Block composite likelihood models for analysis of large spatial datasets

Jo Eidsvik, Benjamin A. Shaby and Brian J. Reich

Abstract Large spatial datasets become more common as a result of automatic sensors, remote sensing and the increase in data storage capacity. But large spatial datasets are hard to analyse. Even in the simplest Gaussian situation, parameter estimation and prediction are troublesome because one requires matrix factorization of a large covariance matrix. We consider a composite likelihood construction built on the joint densities of subsets of variables. This composite model thus splits a datasets in many smaller datasets, each of which can be evaluated separately. These subsets of data are combined through a summation giving the final composite likelihood. Massive datasets can be handled with this approach. In particular, we consider a block composite likelihood model, constructed over pairs of spatial blocks. The blocks can be disjoint, overlapping or at various resolution. The main idea is that the spatial blocking should capture the important correlation effects in the data. Estimates for unknown parameters as well as optimal spatial predictions under the block composite model are obtained. Asymptotic variances for both parameter estimates and predictions are computed using Godambe sandwich matrices. The procedure is demonstrated on 2D and 3D datasets with regular and irregular sampling of data. For smaller data size we compare with optimal predictors, for larger data size we discuss and compare various blocking schemes.

Jo Eidsvik
Department of Mathematical Sciences, NTNU, Norway, e-mail: joeid@math.ntnu.no

Benjamin A. Shaby
Department of Statistics, University of California, Berkeley, US, e-mail: bshaby@gmail.com

Brian J. Reich
Department of Statistics, North Carolina State University, US, e-mail: brian_reich@ncsu.edu

Ninth International Geostatistics Congress, Oslo, Norway, June 11. – 15., 2012

1 Introduction

In recent years there has been a tremendous increase in the magnitude of massive geocoded scientific datasets. These developments have triggered demand for more sophisticated statistical modeling and methodology for such data. The computations required for inference and prediction in spatial Gaussian process models, the central construct in spatial statistics, are challenging for large datasets because they require manipulations of large covariance matrices. In particular, evaluation of the likelihood function necessitates performing inverse and determinant operations, both of which are computationally intractable for large matrices.

In this paper we implement a unified framework for parameter estimation and prediction based on the composite likelihood (Lindsay, 1988). The composite likelihood (CL) is a product of several joint likelihoods of subsets of the data. One important special case is the pairwise likelihood, which is the product of all bivariate marginal likelihoods. Here, we use a form of the CL function defined as the product of joint density functions of pairs of spatial blocks. The motivation behind the spatial blocking strategy is that it captures much of the spatial dependence, while still providing the divide and conquer aspect of the CL.

The block CL model reduces the computational burden to $O(n)$, where n is the number of data, and the hidden constant depends on the block sizes. Moreover, the usual memory restrictions for large datasets are avoided since the blocks of data can be loaded into memory separately.

The asymptotic covariance for maximum CL estimates is given by a sandwich matrix (Godambe, 1960) rather than the usual Fisher information matrix for maximum likelihood estimates (MLEs). We show how to use the CL for the crucial complementary problem of spatial prediction, which has not previously been considered. We demonstrate how to construct predictions at unobserved sites that are optimal under the block CL, the composite analogue to Kriging. This prediction approach allows fast computing and we derive asymptotic prediction variances under the CL model, which have the familiar sandwich form.

The earliest use of the CL for spatial data seems to be Curriero and Lele (1999), who used the pairwise form of the CL to estimate covariance parameters. The main contribution of the current work is to utilize blocking and to provide predictions under the CL model. See also Eidsvik et al. (2011).

2 Composite likelihood

We consider a Gaussian response variable $Y(s)$ along with a $p \times 1$ vector of spatially-referenced explanatory variables $x(s)$ which are associated through a spatial regression model

$$Y(s) = x'(s)\beta + w(s) + \varepsilon(s), \tag{1}$$

where the spatial location $s \in \mathscr{D}$ is in 2, 3 or 4 dimensions. Moreover, $\beta = (\beta_1, \dots, \beta_p)'$ is the regression parameter, and $\varepsilon(s) \sim N(0, \tau^2)$ is independent error. The spatial residual $w(s)$ provides structural dependence, capturing the effect of unobserved covariates with spatial pattern. The covariance structure of the Gaussian process $w(s)$ is typically characterised by a small number of parameters. We denote the collection of all covariance parameters θ, which includes the nugget effect τ^2. One common model is the exponential $\mathrm{Cov}(w(s), w(s')) = C(s', s) = \sigma^2 \exp(-\phi|s' - s|)$, with variance σ^2 and spatial decay parameter ϕ.

e assume data are available at n locations $\{s_1, \dots, s_n\}$, and denote the collection of data $Y = (Y(s_1), \dots, Y(s_n))'$. Then $Y \sim N(X\beta, \Sigma)$, where $\Sigma = \Sigma(\theta) = C + \tau^2 I_n$, with $C(i, j) = \mathrm{Cov}(w(s_i), w(s_j))$. Moreover, row i of matrix X contains the explanatory variables $x'(s_i)$. The log likelihood is

$$\ell(Y; \beta, \theta) = -n/2 \log(2\pi) - \frac{1}{2} \log|\Sigma| - \frac{1}{2}(Y - X\beta)' \Sigma^{-1} (Y - X\beta). \tag{2}$$

Noting that Σ is $n \times n$, the difficulty with the usual maximum likelihood methods is apparent; evaluating the log-likelihood requires computing $|\Sigma|$ and a quadratic form that includes Σ^{-1}, both of which are computationally intractable for large n.

2.1 The block composite model

Instead, we next present a block CL, where we partition the region D into M blocks $D_1, \dots, D_M$, with $\cup_k D_k = D$, $D_k \cap D_l = \emptyset$, for all pairs of blocks k, l. Denote the response in block $k = 1, \dots, M$ as $Y_k = \{Y(s_i); s_i \in D_k\}$. The number of sites in block k is n_k, $\sum_k n_k = n$. Let further $Y_{kl} = (Y_k', Y_l')'$ be the collection of data in block k and l. We define the block composite log likelihood as

$$\begin{aligned} \ell_{CL}(Y; \beta, \theta) &= \sum_{k=1}^{M-1} \sum_{l>k} \ell(Y_{kl}; \beta, \theta) \\ &= \sum_{k=1}^{M-1} \sum_{l \in N_k, l>k} \left[-\frac{1}{2} \log|\Sigma_{kl}| - \frac{1}{2}(Y_{kl} - X_{kl}\beta)' \Sigma_{kl}^{-1} (Y_{kl} - X_{kl}\beta) \right]. \end{aligned} \tag{3}$$

where the second sum goes over all blocks l in a neighborhood N_k of block k, and $l > k$. Here, $X_{kl} = (X_k', X_l')'$ is the collection of all covariates in block k and l, and Σ_{kl} is the $(n_k + n_l) \times (n_k + n_l)$ covariance matrix

$$\Sigma_{kl} = \begin{bmatrix} \Sigma_{kl}(1) & \Sigma_{kl}(1,2) \\ \Sigma_{kl}(2,1) & \Sigma_{kl}(2) \end{bmatrix}. \tag{4}$$

This covariance matrix is thus partitioned into four parts, where $\Sigma_{kl}(1)$ is the $n_k \times n_k$ covariance matrix of Y_k, $\Sigma_{kl}(2)$ is the $n_l \times n_l$ covariance matrix of Y_l, and $\Sigma_{kl}(1,2) = \Sigma_{kl}'(2,1)$ is the $n_k \times n_l$ cross-covariance between Y_k and Y_l. The block

CL approach is a natural compromise for spatial models, as the number of blocks M represents a trade-off between computational and statistical efficiency. Figure 1 illustrates blocking schemes over a two dimensional region.

The maximum CL estimates of θ and β are given by

$$(\hat{\beta}_{CL}, \hat{\theta}_{CL}) = \text{argmax}_{\beta,\theta} \left[\ell_{CL}(Y; \beta, \theta) \right]. \tag{5}$$

In general, the maximum CL estimators are known to be consistent and asymptotically normal under the same conditions as MLEs (Lindsay, 1988). We compute the maximum CL estimates by a Fisher-scoring algorithm. The maximum is attained in about 5 iterations. The computer cost is small, since only small matrices are factorized.

The asymptotic covariance of $\hat{\theta}_{CL}$ has a sandwich form (Godambe, 1960), $\hat{\theta}_{CL} \sim N(\theta, G^{-1})$, where

$$G = G(\theta) = H(\theta) J^{-1}(\theta) H(\theta),$$

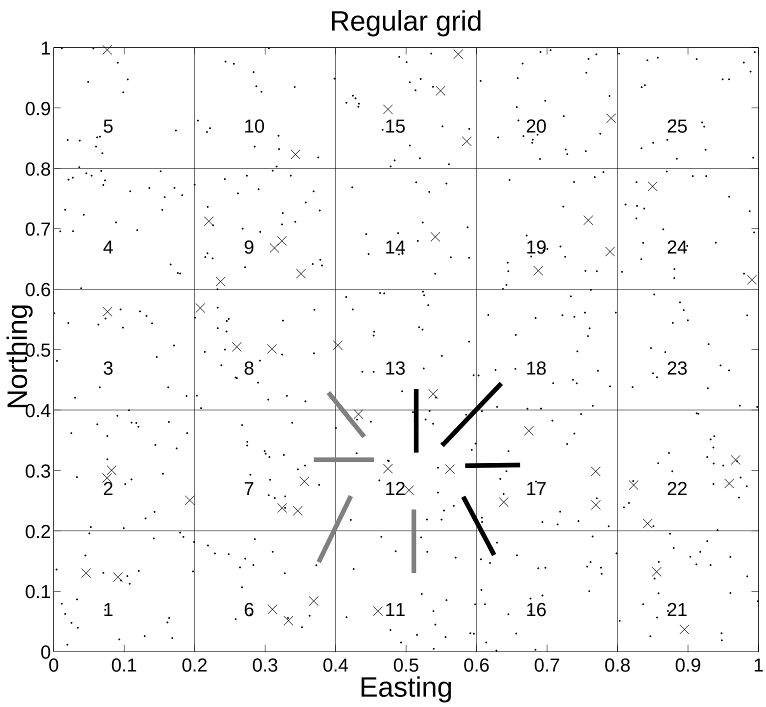

Fig. 1 Observation sites illustrated by '.' and predictions sites by 'x'. A block CL splits the spatial domain, here $\mathscr{D} = (0,1)x(0,1)$, into blocks. Each block communicates pairwise with each of its neighbours. An interior block has eight neighbours. The block $k = 12$ has four neighbours with larger indeces $l > k$ (black edges).

$$H(\theta) = -E\left(\frac{\partial^2 \ell_{CL}(Y;\theta)}{\partial \theta^2}\right), \quad J(\theta) = \text{Var}\left(\frac{\partial \ell_{CL}(Y;\theta)}{\partial \theta}\right).$$

We note that in the case of the full likelihood function, $H(\theta)J^{-1}(\theta) = I$, so $G(\theta)$ is just the Fisher information.

2.2 Prediction

Consider the task of making predictions at $n_{k0} \geq 1$ unobserved sites, all situated within block k. To get the best linear unbiased predictor (BLUP) or kriging surface, we must invert a large $n \times n$ matrix. Instead, we next present predictions derived from the CL model. The first step is to augment the data vector such that $Y_k^a = (Y_{k0}', Y_k')'$. By including Y_{k0} as unobserved data in the CL and setting the derivative of ℓ_{CL} equal to 0, we obtain the composite predictions $\hat{Y}_{k0}$.

The contribution of the unobserved data Y_{k0} to the CL is given by block terms (k,l), $l \in N_k$. We organise these pairs such that block k is always at the top in every (k,l) block-pair, so that the $(n_{k0}+n_k+n_l) \times (n_{k0}+n_k+n_l)$ precision matrix for $(Y_k'^a, Y_l')'$ is

$$\Sigma_{kl}^{-1} = Q_{0kl} = \begin{bmatrix} Q_{0kl}(0) & Q_{0kl}(0,1) & Q_{0kl}(0,2) \\ Q_{0kl}(1,0) & Q_{0kl}(1) & Q_{0kl}(1,2) \\ Q_{0kl}(2,0) & Q_{0kl}(2,1) & Q_{0kl}(2) \end{bmatrix}. \quad (6)$$

The block CL at the unobserved locations is thus

$$\begin{aligned} \ell_{CL}(Y_{k0}) = \sum_{l \in N_k} \Big[& \text{const} - \frac{1}{2}(Y_{k0} - X_{k0}\beta)' Q_{0kl}(0)(Y_{k0} - X_{k0}\beta) \\ & - (Y_{k0} - X_{k0}\beta)' Q_{0kl}(0,1)(Y_k - X_k\beta) \\ & - (Y_{k0} - X_{k0}\beta)' Q_{0kl}(0,2)(Y_l - X_l\beta) \Big], \end{aligned} \quad (7)$$

now regarded as a function of the unobserved data Y_{k0}, and where the $n_{k0} \times p$ matrix X_{k0} collects the explanatory variables at prediction sites in block k.

The first and second derivatives of $\ell_{CL}(Y_{k0})$ are obtained by differentiating the quadratic form. The first derivative is

$$\frac{d\ell_{CL}(Y_{k0})}{dY_{k0}} = -\sum_{l \in N_k} \Big[Q_{0kl}(0)(Y_{k0} - X_{k0}\beta) + Q_{0kl}(0,1)(Y_k - X_k\beta) + Q_{0kl}(0,2)(Y_l - X_l\beta) \Big]. \quad (8)$$

Setting the derivative equal to 0 gives the block composite prediction:

$$\begin{aligned} \hat{Y}_{k0} &= X_{k0}\beta + A_0^{-1} b_0, \\ A_0 &= \sum_{l \in N_k} Q_{0kl}(0), \end{aligned}$$

$$b_0 = -\sum_{l \in N_k} \Big[Q_{0kl}(0,1)(Y_k - X_k\beta) + Q_{0kl}(0,2)(Y_l - X_l\beta) \Big]. \tag{9}$$

The asymptotic variance of the block composite prediction is described by a Godambe sandwich.

2.3 Extensions

The model described above can be extended in various directions. Many applications include multivariate responses at each spatial location. If the locationwise covariances are stationary, this is included by a Kronecker-product in the covariance structure. One could also imagine non-stationarity in the model, where uni- or multivariate variables have block-dependent covariance parameters.

The framework also allows certain hierarchical models. In the example below with seismic data, the goal is to predict the latent variables, given indirect seismic observations. One could use the same approach for spatio-temporal models that develop over space and time coordinates.

The main requirement of the CL construction is a joint model for pairs of variables. In our context this is given by the multivariate Gaussian density, but the pairwise joint models can also be constructed in other situations. For instance, they are directly available in skew-Normal models or for special heavy-tailed distributions. For complex hierarchical models one can approximate the joint model of data by using numerical integration schemes.

3 Numeric example

Several examples of block composite likelihood was presented in Eidsvik et al (2011). Here, we present a seismic example on a regular 2D grid and a joint frequency dataset in 3D with irregular sampling. We use the examplee to compare blocking strategies and to discuss the computational advantages with the CL approach.

3.1 Time lapse seismic

The current example is created to mimic the situation with time-lapse seismic amplitude data, where two seismic surveys collected before and during production are compared to map the fluid and pressure changes in petroleum reservoirs. The reservoir variable of interest here is the time-lapse differences of P-wave impedance. The depth contrasts of P-impedance are related to the response variable of time-lapse stacked seismic amplitude data.

Focus is on estimating the model parameters and to predict the latent P-impedance variables. The prediction is done by augmenting the dataset with the missing latent variables, solving for the distinction of interest, and computing the associated Godambe sandwich matrices to assess the uncertainties.

In order to compare the results with full likelihood modeling, we use a 2D model (northing and depth) of size $n = 30 \cdot 100 = 3000$. We try different blocking schemes and compare the results with the optimal ones using the full likelihood model and the associated prediction (BLUP). We use the model with prior mean 0 for the time-lapse reservoir properties. The spatial correlation function is of the Matern type with smoothness parameter $3/2$. The correlation range is about 10 cells. We set the standard error of the reservoir changes to be twice as large in a reservoir zone, which is specified to be within the stratigraphic zone between depth index 30 and 50. The response variable is convolved in the depth direction with a seismic wavelet of a Mexican hat type, with bandwidth 40 cells. Because of this vertical smoothing in the likelihood, we use a CL blocking scheme which is always of full length in the vertical direction, while it covers sub-blocks in the lateral dimension.

Figure 2 shows one realization of the reservoir properties (top, left) and the synthetic seismic reflection data (top, right). Note the large changes in the reservoir zone, with only small contrasts below and above this zone.

The middle and bottom rows of Figure 2 show the predictions and prediction errors, conditional on the time-lapse seismic data. We display the optimal solutions using the full likelihood model to the left, while the CL solutions with 3 blocks is to the right. For predictions there is hardly any difference between the two solutions. For the CL results, the prediction variances are slightly increased at the block boundaries. This is natural, since the Godambe information corrects for the model simplification.

We note that the computationally aspects are much better for the CL approach. For this model that is stationary in the lateral direction, the CL is here based on matrix computions for only a pair of blocks, and then using these results for all other blocks.

We next compare parameter estimation and prediction mean square errors (MSE) and coverage probabilities compared over 1000 replicates of synthetic data, all generated with the same model. Table 1 shows the MSE and coverage probabilities (0.90 nominal level) of parameter estimates and predictions. The columns correspond to different number of blocks for the CL model. The rightmost column represents the full size model with optimal solutions. We note that prediction covarages and mean square errors are very similar for the CL and the full models. In particular when we use 3 or more blocks in the CL expression. The parameter estimates clearly have larger uncertainty for the CL,1 model. This occurs because important aspects of the correlation structure is missing from the CL,1 model, looking only at pairs of traces.

Table 1 Simulation study showing mean square error and coverage probabilities for 5 models over 1000 replicates. Full refers to the optimal solution in this case. CL, k refers to a blocksize k composite likelihood model.

	CL, 1	CL, 3	CL, 5	CL, 10	Full
MSE (σ)	0.0104	0.0099	0.0099	0.0098	0.0091
MSE (ϕ)	0.0067	0.0033	0.0033	0.0032	0.0030
MSE (τ)	0.00084	0.00084	0.00084	0.00085	0.00083
Coverage (σ)	0.78	0.86	0.89	0.91	0.91
Coverage (ϕ)	0.86	0.89	0.90	0.90	0.91
Coverage (τ)	0.92	0.92	0.92	0.90	0.92
MSPE	0.0366	0.0357	0.0356	0.0355	0.0355
Prediction coverage	0.90	0.90	0.90	0.90	0.90
Computer time (sec)	0.6	3.2	9.8	59	140

3.2 Joint frequency data

We study a joint frequency dataset acquired in an iron mine in Norway. The size is $n = 11,107$, see Ellefmo and Eidsvik (2009) for more details.

We use the block CL model with different block sizes. The blocks are constructed by a Voronoi / Delaunay tessellation adapted to the (north,east) coordinates of the data, with cells extending for all depths. The tessellation is made by random sampling, without replacement, among all data sites, which on average gives smaller area blocks where sampling locations are dense. We compare the composite likelihood results with the predictive process model (Banerjee et al., 2008), a dimension-reduction technique using a fixed set of knots.

Table 2 shows the parameter estimates, the average mean squared prediction error and coverage probabilities for a hold-out set of 1000 prediction sites. We compare three common covariance functions: the exponential model specified by $\Sigma(h) = \sigma^2\exp(-\phi h) + \tau^2 I(h=0)$, Cauchy(3) which is $\Sigma(h) = \sigma^2(1+\phi h)^{-3} + \tau^2 I(h=0)$, and Matérn(3/2) with $\Sigma(h) = \sigma^2(1+\phi h)\exp(-\phi h) + \tau^2 I(h=0)$. The parameter

Table 2 Joint frequency data: Parameter estimates, mean square prediction error (MSPE) and coverage probabilities for prediction distributions. The different columns correspond to different number of blocks for the composite likelihood (CL) model and different knot sizes for the predictive process (PP) models.

		CL, 200	CL, 40	CL, 10	PP, 1000	PP, 1500
	$\hat{\sigma}$	0.42	0.42	0.42	0.44	0.45
Exponential	$\hat{\phi}$	0.031	0.030	0.028	0.013	0.015
	$\hat{\tau}$	0.30	0.30	0.30	0.32	0.31
	MSPE	145	144	144	182	171
	Pred cov (0.95)	0.95	0.95	0.95	0.94	0.94

estimates are very similar for all three CL models, but different between the CL and predictive process models. In particular, the range parameter ϕ is smaller for predictive process models. This implies a larger effective spatial correlation, imposing a smoother process. To some extent, the predictive process models compensate for this with larger estimated variance terms.

The mean squared prediction error is clearly smaller for the CL models than for the predictive process models. Even with 1500 knots, the mean squared prediction error for the predictive process model is 15% larger. The coverage probabilities are very good for all models considered. The computation times required is about 10 seconds for CL, 200, while the predictive process with 1500 knots takes a few minutes.

4 Discussion

In this paper we use a block composite likelihood model for parameter estimation and prediction in large Gaussian spatial models. The properties of the composite likelihood are well-understood in the context of parameter estimation. Here we also present a method for spatial prediction using the block composite likelihood.

The block composite likelihood performs well for reasonably-sized blocks, especially for spatial prediction. The required computation time is reduced considerably relative to likelihood-based calculations using the divide and conquer strategy inherent in the composite likelihood. We recommend testing results with a couple of choices of block sizes (hundreds to thousands sites per block) and blocking designs.

5 References

Banerjee S, Gelfand AE, Finley AO and Sang H (2008) Gaussian Predictive Process Models for Large Spatial Data Sets, Journal of the Royal Statistical Society, Series B, 70(4), 825-848.
Curriero FC, and Lele S (1999) A composite likelihood approach to semivariogram estimation, Journal of Agricultural, Biological, and Environmental Statistics, 4(1), 9-28.
Eidsvik J, Shaby BJ, Reich BJ, Wheeler M, and Niemi J (2011) Estimation and prediction in spatial models with block composite likelihoods, Technical Report 7, 2011, Mathematical Sciences, NTNU.
Ellefmo, SL, and Eidsvik J (2009) Local and Spatial Joint Frequency Uncertainty and its Application to Rock Mass Characterisation, Rock mechanics and rock engineering, 42(4), 667-688.
Godambe VP (1960) An optimum property of regular maximum likelihood estimation, The Annals of Mathematical Statistics, 31(4), 1208-1211
Lindsay BG (1988) Composite likelihood methods, Contemp. Math., 80, 221-239.

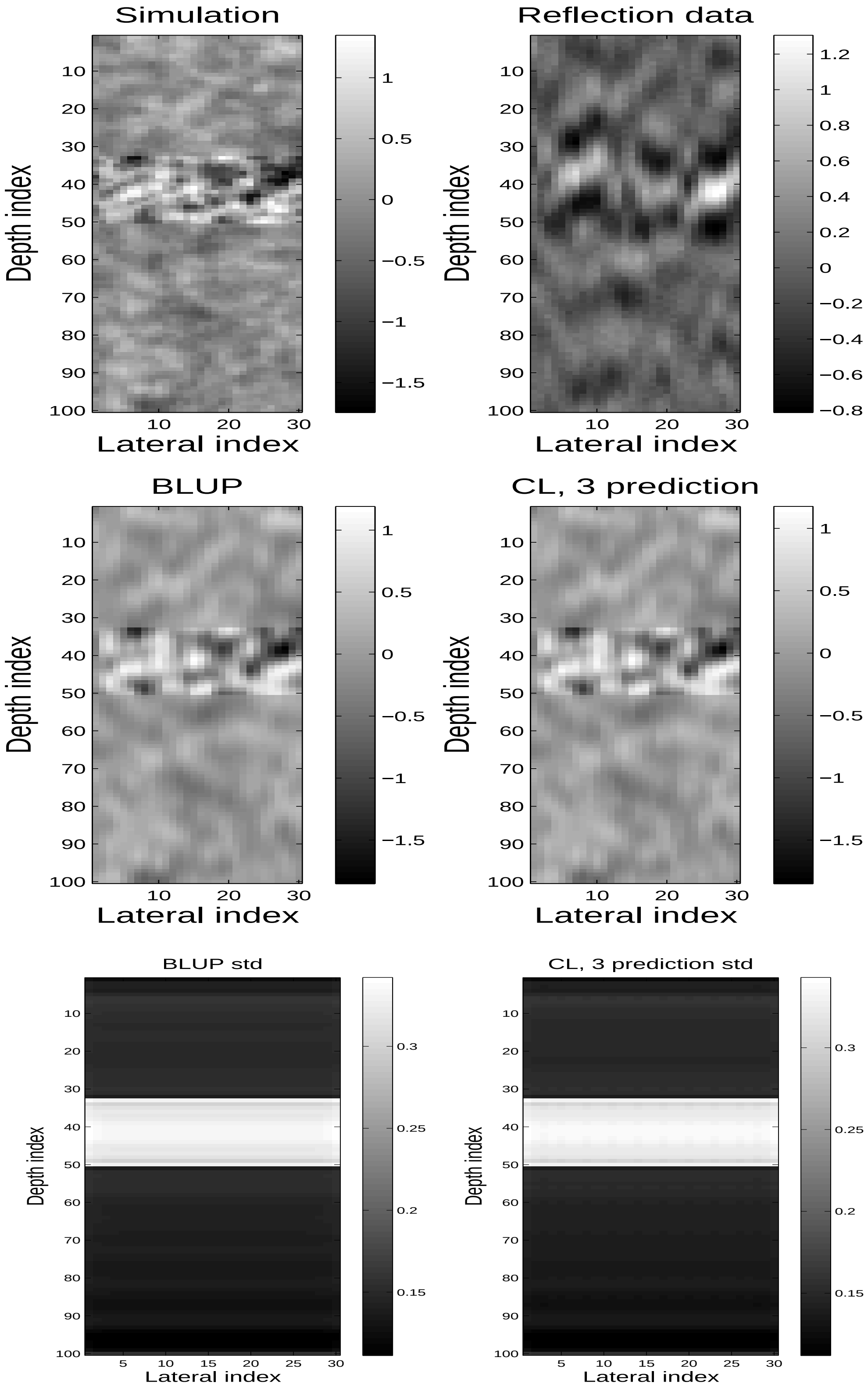

Fig. 2 Synthetic 2D example of size 30×100: Simulated reservoir change (top, left) and associated seismic reflections (top, right). Middle: Predicted reservoir changes using full likelihood model (left) and the composite likelihood model with size-3 blocks (right). Bottom: Prediction standard errors using the full likelihood model (left) and the composite likelihood model with size-3 blocks (right).