On building and fitting a spatio-temporal change-point model for settlement and growth at Bourewa, Fiji Islands: Supplementary material

Geoff K. Nicholls

Department of Statistics, Oxford University, Oxford, UK.

Patrick D. Nunn

Department of Geography, The University of the South Pacific, Suva, Fiji

1. Introduction

Material in this supplement is placed in the same section as the section in the main paper to which it relates.

2. Related literature

2.1. Methodology
2.2. Application

3. Data, and observation model

4. Conditioning on phase structure

The layout in the first model is illustrated in Fig. 1. Age increases up the page, so the oldest events are at the top and calendar time increases down the page.

5. A spatial-temporal onset-field model

5.1. The onset-field process

5.1.1. Minor technical results

In Nicholls and Nunn (2009), Section 5 we observe that $\int_{(-\infty,\psi_M]^C} p(\phi|\alpha,\beta,\psi)d\phi = 1$, so we can differentiate with respect to α_c to get the natural moment identity

$$\mathsf{E}(\phi_c) = \psi_M - \mathsf{E}\left(\frac{1}{\rho(\phi_{\mathcal{N}^+(c)})}\right).$$

Spatial inhomogeneity in the mean $\mathsf{E}(\phi_c) - \mathsf{E}(\phi_{c'}) \neq 0$ is due to different rate expectations, and can be corrected by adjusting the local arrival rate α_c. However, this is just a correction in the mean, and in any case, we felt that the uncorrected onset-field gave a better representation of our subjective prior weighting on spatio-temporal structure.

Address for correspondence: GK Nicholls, Department of Statistics, 1 South Parks Road, Oxford OX1 3TG, UK
E-mail: nicholls@stats.ox.ac.uk

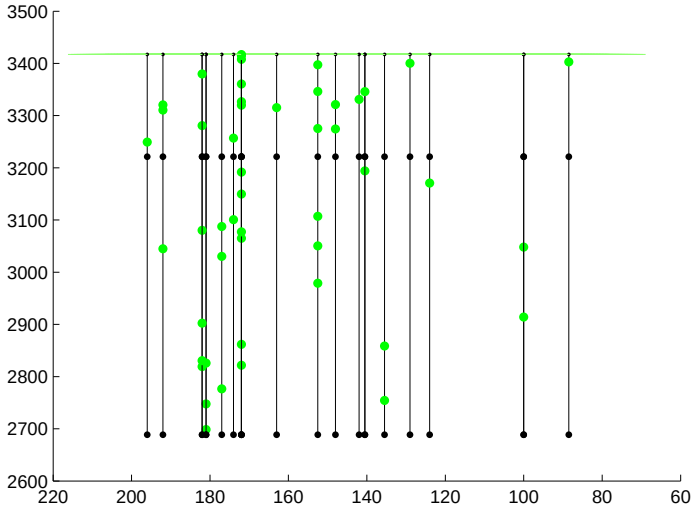

Fig. 1. A realisation of the state of a random phase model. The vertical lines correspond to pits. The green dots are specimen ages θ_i. The black dots indicate phase boundaries. Age increases up the page. The x axis corresponds to the long axis of the box in the second figure of the main paper and is included to improve the display. There is no spatial structure in this model.

Repeating this exercise for $\beta_{c,c'} = \beta_{c',c}$ we find

$$\begin{aligned} \mathsf{E}(|\phi_c - \phi_{c'}|) &= E\left(\frac{\mathbb{I}_{\phi_{c'}>\phi_c}}{\rho(\phi_{\mathcal{N}^+(c)})}\right) + \mathsf{E}\left(\frac{\mathbb{I}_{\phi_c>\phi_{c'}}}{\rho(\phi_{\mathcal{N}^+(c')})}\right) \\ &= \Pr(\phi_{c'} > \phi_c) E\left(\frac{1}{\rho(\phi_{\mathcal{N}^+(c)})} | \mathbb{I}_{\phi_{c'}>\phi_c}\right) + \Pr(\phi_c > \phi_{c'}) \mathsf{E}\left(\frac{1}{\rho(\phi_{\mathcal{N}^+(c')})} | \mathbb{I}_{\phi_c>\phi_{c'}}\right) \\ &< (\alpha + 4\beta)^{-1}, \end{aligned} \tag{1}$$

since ρ is at most $\alpha + 4\beta$ and its distribution is not concentrated there. Now $|\phi_c - \phi_{c'}|$ is the difference between the arrival times at cells c and c', and neighbors are one cell unit apart, so the speed v at which migration spreads from seeds when $\alpha \ll \beta$ is at most around 4β. It is fairly clear from the second line of Equation (1) that the speed will be between two and three times β, as "most" cells will be occupied when between two or three of their neighbors are already occupied.

5.2. The Richardson growth model, and related models

5.2.1. Onset field: mean-field analysis, and comparison with simulation

When $\alpha = 0$, the onset-field model coincides with one of the models which are called Richardson models. A great deal is known about Richardson models and our purpose here is simply to give illustrative simulations, and some intuition from a mean-field analysis of the traveling wave.

Let $p_c(t) = \Pr(\Phi_c > t)$, $\alpha_c = \alpha$ and $\beta_{c,c'} = \beta$. It is straightforward to show, from defining equations for the Poisson process realizing ϕ, that the evolution satisfies

$$\frac{d}{dt}p_c(t) = \alpha(1 - p_c(t)) + \beta \sum_{c' \in \mathcal{N}(c)} \Pr(\Phi_c < t, \Phi_{c'} > t).$$

The mean field approximation $\tilde{p}_c(t)$ has the events $\Phi_c < t$ and $\Phi_{c'} > t$ independent,

$$\frac{d}{dt}\tilde{p}_c(t) = \alpha(1 - \tilde{p}_c(t)) + \beta(1 - \tilde{p}_c(t)) \sum_{c' \in \mathcal{N}(c)} \tilde{p}_{c'}(t). \tag{2}$$

In Fig. 2 (left) we give the solution, $p_c(t)$ for $\alpha = 0$ and starting from $p_c(0) = \mathbb{I}_{c=c_{mid}}$, where cell c_{mid} is a central cell, $\psi_M = 0$, and $t = -0.3$, for $\beta = 1$ in a $M = 256$ by $n = 256$ square grid with a free boundary. This shows the kind of anisotropy we might expect: faster grown along the lattice axes, as diagonal lattice-metric separations are greater for pairs points at equal Euclidean separations. There is clearly some sort of traveling wave (Fig. 2 right). We can check this *via* the naive continuum limit of the mean-field approximation, $\lim_{C\to\infty} \tilde{p}_{\mathbf{c}(x)}(t) = \tilde{p}(x, t)$, which is

$$\frac{\partial \tilde{p}}{\partial t} = (1 - \tilde{p})(\alpha + 4\beta\tilde{p}) + \beta(1 - \tilde{p})\nabla^2\tilde{p}. \tag{3}$$

For the case $\alpha = 0$, a straightforward phase-plane analysis shows that this equation admits a traveling wave solution of the form $\tilde{p}(x - ct)$, for $c \geq 4\beta$. The calculation is similar to the corresponding classical calculation for the Fisher equation. We have not checked traveling waves with speed $c = 4\beta$ are stable in Equation (3), but we expect that it is true. Our numerical solution of Equation (2) has a traveling wave with speed very close to 4β. For example, in Fig. 2 at right, the wave has traveled about 125(118) cells in 30 time units with $\beta = 1$, on the lattice axis (diagonal) for a speed of around $4.17\beta(3.93\beta)$.

Simulation (Fig. 3, left) of stochastically growing clusters gave results which differ in important ways from the mean-field results. First, the result for the mean speed of a growing cluster edge, $v \leq \alpha + 4\beta$ from Equation (1), which makes sense for very small α, effectively contradicts the mean field speed $c \geq 4\beta$. The measured mean speed (from Fig. 2, right) is about 2.33β cells per year, which is natural, in light of the discussion below Equation (1). Secondly, the anisotropy displayed at the left of Fig. 2 may be a consequence of the mean-field approximation, and the cluster growth speed does not equal that of the mean-field approximation. The estimate $\hat{p}_c(t) = \frac{1}{J}\sum_j \mathbb{I}_{\phi_c^{(j)} > t}$ is computed from 85 independent realisations $\phi^{(j)}$, $j = 1, 2, ..., J$ of the field, simulated under the same parameter values $\psi = 0, \alpha = 0, \beta = 1, t = 30, C = 256^2$ and a free boundary) as the mean field solution. Anisotropy in Fig. 3 (right) is clearly slight, but we hesitate to declare it absent: Paiva and Jr (2007) consider this kind of problem for the Eden model and show that some very small amount of anisotropy is present. Hammersley (1977) provides some intuition on this subject.

Note that by homogeneity and isotropy we are interested in *strict* homogeneity and isotropy, that is, invariance of the distribution under the relevant rigid body motion, not simply invariance of the mean and covariance, which we saw for $\phi_c + \rho(\phi_{\mathcal{N}^+(c)})^{-1}$.

5.3. The onset-field as a prior probability density for the Bourewa data

Around seventy pits are left undated. The absence of dates from these pits is not informative, so "no dates" means "didn't look" not "looked and didn't find". The location of undated

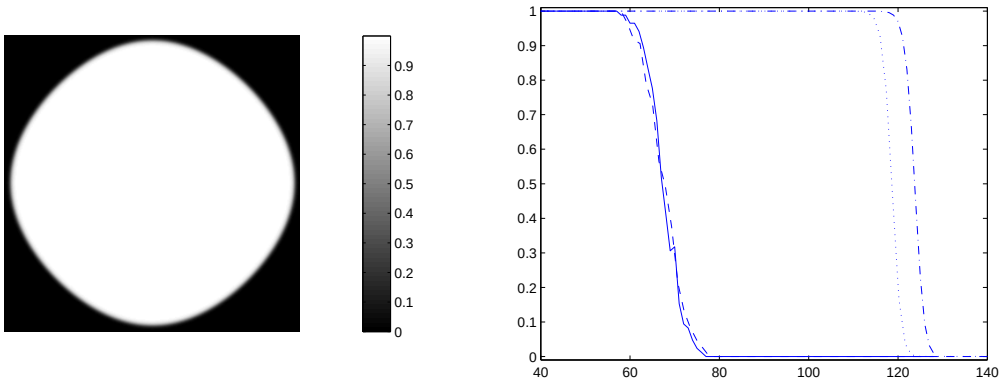

Fig. 2. (left) The mean field solution $\tilde{p}(t)$ for $\tilde{p}_c(t) = \Pr(\Phi_c > t),\ c = 1, 2, ..., C$ at $t = -30$ in Equation (2) for $\alpha = 0$, $\psi_M = 0$ and $\beta = 1$ on a 256^2 lattice, initialised to $\phi_c = \mathbb{I}_{c=c_{mid}}$ at $t = 0$ shows apparent anisotropy. (right) Traveling wave solutions to the mean-field equations: (x-axis) euclidian distance from cell c_{mid}; (y-axis) $p_c(t)$; (dotted and dot-dashed curves) respectively diagonal and on-axis sections of the mean-field solution $\tilde{p}_c(t)$ at $t = 30$; (solid and dashed curves) respectively diagonal and on-axis sections of $\hat{p}_c(t)$ at $t = 30$ estimated by simulation.

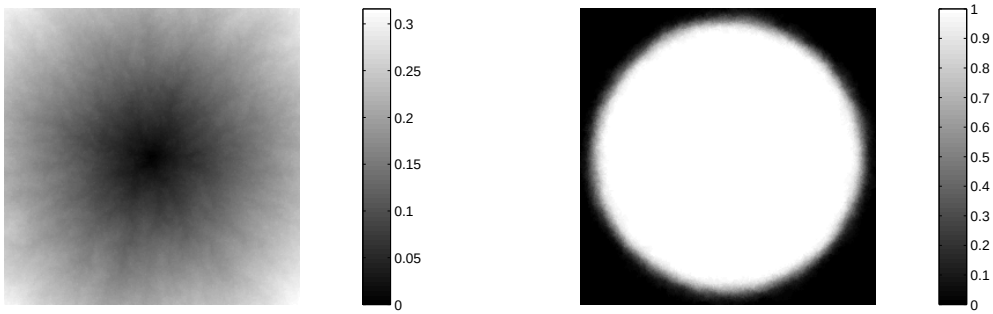

Fig. 3. (left) one realisation of ϕ simulated with $\alpha = 0$, $\beta = 1$, $\psi_M = 0$ and initialised to $\phi_c = \mathbb{I}_{c=c_{mid}}$ at $t = 0$. (right) An image of $\hat{p}(t)$ at $t = 30$ estimated from 85 iid realisations of ϕ as at left.

pits may be informative, since pits are made in places likely on independent grounds to show ancient human presence. However, with the prior assertion of site-wide deposition, these pit placements are uninformative.

When we condition on $\mathbf{m}$ we condition on the assignment of specimen $i = 1, 2, ..., K$ to phases, $m = 1, 2, ..., M$, and when that information is available (it is not at Bourewa, or there is a single phase) it is clearly informative. We might hope that the assignment $\mathbf{h}$ of specimen to pits is informative of the length of time for which the deposition process acted at the pit location. This is to some extent true. Specimen thought likely to belong to earlier phases of occupation are more likely to be selected for dating. However, the number of dated specimen taken from a given pit depends also on quantities which are not part of our data, such as excavation budgets, pit areas, and the chance occurrence of date-worthy specimen in later phases overlying the early occupation. The outcome is, that some pits have many more dates than others, and these well-studied pits are clustered together. Even if the entire site was occupied simultaneously, we might nevertheless find more early dates where there are more dates, spuriously suggesting a center for expansion. Since the spatial distribution of dates across pits is not informative of any process of interest, we have conditioned the whole analysis on the assignment of dated specimen to pits, rather than simply conditioning on the total number of dates across pits.

6. Fitting the onset-field for a single phase

In this this section we give further background on the simulation and fit for the single-phase onset field model of this section in the main paper.

The default setup has $L = 2000$, $U = 3500$, $A = 10$, $B = 1$ on a 32×13 lattice. In some of the runs reported in the main paper A and B are varied to check for sensitivity to prior settings. Here we give brief results for fitting on a 64×26 lattice, fitting synthetic data for the case that the true onset-field is a constant (to check no spurious spread is found) and give further figures for the default prior and posterior.

MCMC simulation traces for the default prior and posterior are given in Fig. 4 and Fig. 5 respectively. The minimum number of immigration events is V, the number of peaks, defined in this section of the main paper. The log-likelihood is $\log(\ell(\theta; y))$ and the log prior is the log of $p(\alpha, \beta, \phi, \psi, \theta|\mathbf{m}, M)$. The 'sum pit totals' is the log of $p(\theta|\phi, \psi, \mathbf{m})$.

Further prior simulation summaries for the default setup are given in Fig. 6 and on the left in each of Figures 8, 9 and 10 (though Fig. 8 is identical to the mean given in the main paper, and is included here to make this self-contained). The graph at right in each of Figures 8, 9 and 10 is the corresponding posterior statistic. We have included (in Fig. 9) the marginal posterior standard deviation of the offset $\psi_M - \phi_c$ in the c'th cell. This is used (in Fig. 10) to show how far the offset is from zero, and how this changes as we move from prior to posterior.

The results of a synthetic study are shown in Fig. 11. The onset-field was flat, $\phi_c = \psi_M$ for $c = 1, 2, \ldots, C$. The true values of ψ_M and ψ_0 are indicated by the dashed horizontal lines. The estimated onset and ending times in each pit are similar, and overlap with the true value.

Further details of the peak-counting statistic, mentioned at the end of this section in the main paper, are given on the left in Fig. 12, and contrasted with the corresponding results for the synthetic data, where the data shows no preference for the number of peaks.

In Figures 13 and 14 we give the results for an analysis on a 64×26 lattice, in the

onset-field model with a single phase of this section. The results are very similar to those obtained on the 32×13 lattice (compare Fig. 13 with the posterior fields in Figures 8 and 9 and compare Fig. 14 left and right).

MCMC simulation of the posterior distribution for the single-phase onset field model (model 2) on a 64×25-cell lattice was fairly time consuming (a couple of days). Output is shown in Fig. 15.

7. Recovering phase structure without spatial structure

7.1. Prior model for assignments of specimens to phases

In order to simulate the parameter model we introduced in this section, we use the following steps:

(a) the number of phases, $M \sim \text{Poisson}(\log(2))$;
(b) the span, $\psi_M - \psi_0 \sim U(0, U - L)$ and the offset, $\psi_0 \sim U(L, U - (\psi_M - \psi_0))$;
(c) intermediate phase boundaries, $(\psi_1, ..., \psi_{M-1}) \sim U(\psi_0, \psi_M)$ (but sorted in increasing order);
(d) scaled deposition rates; $\lambda_m \sim \text{Exp}(1)$;
(e) phase assignments for dated specimen, $\mathbf{m} \sim \prod_{i=1}^{K} p(\mathbf{m}(i)|\psi, \lambda)$ with

$$p(\mathbf{m}(i)|\psi, \lambda) = \left[\frac{\lambda_{\mathbf{m}(i)}(\psi_{\mathbf{m}(i)} - \psi_{\mathbf{m}(i)-1})}{\sum_{m'=1}^{M} \lambda_{m'}(\psi_{m'} - \psi_{m'-1})} \right];$$

(f) specimen ages, $\theta_i \sim U(\psi_{\mathbf{m}(i)-1)}, \psi_{\mathbf{m}(i)})$.

We introduce λ in order to model $p(\mathbf{m}(i)|\psi, \lambda)$, the probability that specimen i is assigned to phase $\mathbf{m}(i)$. We observed that, simulating the above from fixed M we found $(p_{1,2}, ..., p_M)|M \sim \text{Dirichelet}(1/2, 1/2, ..., 1/2)$ is a good approximation for M up to at least 10. It follows that the marginal prior for the probability of allocation to phase m of M is $\text{Beta}(1/2, M/2 - 1/2)$ to a very good approximation. Also, since

$$p(\mathbf{m}|p_1..p_M, M)p(p_1..p_M|M) \propto (p_1 p_2 ... p_M)^{1/2} p_1^{K_1} p_2^{K_2} ... p_M^{K_M}$$

we can use the known normalization of a Dirichelet distribution to compute the probability distribution $p(\mathbf{m}|M)$,

$$p(\mathbf{m}|M) \simeq \frac{\Gamma(M/2)}{\Gamma(1/2)^M} \prod_{m=1}^{M} \frac{\Gamma(K_m(\mathbf{m}) + 1/2)}{\Gamma(K + M/2)}. \tag{4}$$

If we sum this over $\mathbf{m} : K(\mathbf{m}) = (K_1, ..., K_M)$ we get the multivariate Polya distribution, since simply multiply by $K!/\prod_m K_m(\mathbf{m})!$ and replace $K_m(\mathbf{m})$ with K_m.

7.2. Estimated phase structure for the Bourewa and Stud Creek data sets

7.2.1. Bourewa

7.2.2. Stud Creek: hearth deposition rate

Fig. 16 shows the posterior mean value of $\lambda_\theta(t)$ (smooth curve, but on an arbitrary y-axis scale) and the estimated probability that there is at least one specimen in each year interval.

The later is calculated as $1-\prod_k(1-\Pr(\Theta_k \in [t,t+1]|y))$. This is very close to the summed likelihood, so the naive procedure of summing likelihoods would be legitimate and Holdaway et al. (2008) have given grounds to connect this function to the original hearth creation rate. The reason the two rate estimates differ is that the estimate based on $\lambda_\theta(t)$ is for a prior which allows any number of phases, so short phases with high rates may occur in the hiatus. A rate estimate which conditions on the final hiatus model gives rates much closer to the naive rate estimate.

7.2.2. Stud Creek: Bayes-factor details

When we fit this model to the Stud creek data, and come to compute Bayes factors for phase-structure model comparisons, we need to compute the prior probability for the number of phases M and the phase assignment $\mathbf{m}$. We have $p(M,\mathbf{m}) = p(M)p(\mathbf{m}|M)$ with $P(M)$ Poisson and $p(\mathbf{m}|M)$ in Equation (4). For $A = \{M = 3, \mathbf{m} = (1 \times 13, 3 \times 15)\}$ this is

$$\Pr(\mathbf{m}|M) \simeq \frac{\Gamma(3/2)}{\Gamma(28+3/2)}\frac{\Gamma(13+1/2)}{\Gamma(1/2)}\frac{\Gamma(15+1/2)}{\Gamma(1/2)}$$

so $\Pr(\mathbf{m}|M)$ is 9.8×10^{-11}. Notice that, if all specimen assignments to phases were equally likely, $\Pr(\mathbf{m}|M)$ would equal $3^{-28} \simeq 4 \times 10^{-14}$ for all $\mathbf{m}$ and $M = 3$.

7.2.2. Stud Creek: other priors for phase assignment of specimen

We have considered other priors on M and $\mathbf{m}$. In one of these we used a 'library' prior. This prior is simply a look-up table for $p(M,\mathbf{m})$. Let $C = \{M = 2, \mathbf{m} = (1 \times 13, 2 \times 15)\}$. We fit the Stud creek data using the library prior and $p(A) = p(B) = p(C) = 1/3$. Just three specific phase models are allowed, with equal probability. The rate parameters λ are no longer relevant, since the library gives the probability for specimen phase assignments. The posterior density is now $p(M,\psi,\mathbf{m},\theta|y)$ with overall prior

$$p(M,\psi,\mathbf{m},\theta) = p(\psi|M)p(\theta|\psi,\mathbf{m})p(M,\mathbf{m}),$$

and $p(\psi|M)$ and $p(\theta|\psi,\mathbf{m})$ are unchanged. Note that we added the C-state in order to ensure our MCMC phase-boundary birth-death process is irreducible, since it adds and deletes a single phase boundary at a time, and cant jump from $M = 1$ to $M = 3$. Adding a state in this way does not effect the relative prior weighting of states A and B, so if we simply throw out of our final MCMC sample any states in C, the sampled distribution of all variables is the same as it would have been if $p(C) = 0$ from the start.

This prior is useful for doing model comparison for specific models, each given with fixed M and $\mathbf{m}$; this is the case for the A and B model comparison. If we compute the Bayes factor for model A against model B under the Library-model weighting we get 2×10^{10}. This is fairly close to the Bayes factor for A *v.* B (8×10^9) we got for the same phase-structure comparison, but with other aspects of the model different (the multivariate Polya assignment model, where λ and ψ remain connected even when w condition on $\mathbf{m}$ and M).

8. Fitting the onset-field for multiple phases

In this section we give, for the random-phase onset-field model of this section in the main paper, the equivalent summary statistics to those presented in Section 6 for the single phase

onset-field model: MCMC output in Figures 17 and 18; mean fields in Fig. 19; image of mean divided by standard deviation for elapsed time $\psi_M - \phi_c$ to settlement in Fig. 20; threshold probability in 21.

The prior and posterior distributions of the number of phases are plotted in Figures 22 and 23 (top histogram in each Figure). The lower graph in Fig. 23 shows $\widehat{p(m|y)}$ plotted against $m = 1, 2, \ldots, 5$. Support for $M = m$ phases increases with m, as far as we have reliable estimates, though the Bayes factors would be negligibly larger than one.

9. Conclusions

Acknowledgements

References

Hammersley, J. (1977). Comment on "spatial contact models for ecological and epidemic spread" by Denis Mollison. *J. R. Statist. Soc. B 39*, 319.

Holdaway, S., P. Fanning, and E. Rhodes (2008). Challenging intensification: human-environment interactions in the Holocene. *The Holocene 18*, 403–412.

Nicholls, G. and P. Nunn (2009). On building and fitting a spatio-temporal change-point model for settlement and growth at Bourewa, Fiji Islands. *submitted*.

Paiva, L. and S. F. Jr (2007). Universality class of isotropic on-lattice Eden clusters. *J. Phys. A: Math. Theor.*, F43–F49.

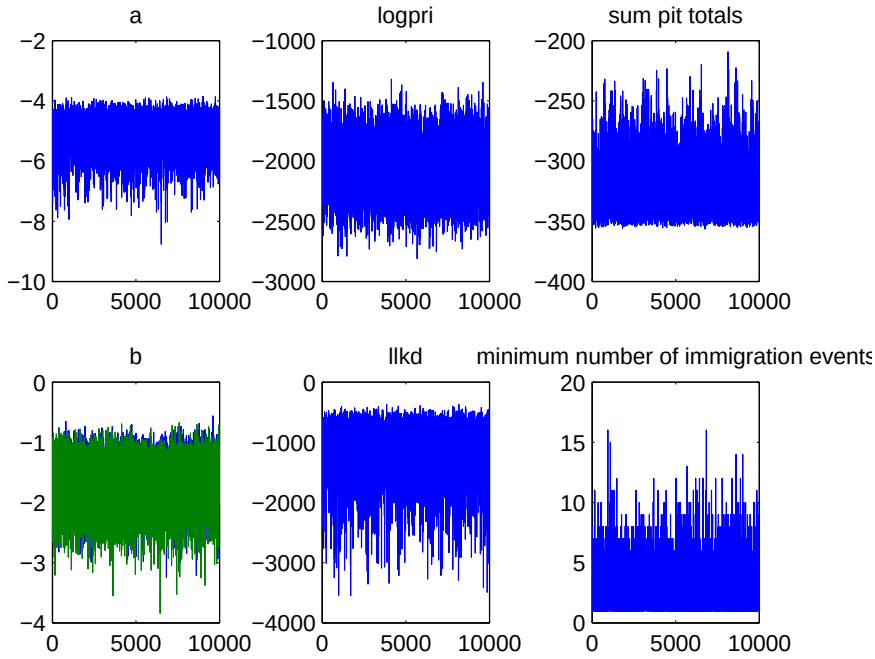

Fig. 4. MCMC run statistics for the prior simulation of the single-phase onset field model (model 2).

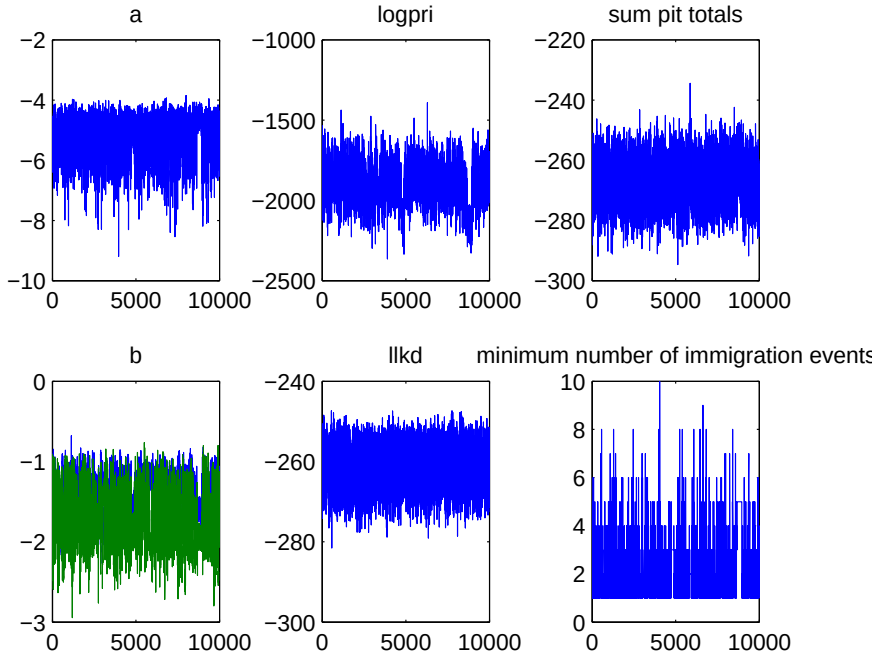

Fig. 5. MCMC run statistics for the posterior simulation of the single-phase onset field model (model 2).

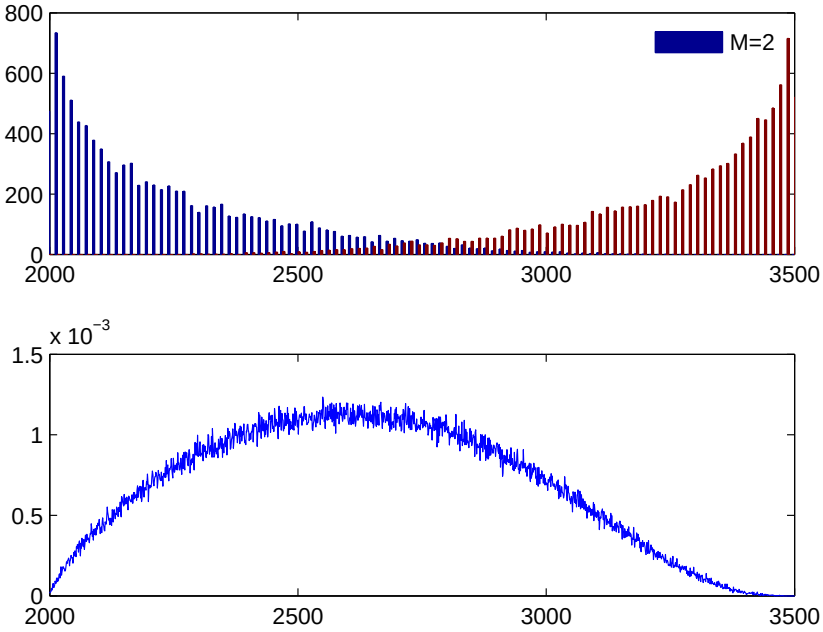

Fig. 6. (top) Prior distributions with ψ_0 at the left and ψ_M at the right end of the interval, with x-axis in years BP. (bottom) the prior probability for there to be at least one specimen in a given year interval. x-axis in years BP, y-axis unnormalised probability.

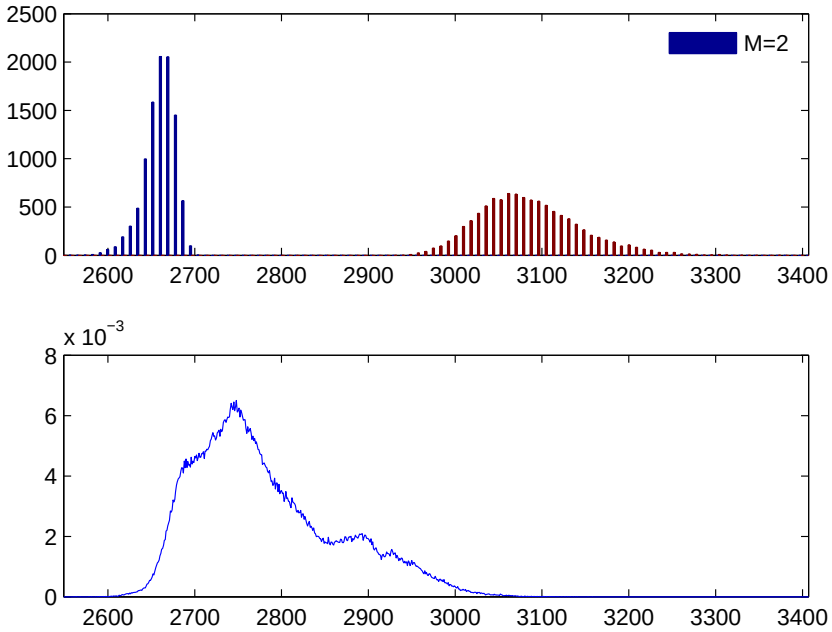

Fig. 7. As Figure 6 for the posterior.

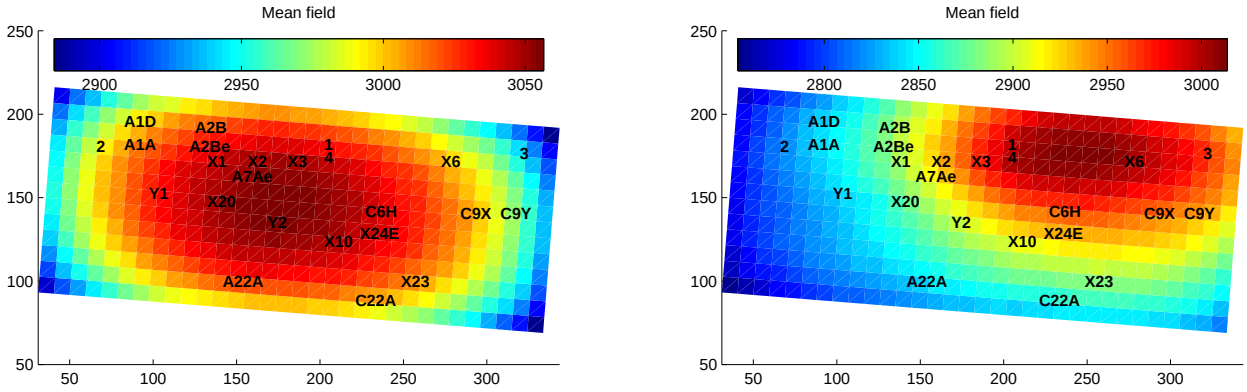

Fig. 8. Second model, single phase onset field default setup (left) prior mean onset field ϕ_c (right) posterior

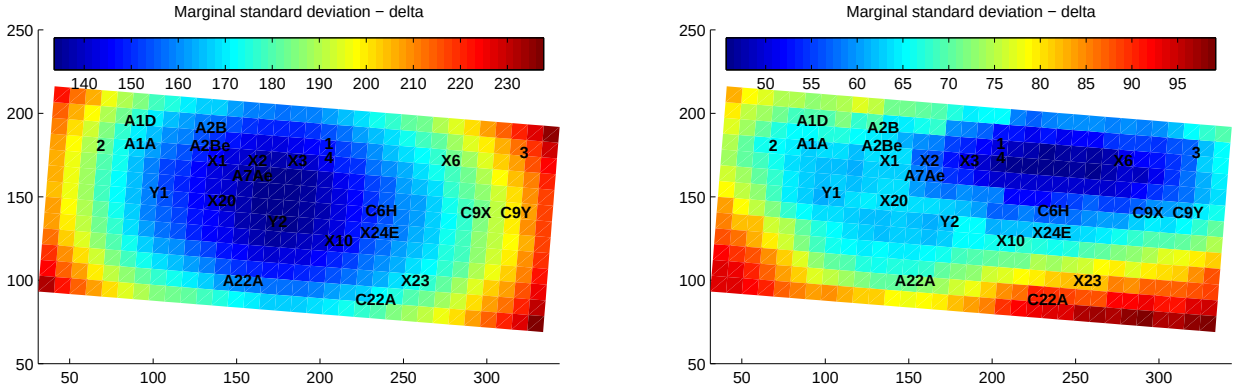

Fig. 9. Second model, single phase onset field default setup (left) prior standard deviation of the elapsed time to settlement, $\psi_m - \phi_c$ (right) posterior

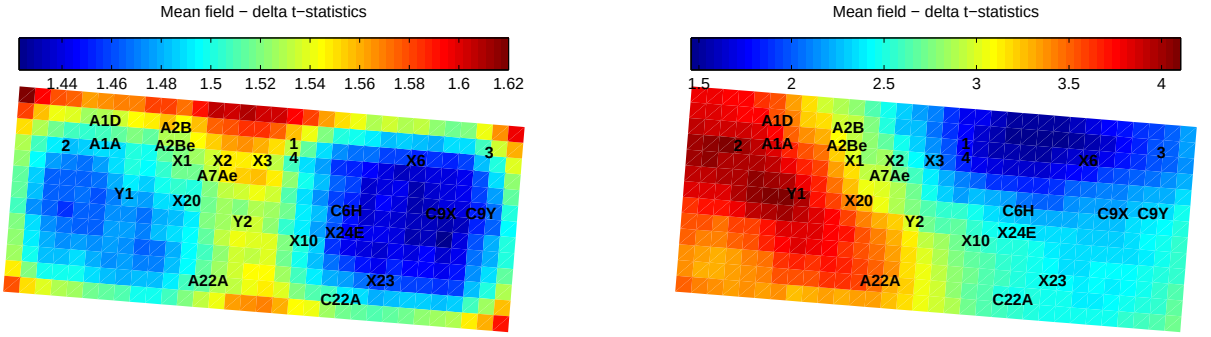

Fig. 10. Second model, single phase onset field default setup (left) mean of $\psi_m - \phi_c$ over its standard deviation in the prior (right) posterior

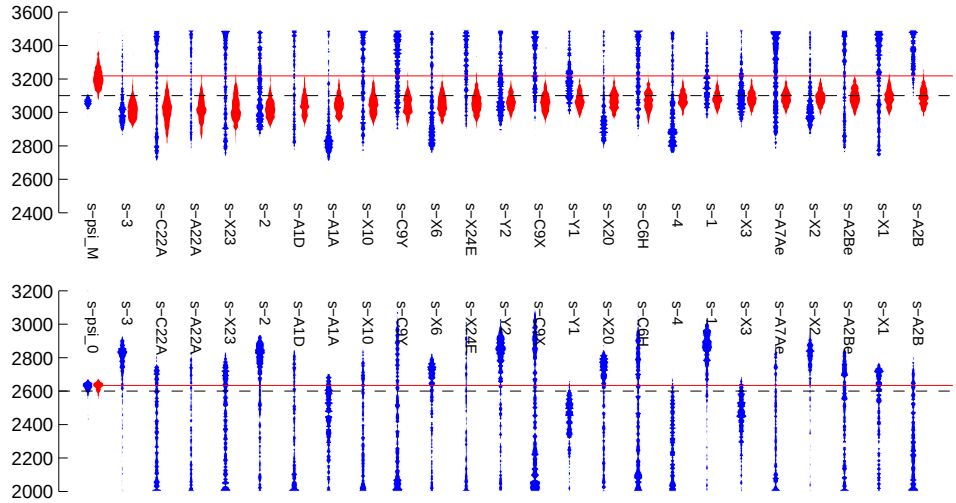

Fig. 11. (top) onset (bottom) ending time. Leftmost blue histogram fits the first single-pit/single phase model. The leftmost red histogram gives the onset-field result (default settings). The other red histograms are the onset field model values in each pit. The blue histograms are the single-phase pit-by-pit analysis.

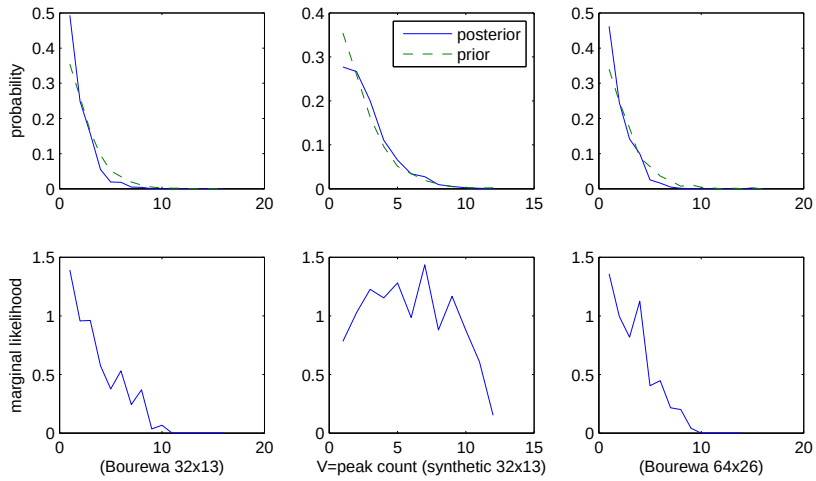

Fig. 12. (top row) the prior (green) and posterior (blue) distributions of V, the peak-count statistic. (bottom row) the marginal likelihood $\Pr(V = v|y)$ estimated as the ratio of green and blue curves on the 32×13 lattice except rightmost, and with the default hyper-parameter setup $(A, B) = (10, 1)$. (left column) Bourewa data (middle column) Synthetic data, (right column) Bourewa data 64×25

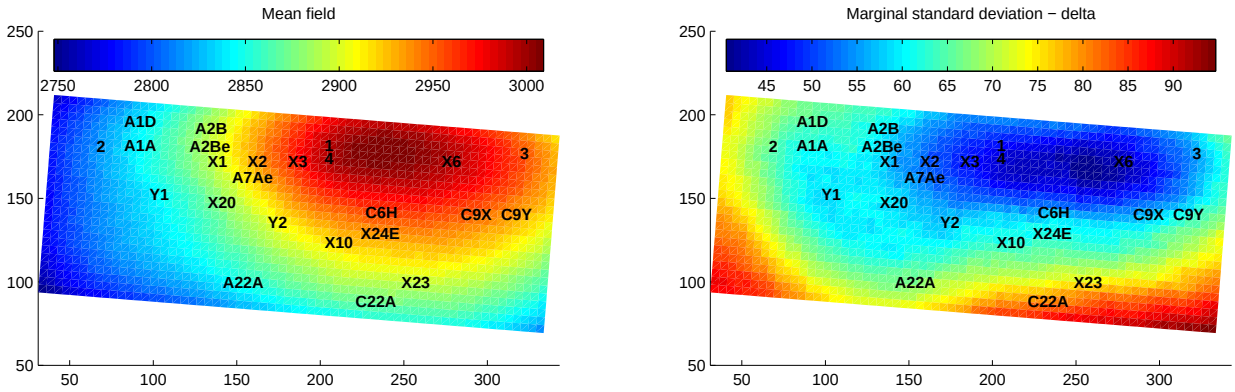

Fig. 13. Analysis on a 64×25 lattice covering the same region as the default setup. (left) Posterior marginal mean field. (right) Posterior marginal standard deviation of the elapsed time to settlement, $\psi_m - \phi_c$.

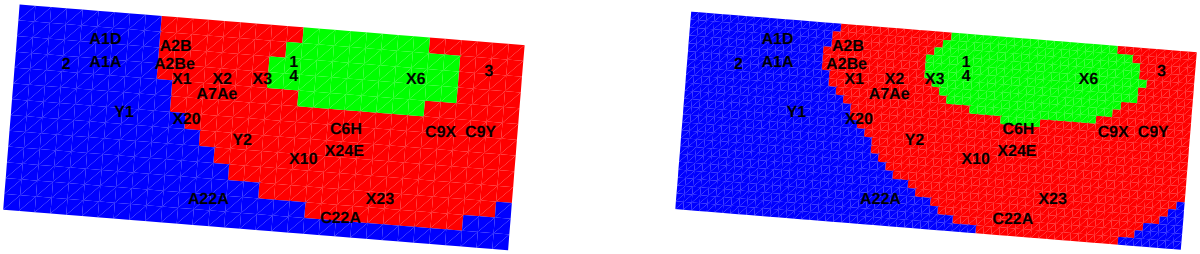

Fig. 14. (right) Analysis on a 64×25 lattice covering the same region as the default setup (left, with 32×13 cells). (green) cells occupied less than 150 years after first arrival with probability at least 0.8 (blue) cells occupied more than 150 years after first arrival with probability at least 0.8.

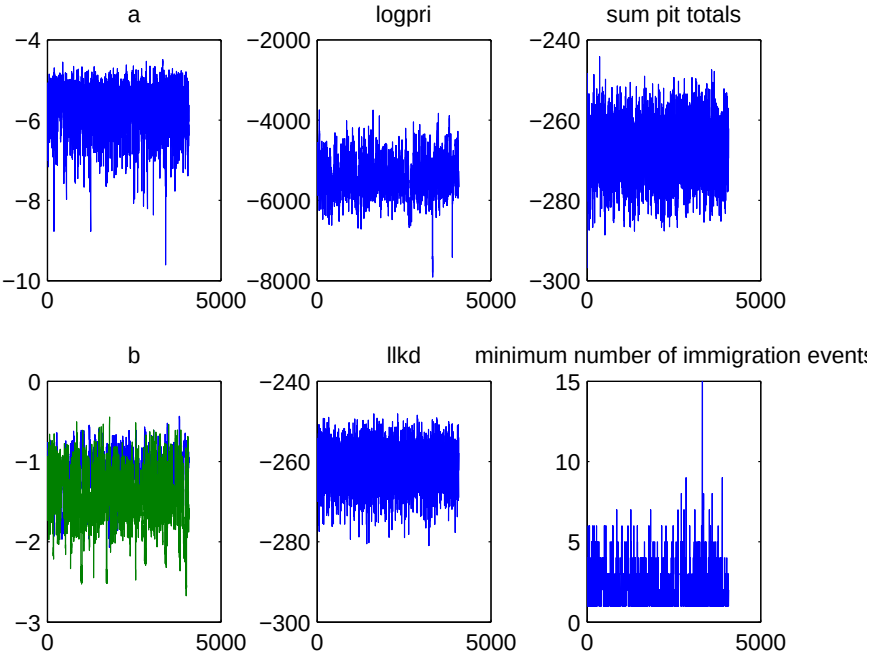

Fig. 15. MCMC run statistics for the posterior simulation of the single-phase onset field model (model 2) on a 64×25-cell lattice.

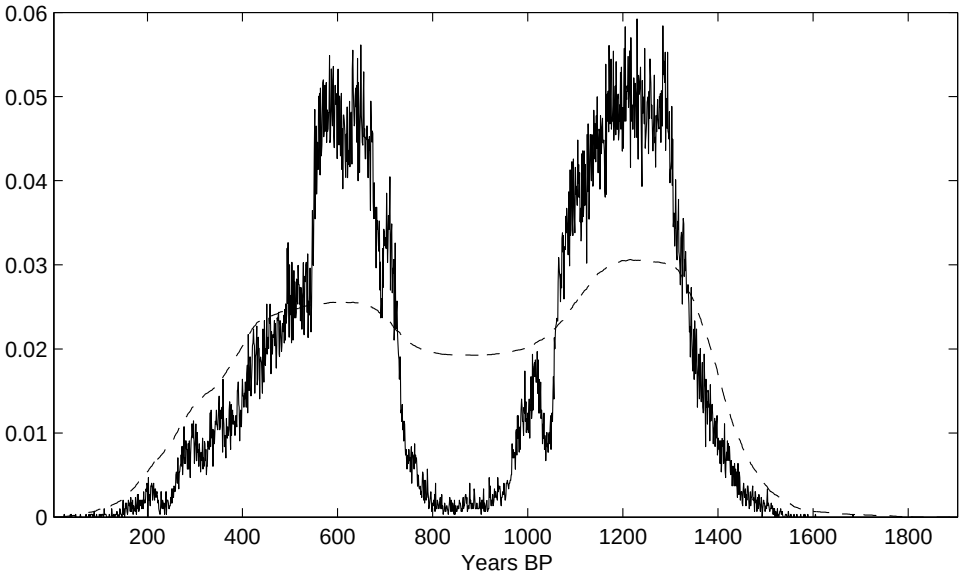

Fig. 16. (histogram) the estimated posterior probability that there is at least one specimen age in any given year interval, which closely approximates the summed likelihoods. (curve) the posterior mean deposition rate, $\mathrm{E}(\Lambda_\theta(t))$ as a function of age.

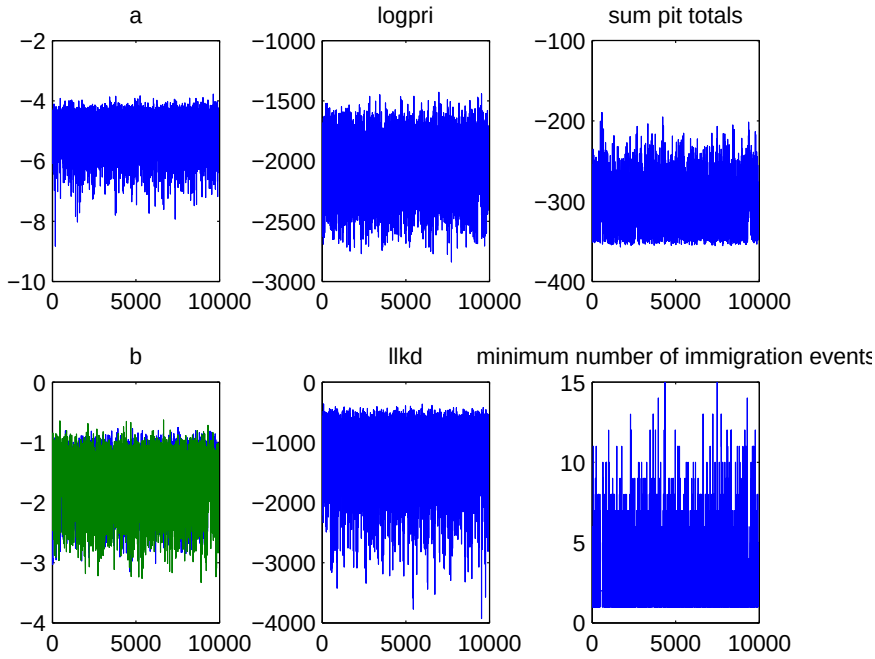

Fig. 17. MCMC run statistics for the prior simulation of the multiple-phase onset field model (model 4).

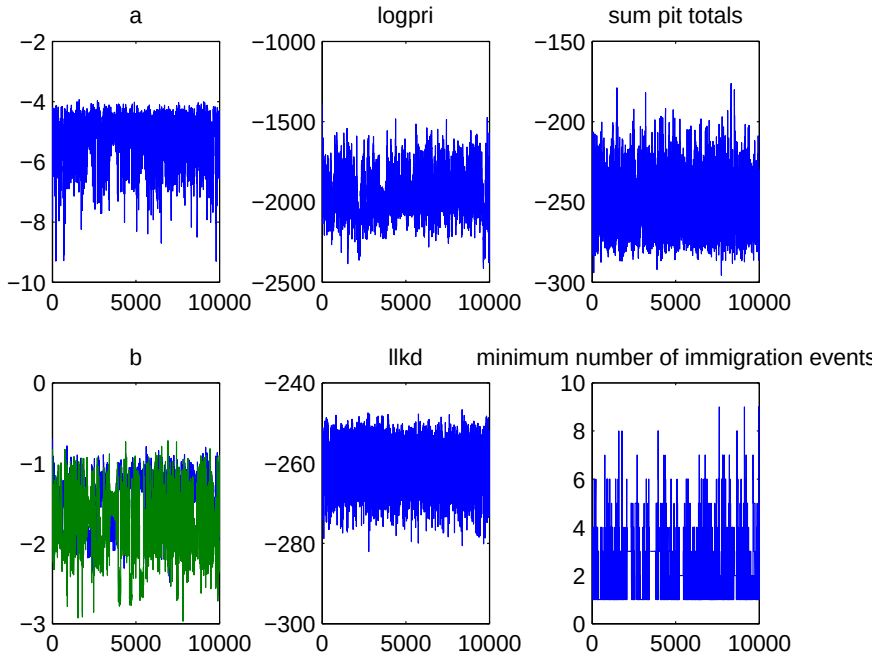

Fig. 18. MCMC run statistics for the posterior simulation of the multiple-phase onset field model (model 4).

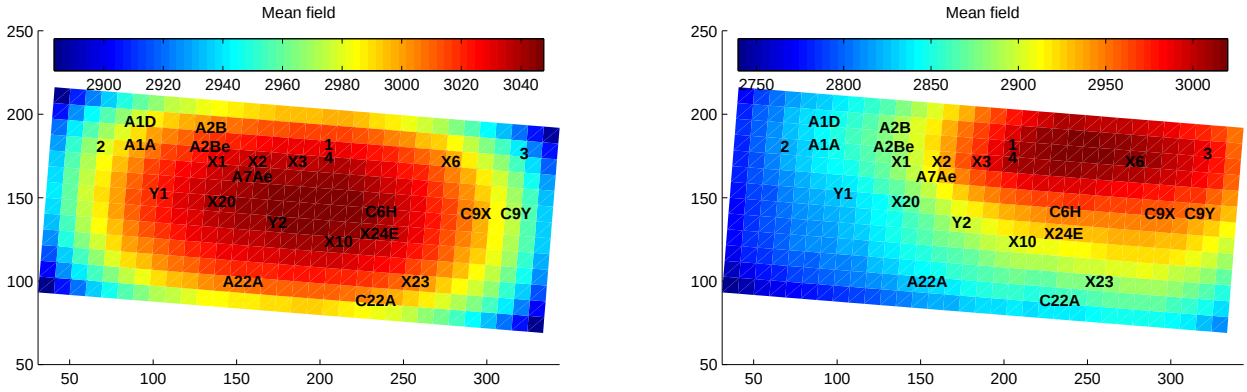

Fig. 19. Final model, random phase onset field (left) prior mean onset field ϕ_c (right) posterior

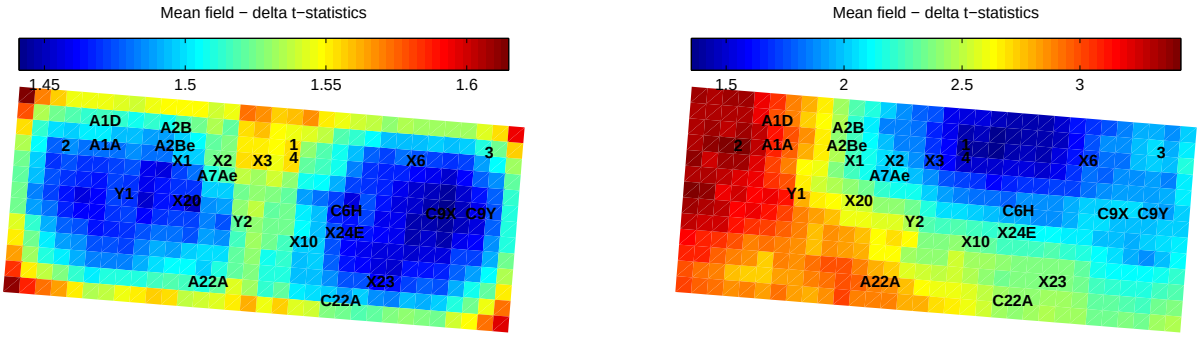

Fig. 20. Final model, random phase onset field (left) mean of $\psi_m - \phi_c$ over its standard deviation in the prior (right) posterior

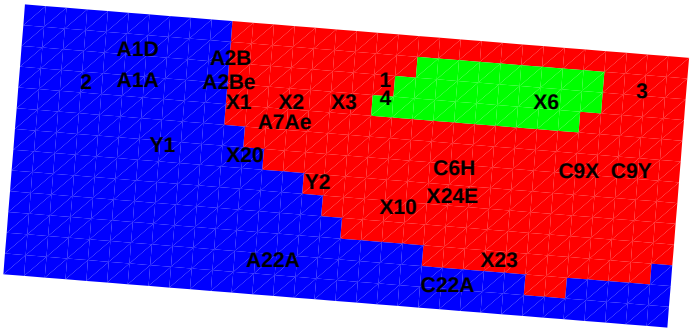

Fig. 21. Support for spread: the posterior probability that $\psi_M - \phi_c < 150$ (respectively, $\psi_M - \phi_c > 150$) is greater than 0.8 for cells c in the blue (green) region, in the multiple-phase onset-field model of Section 8.

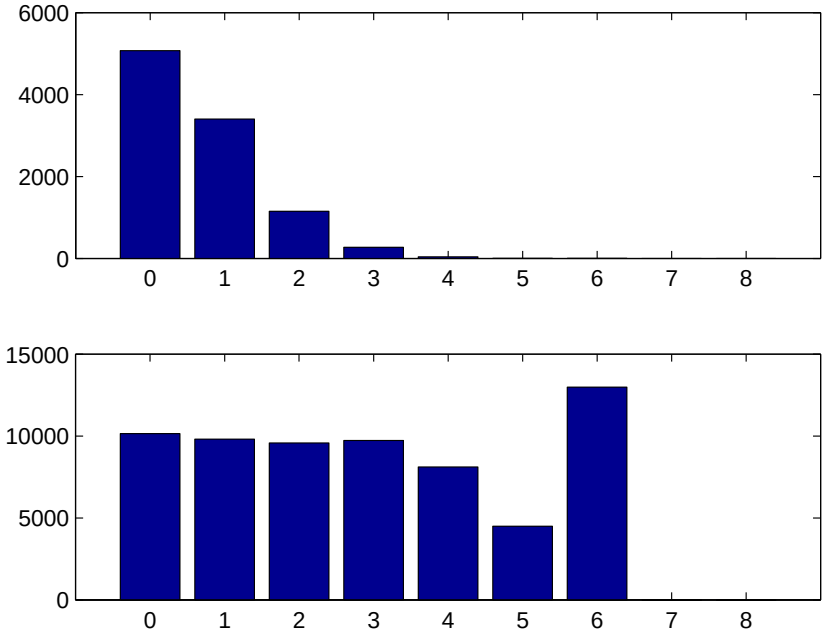

Fig. 22. The prior distribution of the number of phases in the random-phase model with onset-field is shown at top. The corresponding marginal likelihoods are shown below. These should of course be flat. The prior distribution is Poisson($\log(2)$).

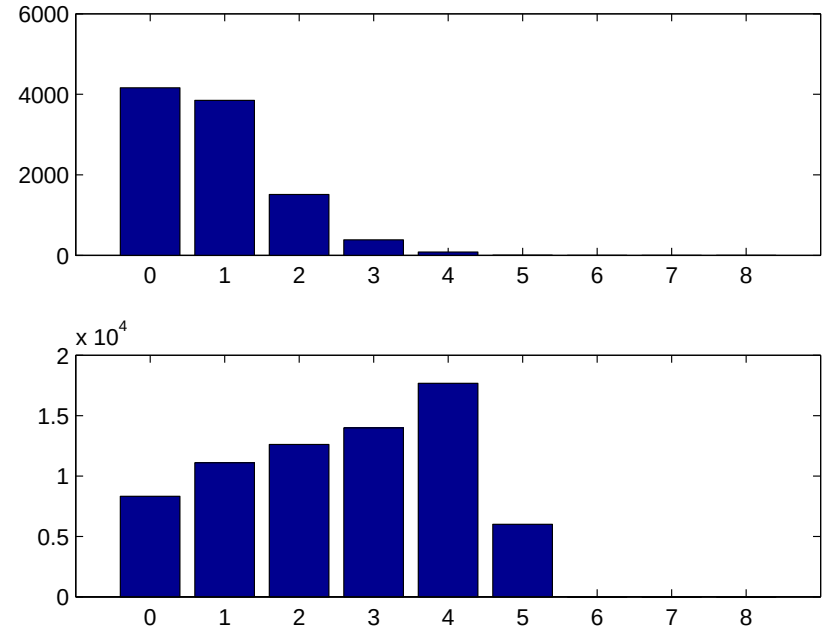

Fig. 23. As for Figure 22 but here the posterior distribution and marginal likelihood.