, elastic strain energy release and moment magnitude for the Tocopilla earthquake.

5.1 Inferred Source Characteristics

Several source parameters such as static stress change are difficult to estimate because they are based not on any actual observation of the quantity being estimated, but instead are derived from other models and observations. The Bayesian approach does nothing to ameliorate these problems. But since Bayesian analysis produces an ensemble of many source models, we can use these probability distributions to calculate the range of possible values for any source parameter. As an example, I consider four quantities.

First, is the scalar seismic moment of the earthquake, which is given directly by our posterior slip distribution. For a fault with rigidity μ, area A and average slip $\bar{u}$, the moment is

$$M_0 = \mu A \bar{u} \tag{5.1}$$

Second, I look at moment magnitude (*Kanamori*, 1977),

$$M_w = \frac{\log M_0}{1.5} - 10.73 \text{ (Moment in dyne*cm)} \tag{5.2}$$

Third, there is static stress drop for a rectangular dip-slip fault of length L and width w in a Poisson medium (*Kanamori and Anderson*, 1975),

$$\Delta\sigma = \frac{8}{3\pi}\mu\frac{\bar{u}}{w} \tag{5.3}$$

This can be rewritten as

$$\Delta\sigma = \frac{8}{3\pi}\frac{M_0}{\frac{A^{\frac{3}{2}}}{\sqrt{R}}} \tag{5.4}$$

where $R = \frac{L}{w}$ is the aspect ratio of the rupture.

There are many ways to attempt to derive an estimate of static stress drop from a slip model, and Equation 5.3 is perhaps the crudest method. The main difficultly arises in trying to geometrically interpret the slip pattern since the stress drop changes with the size and shape of the rupture. *Das* (1988) said that static stress drop is fairly independent of the slip distribution as long as the average stress drop for the different slip distributions were similar, which indicates that the relationship between average slip and average stress drop for smooth slip distributions may be reliable even for faults with multiple asperities. *Madariaga* (1979) argued that when there are multiple asperities, the stress drop scales on the order of the ratio of the area of the asperity to the total fault area, so that for a given moment and rupture area, faults with slip concentrated in multiple asperities have higher stress drop that faults with a smooth distribution of slip. And *Rudnicki and Kanamori* (1981) found that while strong asperities embedded in slip zones can produce locally large stress drops, the average stress drop over the entire slip zone is approximately equal to that inferred for an isolated fault.

Fourth, there is ΔW_0, the energy available to power the earthquake fracture and to radiate seismically. It can also be described as the minimum elastic strain energy drop, or the strain energy released if the stress drop is complete (*Kanamori*, 1977), and thus it can be estimated from the static stress drop by the expression

$$\Delta W_0 = \frac{\Delta\sigma}{2\mu} M_0 \tag{5.5}$$

The posterior slip distribution includes slip near the edge of the fault, as we commonly find in traditional optimization techniques. But because we have employed Baysian sampling to generate the full posterior distribution, we know that for this earthquake, the slip near the edge of the model is within our model uncertainty. This is evidence that constraints which enforce sparsity in the solution (i.e., require a spatially compact rupture) may be useful in at least the context of optimization, if not also Bayesian analysis. I discuss this now because the scalar seismic moment is proportional to the area of the earthquake rupture. If we assume, for a particular sample of the posterior, that the area of the rupture is all patches with slip, even negligible slip, our estimate of the scalar seismic moment will be large.

As a crude a posteriori investigation of the effects of sparseness on the earthquake model and the inferred properties of the source process, I calculated the four parameters listed above for three thresholds of slip: all patches with any slip, only the patches whose slips are least 10% of the maximum slip for each model, and only the patches whose slips are at least 20% of the maximum slip. Inspection of the slip models indicate that the aspect ratio of the rupture is approximately 3, and this is the value that I use for the purposes of calculating the static stress drop.

The distributions for all four of the above source properties for each slip threshold are shown in Figure 5.1 along with the distribution of rupture area for the three slip thresholds. As we include more fault patches in our calculations, the total moment increases and the distribution of rupture area tightens, until, when we allow all patches with any slip, the only possible rupture area is the entire area of the fault. Of course, this is an arbitrary end point. If we has used a fault model that was larger or smaller, we would have a distribution that converged to that area instead. However, changing our slip threshold by as much as 20% does not have much of an effect on the inferred source parameters. So a sparsity constraint might not affect our results very much. Since our posterior slip distributions were normally distributed, each of these derived source parameters is Gaussian. And since our slip distribution was well-constrained, the apparent uncertainties in each of these source characteristics are relatively small.

5.2 Slip Heterogeneity

As illustrated in Section 4.4, different smoothing choices in the traditional optimization approach lead to different solutions, and it is unclear which model is best. For fast optimization methods, the effects of different smoothing weights can be explored using bootstrap techniques. But this is not feasible with simulated annealing because it is computationally expensive. With Bayesian sampling, we can explore the roughness of the possible slip models a posteriori. My metric for slip heterogeneity is

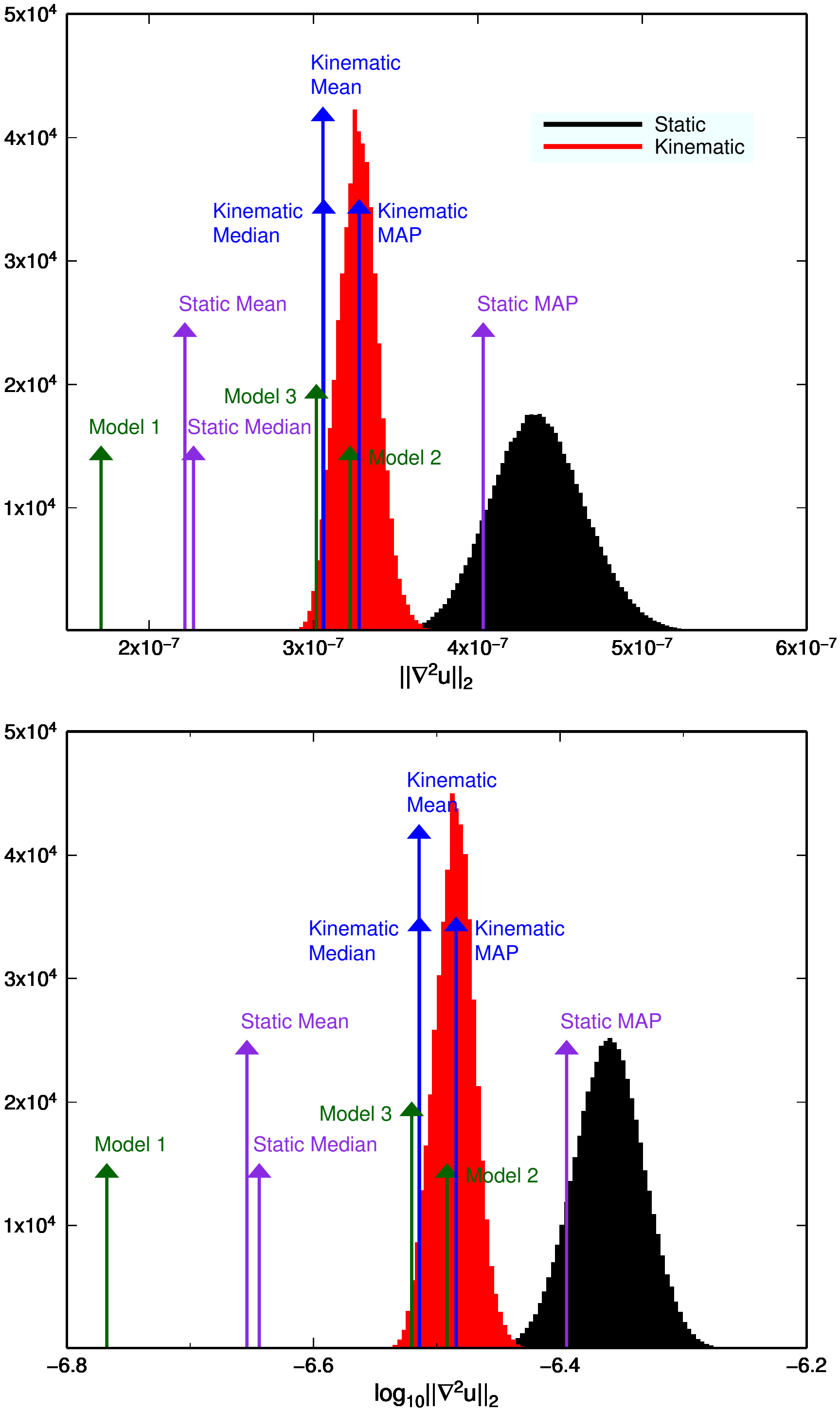
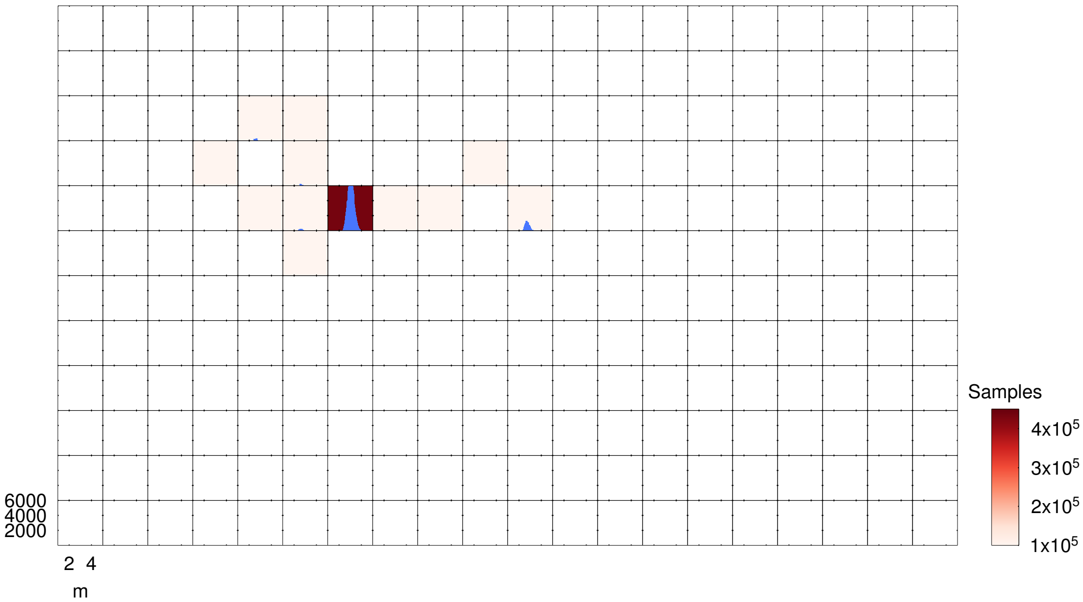


Figure 5.2: *Slip heterogeneity*: The roughness of the posterior slip distributions for the static and kinematic models are compared to the three models presented in Section 4.4. The model roughness is quantified by the L_2 norm of the Laplacian of the spatial distribution of slip $\|\nabla^2 u\|_2$ (top), and $\log_{10}\|\nabla^2 u\|_2$ (bottom).



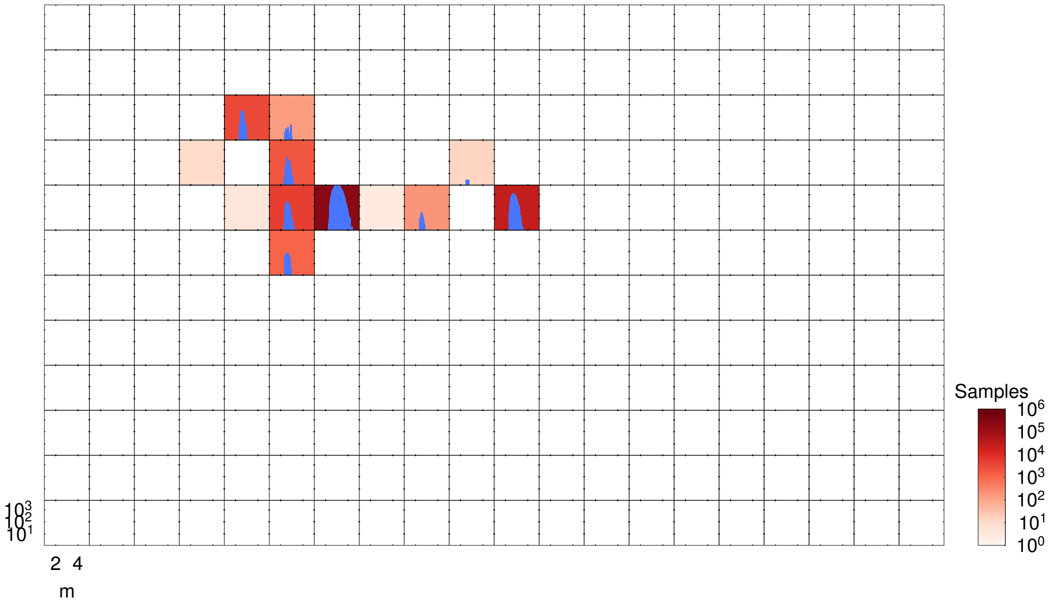


Figure 5.3: *Location of peak slip*: The variability in the magnitude of the maximum dislocation is shown with histograms. Patches are colored by the number of samples for which that patch is the location of the largest slip. (Top) Linear scaling shows that the location of the peak slip is well-constrained to one of two asperities and is most likely in the southern asperity. (Bottom) Logarithmic scaling reveals the existence of a few more possible locations of peak slip, although there are few samples associated with these models, and thus their probability is low.

the L_2 norm of the spatial Laplacian of slip u,

$$\|\nabla^2 u\|_2 = \|\nabla \cdot \nabla u\|_2 = \|\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\|_2 \tag{5.6}$$

This type of smoothing is commonly used in earthquake source inversions (e.g. *Harris and Segall*, 1987; *Du et al.*, 1992). So by plotting the distribution of slip roughness from the posterior distribution, we can see how the models produced by Bayesian sampling compare to those from traditional optimization techniques. The results are shown in Figure 5.2.

Overall, the distribution of slip models from the kinematic sampling are smoother than the distribution of static slip models. The three simulated annealing models are significantly smoother than the distribution of kinematic models from Bayesian sampling. Only Model 2 and Model 3 of Section 4.4 begin to approach the slip distribution seen in the MAP, mean, and median Bayesian models, and thus it is not surprising that only the spatial roughness of those models is consistent with the Bayesian posterior distributions. The mean and median of the static posterior distribution are much smoother than the models which comprise the posterior distribution. In the kinematic case, the roughness of the MAP, mean, and median of the posterior are similar to the samples in the distribution. Relative to the kinematic model, the static data have limited ability to resolve the slip on each patch. This leads to trade-offs between the model parameters, and each individual model is rough. However, when these models are averaged together to compute the mean or median, the resulting model is much smoother than the individual samples without significant decrease in the ability of the model to fit the static data. As we saw in Figure 4.48, the kinematic data have additional resolving power relative to the static data. There are fewer trade-offs between slip on neighboring patches, leading to smoother individual samples of the posterior and a distribution whose mean and median are about as rough as the samples themselves.

5.3 Peak Slip

Another benefit of Bayesian analysis over traditional optimization methods is the ability to assess the robustness of a particular feature in the slip model. As an example, I look at the location and magnitude of the largest dislocation (Figure 5.3). For the posterior distribution in its entirety, the peak slip is $3.015 \pm 0.2733\,\text{m}$. So the magnitude of the maximum slip is quite well-constrained. The location of the peak slip is also fairly well-constrained, although there is more uncertainty here. There are two modes. The less probable mode has the peak slip in the northern asperity. But the vast majority of models has the peak slip in the southern asperity.

5.4 Super-Shear Rupture Velocity

One important question in earthquake seismology is: Which, if any, earthquake ruptures propagate at super-shear velocities? There are several source models for various earthquakes, such as the 1999 Mw 7.2 Duzce, Turkey earthquake, which include super-shear rupture (e.g. *Bouchon et al.*, 2001; *Konca et al.*, 2010). Of course, each of these finite fault models is one possible solution to an under-determined inverse problem. There are other solutions where the rupture velocity remains sub-shear (e.g. *Utkucu et al.*, 2003; *Umutlu et al.*, 2004). The important question is: How likely is it that these earthquake ruptures exceeded the shear-wave velocity and in what part of the rupture process did it occur? Of course, I cannot answer this question here. But I can estimate the likelihood that the rupture velocity of the Tocopilla earthquake exceeded the shear-wave velocity.

Consider Figure 5.4. There are few patches where a significant fraction of the models exceed the shear-wave velocity, β. Nowhere does V_r greatly exceed β, and thus any possible appearace of super-shear behavior is probably well within the uncertainty in the velocity model. I would say that this rupture was not super-shear.

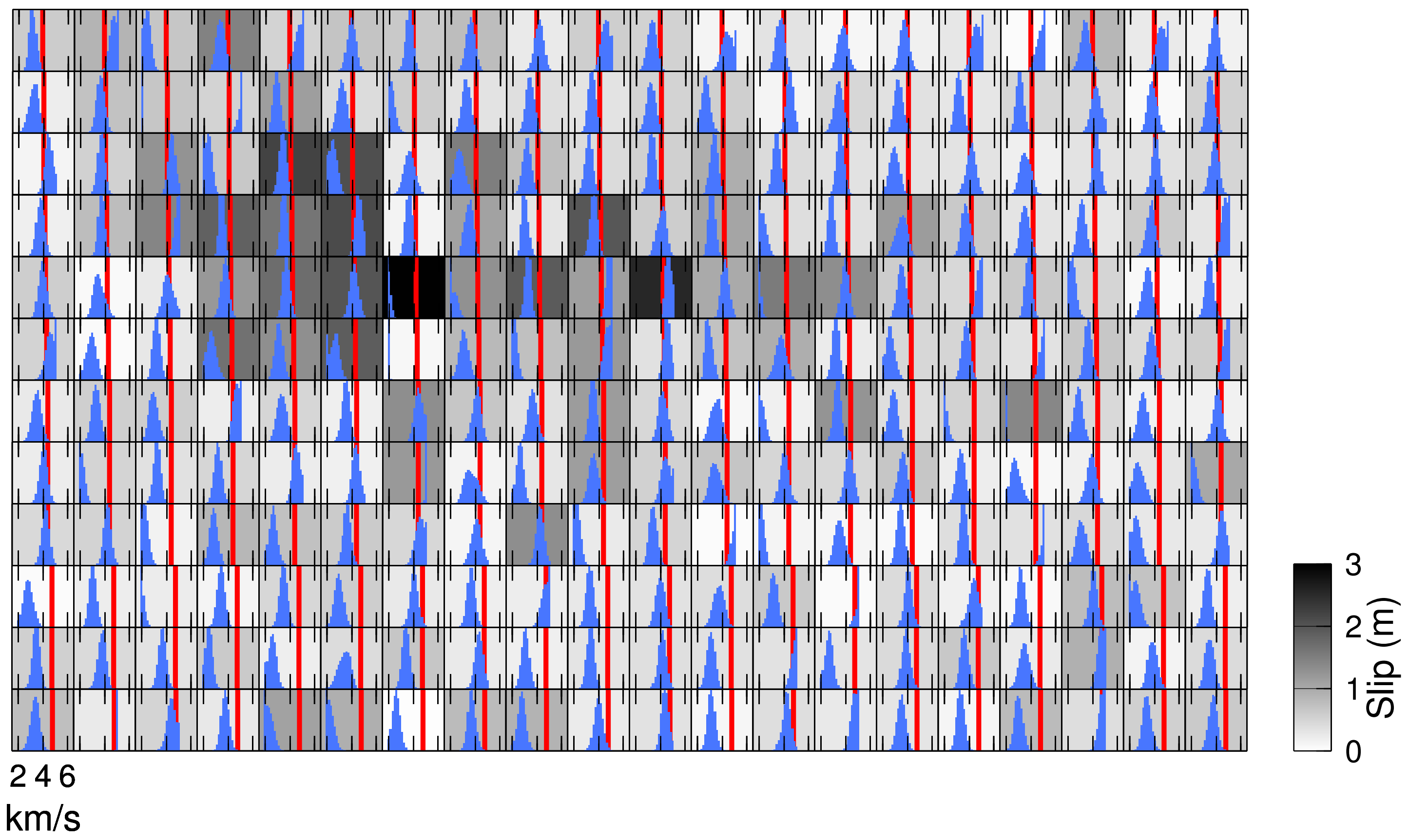


Figure 5.4: *Rupture velocity vs. shear-wave velocity*: The posterior distribution of rupture velocity for each patch is plotted in blue. Red lines indicate the shear-wave velocity β for a given patch from the *Husen et al.* (1999) tomographic velocity model.

Chapter 6

Closing Remarks

In geophysics, and especially in earthquake seismology, most of the inverse problems we deal with are under-determined. If you come away with nothing else from this work, I hope you will appreciate why Bayesian analysis is a more appropriate way to handle these problems than traditional optimization approaches. Regardless of its utility, Bayesian analysis has until now simply not been a practical option for geophysical modeling. It is too computationally intensive for problems with a large number of free parameters or problems whose forward models are themselves expensive to calculate. I therefore designed and implemented a parallel sampling technique, CATMIP, which allows us to tackle problems as complex as kinematic finite fault modeling. It is worth emphasizing that while I have used this algorithm to produce earthquake source models, the sampling algorithm is independent of the forward model, and can be used for any inverse problem.

I have done much to make the CATMIP algorithm as efficient as possible and to take advantage of parallel computing infrastructure. So while my sampling technique makes large-scale Bayesian modeling possible, it is only possible with the use of thousands of processors working in parallel. But as more computing facilities are built, I anticipate that more researchers will be able to utilize these methods.

Applying my new Bayesian sampling technique to the 2007 Tocopilla, Chile earthquake, I find that the dense near-field data lead to a well-constrained source model. The slip distribution is compact and localized in a few asperities. The rupture front expanded fairly homogeneously, but most of slip occurred south of the hypocenter near the Antofagasta peninsula.

The work presented here represents an initial application of Bayesian analysis to finite fault earthquake source models. There is much more work that can be done in this field. The models presented here use a simple parameterization of the earthquake source process. More elaborate models could be made which allow for the propagation of multiple rupture fronts. A multi-scale parameterization could be tried in order to obtain the best spatial resolution permitted by the data. The model could be expanded to additionally solve for the source geometry and location. And, perhaps most importantly of all, a Bayesian analysis of the error structure could be done to learn about the effects of unmodeled dynamics on our posterior distribution.

The goal of any slip model is to explain the physics of the source process. In the future, this is the field that will see the greatest reward from Bayesian analysis of earthquakes. The earthquake source process details needed to determine the dynamics of the rupture are often beyond the resolution of the available data. While this limitation cannot be avoided, with Bayesian analysis we can use the posterior probability distribution to quantify the likelihood of a particular rupture behavior.

The results I have presented are just the beginning. The exciting part comes next, when researchers use the posterior distributions from Bayesian earthquake models to probabilistically constrain the physics of the earthquake source rupture. We will never be able to resolve the details of the source process more than the data allow. But at least with Bayesian techniques, we can determine what we do and do not know, and how well everything is resolved.

Appendix A

CATMIP Implementation and Some Practical Considerations

In the development of the CATMIP algorithm, there were many false starts and deadends. Most of the details are too small, unimportant, and dull to discuss here. But what follows is a catalog of some points to keep in mind if you ever find yourself building a Bayesian sampler.

- While all of the Monte Carlo algorithms discussed here have been written in terms of the probability of various distributions, it is straightforward to frame them in terms of log-likelihoods, ln(P). In high dimensional simulations, you will want to use log-likelihoods. If you are multiplying the probabilities of hundreds of parameters, each of which has a value between zero and one (and usually much closer to zero than one), you will find that their product quickly becomes zero to working precision.

- These numerical issues apply to other computations as well. While Σ_m

(Equation 2.4) is by definition a symmetric positive-definite matrix, the process of calculating Σ_m may yield something which is close to having those properties but does not quite. Any small errors in Σ_m can be handled by replacing Σ_m with $\frac{1}{2}(\Sigma_m + \Sigma_m^T)$, or in more extreme cases by calculating the singular value decomposition of Σ_m and removing insignificant singular values.

- It is rather involved to compute Σ_m as it requires calculating the mean of the ensemble of models and the variance of each model from that mean. Although this calculation is only made once per cooling step, for large problems, efficiency requires that we do this work in parallel.

- In any Bayesian sampling, most of the work is spent on repeatedly evaluating the forward model. Anything you can do to speed up the forward model pays enormous dividends. One simple example: if you have a data covariance matrix, do not invert the matrix every time you evaluate $[d - g(m)]^T * C_d^{-1} * [d - g(m)]$. Pre-compute and store C_d^{-1}.

- The second best target for improving the speed of sampling is the random number generators. If you are generating your candidate samples from a multivariate normal distribution, your random number generator is inverting your covariance matrix Σ_m. For a large-scale problem with hundreds of model parameters, and thus a proposal covariance matrix which has thousands of elements or more, you do not want to invert that matrix for every candidate sample in every Markov chain. It is better to write your own random number generator which accepts Σ_m^{-1} or to pre-compute all of the random numbers you will need.

- One algorithm not discussed here is the modified Metropolis algorithm (*Au and Beck*, 2001), which was developed to draw samples from small failure-probability regions. In this algorithm, rather than compare your current model $x =< x_1 ... x_n >$ to candidate model $y =< y_1 ... y_n >$, you first compare x to $y^{(1)} =< y_1, x_2, ... x_n >$. If you accept this model, your next candidate is $y^{(2)} =< y_1, y_2, x_3 ... x_n >$, $y^{(2)} =< x_1, y_2, x_3 ... x_n >$ otherwise. This approach increases the likelihood that your current model will be updated, but the total number of model evaluations is increased by a factor of n. In high-dimensional simulations, the modified Metropolis algorithm is popular because it increases the acceptance rate. This algorithm should not be mixed with TMCMC, CATMIP, or any algorithm which uses an adaptive proposal PDF. By updating the model one parameter at a time, you erase the increased efficiency that comes from producing models with the correct model covariance.

- For high-dimensional problems, it may be beneficial to replace any probabilities you have (particularly those in your prior distribution) with unnormalized distributions so as to avoid numerical underflow. This change will not affect the sampling process. In the Metropolis algorithm and its progeny, the absolute value of any probability is unimportant. We always evaluate the ratio of probabilities.

The problems you find in Bayesian sampling generally fall into the categories exemplified above. There are issues of numerical underflow and overflow. Insignificant calculations, such as random number generation, may become time drains for very large problems. And none of this will be of any use unless you can make your forward model fast enough that the sampling process will complete in a reasonable length of time.

My CATMIP implementation is written in the C programming language using the GNU Scientific Library. Parallel communication is performed using the Message Passing Interface (MPI) protocols. A schematic of my program is presented in Figure A.1.

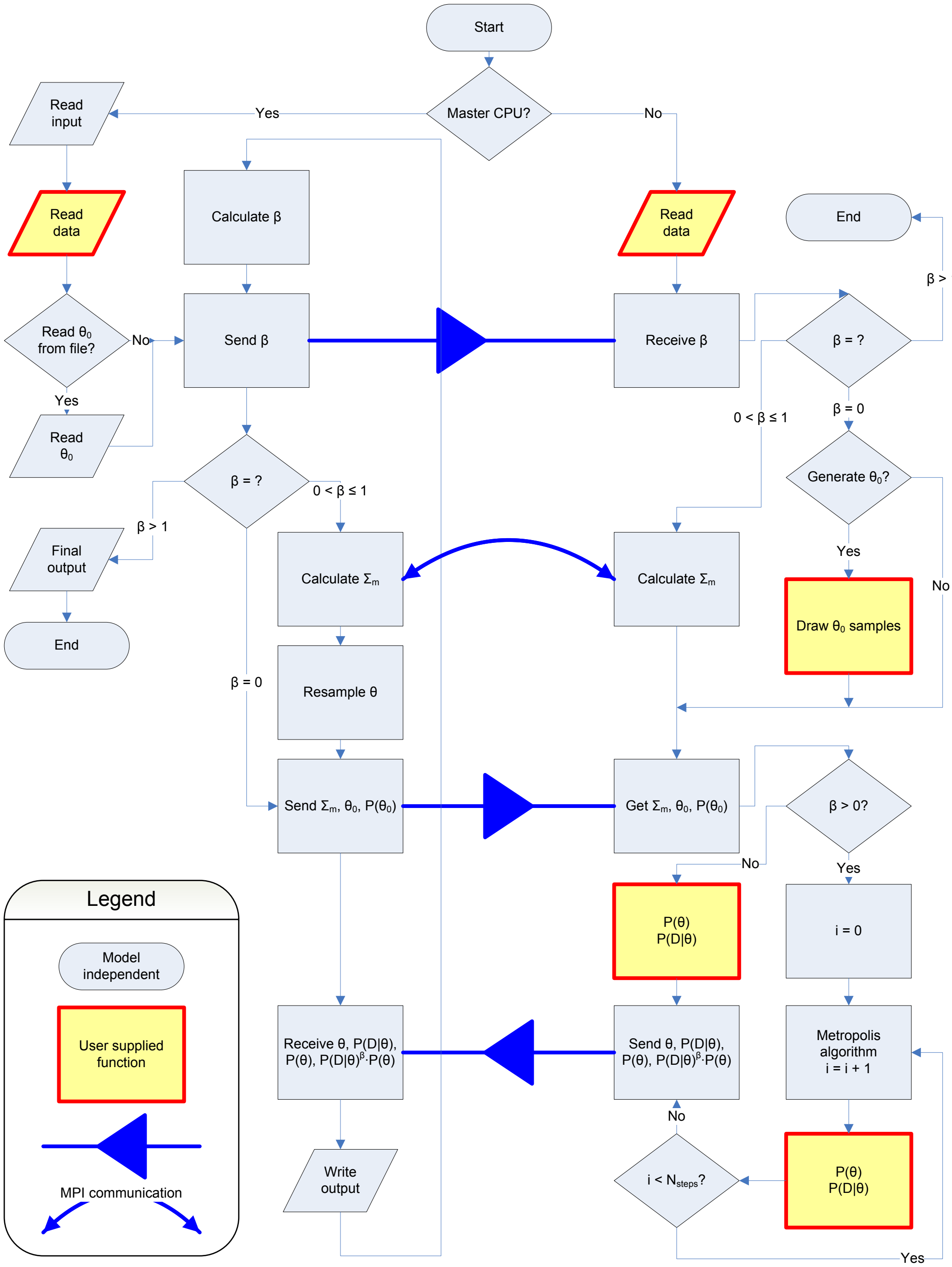


Figure A.1: *Flowchart illustrating operation of the CATMIP algorithm*: The left column represents actions taken on the master node. The right column represents actions on the workers. Thick arrows denote inter-processor communication.

Bibliography

Akaike, H. (1974), A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Au, S., and J. Beck (2001), Estimation of small failure probabilities in high dimensions by subset simulation, *Probabilistic Engineering Mechanics*, *16*(4), 263–277.

Barazangi, M., and B. Isacks (1976), Spatial distribution of earthquakes and subduction of the Nazca plate beneath South America, *Geology*, *4*(11), 686–692.

Bayes, T. (1763), An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S., *Philosophical Transactions of the Royal Society*, *53*, 370–418, doi:10.1098/rstl.1763.0053.

Beck, J., and L. Katafygiotis (1998), Updating models and their uncertainties. Part I: Bayesian statistical framework, *Journal of Engineering Mechanics*, *124*(4), 455–461.

Beck, J., and K. Yuen (2004), Model Selection using Response Measurements:

Bayesian Probabilistic Approach, *Journal of Engineering Mechanics*, *130*(2), 192–203.

Bilich, A., J. Cassidy, and K. Larson (2008), GPS Seismology: Application to the 2002 MW 7.9 Denali fault earthquake, *Bulletin of the Seismological Society of America*, *98*(2), 593–606.

Bouchon, M., M. Bouin, H. Karabulut, M. Toksoz, M. Dietrich, and A. Rosakis (2001), How fast is rupture during an earthquake? New insights from the 1999 Turkey earthquakes, *Geophysical Research Letters*, *28*(14), 2723–2726.

Cerny, V. (1985), Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm, *Journal of Optimization Theory and Applications*, *45*(1), 41–51.

Cheung, S. (2009), Stochastic analysis, model and reliability updating of complex systems with applications to structural dynamics, Ph.D. thesis, California Institute of Technology.

Chib, S., and E. Greenberg (1995), Understanding the Metropolis-Hastings algorithm, *American Statistician*, *49*(4), 327–335.

Ching, J., and Y.-C. Chen (2007), Transitional Markov chain Monte Carlo method for Bayesian model updating, model class selection, and model averaging, *Journal of Engineering Mechanics*, *133(7)*, 816–832, doi:10.1061/(ASCE)0733-9399(2007)133:7(816).

Cohee, B., and G. Beroza (1994a), Slip distribution of the 1992 Landers earthquake and its implications for earthquake source mechanics, *Bulletin of the Seismological Society of America*, *84*(3), 692–712.

Cohee, B., and G. Beroza (1994b), A comparison of two methods for earthquake source inversion using strong motion seismograms, *Annals of Geophysics*, *37*(6), 1515–1538.

Cotton, F., and M. Campillo (1995), Frequency domain inversion of strong motions: Application to the 1992 Landers earthquake, *Journal of Geophysical Research*, *100*(B3), 3961–3975.

Cox, R. (1961), *The Algebra of Probable Inference*, The Johns Hopkins University Press, Baltimore, MD, US.

Das, S. (1988), Relation between average slip and average stress drop for rectangular faults with multiple asperities, *Bulletin of the Seismological Society of America*, *78*(2), 924–930.

Delouis, B., M. Pardo, D. Legrand, and T. Monfret (2009), The Mw 7.7 Tocopilla earthquake of 14 November 2007 at the southern edge of the northern Chile seismic gap: Rupture in the deep part of the coupled plate interface, *Bulletin of the Seismological Society of America*, *99*(1), 87–94.

Dow, J., R. Neilan, and C. Rizos (2009), The International GNSS Service in a changing landscape of Global Navigation Satellite Systems, *Journal of Geodesy*, *83*(3), 191–198.

Du, Y., A. Aydin, and P. Segall (1992), Comparison of various inversion techniques as applied to the determination of a geophysical deformation model for the 1983 Borah Peak earthquake, *Bulletin of the Seismological Society of America*, *82*(4), 1840–1866.

Gelman, A., H. Stern, and D. Rubin (2004), *Bayesian data analysis.*, Chapman and Hall, Boca Raton, Florida.

Harris, R., and P. Segall (1987), Detection of a locked zone at depth on the Parkfield, California, segment of the San Andreas fault, *Journal of Geophysical Research*, *92*(8), 7945–7962.

Hartzell, S., and T. Heaton (1983), Inversion of strong ground motion and teleseismic waveform data for the fault rupture history of the 1979 Imperial Valley, California, earthquake, *Bulletin of the Seismological Society of America*, *73*(6), 1153–1184.

Hartzell, S., P. Liu, C. Mendoza, C. Ji, and K. Larson (2007), Stability and uncertainty of finite-fault slip inversions: application to the 2004 Parkfield, California, earthquake, *Bulletin of the Seismological Society of America*, *97*(6), 1911–1934.

Hastings, W. (1970), Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, pp. 97–109.

Hernandez, B., F. Cotton, and M. Campillo (1999), Contribution of radar interferometry to a two-step inversion of the kinematic process of the 1992 Landers earthquake, *Journal of Geophysical Research*, *104*(B6), 13,083–13,099.

Holland, J. H. (1992), *Adaptation in natural and artificial systems*, MIT Press, Cambridge, MA, USA.

Husen, S., E. Kissling, E. Flueh, and G. Asch (1999), Accurate hypocentre determination in the seismogenic zone of the subducting Nazca Plate in northern Chile using a combined on-/offshore network, *Geophysical Journal International*, *138*(3), 687–701.

Ide, S. (2007), Slip Inversion, in *Earthquake Seismology*, *Treatise on Geophysics*, vol. 4, edited by H. Kanamori, Elsevier, Amsterdam, The Netherlands.

Jaynes, E. (1957a), Information theory and statistical mechanics, *Physical Review*, *106*(4), 620–630, doi:10.1103/PhysRev.106.620.

Jaynes, E. (1957b), Information theory and statistical mechanics II, *Physical Review*, *108*(2), 171–190, doi:10.1103/PhysRev.108.171.

Jaynes, E. (2003), *Probability Theory: The Logic of Science*, Cambridge University Press, UK.

Jeffreys, H. (1931), *Scientific Inference*, Cambridge University Press, UK.

Jeffreys, H. (1939), *Theory of Probability*, Cambridge University Press, UK.

Ji, C., D. Wald, and D. Helmberger (2002), Source Description of the 1999 Hector Mine, California, Earthquake, Part I: Wavelet Domain Inversion Theory and Resolution Analysis, *Bulletin of the Seismological Society of America*, *92*(4), 1192–1207.

Ji, C., K. Larson, Y. Tan, K. Hudnut, and K. Choi (2004), Slip history of the 2003 San Simeon earthquake constrained by combining 1-Hz GPS, strong motion, and teleseismic data, *Geophysical Research Letters*, *31*(L17608).

Kanamori, H. (1977), The energy release in great earthquakes, *Journal of Geophysical Research*, *82*(20), 2981–2987.

Kanamori, H., and D. L. Anderson (1975), Theoretical basis of some empirical relations in seismology, *Bulletin of the Seismological Society of America*, *65*(5), 1073–1095.

Kikuchi, M., and H. Kanamori (1982), Inversion of complex body waves, *Bulletin of the Seismological Society of America*, *72*(2), 491–506.

Kirkpatrick, S., C. Gelatt, and M. Vecchi (1983), Optimization by simulated annealing, *Science*, *220*(4598), 671–680.

Konca, A., S. Leprince, J. Avouac, and D. Helmberger (2010), Rupture Process of the 1999 Mw 7.1 Duzce Earthquake from Joint Analysis of SPOT, GPS, InSAR, Strong-Motion, and Teleseismic Data: A Supershear Rupture with Variable Rupture Velocity, *Bulletin of the Seismological Society of America*, *100*(1), 267–288.

Lin, Y., M. Simons, E. Hetland, P. Muse, and C. Dicaprio (2010), A multi-scale approach to estimating topographically-correlated propagation delays in radar interferograms, *submitted to Geochemistry, Geophysics, Geosystems.*

Liu, J. (2008), *Monte Carlo Strategies in Scientific Computing*, Springer-Verlag, Springer, NY, US.

Lohman, R., and M. Simons (2005), Some thoughts on the use of InSAR data to constrain models of surface deformation: Noise structure and data downsampling, *Geochemistry Geophysics Geosystems*, *6*(1).

Lomnitz, C. (2004), Major earthquakes of Chile: a historical survey, 1535-1960, *Seismological Research Letters*, *75*(3), 368–378.

Loredo, T. (1990), From Laplace to Supernova SN 1987A: Bayesian Inference in Astrophysics, in *Maximum-Entropy and Bayesian Methods*, edited by P. Fougere, pp. 81–142, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Loveless, J., M. Pritchard, and N. Kukowski (2009), Testing mechanisms of subduction zone segmentation and seismogenesis with slip distributions from recent Andean earthquakes, *submitted to Tectonophysics.*

Madariaga, R. (1979), On the relation between seismic moment and stress drop in the presence of stress and strength heterogeneity, *Journal of Geophysical Research*, *84*(B5), 2243–2250.

Marinari, E., and G. Parisi (1992), Simulated tempering: A new Monte Carlo scheme, *Europhysics Letters*, *19*, 451–458.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953), Equation of state calculations by fast computing machines, *Journal of Chemical Physics*, *21*(6), 1087–1092.

Miyazaki, S., K. Larson, K. Choi, K. Hikima, K. Koketsu, P. Bodin, J. Haase, G. Emore, and A. Yamagiwa (2004), Modeling the rupture process of the 2003 September 25 Tokachi-Oki (Hokkaido) earthquake using 1-Hz GPS data, *Geophysical Research Letters*, *31*(21).

Muto, M., and J. Beck (2008), Bayesian updating and model class selection for hysteretic structural models using stochastic simulation, *Journal of Vibration and Control*, *14*(1-2), 7.

Norabuena, E., L. Leffler-Griffin, A. Mao, T. Dixon, S. Stein, I. Sacks, L. Ocola, and M. Ellis (1998), Space geodetic observations of Nazca-South America convergence across the central Andes, *Science*, *279*(5349), 358–362.

Pritchard, M., and M. Simons (2006), An aseismic slip pulse in northern Chile and along-strike variations in seismogenic behavior, *Journal of Geophysical Research*, *111*(B08405), doi:10.1029/2006JB004258.

Robert, C., and G. Casella (2004), *Monte Carlo Statistical Methods.*, Springer-Verlag, Springer, NY, US.

Rouy, S., and A. Tourin (1992), A viscosity solutions approach to shape-from-shading, *SIAM Journal on Numerical Analysis*, *29*(3), 867–884, doi:10.1137/0729053.

Rudnicki, J., and H. Kanamori (1981), Effects of fault interaction on moment, stress drop, and strain energy release, *Journal of Geophysical Research*, *86*(B3), 1785–1793.

Sambridge, M. (1999a), Geophysical inversion with a neighbourhood algorithm-I. Searching a parameter space, *Geophysical Journal International*, *138*(2), 479–494.

Sambridge, M. (1999b), Geophysical inversion with a neighbourhood algorithm-II. Appraising the ensemble, *Geophysical Journal International*, *138*(3), 727–746.

Schwarz, G. (1978), Estimating the dimension of a model, *The Annals of Statistics*, pp. 461–464.

Stigler, S. M. (1990), *The History of Statistics: The Measurement of Uncertainty Before 1900*, Belknap Press, Cambridge, MA, US.

Tarantola, A. (2005), *Inverse Problem Theory and Methods for Model Parameter Estimation*, Society for Industrial and Applied Mathematics, Philadelphia, PA, US.

Umutlu, N., K. Koketsu, and C. Milkereit (2004), The rupture process during the 1999 Duzce, Turkey, earthquake from joint inversion of teleseismic and strong-motion data, *Tectonophysics*, *391*(1-4), 315–324.

Utkucu, M., S. Nalbant, J. McCloskey, S. Steacy, and O. Alptekin (2003), Slip distribution and stress changes associated with the 1999 November 12, Duzce (Turkey) earthquake (Mw= 7.1), *Geophysical Journal International*, *153*(1), 229–241.

Von Huene, R., and C. Ranero (2003), Subduction erosion and basal friction along the sediment-starved convergent margin off Antofagasta, Chile, *Journal of Geophysical Research*, *108*(B2), 2079.

von Neumann, J. (1951), Various techniques used in connection with random digits, *National Bureau of Standards, Applied Math Series*, *12*, 36–38.

Wald, D., and T. Heaton (1994), Spatial and Temporal Distribution of Slip for the 1992 Landers, California, Earthquake, *Bulletin of the Seismological Society of America*, *84*(3), 668–691.

Zhao, H. (2005), A fast sweeping method for eikonal equations, *Mathematics of Computation*, *74*(250), 603–628.

Zhu, L., and L. Rivera (2002), A note on the dynamic and static displacements from a point source in multilayered media, *Geophysical Journal International*, *148*(3), 619–627.