

DEPARTMENT OF STATISTICS  
University of Wisconsin – Madison  
1210 West Dayton Street  
Madison, WI 53706

TECHNICAL REPORT NO. 1077

April 28, 2003

## A Multiresolution Tree-Structured Spatial Linear Model

Jun Zhu and Wei Yue

[jzhu@stat.wisc.edu](mailto:jzhu@stat.wisc.edu)

<http://www.stat.wisc.edu/~jzhu>

## Abstract

Multiresolution spatial models are able to capture complex dependence correlation in spatial data and are excellent alternatives to the traditional random field models for mapping spatial processes. Because of the multiresolution structures, spatial process prediction can be obtained by direct and fast computation algorithms. However the existing multiresolution models usually assume a simple constant mean structure, which may not be suitable in practice. In this article, we focus on a multiresolution tree-structured spatial model developed by Huang et al. (2002) and extend the model to incorporate a linear regression mean. We explore the properties of the multiresolution tree-structured spatial linear model in depth and estimate the parameters in the linear regression mean and the spatial dependence simultaneously. An Expectation-Maximization algorithm is adopted to obtain the maximum likelihood estimates of the model parameters and the corresponding information matrix. Given the estimated parameters, a one-pass change-of-resolution Kalman-filter algorithm is implemented to obtain the best linear unbiased predictor of the true underlying spatial process. For illustration, the methodology is applied to map optimally crop yield in a Wisconsin field, after accounting for the field conditions by a linear regression.

*Keywords and Phrases:* Change-of-resolution Kalman filter; Random field model; Spatial prediction.

## 1 Introduction

Over the spatial region of interest  $D$  in two dimensions, consider a spatial random process of interest,  $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$ , which can be modeled as the totality of an underlying mean process  $\{\mu(\mathbf{s}) : \mathbf{s} \in D\}$ , a random process  $\{\eta(\mathbf{s}) : \mathbf{s} \in D\}$  that captures spatially correlated variability, and a measurement-error process  $\{\epsilon(\mathbf{s}) : \mathbf{s} \in D\}$  that represents spatially uncorrelated exogenous variability. That is, for each location  $\mathbf{s}$  in  $D$ , we can write the  $Z(\cdot)$  process as,

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \eta(\mathbf{s}) + \epsilon(\mathbf{s}). \quad (1)$$

One main interest in application has been to predict optimally the underlying process,

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + \eta(\mathbf{s}); \quad \mathbf{s} \in D, \quad (2)$$

which consists of both the fixed underlying mean process and the spatial variability over the region  $D$ . A traditional approach is to model  $Y(\cdot)$  by a Gaussian process with a zero mean and a

semivariogram that captures the spatial dependence among  $Y(\cdot)$ . Under the model, the best linear unbiased predictor (BLUP) of  $Y(\cdot)$  can be computed, given the measurements  $\{Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n)\}$  taken at spatial locations  $\mathbf{s}_1, \dots, \mathbf{s}_n$ , a procedure also called kriging. Well-known examples of kriging include ordinary kriging where the mean process is assumed to be constant (i.e.,  $\mu(\cdot) \equiv \mu$ ) and universal kriging where the mean process is assumed to have a regression form.

However, the traditional kriging methods become inefficient in processing vast amounts of complex spatial data. More recently, several alternative modeling approaches accompanied by fast prediction algorithms have been developed, which we categorize as “multiresolution models”. In this approach, the spatial region  $D$  is partitioned into cells  $\{D_{j,k} : k = 1, \dots, N_j\}$  on the  $j$ -th resolution, where  $N_j$  is the total number of cells on the  $j$ -th resolution,  $j = 1, \dots, J$ , and  $J$  is the total number of resolutions. The partition is nested from the coarsest resolution  $j = 1$  to the finest resolution  $j = J$ , such that each cell  $D_{j,k}$  on the  $j$ -th resolution is partitioned into cells  $D_{j+1,k'}$  on the  $(j + 1)$ -th resolution, for  $j = 1, \dots, J - 1$ . The true process of interest in this approach is an aggregation of  $Y(\cdot)$  over each cell  $D_{j,k}$ :  $y_{j,k} \equiv |D_{j,k}|^{-1} \int_{D_{j,k}} Y(\mathbf{s}) d\mathbf{s}$ . Let  $\mu_{j,k} \equiv |D_{j,k}|^{-1} \int_{D_{j,k}} \mu(\mathbf{s}) d\mathbf{s}$  and  $\eta_{j,k} \equiv |D_{j,k}|^{-1} \int_{D_{j,k}} \eta(\mathbf{s}) d\mathbf{s}$ . Then, we have  $y_{j,k} = \mu_{j,k} + \eta_{j,k}$ . Associated with each cell, the datum is denoted by  $z_{j,k}$  and is modeled as  $z_{j,k} = y_{j,k} + \epsilon_{j,k}$ , where  $\epsilon_{j,k}$  represents measurement error;  $k = 1, \dots, N_j$ ,  $j = 1, \dots, J$ . Kolaczyk (1999) and Nowak (1999) factorize the joint distribution of  $\{z_{j,k}\}$  on the finest resolution across the coarser resolutions. Prediction of  $\{\mu_{j,k}\}$  is then obtained in a Bayesian hierarchical modeling framework. Extending the work by Chou et al. (1994), Huang and Cressie (1997) and Huang et al. (2002) specify the underlying true process  $y_{j,k}$  via a multiresolution autoregressive tree-structured model, which consists of a set of nodes inter-connected by a set of edges directed from the coarser resolutions to the finer resolutions. Let  $(j, k)$  denote the  $k$ -th node on the  $j$ -th resolution located in the cell  $D_{j,k}$ . The parent node of  $(j, k)$ ,  $pa(j, k)$ , is the node on the  $(j - 1)$ -th resolution which has an edge directed from  $pa(j, k)$  to  $(j, k)$ , whereas the child nodes of  $(j, k)$ ,  $ch(j, k)$ , is the set of nodes on the  $(j + 1)$ -th resolution which have an edge directed from  $(j, k)$  to  $ch(j, k)$ :  $\{ch(j, k, 1), \dots, ch(j, k, n_{j,k})\}$ , where  $ch(j, k, i)$  denotes the  $i$ -th child node and  $n_{j,k}$  denotes the number of child nodes of  $(j, k)$ . Moreover, when a node does not have a parent, it is assumed to come from the coarsest resolution  $j = 1$  and is called a root node. When a node does not have any children, it is assumed to come from the finest resolution  $j = J$  and is called a leaf node (Figure 1). Huang et al. (2002) develop a fast algorithm for computing the BLUP of

$\{y_{j,k}\}$  and Huang and Cressie (2001) extend the tree-structured model to more general graphical models.

---

Figure 1 here

---

Nevertheless the model by Huang et al. (2002) assumes that the mean process  $\mu(\cdot)$  is constant, which may not hold in many applications. The main purpose of this article is to extend the model to a multiresolution tree-structured spatial linear model (MTSLM), which allows for a spatial linear regression mean structure. Hence our main contribution is to generalize the multiresolution tree-structured spatial model, in analogy to extending ordinary kriging to universal kriging. Moreover, we obtain the maximum likelihood estimates (MLE) of the model parameters in the linear regression mean and spatial dependence structure simultaneously using an Expectation-Maximization (EM) algorithm. As a by-product, the information matrix of the parameter estimates are derived, which was not addressed previously. Given the estimated model parameters, we then use a one-pass change-of-resolution Kalman-filter algorithm to compute the (empirical) BLUP of  $\{y_{j,k}\}$ .

In Section 2, we describe the MTSLM and its properties. In Section 3, we use an EM algorithm to obtain the maximum likelihood estimates of the model parameters and their corresponding information matrix. We describe a change-of-resolution Kalman-filter for obtaining the optimal prediction of the true underlying processes in Section 4. Issues related to missing observations are addressed in Section 5. For illustration, an example of mapping crop yield in a Wisconsin field is given in Section 6 and conclusions are given in Section 7. Technical details of the statistical inference are given in the Appendix.

## 2 Multiresolution Tree-Structured Spatial Linear Model

Similar to (1), we use a measurement error model for the observations,

$$z_{j,k} = y_{j,k} + \epsilon_{j,k}, \quad k = 1, \dots, N_j, j = 1, \dots, J, \quad (3)$$

where  $z_{j,k}$  denotes the observation,  $y_{j,k}$  denotes the true underlying process, and  $\epsilon_{j,k}$  denotes the measurement error that captures exogenous variability independent of  $y_{j,k}$ , for the  $k$ -th location on the  $j$ -th resolution;  $k = 1, \dots, N_j, j = 1, \dots, J$ . We further assume that the measurement errors

$\{\epsilon_{j,k}\}$  have an independent Gaussian distribution with mean 0 and variance  $\phi_j$ :

$$\epsilon_{j,k} \sim N(0, \phi_j), \quad k = 1, \dots, N_j, j = 1, \dots, J. \quad (4)$$

For the underlying true process, we generalize the multiresolution tree-structured model developed by Huang et al. (2002) to have a linear regression mean structure, so that potential covariates can be included in the models. In particular, let

$$y_{j,k} = \mathbf{x}'_{j,k} \boldsymbol{\beta} + u_{j,k}, \quad k = 1, \dots, N_j, j = 1, \dots, J, \quad (5)$$

where  $\mathbf{x}_{j,k} \in \mathbb{R}^p$  denotes a vector of covariates associated with  $z_{j,k}$  and  $\boldsymbol{\beta} \in \mathbb{R}^p$  denotes a vector of regression coefficients, both in  $p$  dimensions. Further, we assume that the cells on a given resolution have equal sizes (i.e.,  $|D_j| \equiv |D_{j,k}|$ , for  $k = 1, \dots, N_j$ ) and the number of children on a given resolution is homogeneous (i.e.,  $n_j \equiv n_{j,k}$ , for  $k = 1, \dots, N_j$ ). Hence, the total number of nodes on the  $j$ -th resolution is  $N_j = N_1 n_1 \cdots n_{j-1}$ ,  $j = 2, \dots, J$ . We model the residual process  $\{u_{j,k}\}$  by a multiresolution tree-structure:

$$\begin{aligned} u_{1,k} &\sim N(0, \sigma_1^2), \quad k = 1, \dots, N_1, \\ \mathbf{u}_{ch(j,k)} &= u_{j,k} \mathbf{1}_{n_j} + \boldsymbol{\omega}_{ch(j,k)}, \quad k = 1, \dots, N_j, j = 1, \dots, J-1, \end{aligned} \quad (6)$$

where  $\mathbf{u}_{ch(j,k)} \equiv (u_{ch(j,k,1)}, \dots, u_{ch(j,k,n_j)})'$  denotes the  $n_j$  children of  $u_{j,k}$ ,  $ch(j,k,i)$  is the  $i$ -th child node of  $(j,k)$ ,  $\mathbf{1}_{n_j} \equiv (1, \dots, 1)'$  and  $\boldsymbol{\omega}_{ch(j,k)} \equiv (\omega_{ch(j,k,1)}, \dots, \omega_{ch(j,k,n_j)})'$  denotes the  $n_j$  children of  $\omega_{j,k}$  that captures random fluctuations independent of  $u_{j,k}$ . More specifically, the child nodes of  $(j,k)$  are  $\{(j+1, (k-1)n_j+1), \dots, (j+1, (k-1)n_j+n_j)\}$  and hence the  $i$ -th child node is  $ch(j,k,i) \equiv (j+1, (k-1)n_j+i)$ , for  $i = 1, \dots, n_j$ .

Huang et al. (2002) point out that, to match the physical conditions, it is necessary that an average of all the children's values be equal to their parent's value. That is,  $n_j^{-1} (\mathbf{y}'_{ch(j,k)} \mathbf{1}_{n_j}) = y_{j,k}$ , where  $\mathbf{y}_{ch(j,k)} \equiv (y_{ch(j,k,1)}, \dots, y_{ch(j,k,n_j)})'$  denotes the children processes of  $y_{j,k}$ . This property is referred to as the ‘‘mass-balanced property’’. Here we assume the following conditions under which the mass-balanced property holds:

$$n_j^{-1} (\mathbf{u}'_{ch(j,k)} \mathbf{1}_{n_j}) = u_{j,k}, \quad k = 1, \dots, N_j, j = 1, \dots, J-1. \quad (7)$$

and

$$n_j^{-1} (\mathbf{X}'_{ch(j,k)} \mathbf{1}_{n_j}) = \mathbf{x}_{j,k}, \quad k = 1, \dots, N_j, j = 1, \dots, J-1. \quad (8)$$

where the rows of the matrix  $\mathbf{X}_{ch(j,k)} \equiv [\mathbf{x}_{ch(j,k,1)}, \dots, \mathbf{x}_{ch(j,k,n_j)}]'$  correspond to the children covariates of  $\mathbf{x}_{j,k}$ . It follows from (6) and (7) that  $\mathbf{1}'_{n_j} \boldsymbol{\omega}_{ch(j,k)} = 0$ . In addition, we assume that the noises  $\{\omega_{j,k}\}$  have a Gaussian distribution with mean 0 and variance  $\sigma_j^2$ . Consequently, the variance-covariance matrix of  $\boldsymbol{\omega}_{ch(j,k)}$  is  $\sigma_{j+1}^2 \mathbf{H}_{n_j}$ , where  $\mathbf{H}_{n_j} \equiv \frac{n_j}{n_j-1} (\mathbf{I}_{n_j} - \frac{1}{n_j} \mathbf{1}_{n_j} \mathbf{1}'_{n_j})$ . That is,

$$\boldsymbol{\omega}_{ch(j,k)} \sim N(\mathbf{0}, \sigma_{j+1}^2 \mathbf{H}_{n_j}); \quad k = 1, \dots, N_j, j = 1, \dots, J-1. \quad (9)$$

Given the multiresolution tree-structured spatial linear model (MTSLM) defined in (3)–(9), we now explore the mean, variance, and covariance structure of the variables in the model. For this purpose, we denote the  $N_j$  variables on a given  $j$ -th resolution by  $\mathbf{z}_j \equiv (z_{j,1}, \dots, z_{j,N_j})'$ ,  $\mathbf{y}_j \equiv (y_{j,1}, \dots, y_{j,N_j})'$ ,  $\mathbf{u}_j \equiv (u_{j,1}, \dots, u_{j,N_j})'$ , and  $\boldsymbol{\omega}_j \equiv (\omega_{j,1}, \dots, \omega_{j,N_j})'$  for the observations and the underlying processes. Also let  $\mathbf{X}_j \equiv [\mathbf{x}_{j,1}, \dots, \mathbf{x}_{j,N_j}]'$  denote the covariates on the  $j$ -th resolution.

By (6) and (9),  $E(\mathbf{u}_j) = \mathbf{0}$ , and  $var(\mathbf{u}_1) = \sigma_1^2 \mathbf{I}_{N_1}$ ,  $var(\mathbf{u}_{j+1}) = (\mathbf{I}_{N_j} \otimes \mathbf{1}_{n_j}) var(\mathbf{u}_j) (\mathbf{I}_{N_j} \otimes \mathbf{1}_{n_j})' + var(\boldsymbol{\omega}_{j+1})$ , where  $var(\boldsymbol{\omega}_{j+1}) = \sigma_{j+1}^2 \mathbf{I}_{N_j} \otimes \mathbf{H}_{n_j}$ ;  $j = 1, \dots, J-1$ . A simplification of  $var(\mathbf{u}_j)$  gives:

$$\begin{aligned} var(\mathbf{u}_1) &= \sigma_1^2 \mathbf{I}_{N_1}, \\ var(\mathbf{u}_j) &= \frac{n_j}{n_j-1} \sigma_j^2 \mathbf{I}_{N_j} + \left( \frac{n_{j-2}}{n_{j-2}-1} \sigma_{j-1}^2 - \frac{1}{n_j-1} \sigma_j^2 \right) \mathbf{I}_{N_{j-1}} \otimes (\mathbf{1}_{n_{j-1}} \mathbf{1}'_{n_{j-1}}) \\ &\quad + \dots + \left( \frac{n_1}{n_1-1} \sigma_2^2 - \frac{1}{n_2-1} \sigma_3^2 \right) \mathbf{I}_{N_2} \otimes (\mathbf{1}_{n_2 \dots n_{j-1}} \mathbf{1}'_{n_2 \dots n_{j-1}}) \\ &\quad + \left( \sigma_1^2 - \frac{1}{n_1-1} \sigma_2^2 \right) \mathbf{I}_{N_1} \otimes (\mathbf{1}_{n_1 \dots n_{j-1}} \mathbf{1}'_{n_1 \dots n_{j-1}}), \quad j = 2, \dots, J. \end{aligned} \quad (10)$$

Further,  $cov(\mathbf{u}_j, \mathbf{u}_{j'}) = var(\mathbf{u}_j) \otimes \mathbf{1}'_{n_j \dots n_{j'-1}}$ ;  $1 \leq j < j' \leq J$ .

Although it is easier to summarize the covariances among  $\{u_{j,k}\}$  by the variance-covariance matrix  $var(\mathbf{u}_j)$  in (10), we now focus on the covariance between any pair of variables on the same resolution to gain more insight of the spatial dependence structure. While  $E(u_{j,k}) = 0$ , the usual second-order stationarity or intrinsic stationarity does not hold here, because of the special feature of the multiresolution tree structure. However, the model is stationary and isotropic in the following sense. Given two locations  $(j, k)$  and  $(j, k')$  on the  $j$ -th resolution, we denote the first common ancestor as  $an(j, k, k')$  and denote its corresponding resolution as  $j_{an}$ . When two locations are identical ( $k = k'$ ), then  $j_{an} = j$  and  $cov(u_{j,k}, u_{j,k'}) = var(u_{j,k}) = \sum_{j'=1}^j \sigma_{j'}^2$ . When two locations do not have a common ancestor, then they must be in two independent subtrees with two different root nodes, in which case we define  $j_{an} = 0$ . If  $j_{an} = 0$ , then  $cov(u_{j,k}, u_{j,k'}) = 0$ . When two

locations have a first common ancestor on a resolution  $j_{an} = 1, \dots, j-1$ , then it can be shown that  $cov(u_{j,k}, u_{j,k'}) = cov(u_{an(j,k,k')} + \omega_{j_{an}+1,\ell}, u_{an(j,k,k')} + \omega_{j_{an}+1,\ell'})$ , where  $(j_{an} + 1, \ell)$  and  $(j_{an} + 1, \ell')$  are the children of  $an(j, k, k')$  and the (not shared) ancestors of  $(j, k)$  and  $(j, k')$ , by the fact that these are the only two descendants of  $an(j, k, k')$  which are correlated via the matrix  $\sigma_{j_{an}+1}^2 \mathbf{H}_{n_{j_{an}}}$  in (9). Hence  $cov(u_{j,k}, u_{j,k'}) = var(u_{an(j,k,k')}) + cov(\omega_{j_{an}+1,\ell}, \omega_{j_{an}+1,\ell'}) = \sum_{j'=1}^{j_{an}} \sigma_{j'}^2 - \frac{1}{n_{j_{an}}-1} \sigma_{j_{an}+1}^2$ . That is,

$$cov(u_{j,k}, u_{j,k'}) = \begin{cases} 0 & ; \quad j_{an} = 0, \\ \sum_{j'=1}^{j_{an}} \sigma_{j'}^2 - \frac{1}{n_{j_{an}}-1} \sigma_{j_{an}+1}^2 & ; \quad j_{an} = 1, \dots, j-1, \\ \sum_{j'=1}^j \sigma_{j'}^2 & ; \quad j_{an} = j. \end{cases} \quad (11)$$

From (11), the covariance between two locations on the  $j$ -th resolution decreases as the resolution  $j_{an}$  of their first common ancestor decreases. Moreover, the covariance function  $cov(u_{j,k}, u_{j,k'})$  is invariant to the actual locations of  $(j, k)$  and  $(j, k')$  as long as the first common ancestor  $an(j, k, k')$  is on the same  $j_{an}$ -th resolution.

Now, by (5),  $E(\mathbf{y}_j) = \mathbf{X}_j \boldsymbol{\beta}$ ,  $var(\mathbf{y}_j) = var(\mathbf{u}_j)$ , and  $cov(\mathbf{y}_j, \mathbf{y}_{j'}) = cov(\mathbf{u}_j, \mathbf{u}_{j'}) = var(\mathbf{u}_j) \otimes \mathbf{1}'_{n_j \dots n_{j'-1}}$ ;  $1 \leq j < j' \leq J$ . Hence the underlying process  $\{y_{j,k}\}$  features a linear regression mean. Moreover, the variance-covariance structure of  $\{y_{j,k}\}$  is identical to that of  $\{u_{j,k}\}$  shown in (10) and (11). By (3),  $E(\mathbf{z}_j) = \mathbf{X}_j \boldsymbol{\beta}$ ,  $var(\mathbf{z}_j) = var(\mathbf{u}_j) + \phi_j \mathbf{I}_{N_j}$ ;  $j = 1, \dots, J$ . That is,

$$\begin{aligned} var(\mathbf{z}_1) &= (\sigma_1^2 + \phi_1) \mathbf{I}_{N_1}, \\ var(\mathbf{z}_j) &= \left( \frac{n_j}{n_j-1} \sigma_j^2 + \phi_j \right) \mathbf{I}_{N_j} + \left( \frac{n_{j-2}}{n_{j-2}-1} \sigma_{j-1}^2 - \frac{1}{n_j-1} \sigma_j^2 \right) \mathbf{I}_{N_{j-1}} \otimes (\mathbf{1}_{n_{j-1}} \mathbf{1}'_{n_{j-1}}) \\ &\quad + \dots + \left( \frac{n_1}{n_1-1} \sigma_2^2 - \frac{1}{n_2-1} \sigma_3^2 \right) \mathbf{I}_{N_2} \otimes (\mathbf{1}_{n_2 \dots n_{j-1}} \mathbf{1}'_{n_2 \dots n_{j-1}}) \\ &\quad + \left( \sigma_1^2 - \frac{1}{n_1-1} \sigma_2^2 \right) \mathbf{I}_{N_1} \otimes (\mathbf{1}_{n_1 \dots n_{j-1}} \mathbf{1}'_{n_1 \dots n_{j-1}}), \quad j = 2, \dots, J. \end{aligned} \quad (12)$$

Also  $cov(\mathbf{z}_j, \mathbf{z}_{j'}) = cov(\mathbf{y}_j, \mathbf{y}_{j'}) = cov(\mathbf{u}_j, \mathbf{u}_{j'}) = var(\mathbf{u}_j) \otimes \mathbf{1}'_{n_j \dots n_{j'-1}}$ ;  $1 \leq j < j' \leq J$ . The MTSLM is parameterized by the regression coefficients  $\boldsymbol{\beta}$ , the measurement error variances  $\{\phi_j : j = 1, \dots, J\}$ , and the noise variances  $\{\sigma_j^2 : j = 1, \dots, J\}$ . Since  $E(\mathbf{z}_j) = \mathbf{X}_j \boldsymbol{\beta}$ , (12), and  $cov(\mathbf{z}_j, \mathbf{z}_{j'}) = var(\mathbf{u}_j) \otimes \mathbf{1}'_{n_j \dots n_{j'-1}}$ , where  $var(\mathbf{u}_j)$  is in (10), it follows directly that the parameters are identifiable in the MTSLM.

### 3 Parameter Estimation by Maximum Likelihood

Because of the “mass-balanced” property, statistical inference of the MTSLM operates on the basis of groups of nodes instead of singleton nodes, where the child nodes of a common parent node form a group. Notationwise, for the resolution  $j = 2, \dots, J$ , let  $\{j, k\} \equiv \{(j, (k-1)n_{j-1} + 1), \dots, (j, (k-1)n_{j-1} + n_{j-1})\}$  denote the group of nodes that have a common parent, where  $k = 1, \dots, N_{j-1}$ . In fact, the common parent of  $\{j, k\}$  is the scalar node  $(j-1, k)$  on the  $(j-1)$ -th resolution. Since the number of groups on the  $j$ -th resolution is the number of (scalar) parent nodes on the  $(j-1)$ -th resolution (i.e.  $N_{j-1}$ ), we have  $k = 1, \dots, N_{j-1}$ . Now let  $\mathbf{z}_{j,k} \equiv (z_{j,(k-1)n_{j-1}+1}, \dots, z_{j,(k-1)n_{j-1}+n_{j-1}})'$ ,  $\mathbf{y}_{j,k} \equiv (y_{j,(k-1)n_{j-1}+1}, \dots, y_{j,(k-1)n_{j-1}+n_{j-1}})'$ ,  $\mathbf{u}_{j,k} \equiv (u_{j,(k-1)n_{j-1}+1}, \dots, u_{j,(k-1)n_{j-1}+n_{j-1}})'$ ,  $\mathbf{X}_{j,k} \equiv (\mathbf{x}_{j,(k-1)n_{j-1}+1}, \dots, \mathbf{x}_{j,(k-1)n_{j-1}+n_{j-1}})'$ ,  $\boldsymbol{\epsilon}_{j,k} \equiv (\epsilon_{j,(k-1)n_{j-1}+1}, \dots, \epsilon_{j,(k-1)n_{j-1}+n_{j-1}})'$ ,  $\boldsymbol{\omega}_{j,k} \equiv (\omega_{j,(k-1)n_{j-1}+1}, \dots, \omega_{j,(k-1)n_{j-1}+n_{j-1}})'$ , denote the vector of observations, the vector of underlying processes (original and detrended), the matrix of covariates, the vector of random error processes, and the vector of noise processes associated with  $\{j, k\}$ . Note that previously we used  $a_{j,k}$  to denote a scalar  $a$  on a scalar node  $(j, k)$  and now we use  $\mathbf{a}_{j,k}$  to denote a vector  $\mathbf{a}$  on a group of nodes  $\{j, k\}$ . On the other hand, for the coarsest resolution  $j = 1$ , the root nodes do not share a common parent. We let  $N_0 \equiv 1$ ,  $n_0 \equiv N_1$  and hence,  $\{1, 1\} \equiv \{(1, 1), \dots, (1, N_1)\}$ .

The MTSLM can now be rewritten in the vector form as:

$$\mathbf{z}_{j,k} = \mathbf{X}_{j,k}\boldsymbol{\beta} + \mathbf{u}_{j,k} + \boldsymbol{\epsilon}_{j,k}, \quad \boldsymbol{\epsilon}_{j,k} \sim N(\mathbf{0}, \boldsymbol{\Phi}_j); \quad k = 1, \dots, N_{j-1}, j = 1, \dots, J, \quad (13)$$

$$\mathbf{u}_{1,1} \sim N(\mathbf{0}, \sigma_1^2 \mathbf{I}_{N_1}), \quad (14)$$

$$\mathbf{u}_{j,k} = \mathbf{A}_{j,k} \mathbf{u}_{pa\{j,k\}} + \mathbf{w}_{j,k}, \quad \mathbf{w}_{j,k} \sim N(\mathbf{0}, \mathbf{W}_j); \quad k = 1, \dots, N_{j-1}, j = 2, \dots, J, \quad (15)$$

where  $\boldsymbol{\Phi}_j \equiv \text{diag}\{\phi_j \mathbf{1}_{n_{j-1}}\}$ ,  $pa\{j, k\}$  is the group of nodes that contain the scalar node  $(j-1, k)$  (parent of  $\{j, k\}$ ) on the  $(j-1)$ -th resolution, or equivalently, the group of nodes that share a common parent with  $(j-1, k)$ . Further  $\mathbf{A}_{j,k}$  is an  $n_{j-1} \times n_{j-2}$  matrix consisting of  $\mathbf{1}_{n_{j-1}}$  in the  $i$ -th column and  $\mathbf{0}_{n_{j-1}}$  in the other columns if the scalar parent node  $(j-1, k)$  is the  $i$ -th node within the group of nodes  $pa\{j, k\}$ , along with  $\mathbf{W}_j = \sigma_j^2 \mathbf{H}_{n_{j-1}}$  where  $\mathbf{H}_{n_{j-1}} \equiv \frac{n_{j-1}}{n_{j-1}-1} (\mathbf{I}_{n_{j-1}} - \frac{1}{n_{j-1}} \mathbf{1}_{n_{j-1}} \mathbf{1}'_{n_{j-1}})$  is defined shortly after (8).

Let  $\mathbf{z} \equiv (\mathbf{z}'_{1,1}, \mathbf{z}'_{2,1}, \dots, \mathbf{z}'_{2,N_1}, \dots, \mathbf{z}'_{J,1}, \dots, \mathbf{z}'_{J,N_{J-1}})'$  denote the complete set of observed values, let  $\mathbf{u} \equiv (\mathbf{u}'_{1,1}, \mathbf{u}'_{2,1}, \dots, \mathbf{u}'_{2,N_1}, \dots, \mathbf{u}'_{J,1}, \dots, \mathbf{u}'_{J,N_{J-1}})'$  denote the complete set of the latent values.

Now let  $\mathbf{u}_{j,k}^* \equiv (u_{j,(k-1)n_{j-1}+2}, \dots, u_{j,(k-1)n_{j-1}+n_{j-1}})'$ , which is essentially  $\mathbf{u}_{j,k}$  with the first entry omitted. By the mass-balanced property specified in (7), it suffices to work with a reduced version of  $\mathbf{u}$ :

$$\mathbf{u}^* \equiv (\mathbf{u}'_{1,1}, \mathbf{u}'_{2,1}, \dots, \mathbf{u}'_{2,N_1}, \dots, \mathbf{u}'_{J,1}, \dots, \mathbf{u}'_{J,N_{J-1}})'$$

because the values of the first child of each parent can be computed as the difference between  $n_{j-1}u_{j-1,k}$  and the sum of its sibling values  $\mathbf{1}'_{n_{j-1}-1}\mathbf{u}_{j,k}^*$ , for each  $j = 2, \dots, J$ . For ease of presentation, we assume that all of the observations are available and will discuss the case of missing data in Section 5.

From (13), the probability density of  $\mathbf{z}$  conditional on  $\mathbf{u}$  is

$$f(\mathbf{z}|\mathbf{u}) = \prod_{j=1}^J \prod_{k=1}^{N_{j-1}} f(\mathbf{z}_{j,k}|\mathbf{u}_{j,k}), \quad \text{where } \mathbf{z}_{j,k}|\mathbf{u}_{j,k} \sim N(\mathbf{X}_{j,k}\boldsymbol{\beta} + \mathbf{u}_{j,k}, \boldsymbol{\Phi}_j). \quad (16)$$

From (14) and (15), the probability density of  $\mathbf{u}^*$  is

$$\begin{aligned} f(\mathbf{u}^*) &\equiv f(\mathbf{u}_{1,1}) \prod_{j=2}^J \prod_{k=1}^{N_{j-1}} f(\mathbf{u}_{j,k}^*|\mathbf{u}_{pa\{j,k\}}), \quad \text{where} \\ \mathbf{u}_{1,1} &\sim N(\mathbf{0}, \sigma_1^2 \mathbf{I}_{N_1}), \quad \mathbf{u}_{j,k}^*|\mathbf{u}_{pa\{j,k\}} \sim N(\mathbf{A}_{j,k}^* \mathbf{u}_{pa\{j,k\}}, \mathbf{W}_j^*), \quad j = 1, \dots, J-1, \end{aligned} \quad (17)$$

where  $\mathbf{W}_j^* \equiv \sigma_j^2 \mathbf{H}_{n_{j-1}}^*$ ,  $\mathbf{H}_{n_{j-1}-1}^* \equiv \frac{n_{j-1}}{n_{j-1}-1}(\mathbf{I}_{n_{j-1}-1} - \frac{1}{n_{j-1}}\mathbf{1}_{n_{j-1}-1}\mathbf{1}'_{n_{j-1}-1})$  is essentially  $\mathbf{H}_{n_{j-1}}$  without the first row and the first column, and  $\mathbf{A}_{j,k}^*$  is  $\mathbf{A}_{j,k}$  without the first row.

## Maximum likelihood by EM algorithm

Recall that the MTSLM is parameterized by the regression coefficients  $\boldsymbol{\beta}$ , the measurement error variances  $\{\phi_j : j = 1, \dots, J\}$ , and the noise variances  $\{\sigma_j^2 : j = 1, \dots, J\}$ . Let  $\boldsymbol{\theta} \equiv (\sigma_1^2, \dots, \sigma_J^2, \phi_1, \dots, \phi_J, \boldsymbol{\beta}')$ . In theory, for a given  $\boldsymbol{\theta}$ , direct evaluation of  $f(\mathbf{z}; \boldsymbol{\theta})$  can be obtained via Monte Carlo integration  $f(\mathbf{z}; \boldsymbol{\theta}) = \int f(\mathbf{z}, \mathbf{u}^*; \boldsymbol{\theta}) d\mathbf{u}^*$  or more realistically, from the Kalman-filter iterations (see e.g. Section 3.4, Harvey (1989)). Then a Newton-Raphson type of algorithm can be used for obtaining the MLEs of the model parameters. Following Huang et al. (2002), we consider an Expectation-Maximization (EM) algorithm, because it tends to be more stable numerically.

Viewing the latent values as observable but missing, we have an explicit form of the complete ‘‘data’’ model from  $f(\mathbf{z}, \mathbf{u}; \boldsymbol{\theta}) = f(\mathbf{z}|\mathbf{u}; \boldsymbol{\beta}, \phi_1, \dots, \phi_J) f(\mathbf{u}^*; \sigma_1^2, \dots, \sigma_J^2)$ , with the following

log-likelihood function,

$$\begin{aligned} \ell(\boldsymbol{\theta}; \mathbf{z}, \mathbf{u}) &\propto - \sum_{j=1}^J \sum_{k=1}^{N_{j-1}} \left\{ \frac{n_{j-1}}{2} \log \phi_j + \frac{1}{2\phi_j} (\mathbf{z}_{j,k} - \mathbf{X}_{j,k}\boldsymbol{\beta} - \mathbf{u}_{j,k})' (\mathbf{z}_{j,k} - \mathbf{X}_{j,k}\boldsymbol{\beta} - \mathbf{u}_{j,k}) \right\} \\ &\quad - \left( \frac{N_1}{2} \log \sigma_1^2 + \frac{1}{2\sigma_1^2} \mathbf{u}'_{1,1} \mathbf{u}_{1,1} \right) - \sum_{j=2}^J \sum_{k=1}^{N_{j-1}} \left( \frac{n_{j-1}-1}{2} \log \sigma_j^2 + \frac{1}{2\sigma_j^2} \frac{n_{j-1}-1}{n_{j-1}} Q_{j,k} \right), \end{aligned} \quad (18)$$

where  $Q_{j,k} = \frac{n_{j-1}}{n_{j-1}-1} (\mathbf{u}_{j,k}^* - \mathbf{A}_{j,k}^* \mathbf{u}_{pa\{j,k\}})' \mathbf{H}_{n_{j-1}}^{*-1} (\mathbf{u}_{j,k}^* - \mathbf{A}_{j,k}^* \mathbf{u}_{pa\{j,k\}})$ , which by simple algebra is  $(\mathbf{u}_{j,k} - \mathbf{A}_{j,k} \mathbf{u}_{pa\{j,k\}})' (\mathbf{u}_{j,k} - \mathbf{A}_{j,k} \mathbf{u}_{pa\{j,k\}})$ . Given the  $(t-1)$ -th parameter estimates  $\hat{\boldsymbol{\theta}}$  (with a superscript of  $(t-1)$  omitted for ease of notation), compute the following conditional expectations:

$$\begin{aligned} E\left( (\mathbf{z}_{j,k} - \mathbf{X}_{j,k}\boldsymbol{\beta} - \mathbf{u}_{j,k})' (\mathbf{z}_{j,k} - \mathbf{X}_{j,k}\boldsymbol{\beta} - \mathbf{u}_{j,k}) \mid \mathbf{z}; \hat{\boldsymbol{\theta}} \right) \\ &= (\mathbf{z}_{j,k} - \mathbf{X}_{j,k}\boldsymbol{\beta} - \hat{\mathbf{u}}_{j,k})' (\mathbf{z}_{j,k} - \mathbf{X}_{j,k}\boldsymbol{\beta} - \hat{\mathbf{u}}_{j,k}) + \text{tr}(\hat{\mathbf{V}}_{j,k}), \\ E(\mathbf{u}'_{1,1} \mathbf{u}_{1,1} \mid \mathbf{z}; \hat{\boldsymbol{\theta}}) &= \hat{\mathbf{u}}'_{1,1} \hat{\mathbf{u}}_{1,1} + \text{tr}(\hat{\mathbf{V}}_{1,1}), \\ \hat{Q}_{j,k} \equiv E(Q_{j,k} \mid \mathbf{z}; \hat{\boldsymbol{\theta}}) &= (\hat{\mathbf{u}}_{j,k} - \mathbf{A}_{j,k} \hat{\mathbf{u}}_{pa\{j,k\}})' (\hat{\mathbf{u}}_{j,k} - \mathbf{A}_{j,k} \hat{\mathbf{u}}_{pa\{j,k\}}) \\ &\quad + \text{tr}(\hat{\mathbf{V}}_{j,k} - 2\mathbf{A}_{j,k} \hat{\mathbf{V}}_{pa\{j,k\},j,k} + \mathbf{A}_{j,k} \hat{\mathbf{V}}_{pa\{j,k\}} \mathbf{A}'_{j,k}), \end{aligned}$$

where  $\hat{\mathbf{u}}_{j,k} \equiv E(\mathbf{u}_{j,k} \mid \mathbf{z}; \hat{\boldsymbol{\theta}})$ ,  $\hat{\mathbf{V}}_{j,k} \equiv \text{var}(\mathbf{u}_{j,k} \mid \mathbf{z}; \hat{\boldsymbol{\theta}})$ ,  $\hat{\mathbf{u}}_{pa\{j,k\}} \equiv E(\mathbf{u}_{pa\{j,k\}} \mid \mathbf{z}; \hat{\boldsymbol{\theta}})$ ,  $\hat{\mathbf{V}}_{pa\{j,k\}} \equiv \text{var}(\mathbf{u}_{pa\{j,k\}} \mid \mathbf{z}; \hat{\boldsymbol{\theta}})$ , and  $\hat{\mathbf{V}}_{pa\{j,k\},j,k} \equiv \text{cov}(\mathbf{u}_{pa\{j,k\}}, \mathbf{u}_{j,k} \mid \mathbf{z}; \hat{\boldsymbol{\theta}})$  can be obtained from a Kalman-filter algorithm, which will be described in Section 4. In the E-step, the conditional expectation of the complete data log-likelihood function (18) is computed:

$$\begin{aligned} E(\ell(\boldsymbol{\theta}; \mathbf{z}, \mathbf{u}) \mid \mathbf{z}, \hat{\boldsymbol{\theta}}) &\propto \\ &- \sum_{j=1}^J \sum_{k=1}^{N_{j-1}} \left[ \frac{n_{j-1}}{2} \log \phi_j + \frac{1}{2\phi_j} \left\{ (\mathbf{z}_{j,k} - \mathbf{X}_{j,k}\boldsymbol{\beta} - \hat{\mathbf{u}}_{j,k})' (\mathbf{z}_{j,k} - \mathbf{X}_{j,k}\boldsymbol{\beta} - \hat{\mathbf{u}}_{j,k}) + \text{tr}(\hat{\mathbf{V}}_{j,k}) \right\} \right] \\ &- \left[ \frac{N_1}{2} \log \sigma_1^2 + \frac{1}{2\sigma_1^2} \left\{ \hat{\mathbf{u}}'_{1,1} \hat{\mathbf{u}}_{1,1} + \text{tr}(\hat{\mathbf{V}}_{1,1}) \right\} \right] - \sum_{j=2}^J \sum_{k=1}^{N_{j-1}} \left( \frac{n_{j-1}-1}{2} \log \sigma_j^2 + \frac{1}{2\sigma_j^2} \frac{n_{j-1}-1}{n_{j-1}} \hat{Q}_{j,k} \right). \end{aligned} \quad (19)$$

In the M-step, we maximize (19) and obtain the  $t$ -th parameter estimates  $\hat{\boldsymbol{\theta}}^{(t)}$ :

$$(\hat{\sigma}_1^2)^{(t)} = \frac{1}{N_1} \left\{ \hat{\mathbf{u}}'_{1,1} \hat{\mathbf{u}}_{1,1} + \text{tr}(\hat{\mathbf{V}}_{1,1}) \right\}, \quad (\hat{\sigma}_j^2)^{(t)} = \frac{1}{N_{j-1} n_{j-1}} \sum_{k=1}^{N_{j-1}} \hat{Q}_{j,k}, \quad j = 2, \dots, J,$$

and the simultaneous solutions to the equations,

$$\begin{aligned}\phi_j &= \frac{1}{N_{j-1}n_{j-1}} \sum_{k=1}^{N_{j-1}} \left\{ (\mathbf{z}_{j,k} - \mathbf{X}_{j,k}\boldsymbol{\beta} - \hat{\mathbf{u}}_{j,k})' (\mathbf{z}_{j,k} - \mathbf{X}_{j,k}\boldsymbol{\beta} - \hat{\mathbf{u}}_{j,k}) + \text{tr}(\hat{\mathbf{V}}_{j,k}) \right\}, \quad j = 1, \dots, J, \\ \boldsymbol{\beta} &= \left( \sum_{j=1}^J \sum_{k=1}^{N_{j-1}} \frac{1}{\phi_j} \mathbf{X}'_{j,k} \mathbf{X}_{j,k} \right)^{-1} \left( \sum_{j=1}^J \sum_{k=1}^{N_{j-1}} \frac{1}{\phi_j} \mathbf{X}'_{j,k} (\mathbf{z}_{j,k} - \hat{\mathbf{u}}_{j,k}) \right).\end{aligned}$$

At the convergence step of the Kalman-filter  $t = T$ , the parameter estimates  $\boldsymbol{\theta}^{(T)}$  are taken as the MLEs  $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ .

### Observed information matrix

Using the formula proposed by Louis (1982), the observed information matrix is,

$$\mathcal{I}(\boldsymbol{\theta}) = E \left( - \frac{\partial^2 \ell(\boldsymbol{\theta}; \mathbf{z}, \mathbf{u})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \middle| \mathbf{z} \right) - E \left\{ \left( \frac{\partial \ell(\boldsymbol{\theta}; \mathbf{z}, \mathbf{u})}{\partial \boldsymbol{\theta}} \right) \left( \frac{\partial \ell(\boldsymbol{\theta}; \mathbf{z}, \mathbf{u})}{\partial \boldsymbol{\theta}} \right)' \middle| \mathbf{z} \right\}.$$

The matrix  $E \left( - \frac{\partial^2 \ell(\boldsymbol{\theta}; \mathbf{z}, \mathbf{u})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \middle| \mathbf{z}; \hat{\boldsymbol{\theta}}_{\text{MLE}} \right)$  evaluated at  $\hat{\boldsymbol{\theta}}_{\text{MLE}}$  has diagonal elements:

$$\frac{N_1}{2(\sigma_1^2)^2}, \quad \left\{ \frac{N_{j-1}(n_{j-1} - 1)}{2(\sigma_j^2)^2} : j = 2, \dots, J \right\}, \quad \left\{ \frac{N_{j-1}n_{j-1}}{2\phi_j^2} : j = 1, \dots, J \right\}, \quad \mathbf{X}'\boldsymbol{\Phi}^{-1}\mathbf{X}$$

and non-zero off-diagonal elements corresponding to  $\phi_j$  and  $\boldsymbol{\beta}$ :  $\frac{1}{\phi_j^2} \sum_{k=1}^{N_{j-1}} (\mathbf{z}_{j,k} - \mathbf{X}_{j,k}\boldsymbol{\beta} - \hat{\mathbf{u}}_{j,k})' \mathbf{X}_{j,k}$ . The non-zero first and second derivatives of the log-likelihood function (18) and hence  $\mathcal{I}(\hat{\boldsymbol{\theta}}_{\text{MLE}})$  are derived in Appendix A. It follows that the inverse of  $\mathcal{I}(\hat{\boldsymbol{\theta}}_{\text{MLE}})$  gives a variance-covariance estimate of  $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ .

### Model selection

To use the MTSLM described above for a given data set  $\{\mathbf{z}_{j,k} : k = 1, \dots, N_{j-1}, j = 1, \dots, J\}$ , various choices of the model need to be made. The linear regression mean structure depends on which covariates are incorporated in the model. In addition, the spatial dependence is specified via the tree structure and depends on several factors such as the number of resolutions  $J$ , the partition on the coarsest resolution  $N_1$ , the number of children for each parent node on the  $j$ -th resolution  $n_j$ . We use Akaike's Information Criteria (AIC) to rank the candidate models, where the log-likelihood is computed by a Monte Carlo integration.

## 4 Optimal Prediction by Kalman-filter Algorithm

We adopt a fast change-of-resolution Kalman-filter algorithm by Huang et al. (2002) for an optimal prediction of  $\mathbf{u}_{j,k}$ . There are two steps in the change-of-resolution computing algorithm to obtain  $\{E(\mathbf{u}_{j,k}|\mathbf{z}) : k = 1, \dots, N_{j-1}, j = 1, \dots, J\}$ : a high-to-low-resolution filtering step followed by a low-to-high-resolution smoothing step. In the filtering step, the algorithm moves from finer resolutions to coarser resolutions, recursively computing the optimal predictor of the underlying process, based on the data on the relevant higher resolutions. Once the coarsest resolution is reached, the algorithm goes back from the coarser resolutions to finer resolutions, recursively computing the optimal predictor of  $\mathbf{u}_{j,k}$  on each resolution based on all the data. In the final step of the recursion, the optimal prediction of  $\{\mathbf{u}_{j,k}\}$ , given all the data, is achieved.

In the high-to-low-resolution filtering step, we start with the finest resolution  $J$  and compute for  $k = 1, \dots, N_{J-1}$ :

$$\begin{aligned}\hat{\mathbf{u}}_{J,k|J,k} &\equiv E(\mathbf{u}_{J,k}|\mathbf{z}_{J,k}) = \mathbf{V}_{J,k}(\mathbf{V}_{J,k} + \mathbf{\Phi}_J)^{-1}(\mathbf{z}_{J,k} - \mathbf{X}_{J,k}\boldsymbol{\beta}), \\ \hat{\mathbf{V}}_{J,k|J,k} &\equiv var(\mathbf{u}_{J,k}|\mathbf{z}_{J,k}) = \mathbf{V}_{J,k} - \mathbf{V}_{J,k}(\mathbf{V}_{J,k} + \mathbf{\Phi}_J)^{-1}\mathbf{V}_{J,k},\end{aligned}$$

where  $\mathbf{V}_{J,k} = var(\mathbf{u}_{J,k})$  can be obtained from  $var(\mathbf{u}_J)$  in (10) (as the  $(J, (k-1)n_{J-1}+1), \dots, (J, (k-1)n_{J-1} + n_{J-1})$  elements of  $var(\mathbf{u}_J)$ ). As we move from the resolution  $j = J - 1$  to the coarsest resolution  $j = 1$ , we compute for a given vector node  $\{j, k\}$ ,

$$\begin{aligned}\hat{\mathbf{u}}_{j,k|ch\{j,k,i\}} &\equiv E(\mathbf{u}_{j,k}|\mathbf{z}_{de\{ch\{j,k,i\}\}}) = \mathbf{B}_{ch\{j,k,i\}}\hat{\mathbf{u}}_{ch\{j,k,i\}|ch\{j,k,i\}}, \\ \hat{\mathbf{V}}_{j,k|ch\{j,k,i\}} &\equiv var(\mathbf{u}_{j,k}|\mathbf{z}_{de\{ch\{j,k,i\}\}}) = \mathbf{B}_{ch\{j,k,i\}}\hat{\mathbf{V}}_{ch\{j,k,i\}|ch\{j,k,i\}}\mathbf{B}'_{ch\{j,k,i\}} + \mathbf{R}_{ch\{j,k,i\}},\end{aligned}$$

where  $de\{ch\{j, k, i\}\}$  denotes the descendant vector nodes of the  $i$ -th vector child node  $ch\{j, k, i\}$  (child of  $\{j, k\}$ ) including  $ch\{j, k, i\}$ ,  $\mathbf{B}_{ch\{j,k,i\}} \equiv \mathbf{V}_{j,k}\mathbf{A}'_{ch\{j,k,i\}}\mathbf{V}_{ch\{j,k,i\}}^{-1}$ ,  $\mathbf{R}_{ch\{j,k,i\}} \equiv \mathbf{V}_{j,k} - \mathbf{V}_{j,k}\mathbf{A}'_{ch\{j,k,i\}}\mathbf{V}_{ch\{j,k,i\}}^{-1}\mathbf{A}_{ch\{j,k,i\}}\mathbf{V}_{j,k}$ ;  $i = 1, \dots, n_{j-1}$ . Here  $\mathbf{V}_{j,k} = var(\mathbf{u}_{j,k})$  and  $\mathbf{V}_{ch\{j,k,i\}} = var(\mathbf{u}_{ch\{j,k,i\}})$  can be obtained from  $var(\mathbf{u}_j)$  and  $var(\mathbf{u}_{j+1})$  as defined in (10). Further,

$$\begin{aligned}\hat{\mathbf{u}}_{j,k|j,k}^* &\equiv E(\mathbf{u}_{j,k}|\mathbf{z}_{de\{j,k\}\setminus\{j,k\}}) = \hat{\mathbf{V}}_{j,k|j,k}^* \left( \sum_{i=1}^{n_j} \hat{\mathbf{V}}_{j,k|ch\{j,k,i\}}^{-1} \hat{\mathbf{u}}_{j,k|ch\{j,k,i\}} \right), \\ \hat{\mathbf{V}}_{j,k|j,k}^* &\equiv var(\mathbf{u}_{j,k}|\mathbf{z}_{de\{j,k\}\setminus\{j,k\}}) = \left\{ \mathbf{V}_{j,k}^{-1} + \sum_{i=1}^{n_j} (\hat{\mathbf{V}}_{j,k|ch\{j,k,i\}}^{-1} - \mathbf{V}_{j,k}^{-1}) \right\}^{-1}, \\ \hat{\mathbf{u}}_{j,k|j,k} &\equiv E(\mathbf{u}_{j,k}|\mathbf{z}_{de\{j,k\}}) = \hat{\mathbf{V}}_{j,k|j,k} \{ \mathbf{\Phi}_j^{-1}(\mathbf{z}_{j,k} - \mathbf{X}_{j,k}\boldsymbol{\beta}) + (\hat{\mathbf{V}}_{j,k|j,k}^*)^{-1} \hat{\mathbf{u}}_{j,k|j,k}^* \},\end{aligned}$$

$$\hat{\mathbf{V}}_{j,k|j,k} \equiv \text{var}(\mathbf{u}_{j,k} | \mathbf{z}_{de\{j,k\}}) = \hat{\mathbf{V}}_{j,k|j,k}^* - \hat{\mathbf{V}}_{j,k|j,k}^* (\hat{\mathbf{V}}_{j,k|j,k}^* + \Phi_j)^{-1} \hat{\mathbf{V}}_{j,k|j,k}^*$$

Here  $de\{j,k\}$  denotes the descendant vector nodes of  $\{j,k\}$  including the vector node  $\{j,k\}$ . At the end of the filtering step, the root nodes are reached and hence the BLUP's for  $\{1,1\}$  are:

$$\hat{\mathbf{u}}_{1,1} \equiv E(\mathbf{u}_{1,1} | \mathbf{z}) = \hat{\mathbf{u}}_{1,1|1,1}, \quad \hat{\mathbf{V}}_{1,1} \equiv \text{var}(\mathbf{u}_{1,1} | \mathbf{z}) = \hat{\mathbf{V}}_{1,1|1,1},$$

where  $\mathbf{z}$  consists of all the observations.

In the low-to-high-resolution smoothing step, we move from the coarsest resolution  $j = 2$  to the finest resolution  $j = J$  and compute for a given node  $\{j,k\}$  where  $k = 1, \dots, N_{j-1}$ :

$$\hat{\mathbf{u}}_{j,k} \equiv E(\mathbf{u}_{j,k} | \mathbf{z}) = \hat{\mathbf{u}}_{j,k|j,k} + \hat{\mathbf{V}}_{j,k|j,k} \mathbf{B}'_{j,k} \hat{\mathbf{V}}_{pa\{j,k\}|j,k}^{-1} (\hat{\mathbf{u}}_{pa\{j,k\}} - \hat{\mathbf{u}}_{pa\{j,k\}|j,k}), \quad (20)$$

$$\begin{aligned} \hat{\mathbf{V}}_{j,k} \equiv \text{var}(\mathbf{u}_{j,k} | \mathbf{z}) &= \hat{\mathbf{V}}_{j,k|j,k} + \\ &\hat{\mathbf{V}}_{j,k|j,k} \mathbf{B}'_{j,k} \hat{\mathbf{V}}_{pa\{j,k\}|j,k}^{-1} (\hat{\mathbf{V}}_{pa\{j,k\}} - \hat{\mathbf{V}}_{pa\{j,k\}|j,k}) \hat{\mathbf{V}}_{pa\{j,k\}|j,k}^{-1} \mathbf{B}_{j,k} \hat{\mathbf{V}}_{j,k|j,k}, \end{aligned} \quad (21)$$

where  $\mathbf{B}_{j,k} \equiv \mathbf{V}_{pa\{j,k\}} \mathbf{A}'_{j,k} \mathbf{V}_{j,k}^{-1}$ . For derivations of the Kalman-filter formulas, see Huang and Cressie (1997).

For computing  $\hat{\mathbf{V}}_{pa\{j,k\},j,k}$  in the EM algorithm, we give a general formula of a covariance between  $\{j,k\}$  and  $\{j',k'\}$ . Let  $an\{j,k,j',k'\}$  denote the group of nodes on the  $j_{an}$ -th resolution which contains the first common (scalar) ancestor of  $\{j,k\}$  and  $\{j',k'\}$ , if there is a common ancestor. If there is no common ancestor, then  $j_{an} \equiv 0$ . Suppose the path from  $\{j,k\}$  to  $an\{j,k,j',k'\}$  consists of vector nodes  $\{j,k\}, \dots, \{j_{an}-1, k_1\}$ ,  $an\{j,k,j',k'\}$ , and similarly, the path from  $\{j',k'\}$  to  $an\{j,k,j',k'\}$  consists of vector nodes  $\{j',k'\}, \dots, \{j_{an}-1, k'_1\}$ ,  $an\{j,k,j',k'\}$ . Let  $\mathbf{G}_{j,k} \equiv \hat{\mathbf{V}}_{j,k|j,k} \mathbf{B}'_{j,k} \hat{\mathbf{V}}_{pa\{j,k\}|j,k}$ . Then we can compute the conditional covariance between  $\{j,k\}$  and  $\{j',k'\}$  by

$$\hat{\mathbf{V}}_{j,k,j',k'} \equiv \text{cov}(\mathbf{u}_{j,k}, \mathbf{u}_{j',k'} | \mathbf{z}) = \begin{cases} \mathbf{0}_{n_{j-1}} \mathbf{0}'_{n_{j'-1}} & ; \quad j_{an} = 0, \\ \mathbf{G}_{j,k} \cdots \mathbf{G}_{j_{an}-1, k_1} \hat{\mathbf{V}}_{an\{j,k,j',k'\}} \mathbf{G}'_{j_{an}-1, k'_1} \cdots \mathbf{G}'_{j',k'} & ; \quad j_{an} \geq 1. \end{cases} \quad (22)$$

By (20)–(22), we can now compute  $\hat{\mathbf{V}}_{j,k}$ ,  $\hat{\mathbf{V}}_{pa\{j,k\}}$ , and  $\hat{\mathbf{V}}_{pa\{j,k\},j,k}$  in the EM algorithm.

The optimal prediction of the underlying true processes  $\{\mathbf{y}_{j,k}\}$  is achieved by combining the MLE of the regression coefficients and the BLUP of  $\{\mathbf{u}_{j,k}\}$ :

$$\hat{\mathbf{y}}_{j,k} = \mathbf{X}_{j,k} \hat{\beta}_{\text{MLE}} + \hat{\mathbf{u}}_{j,k},$$

where  $\hat{\boldsymbol{\beta}}$  is the MLE of  $\boldsymbol{\beta}$  from the EM algorithm and  $\hat{\mathbf{u}}_{j,k}$  is the optimal predictor of  $\mathbf{u}_{j,k}$  given the observations  $\mathbf{z}$  evaluated at  $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ . We plug  $\hat{\boldsymbol{\theta}}_{\text{MLE}}$  into the formula  $\hat{V}_{j,k}$  to obtain the (empirical) prediction error variance of  $\hat{\mathbf{u}}_{j,k}$ .

## 5 Statistical Inference with Incomplete Data

The MTSLM has been presented in Sections 2 and 3 under the assumption that the observations at all the locations of all the resolutions are available. In this section, we examine situations where observations at some of the locations of some of the resolutions are not available. Because of the mass-balanced property, the change-of-resolution Kalman-filter algorithm operates on the basis of vectors  $\mathbf{u}_{j,k}$  for  $k = 1, \dots, N_{j-1}, j = 1, \dots, J$ . Thus if a scalar observation within  $\{j, k\}$  is missing, then all the observations at  $\{j, k\}$  are regarded as missing. Implicitly, there ought to be observations associated with at least one  $\{J, k\}$  on the finest resolution, for otherwise the finest resolution would be no more than  $J - 1$ . As it turns out, we need to distinguish two cases and address them separately. The first case has at least one vector node of observations available on at least one of the coarser resolutions  $j = 1, \dots, J - 1$ , while the second case does not have any complete vector node of observations on any of the coarser resolutions  $j = 1, \dots, J - 1$ .

The first case is suitable for those data sets that have multiple sources of observations on different resolutions. For example, observations about a same variable are made in  $J$  different ways, each with a different resolution, ranging from the coarsest 1st resolution to the finest  $J$ -th resolution. For this case, minor modifications of the formulas in Sections 3–4 would suffice. We use  $\gamma_{j,k} \equiv \mathcal{I}\{\mathbf{z}_{j,k} \text{ is observed}\}$  to indicate whether all the observation at  $\{j, k\}$  are observed. Let  $\mathbf{z}$  denote the vector of data and continue to let  $\mathbf{u}$  denote the latent variables regardless of missing observations.

Similar as (16), the probability density of  $\mathbf{z}$  conditional on  $\mathbf{u}$  is,

$$f(\mathbf{z}|\mathbf{u}) = \prod_{j=1}^J \prod_{k=1}^{N_{j-1}} f(\mathbf{z}_{j,k}|\mathbf{u}_{j,k})^{\gamma_{j,k}}, \quad \text{where } \mathbf{z}_{j,k}|\mathbf{u}_{j,k} \sim N(\mathbf{X}_{j,k}\boldsymbol{\beta} + \mathbf{u}_{j,k}, \boldsymbol{\Phi}_j).$$

Together with (17), the complete “data” log-likelihood function is,

$$\ell(\boldsymbol{\theta}; \mathbf{z}, \mathbf{u}) \propto - \sum_{j=1}^J \sum_{\gamma_{j,k}=1} \left\{ \frac{n_{j-1}}{2} \log \phi_j + \frac{1}{2\phi_j} (\mathbf{z}_{j,k} - \mathbf{X}_{j,k}\boldsymbol{\beta} - \mathbf{u}_{j,k})' (\mathbf{z}_{j,k} - \mathbf{X}_{j,k}\boldsymbol{\beta} - \mathbf{u}_{j,k}) \right\}$$

$$-\left(\frac{N_1}{2} \log \sigma_1^2 + \frac{1}{2\sigma_1^2} \mathbf{u}'_{1,1} \mathbf{u}_{1,1}\right) - \sum_{j=2}^J \sum_{k=1}^{N_{j-1}} \left( \frac{n_{j-1}-1}{2} \log \sigma_j^2 + \frac{1}{2\sigma_j^2} \frac{n_{j-1}-1}{n_{j-1}} Q_{j,k} \right). \quad (23)$$

We show the EM algorithm based on (23) and Kalman-filter results in Appendix A.

On the other hand, the second case is suitable for single-source observations on the finest  $J$ -th resolution. The purpose here is to predict the underlying true process optimally and hence the multiresolutions are mere artifacts for modeling the spatial dependence and processing the data in a computationally efficient manner. In this case, the MTSLM defined in (3)–(9) feature model parameters  $\boldsymbol{\theta} = (\sigma_1^2, \dots, \sigma_J^2, \phi_J, \boldsymbol{\beta}')'$ . Since  $E(\mathbf{z}_J) = \mathbf{X}_J \boldsymbol{\beta}$ , (12) holds for  $j = J$  only, and there is no  $\text{cov}(\mathbf{z}_j, \mathbf{z}_{j'})$  for  $1 \leq j < j' \leq J$ , the parameters are no longer identifiable. To ensure that the MTSLM is identifiable, at least one of the parameters needs to be constrained. We choose to fix the measurement error variance  $\phi_J$ , because it can oftentimes be estimated from external data (see, e.g., Huang et al., 2002). For the constrained MTSLM, the model parameters are denoted as  $\boldsymbol{\theta}^* \equiv (\sigma_1^2, \dots, \sigma_J^2, \boldsymbol{\beta}')'$  and can be estimated by an EM algorithm based on the observations on the finest resolution.

Let  $\mathbf{z}_J \equiv (\mathbf{z}_{J,k} : k = 1, \dots, N_{J-1})'$  denote the data on the finest  $J$ -th resolution. Similar as (16), the probability density of  $\mathbf{z}_J$  conditional on  $\mathbf{u}$  is,

$$f(\mathbf{z}_J | \mathbf{u}) = \prod_{k=1}^{N_{J-1}} f(\mathbf{z}_{J,k} | \mathbf{u}_{J,k})^{\gamma_{J,k}}, \quad \text{where } \mathbf{z}_{J,k} | \mathbf{u}_{J,k} \sim N(\mathbf{X}_{J,k} \boldsymbol{\beta} + \mathbf{u}_{J,k}, \boldsymbol{\Phi}_J).$$

Together with (17), the complete “data” log-likelihood function is,

$$\begin{aligned} \ell(\boldsymbol{\theta}^*; \mathbf{z}_J, \mathbf{u}) &\propto - \sum_{\gamma_{J,k}=1} \left\{ \frac{1}{2\phi_J} (\mathbf{z}_{J,k} - \mathbf{X}_{J,k} \boldsymbol{\beta} - \mathbf{u}_{J,k})' (\mathbf{z}_{J,k} - \mathbf{X}_{J,k} \boldsymbol{\beta} - \mathbf{u}_{J,k}) \right\} \\ &- \left( \frac{N_1}{2} \log \sigma_1^2 + \frac{1}{2\sigma_1^2} \mathbf{u}'_{1,1} \mathbf{u}_{1,1} \right) - \sum_{j=2}^J \sum_{k=1}^{N_{j-1}} \left( \frac{n_{j-1}-1}{2} \log \sigma_j^2 + \frac{1}{2\sigma_j^2} \frac{n_{j-1}-1}{n_{j-1}} Q_{j,k} \right). \end{aligned} \quad (24)$$

We show the EM algorithm based on (24) and Kalman-filter results in Appendix B.

## 6 Example: Optimal Prediction of Crop Yield

For illustration, we use the MTSLM to map crop yield in an experimental field at the Arlington Research Station in Columbia County, Wisconsin. The field size is 160 m by 120 m and the

landscape includes one closed depression and less than a one percent slope. The raw corn yield data, recorded by a yield monitor attached to a harvest combine, are gridded to 240 cells, each of size 8 m by 10 m (Figure 2(a)). Within each cell, an average of the yield is used as a datum. A 3D perspective plot (Figures 2(b)) suggests that lower yield tends to occur near the edge of the field and in the depression area toward the center of the field.

---

Figure 2 here

---

Crop yield is influenced by various factors, including field topography, soil types, weather conditions, fertilization, and pest control. In this study, information of the field topography (elevation, slope, and aspect) is available at 2-m resolution. Moisture levels of the crops were measured and recorded, at the same time as the yield monitor measured the crop yield. As we do with the yield data, these covariates are averaged within the 240 8 m by 10 m cells. In addition, we create an indicator variable for the cells that are along the edges of the field. Thus we consider five possible covariates, namely, moisture level of the crops, elevation, slope, aspect of the field, and edge effect. A scatter plot of the crop yield, all the covariates, and the Global Positioning System (GPS) measures of easting and northing is shown in Figure 3. The plot between crop yield and the covariates seem to suggest that yield is correlated with elevation and aspect, while it is not clear how yield relates to moisture level and slope. Among the covariates, slope and aspect are the first and second derivatives of elevation, while moisture level seems to correlate with all three topographic variables.

---

Figure 3 here

---

We now consider aggregating the cells on the finest 8 m  $\times$  10 m resolution, to a coarser 16 m  $\times$  20 m resolution, and to the coarsest 32 m  $\times$  40 m resolution of cells. Associated with the partition, we use a tree structure consisting of  $J = 3$  resolutions,  $N_1 = 15$  roots, and  $n_1 = n_2 = 4$  children for each of the node on the first two resolutions. Hence there are  $N_2 = 60$  nodes on the second resolution and  $N_3 = 240$  nodes on the finest resolution.

We use an MTSLM to analyze the yield data along with the covariates. Since there was only one source of data, there is no measurement error variance on the first two resolutions and we

need to specify the measurement error variance  $\phi_3$  on the finest resolution to ensure identifiability of the model. As a conservative estimate,  $\phi_3$  is chosen to be 200. The model parameters of the noise variances are  $(\sigma_1^2, \sigma_2^2, \sigma_3^2)'$ , for the given 3-resolution 15-root quad-tree. On the other hand, the model parameters of the regression coefficients  $\boldsymbol{\beta} = (\beta_0, \dots, \beta_5)'$  are for the intercept, moisture level, elevation, slope, aspect, and edge effect. We consider the full regression model and all its reduced models.

For a given choice of regression terms, the EM algorithm in Section 3 is implemented to obtain the MLE's of the model parameters  $\hat{\boldsymbol{\theta}}^*$  and their standard errors, where  $\boldsymbol{\theta}^* = (\sigma_1^2, \sigma_2^2, \sigma_3^2, \boldsymbol{\beta}')'$ . At the MLE  $\hat{\boldsymbol{\theta}}^*$ , the loglikelihood function is evaluated by Monte Carlo integration (with 10000 iterations). We use AIC to rank the models with different inclusions of the explanatory variables. The best model has covariates elevation, aspect, and edge effect:  $\text{yield} = \hat{\beta}_0 + \hat{\beta}_2 \times \text{elevation} + \hat{\beta}_4 \times \text{aspect} + \hat{\beta}_5 \times \text{edge effect}$ , where  $\hat{\beta}_0 = 121.08$ ,  $\hat{\beta}_2 = -7.44$ , and  $\hat{\beta}_4 = 19.38$  and  $\hat{\beta}_5 = -14.03$ , with standard errors 11.79, 1.99, 13.59, and 2.64, respectively. Further,  $\hat{\sigma}_1^2 = 49.35$ ,  $\hat{\sigma}_2^2 = 131.92$ , and  $\hat{\sigma}_3^2 = 6.92$  with standard errors 22.58, 35.71, and 16.54, respectively.

Now given the best model and the corresponding MLE's, the method described in Section 4 is applied to obtain the best linear unbiased predictor (BLUP) of crop yield. In particular, the Kalman-filter algorithm is applied to the detrended data. The BLUP of the detrended values along with the prediction standard errors are computed. Then the estimated trend is added back to obtain the final prediction of crop yield. Figure 2(c) shows the best predicted yield and the prediction standard errors are 3.16. The prediction surface is smoother than the raw data surface, but the main features of the original data remain such that the lower yield occurs along the boundary and in the center of the field.

## 7 Conclusions

In this article, we have extended an existing multi-resolution tree-structured model to include a linear regression mean. The maximum likelihood estimates of the model parameters and their standard errors are obtained from an EM algorithm. Hence it is possible to draw conclusions about the relationship between a response variable and potential explanatory variables. Further, optimal prediction of the true underlying process is obtained via a fast change-of-resolution algorithm. We

illustrate the method by mapping the yield in an agricultural land, but the method is general for mapping spatial variables.

## References

- Chou, K.C., Willsky, A.S., and Nikoukhah, R. (1994). Multiscale systems, Kalman filters, and Riccati equations. *IEEE Transaction on Automatic Control*, **39**, 479–492.
- Harvey, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge.
- Huang, H.-C. and Cressie, N. (1997). Multiscale spatial modeling. *Proceedings of the Section on Statistics and the Environment*, Alexandria, VA: American Statistical Association, pp.49–54.
- Huang, H.-C. and Cressie, N. (2001). Multiscale graphical modeling in space: Applications to command and control. In *Spatial Statistics: Methodological Aspects and Applications*, Ed. Moore M., Springer Lecture Notes in Statistics, **159**, Springer-Verlag, New York, 83–113.
- Huang, H.-C., Cressie, N., and Gabrosek, J. (2002). Fast, resolution-consistent spatial prediction of global processes from satellite data. *Journal of Computational and Graphical Statistics*, **11**, 1–26.
- Kolaczyk, E.D. (1999). Bayesian multi-scale models for Poisson processes. *Journal of the American Statistical Association*, **94**, 920–933.
- Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society B*, **44**, 226–233.
- Nowak, R.D. (1999). Multiscale hidden Markov models for Bayesian image analysis. *Bayesian Inference in Wavelet-Based Models*, Eds. Müller, P. and Vidakovic, B., Springer, New York.

## Appendix A: Inference with Partially Incomplete Data

In this section, we derive formulas for parameter estimation and prediction when observations are missing at some locations on some of the resolutions. Let  $\mathbf{\Gamma} \equiv \text{diag}\{\gamma_{j,k} \otimes \mathbf{I}_{n_{j-1}} : k =$

$1, \dots, N_{j-1}, j = 1, \dots, J\}$  denote a diagonal matrix of the indicators  $\gamma_{j,k} = \mathcal{I}\{z_{j,k} \text{ is observed}\}$ . Recall that  $\mathbf{z}$  denotes the vector of data and that  $\mathbf{u}$  denotes the latent variables, regardless of missing observations.

In the EM algorithm, the E-step based on (23) gives,

$$\begin{aligned} E(\ell(\boldsymbol{\theta}; \mathbf{z}, \mathbf{u}) | \mathbf{z}, \hat{\boldsymbol{\theta}}) &\propto \\ & - \sum_{j=1}^J \sum_{\gamma_{j,k}=1} \left[ \frac{n_{j-1}}{2} \log \phi_j + \frac{1}{2\phi_j} \{(\mathbf{z}_{j,k} - \mathbf{X}_{j,k}\boldsymbol{\beta} - \hat{\mathbf{u}}_{j,k})'(\mathbf{z}_{j,k} - \mathbf{X}_{j,k}\boldsymbol{\beta} - \hat{\mathbf{u}}_{j,k}) + \text{tr}(\hat{\mathbf{V}}_{j,k})\} \right] \\ & - \left[ \frac{N_1}{2} \log \sigma_1^2 + \frac{1}{2\sigma_1^2} \{\hat{\mathbf{u}}'_{1,1} \hat{\mathbf{u}}_{1,1} + \text{tr}(\hat{\mathbf{V}}_{1,1})\} \right] - \sum_{j=2}^J \sum_{k=1}^{N_{j-1}} \left( \frac{n_{j-1}-1}{2} \log \sigma_j^2 + \frac{1}{2\sigma_j^2} \frac{n_{j-1}-1}{n_{j-1}} \hat{Q}_{j,k} \right). \end{aligned} \quad (25)$$

In the M-step, we maximize (25) and obtain the  $t$ -th parameter estimates  $\hat{\boldsymbol{\theta}}^{(t)}$ :

$$(\hat{\sigma}_1^2)^{(t)} = \frac{1}{N_1} \{\hat{\mathbf{u}}'_{1,1} \hat{\mathbf{u}}_{1,1} + \text{tr}(\hat{\mathbf{V}}_{1,1})\}, \quad (\hat{\sigma}_j^2)^{(t)} = \frac{1}{N_{j-1}n_{j-1}} \sum_{k=1}^{N_{j-1}} \hat{Q}_{j,k}, \quad j = 2, \dots, J,$$

and the simultaneous solutions to the equations,

$$\begin{aligned} \phi_j &= \frac{1}{n_{j-1} \sum_{k=1}^{N_{j-1}} \gamma_{j,k}} \sum_{k=1}^{N_{j-1}} \gamma_{j,k} \{(\mathbf{z}_{j,k} - \mathbf{X}_{j,k}\boldsymbol{\beta} - \hat{\mathbf{u}}_{j,k})'(\mathbf{z}_{j,k} - \mathbf{X}_{j,k}\boldsymbol{\beta} - \hat{\mathbf{u}}_{j,k}) + \text{tr}(\hat{\mathbf{V}}_{j,k})\}, \quad j = 1, \dots, J, \\ \boldsymbol{\beta} &= (\mathbf{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{\Gamma}\mathbf{X})^{-1} \mathbf{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{\Gamma}(\mathbf{z} - \hat{\mathbf{u}}), \end{aligned}$$

where  $\mathbf{X} \equiv [\mathbf{X}'_{j,k} : k = 1, \dots, N_{j-1}, j = 1, \dots, J]'$  is the overall regression design matrix,  $\boldsymbol{\Phi} \equiv \text{diag}\{\boldsymbol{\Phi}_j \otimes \mathbf{I}_{N_{j-1}} : j = 1, \dots, J\}$  consists of the measurement error variances, and  $\hat{\mathbf{u}} \equiv E(\mathbf{u} | \mathbf{z}, \hat{\boldsymbol{\theta}}^{(t-1)})$ .

By the Louis' formula (1982), the observed information matrix is,

$$\mathcal{I}(\boldsymbol{\theta}) = E \left( - \frac{\partial^2 \ell(\boldsymbol{\theta}; \mathbf{z}, \mathbf{u})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \bigg| \mathbf{z} \right) - E \left\{ \left( \frac{\partial \ell(\boldsymbol{\theta}; \mathbf{z}, \mathbf{u})}{\partial \boldsymbol{\theta}} \right) \left( \frac{\partial \ell(\boldsymbol{\theta}; \mathbf{z}, \mathbf{u})}{\partial \boldsymbol{\theta}} \right)' \bigg| \mathbf{z} \right\}.$$

The matrix  $E \left( - \frac{\partial^2 \ell(\boldsymbol{\theta}; \mathbf{z}, \mathbf{u})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \bigg| \mathbf{z}; \hat{\boldsymbol{\theta}}_{\text{MLE}} \right)$  evaluated at  $\hat{\boldsymbol{\theta}}_{\text{MLE}}$  has diagonal elements:

$$\frac{N_1}{2(\sigma_1^2)^2}, \quad \left\{ \frac{N_{j-1}(n_{j-1}-1)}{2(\sigma_j^2)^2} : j = 2, \dots, J \right\}, \quad \left\{ \frac{n_{j-1}}{2\phi_j^2} \sum_{k=1}^{N_{j-1}} \gamma_{j,k} : j = 1, \dots, J \right\}, \quad \mathbf{X}'\boldsymbol{\Phi}^{-1}\boldsymbol{\Gamma}\mathbf{X}$$

and non-zero off-diagonal elements corresponding to  $\phi_j$  and  $\boldsymbol{\beta}$ :  $\frac{1}{\phi_j^2} \sum_{k=1}^{N_{j-1}} \gamma_{j,k} (\mathbf{z}_{j,k} - \mathbf{X}_{j,k}\boldsymbol{\beta} - \hat{\mathbf{u}}_{j,k})' \mathbf{X}_{j,k}$ . The non-zero first and second derivatives of the log-likelihood function (23) are:

$$a_1 \equiv \frac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma_1^2} = -\frac{N_1}{2\sigma_1^2} + \frac{1}{2(\sigma_1^2)^2} \mathbf{u}'_{1,1} \mathbf{u}_{1,1}, \quad \frac{\partial^2 \ell(\boldsymbol{\theta})}{(\partial \sigma_1^2)^2} = \frac{N_1}{2(\sigma_1^2)^2} - \frac{1}{(\sigma_1^2)^3} \mathbf{u}'_{1,1} \mathbf{u}_{1,1};$$

$$\begin{aligned}
a_j &\equiv \frac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma_j^2} = -\frac{N_{j-1}(n_{j-1}-1)}{2\sigma_j^2} + \frac{1}{2(\sigma_j^2)^2} \frac{n_{j-1}-1}{n_{j-1}} \sum_{k=1}^{N_{j-1}} Q_{j,k}, \\
\frac{\partial^2 \ell(\boldsymbol{\theta})}{(\partial \sigma_j^2)^2} &= \frac{N_{j-1}(n_{j-1}-1)}{2(\sigma_j^2)^2} - \frac{1}{(\sigma_j^2)^3} \frac{n_{j-1}-1}{n_{j-1}} \sum_{k=1}^{N_{j-1}} Q_{j,k}, \quad j = 2, \dots, J; \\
b_j &\equiv \frac{\partial \ell(\boldsymbol{\theta})}{\partial \phi_j} = -\frac{n_{j-1}}{2\phi_j} \sum_{k=1}^{N_{j-1}} \gamma_{j,k} + \frac{1}{2\phi_j^2} \sum_{k=1}^{N_{j-1}} \gamma_{j,k} (\mathbf{z}_{j,k} - \mathbf{X}_{j,k}\boldsymbol{\beta} - \mathbf{u}_{j,k})' (\mathbf{z}_{j,k} - \mathbf{X}_{j,k}\boldsymbol{\beta} - \mathbf{u}_{j,k}), \\
\frac{\partial^2 \ell(\boldsymbol{\theta})}{(\partial \phi_j)^2} &= \frac{n_{j-1}}{2\phi_j^2} \sum_{k=1}^{N_{j-1}} \gamma_{j,k} - \frac{1}{\phi_j^3} \sum_{k=1}^{N_{j-1}} \gamma_{j,k} (\mathbf{z}_{j,k} - \mathbf{X}_{j,k}\boldsymbol{\beta} - \mathbf{u}_{j,k})' (\mathbf{z}_{j,k} - \mathbf{X}_{j,k}\boldsymbol{\beta} - \mathbf{u}_{j,k}), \\
\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \phi_j \partial \boldsymbol{\beta}} &= -\frac{1}{\phi_j^2} \sum_{k=1}^{N_{j-1}} \gamma_{j,k} (\mathbf{z}_{j,k} - \mathbf{X}_{j,k}\boldsymbol{\beta} - \mathbf{u}_{j,k})' \mathbf{X}_{j,k}, \quad j = 1, \dots, J; \\
\mathbf{c} &\equiv \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = \mathbf{X}' \boldsymbol{\Phi}^{-1} \boldsymbol{\Gamma} (\mathbf{z} - \mathbf{X}\boldsymbol{\beta} - \mathbf{u}), \quad \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\mathbf{X}' \boldsymbol{\Phi}^{-1} \boldsymbol{\Gamma} \mathbf{X}.
\end{aligned}$$

Thus the elements of the matrix  $E(\Delta \ell \Delta \ell' | \mathbf{z})$  evaluated at  $\boldsymbol{\theta}_{\text{MLE}}$  are,

$$\begin{aligned}
E(a_1^2 | \mathbf{z}) &= \frac{1}{2(\sigma_1^2)^4} \{ \text{tr}(\hat{\mathbf{V}}_{1,1}^2) + 2\hat{\mathbf{u}}'_{1,1} \hat{\mathbf{V}}_{1,1} \hat{\mathbf{u}}_{1,1} \}, \\
E(a_1 a_j | \mathbf{z}) &= \frac{n_{j-1}-1}{2n_{j-1}(\sigma_1^2)^2(\sigma_j^2)^2} \sum_{k=1}^{N_{j-1}} \{ \text{tr}(\hat{W}_{1jk} \hat{W}'_{1jk}) + 2\hat{\mathbf{u}}'_{1,1} \hat{W}_{1jk} \hat{\mathbf{w}}_{j,k} \}; \quad j = 2, \dots, J, \\
E(a_1 b_j | \mathbf{z}) &= \frac{1}{2(\sigma_1^2)^2 \phi_j^2} \{ \text{tr}(\hat{V}_{11,j} \boldsymbol{\Gamma}_j \hat{V}'_{11,j}) - 2\hat{\mathbf{u}}'_{1,1} \hat{V}_{11,j} \boldsymbol{\Gamma}_j (\mathbf{z}_j - \mathbf{X}_j \boldsymbol{\beta} - \hat{\mathbf{u}}_j) \} \quad j = 2, \dots, J, \\
E(a_1 \mathbf{c} | \mathbf{z}) &= -\frac{1}{(\sigma_1^2)^2} \mathbf{X}' \boldsymbol{\Phi}^{-1} \boldsymbol{\Gamma} \hat{V}_{\cdot,11} \hat{\mathbf{u}}_{1,1}, \\
E(a_j a_{j'} | \mathbf{z}) &= \frac{(n_{j-1}-1)(n_{j'-1}-1)}{2n_{j-1}n_{j'-1}(\sigma_j^2)^2(\sigma_{j'}^2)^2} \sum_{k=1}^{N_{j-1}} \sum_{k'=1}^{N_{j'-1}} \{ \text{tr}(\hat{X}_{jkj'k'} \hat{X}'_{jkj'k'}) + 2\hat{\mathbf{w}}'_{j,k} \hat{X}_{jkj'k'} \hat{\mathbf{w}}_{j',k'} \} \quad j, j' = 2, \dots, J, \\
E(a_j b_{j'} | \mathbf{z}) &= \frac{n_{j-1}-1}{2n_{j-1}(\sigma_j^2)^2 \phi_{j'}^2} \sum_{k=1}^{N_{j-1}} \{ \text{tr}(\hat{Y}_{jkj'} \boldsymbol{\Gamma}_{j'} \hat{Y}'_{jkj'}) + 2\hat{\mathbf{w}}'_{j,k} \hat{Y}_{jkj'} \boldsymbol{\Gamma}_{j'} (\mathbf{z}_{j'} - \mathbf{X}_{j'} \boldsymbol{\beta} - \hat{\mathbf{u}}_{j'}) \}, \\
E(a_j \mathbf{c} | \mathbf{z}) &= -\frac{n_{j-1}-1}{n_{j-1}(\sigma_j^2)^2} \mathbf{X}' \boldsymbol{\Phi}^{-1} \boldsymbol{\Gamma} \sum_{k=1}^{N_{j-1}} \{ \hat{V}_{\cdot,jk} - \hat{V}_{\cdot,pa\{j,k\}} \mathbf{A}'_{j,k} \} \hat{\mathbf{w}}_{j,k}, \\
E(b_j b_{j'} | \mathbf{z}) &= \frac{1}{2\phi_j^2 \phi_{j'}^2} \{ \text{tr}(\boldsymbol{\Gamma}_j \hat{V}_{j,j'} \boldsymbol{\Gamma}_{j'} \hat{V}'_{j,j'}) + 2(\mathbf{z}_j - \mathbf{X}_j \boldsymbol{\beta} - \hat{\mathbf{u}}_j)' \boldsymbol{\Gamma}_j \hat{V}_{j,j'} \boldsymbol{\Gamma}_{j'} (\mathbf{z}_{j'} - \mathbf{X}_{j'} \boldsymbol{\beta} - \hat{\mathbf{u}}_{j'}) \}, \\
E(b_j \mathbf{c} | \mathbf{z}) &= \frac{1}{\phi_j^2} \mathbf{X}' \boldsymbol{\Phi}^{-1} \boldsymbol{\Gamma} \hat{V}_{\cdot,j} \boldsymbol{\Gamma}_j (\mathbf{z}_j - \mathbf{X}_j \boldsymbol{\beta} - \hat{\mathbf{u}}_j), \\
E(\mathbf{c} \mathbf{c}' | \mathbf{z}) &= \mathbf{X}' \boldsymbol{\Phi}^{-1} \boldsymbol{\Gamma} \hat{V}_{\cdot,\cdot} \boldsymbol{\Gamma} \boldsymbol{\Phi}^{-1} \mathbf{X}.
\end{aligned}$$

Here  $\mathbf{z}_j \equiv (\mathbf{z}'_{j,1}, \dots, \mathbf{z}'_{j,N_{j-1}})'$ ,  $\mathbf{u}_j \equiv (\mathbf{u}'_{j,1}, \dots, \mathbf{u}'_{j,N_{j-1}})'$ ,  $\boldsymbol{\Gamma}_j \equiv \text{diag}\{\gamma_{j,1}, \dots, \gamma_{j,N_{j-1}}\} \otimes \mathbf{I}_{n_{j-1}}$ ,  $\hat{X}_{jkj'k'} \equiv \text{cov}(\mathbf{w}_{j,k}, \mathbf{w}_{j',k'} | \mathbf{z}) = \hat{V}_{j,k,j',k'} - \mathbf{A}_{j,k} \hat{V}_{pa\{j,k\},j',k'} - \hat{V}_{j,k,pa\{j',k'\}} \mathbf{A}'_{j',k'} + \mathbf{A}_{j,k} \hat{V}_{pa\{j,k\},pa\{j',k'\}} \mathbf{A}'_{j',k'}$ ,

$\hat{Y}_{jkj'} \equiv \text{cov}(\mathbf{w}_{j,k}, -\mathbf{u}_{j'}|\mathbf{z}) = -\hat{V}_{jk,j'} + \mathbf{A}_{j,k}\hat{V}_{pa\{j,k\},j'}$ ,  $\hat{W}_{1jk} \equiv \text{cov}(\mathbf{u}_{1,1}, \mathbf{w}_{j,k}|\mathbf{z}) = \hat{V}_{1,1,j,k} - \hat{V}_{1,1,pa\{j,k\}}\mathbf{A}'_{j,k}$ ,  $\hat{V}_{jk,j'k'} \equiv \text{cov}(\mathbf{u}_{j,k}, \mathbf{u}_{j'k'}|\mathbf{z})$ , We use a single  $\cdot$  to denote covariances involving  $\mathbf{u}_j$  and a double  $\cdot\cdot$  to denote covariances involving  $\mathbf{u}$ . When there are no missing data as in Section 2–3, we have  $\gamma_{j,k} \equiv 1$ ,  $\mathbf{\Gamma}_j \equiv \mathbf{I}_{N_j}$ , and  $\mathbf{\Gamma} \equiv \mathbf{I}_{N_1+\dots+N_J}$ .

As for the Kalman-filter algorithm, the Kalman-filter formulas in Section 4 remain the same, except for the following changes:

$$\begin{aligned}\hat{\mathbf{u}}_{J,k|J,k} &\equiv E(\mathbf{u}_{J,k}|\mathbf{z}_{J,k}) = \gamma_{J,k}\mathbf{V}_{J,k}(\mathbf{V}_{J,k} + \mathbf{\Phi}_J)^{-1}(\mathbf{z}_{J,k} - \mathbf{X}_{J,k}\boldsymbol{\beta}), \\ \hat{\mathbf{V}}_{J,k|J,k} &\equiv \text{var}(\mathbf{u}_{J,k}|\mathbf{z}_{J,k}) = \mathbf{V}_{J,k} - \gamma_{J,k}\mathbf{V}_{J,k}(\mathbf{V}_{J,k} + \mathbf{\Phi}_J)^{-1}\mathbf{V}_{J,k}, \\ \hat{\mathbf{u}}_{j,k|j,k} &\equiv E(\mathbf{u}_{j,k}|\mathbf{z}_{de\{j,k\}}) = \hat{\mathbf{V}}_{j,k|j,k}(\gamma_{j,k}\mathbf{\Phi}_j^{-1}(\mathbf{z}_{j,k} - \mathbf{X}_{j,k}\boldsymbol{\beta}) + (\hat{\mathbf{V}}_{j,k|j,k}^*)^{-1}\hat{\mathbf{u}}_{j,k|j,k}^*), \\ \hat{\mathbf{V}}_{j,k|j,k} &\equiv \text{var}(\mathbf{u}_{j,k}|\mathbf{z}_{de\{j,k\}}) = \hat{\mathbf{V}}_{j,k|j,k}^* - \gamma_{j,k}\hat{\mathbf{V}}_{j,k|j,k}^*(\hat{\mathbf{V}}_{j,k|j,k}^* + \mathbf{\Phi}_j)^{-1}\hat{\mathbf{V}}_{j,k|j,k}^*.\end{aligned}$$

## Appendix B: Inference with Single-Source Data

In this section, we derive formulas for parameter estimation and prediction when observations are missing on all the coarser resolutions. That is, there is only one source of observations on the finest  $J$ -th resolution. Recall that  $\mathbf{z}_J \equiv (\mathbf{z}_{J,k} : k = 1, \dots, N_{J-1})'$  denotes the data on the finest  $J$ -th resolution and let  $\mathbf{\Gamma}_J \equiv \text{diag}\{\gamma_{J,k}\mathbf{I}_{n_{J-1}}\}$  indicate which data among  $\mathbf{z}_J$  are observed.

In the EM algorithm, the E-step based on (24) gives,

$$\begin{aligned}E(\ell(\boldsymbol{\theta}^*; \mathbf{z}_J, \mathbf{u})|\mathbf{z}_J, \hat{\boldsymbol{\theta}}^*) &\propto \\ &- \sum_{\gamma_{J,k}=1} \left[ \frac{1}{2\phi_J} \{(\mathbf{z}_{J,k} - \mathbf{X}_{J,k}\boldsymbol{\beta} - \hat{\mathbf{u}}_{J,k})'(\mathbf{z}_{J,k} - \mathbf{X}_{J,k}\boldsymbol{\beta} - \hat{\mathbf{u}}_{J,k}) + \text{tr}(\hat{\mathbf{V}}_{J,k})\} \right] \\ &- \left[ \frac{N_1}{2} \log \sigma_1^2 + \frac{1}{2\sigma_1^2} \{\hat{\mathbf{u}}'_{1,1}\hat{\mathbf{u}}_{1,1} + \text{tr}(\hat{\mathbf{V}}_{1,1})\} \right] - \sum_{j=2}^J \sum_{k=1}^{N_{j-1}} \left( \frac{n_{j-1}-1}{2} \log \sigma_j^2 + \frac{1}{2\sigma_j^2} \frac{n_{j-1}-1}{n_{j-1}} \hat{Q}_{j,k} \right).\end{aligned}\tag{26}$$

In the M-step, we maximize (26) and obtain the  $t$ -th parameter estimates  $\hat{\boldsymbol{\theta}}^{*(t)}$ :

$$\begin{aligned}(\hat{\sigma}_1^2)^{(t)} &= \frac{1}{N_1} \{\hat{\mathbf{u}}'_{1,1}\hat{\mathbf{u}}_{1,1} + \text{tr}(\hat{\mathbf{V}}_{1,1})\}, \quad (\hat{\sigma}_j^2)^{(t)} = \frac{1}{N_{j-1}n_{j-1}} \sum_{k=1}^{N_{j-1}} \hat{Q}_{j,k}, \quad j = 2, \dots, J, \\ \hat{\boldsymbol{\beta}}^{(t)} &= (\mathbf{X}'_J \mathbf{\Phi}_J^{-1} \mathbf{\Gamma}_J \mathbf{X}_J)^{-1} \mathbf{X}'_J \mathbf{\Phi}_J^{-1} \mathbf{\Gamma}_J (\mathbf{z}_J - \hat{\mathbf{u}}_J).\end{aligned}$$

Again by the Louis' formula (1982), the observed information matrix is,

$$\mathcal{I}(\boldsymbol{\theta}^*) = E \left( -\frac{\partial^2 \ell(\boldsymbol{\theta}^*; \mathbf{z}_J, \mathbf{u})}{\partial \boldsymbol{\theta}^* \partial \boldsymbol{\theta}^{*'}} \Big| \mathbf{z}_J \right) - E \left\{ \left( \frac{\partial \ell(\boldsymbol{\theta}^*; \mathbf{z}_J, \mathbf{u})}{\partial \boldsymbol{\theta}^*} \right) \left( \frac{\partial \ell(\boldsymbol{\theta}^*; \mathbf{z}_J, \mathbf{u})}{\partial \boldsymbol{\theta}^*} \right)' \Big| \mathbf{z}_J \right\}.$$

The matrix  $E \left( -\frac{\partial^2 \ell(\boldsymbol{\theta}^*; \mathbf{z}_J, \mathbf{u})}{\partial \boldsymbol{\theta}^* \partial \boldsymbol{\theta}^{*'}} \Big| \mathbf{z}_J; \hat{\boldsymbol{\theta}}_{\text{MLE}}^* \right)$  evaluated at  $\hat{\boldsymbol{\theta}}_{\text{MLE}}^*$  has diagonal elements:

$$\frac{N_1}{2(\sigma_1^2)^2}, \quad \left\{ \frac{N_{j-1}(n_{j-1} - 1)}{2(\sigma_j^2)^2} : j = 2, \dots, J \right\}, \quad \mathbf{X}'_J \boldsymbol{\Phi}_J^{-1} \boldsymbol{\Gamma}_J \mathbf{X}_J$$

The non-zero first and second derivatives of the log-likelihood function (26) are the same as in Appendix A, except

$$\mathbf{c} \equiv \frac{\partial \ell(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\beta}} = \mathbf{X}'_J \boldsymbol{\Phi}_J^{-1} \boldsymbol{\Gamma}_J (\mathbf{z}_J - \mathbf{X}_J \boldsymbol{\beta} - \mathbf{u}_J), \quad \frac{\partial^2 \ell(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\mathbf{X}'_J \boldsymbol{\Phi}_J^{-1} \boldsymbol{\Gamma}_J \mathbf{X}_J.$$

Thus the elements of the matrix  $E(\Delta \ell \Delta \ell' | \mathbf{z}_J)$  evaluated at  $\hat{\boldsymbol{\theta}}_{\text{MLE}}^*$  are the same as in Appendix A, except

$$\begin{aligned} E(a_1 \mathbf{c} | \mathbf{z}_J) &= -\frac{1}{(\sigma_1^2)^2} \mathbf{X}'_J \boldsymbol{\Phi}_J^{-1} \boldsymbol{\Gamma}_J \hat{V}_{J,11} \hat{\mathbf{u}}_{1,1}, \\ E(a_j \mathbf{c} | \mathbf{z}_J) &= -\frac{n_{j-1} - 1}{n_{j-1} (\sigma_j^2)^2} \mathbf{X}'_J \boldsymbol{\Phi}_J^{-1} \boldsymbol{\Gamma}_J \sum_{k=1}^{N_{j-1}} \{ \hat{V}_{J, \cdot, jk} - \hat{V}_{J, pa\{j,k\}} \mathbf{A}'_{j,k} \} \hat{\mathbf{w}}_{j,k}, \\ E(\mathbf{c} \mathbf{c}' | \mathbf{z}_J) &= \mathbf{X}'_J \boldsymbol{\Phi}_J^{-1} \boldsymbol{\Gamma}_J \hat{V}_{J, \cdot, J} \boldsymbol{\Gamma}_J \boldsymbol{\Phi}_J^{-1} \mathbf{X}_J. \end{aligned}$$

The change-of-resolution Kalman-filter algorithm remains the same, except that  $\gamma_{j,k} = 0$  whenever  $j < J$ .

Figure 1: A square region is partitioned into 4 cells, each partitioned into 4 subcells. The multiresolution tree structure has 3 resolutions, 1 root, and 4 child nodes for each node that is a parent node.

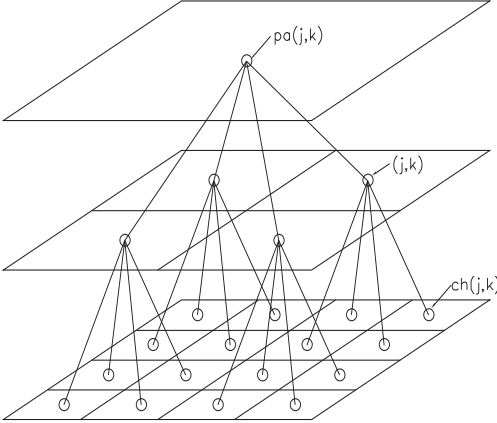


Figure 2: (a) A square field (160 m by 120 m) is partitioned into 240 cells on the finest resolution. (b) 3D smoothed surface plot of yield data over the field. (c) 3D smoothed surface plot of predicted yield over the field.

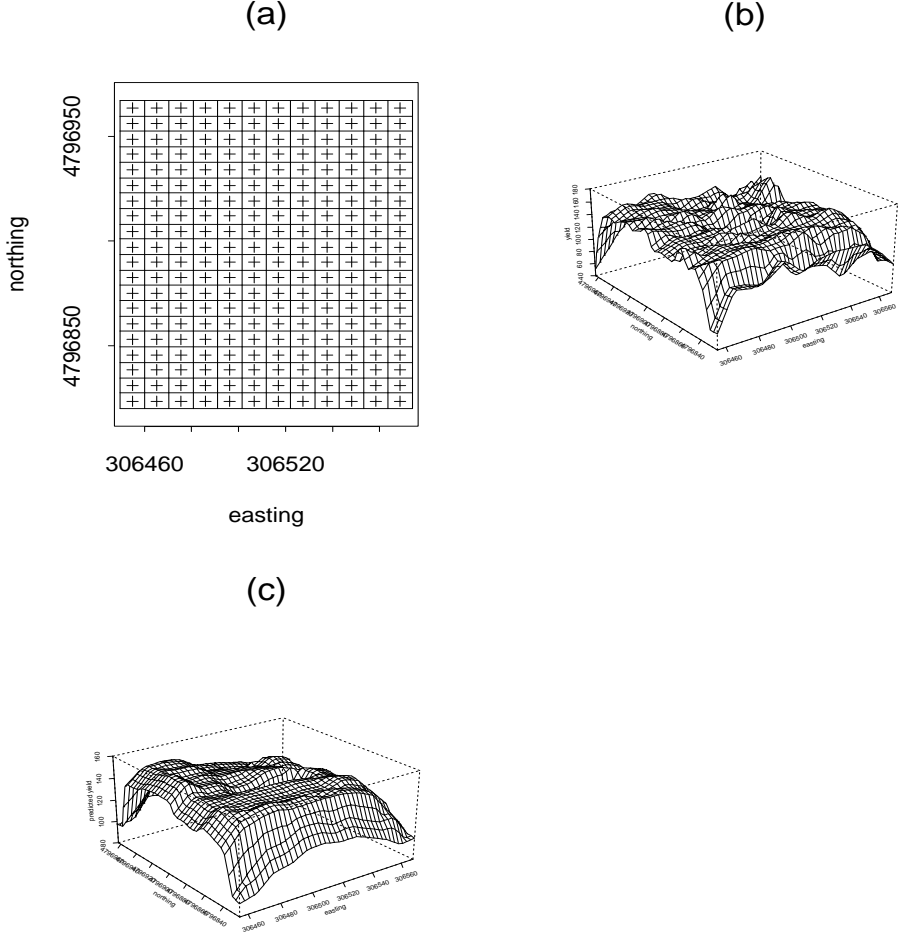


Figure 3: All pairwise scatter plots among the variables of yield, crop moisture, elevation, slope, cosine of aspect, edge effect, easting, and northing.

