Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2009)
71, *Part* 4, *pp.* 859–880

Adaptively varying-coefficient spatiotemporal models

Zudi Lu,

London School of Economics and Political Science, UK, Curtin University of Technology, Perth, and University of Adelaide, Australia

Dag Johan Steinskog,

Nansen Environmental and Remote Sensing Center, and Bjerknes Centre for Climate Research, Bergen, Norway

Dag Tjøstheim

University of Bergen, Norway

and Qiwei Yao

London School of Economics and Political Science, UK, and Peking University, Beijing, People's Republic of China

[Received October 2007. Final revision January 2009]

Summary. We propose an adaptive varying-coefficient spatiotemporal model for data that are observed irregularly over space and regularly in time. The model is capable of catching possible non-linearity (both in space and in time) and non-stationarity (in space) by allowing the auto-regressive coefficients to vary with both spatial location and an unknown index variable. We suggest a two-step procedure to estimate both the coefficient functions and the index variable, which is readily implemented and can be computed even for large spatiotemporal data sets. Our theoretical results indicate that, in the presence of the so-called nugget effect, the errors in the estimation may be reduced via the spatial smoothing—the second step in the estimation procedure proposed. The simulation results reinforce this finding. As an illustration, we apply the methodology to a data set of sea level pressure in the North Sea.

Keywords: β-mixing; Kernel smoothing; Local linear regression; Nugget effect; Spatial smoothing; Unilateral order

1. Introduction

The wide availability of data that are observed over time and space, in particular through inexpensive geographical information systems, has stimulated many studies in a variety of disciplines such as environmental science, epidemiology, political science, demography, economics and geography. In these studies, the geographical areas (e.g. counties and census tracts) are taken as units of analysis, and specific statistical methods have been developed to deal with the spatial structure that is reflected in the distribution of the dependent variable;

Address for correspondence: Dag Tjøstheim, Department of Mathematics, University of Bergen, 5007 Bergen, Norway.
E-mail: Dag.Tjostheim@math.uib.no

 1369–7412/09/71859

see for example Ripley (1981), Anselin (1988), Cressie (1993), Guyon (1995), Stein (1999) and Diggle (2003) for systematic reviews of these and related topics.

In this paper we are concerned with spatiotemporal data which are measured or transformed to a continuous scale and observed irregularly over space but regularly over time. Data sets of this form exist extensively. They may be environmental monitoring stations located irregularly in some region, but with measurements taken daily; see for instance Fotheringham *et al.* (1998), Brunsdon *et al.* (2001), Fuentes (2001) and Zhang *et al.* (2003). Applications in spatial disease modelling may be found in Knorr-Held (2000) and Lagazio *et al.* (2001).

Our model is of the form

$$Y_t(\mathbf{s}) = a\{\mathbf{s}, \boldsymbol{\alpha}(\mathbf{s})^{\mathrm{T}} \mathbf{X}_t(\mathbf{s})\} + \mathbf{b}_1\{\mathbf{s}, \boldsymbol{\alpha}(\mathbf{s})^{\mathrm{T}} \mathbf{X}_t(\mathbf{s})\}^{\mathrm{T}} \mathbf{X}_t(\mathbf{s}) + \varepsilon_t(\mathbf{s}), \tag{1}$$

where $Y_t(\mathbf{s})$ is the spatiotemporal variable of interest, and t is time, and $\mathbf{s} = (u, v) \in \mathcal{S} \subset \mathbb{R}^2$ is a spatial location. Moreover, $a(\mathbf{s}, z)$ and $\mathbf{b}_1(\mathbf{s}, z)$ are unknown scalar and $d \times 1$ functions, $\boldsymbol{\alpha}(\mathbf{s})$ is an unknown $d \times 1$ index vector, $\{\varepsilon_t(\mathbf{s})\}$ is a noise process which, for each fixed $\mathbf{s}$, forms a sequence of independent and identically distributed random variables over time and $\mathbf{X}_t(\mathbf{s}) = \{X_{t1}(\mathbf{s}), \ldots, X_{td}(\mathbf{s})\}^{\mathrm{T}}$ consists of time-lagged values of $Y_t(\cdot)$ in a neighbourhood of $\mathbf{s}$ and, possibly, some exogenous variables. Throughout the paper we use 'T' to denote the transpose of a vector or a matrix. We let both the regression coefficient $\mathbf{b}_1$ and the intercept a depend on the location $\mathbf{s}$, as well as on the index variable $\boldsymbol{\alpha}(\mathbf{s})^{\mathrm{T}} \mathbf{X}_t(\mathbf{s})$, to catch possible non-stationary features over space. Model (1) is unilateral in time and, therefore, it can be easily simulated in a Monte Carlo study. Furthermore it is readily applicable to model practical problems. Those features make the model radically different from purely spatial auto-regressive models (Yao and Brockwell, 2006). Also note that, at a given location $\mathbf{s}$, the coefficient functions are univariate functions of $\boldsymbol{\alpha}(\mathbf{s})^{\mathrm{T}} \mathbf{X}_t(\mathbf{s})$. Therefore only one-dimensional smoothing is required in estimating those coefficients. (See Section 2.2.1 below.)

Model (1) is proposed to model non-linear and non-stationary spatial data in a flexible manner. Note that the literature on spatial processes is overwhelmingly dominated by linear models. In contrast, many practical data exhibit clearly non-linear and non-stationary features. For example, in our analysis of the daily mean sea level pressure MSLP readings in Section 5, the original time series at each location is non-linear. Fig. 3 in Section 5 displays the plots for the MSLP data at five randomly selected locations. There are roughly three high pressure periods within the time span concerned. In Fig. 3(b), the correlation structure of the differenced series in the high pressure periods is different from that of the low pressure periods and the transition periods. This naturally calls for a non-linear model with varying coefficients depending also on the past values of the variables from the neighbourhood locations. Such dynamics cannot result in a Gaussian marginal distribution, as indicated in Fig. 3(c). The estimated densities are more peaked than the normal distribution, the accumulation of density mass around zero being due to the high pressure activity.

Model (1) is a spatial extension of the adaptive varying-coefficient linear model of Fan *et al.* (2003); see also Xia and Li (1999). Owing to the spatial non-stationarity of the model, we adopt an estimation strategy involving two steps, which also facilitates the *parallel computation* over different locations. First, for each fixed location $\mathbf{s}$, we estimate the varying-coefficient functions a and $\mathbf{b}_1$ and the index vector $\boldsymbol{\alpha}$ on the basis of the observations at location $\mathbf{s}$ only. This is a time series estimation problem. Our estimation is based on local linear regression. The second step is to apply spatial smoothing to pool the information from neighbourhood locations. Asymptotic properties of our estimators have been established which indicate that, in the presence of the so-called 'nugget effect', the spatial smoothing will reduce the estimation errors. (See also Lu *et al.* (2008).) Our simulation study reinforces this finding.

As far as we are aware, this is the first paper to address spatiotemporal non-linear dependence structures with spatial non-stationarity and non-Gaussian distributions. Earlier work on non-linear and/or non-Gaussian spatial models includes, among others, Matheron (1976), Rivoirard (1994), Chilès and Delfiner (1999), chapter 6, Hallin *et al.* (2004a, b), Biau and Cadre (2004), Gao *et al.* (2006) and Lu *et al.* (2007a).

The structure of the paper is as follows. We introduce our methodology in Section 2. Illustrations with simulated examples are reported in Section 3. In Section 5 there is an application to a data set of sea level pressure in the North Sea, for which the index variable may be viewed as an analogue of a spatial principal component used in the so-called empirical orthogonal functions analysis for climate time series. The asymptotic properties are reported in Section 4, with the regularity conditions deferred to Appendix A.

2. Methodology

2.1. Identification

Model (1) is not identifiable. In fact, it may be equivalently expressed as

$$\begin{aligned} Y_t(\mathbf{s}) = a\{\mathbf{s}, \boldsymbol{\alpha}(\mathbf{s})^{\mathrm{T}} \mathbf{X}_t(\mathbf{s})\} + \kappa\{\mathbf{s}, \boldsymbol{\alpha}(\mathbf{s})^{\mathrm{T}} \mathbf{X}_t(\mathbf{s})\}\boldsymbol{\alpha}(\mathbf{s})^{\mathrm{T}}\mathbf{X}_t(\mathbf{s}) + (\mathbf{b}_1\{\mathbf{s}, \boldsymbol{\alpha}(\mathbf{s})^{\mathrm{T}} \mathbf{X}_t(\mathbf{s})\} \\ - \kappa\{\mathbf{s}, \boldsymbol{\alpha}(\mathbf{s})^{\mathrm{T}} \mathbf{X}_t(\mathbf{s})\}\, \boldsymbol{\alpha}(\mathbf{s}))^{\mathrm{T}}\, \mathbf{X}_t(\mathbf{s}) + \varepsilon_t(\mathbf{s}), \end{aligned} \tag{2}$$

for any real-valued function $\kappa(\cdot)$. To overcome this problem, we may assume that the last component of $\boldsymbol{\alpha}(\mathbf{s})$ is non-zero, the first non-zero component of $\boldsymbol{\alpha}(\mathbf{s})$ is positive and $\|\boldsymbol{\alpha}(\mathbf{s})\|^2 = 1$. On the basis of those assumptions, Fan *et al.* (2003) imposed the condition that the last component of $\mathbf{b}_1$ is 0 in model (1). Put $\mathbf{b}_1 = (\mathbf{b}^{\mathrm{T}}, 0)^{\mathrm{T}}$. This effectively reduces model (1) to the form

$$Y_t(\mathbf{s}) = a\{\mathbf{s}, \boldsymbol{\alpha}(\mathbf{s})^{\mathrm{T}} \mathbf{X}_t(\mathbf{s})\} + (\mathbf{b}\{\mathbf{s}, \boldsymbol{\alpha}(\mathbf{s})^{\mathrm{T}}\mathbf{X}_t(\mathbf{s})\})^{\mathrm{T}}\, \check{\mathbf{X}}_t(\mathbf{s}) + \varepsilon_t(\mathbf{s}), \tag{3}$$

where $\check{\mathbf{X}}_t(\mathbf{s}) = (X_{t,1}(\mathbf{s}) \ldots X_{t,d-1}(\mathbf{s}))^{\mathrm{T}}$. In fact, $\boldsymbol{\alpha}$, a and $\mathbf{b}$ in model (3) are now identifiable as long as the regression function $E\{Y_t(\mathbf{s})|\mathbf{X}_t(\mathbf{s}) = \mathbf{x}\}$ is not of the form $\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x} + c$, where $\boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathbb{R}^d$ and $c \in \mathbb{R}^1$ are constants related only to $\mathbf{s}$; see theorem 1, part (b), of Fan *et al.* (2003). Note that model (2) reduces to model (3) by setting $\kappa(\mathbf{s}, z) = b_{1d}/\alpha_d$, where b_{1j} and α_j denote respectively the jth component of $\mathbf{b}_1(\mathbf{s}, z)$ and $\boldsymbol{\alpha}(\mathbf{s})$.

A disadvantage of the form (3) is that all the components of $\mathbf{X}_t(\mathbf{s})$ are not on an equal footing, which may cause difficulties in model interpretation. This may be particularly problematic when $\mathbf{X}_t(\mathbf{s})$ consists of neighbour variables of $Y_t(\mathbf{s})$. One possible remedy is to impose an orthogonal constraint in the model instead, i.e.

$$\begin{gathered} Y_t(\mathbf{s}) = a^*\{\mathbf{s}, \boldsymbol{\alpha}(\mathbf{s})^{\mathrm{T}} \mathbf{X}_t(\mathbf{s})\} + \mathbf{b}_1^*\{\mathbf{s}, \boldsymbol{\alpha}(\mathbf{s})^{\mathrm{T}} \mathbf{X}_t(\mathbf{s})\}^{\mathrm{T}}\, \mathbf{X}_t(\mathbf{s}) + \varepsilon_t(\mathbf{s}), \\ \boldsymbol{\alpha}(\mathbf{s})^{\mathrm{T}}\, \mathbf{b}_1^*(\mathbf{s}, z) = 0. \end{gathered} \tag{4}$$

In fact such an orthogonal representation may be obtained from model (3) with

$$\begin{aligned} a^*(\mathbf{s}, z) &= a(\mathbf{s}, z) + (\mathbf{b}(\mathbf{s}, z)^{\mathrm{T}}, 0)\, \boldsymbol{\alpha}(\mathbf{s}) z, \\ \mathbf{b}_1^*(\mathbf{s}, z) &= (\mathbf{b}(\mathbf{s}, z)^{\mathrm{T}}, 0)^{\mathrm{T}} - \boldsymbol{\alpha}(\mathbf{s})(\mathbf{b}(\mathbf{s}, z)^{\mathrm{T}}, 0)\boldsymbol{\alpha}\,(\mathbf{s}) \end{aligned} \tag{5}$$

whereas $\boldsymbol{\alpha}(\mathbf{s})$ remains unchanged. Note that the condition $\boldsymbol{\alpha}(\mathbf{s})^{\mathrm{T}}\, \mathbf{b}_1^*(\mathbf{s}, z) = 0$ is fulfilled as $\|\boldsymbol{\alpha}(\mathbf{s})\|^2 = 1$. Furthermore, model (4) is an equivalent representation of model (3), and it is identifiable as long as model (3) is identifiable. To see this, we first note that model (3) may be deduced from model (4) by letting

$$
\begin{aligned}
a(\mathbf{s}, z) &= a^*(\mathbf{s}, z) + \frac{b^*_{1d}}{\alpha_d} z, \\
\mathbf{b}(\mathbf{s}, z) &= (b^*_{11}, \ldots, b^*_{1,d-1})^{\mathrm{T}} - \frac{b^*_{1d}}{\alpha_d} (\alpha_1, \ldots, \alpha_{d-1})^{\mathrm{T}},
\end{aligned}
\tag{6}
$$

where b^*_{1j} denotes the jth component of $\mathbf{b}^*_1(\mathbf{s}, z)$. This is effectively to write the first equation in model (4) in the form of model (2) with $(a, \mathbf{b}_1, \kappa)$ replaced by $(a^*, \mathbf{b}^*_1, b^*_{1d}/\alpha_d)$. Moreover, as $\boldsymbol{\alpha}(\mathbf{s})^{\mathrm{T}} \mathbf{b}^*_1(\mathbf{s}, z) = 0$ and $\|\boldsymbol{\alpha}(\mathbf{s})\|^2 = 1$, it follows from the second equality in expression (6) that

$$
(\mathbf{b}(\mathbf{s}, z)^{\mathrm{T}}, 0)\, \boldsymbol{\alpha}(\mathbf{s}) = \sum_{j=1}^{d-1} \alpha_j b^*_{1j} - \frac{b^*_{1d}}{\alpha_d} \sum_{j=1}^{d-1} \alpha_j^2 = -\alpha_d b^*_{1d} - \frac{b^*_{1d}}{\alpha_d} (1 - \alpha_d^2) = -\frac{b^*_{1d}}{\alpha_d}.
$$

Hence the d-component b^*_{1d} of $\mathbf{b}^*_1(\mathbf{s}, z)$ is identifiable when model (3) is identifiable. This, together with expression (6), implies that all the other components of $\mathbf{b}^*_1(\mathbf{s}, z)$ as well as $a^*(\mathbf{s}, z)$ are also identifiable.

Since the orthogonal constraint $\boldsymbol{\alpha}^{\mathrm{T}} \mathbf{b}^*_1 = 0$ in model (4) poses further technical complication in inference, in what follows we proceed with the identifiable model (3), assuming that the first non-zero element of $\boldsymbol{\alpha}(\mathbf{s})$ is positive. Only for the real data example in Section 5 do we present the fitted model in the form (4) via the transformation (5).

2.2. Estimation

With the observations $\{Y_t(\mathbf{s}_i), \mathbf{X}_t(\mathbf{s}_i)\}$ for $t = 1, \ldots, T$ and $i = 1, \ldots, N$, we estimate a, $\mathbf{b}$ and $\boldsymbol{\alpha}$ in model (3) in two steps.

(a) For each fixed $\mathbf{s} = \mathbf{s}_i$, we estimate them by using the time series data $\{Y_t(\mathbf{s}), \mathbf{X}_t(\mathbf{s}), t = 1, \ldots, T\}$ only.
(b) By pooling the information from neighbour locations via spatial smoothing, we improve the estimators that are obtained in step (a).

The second step also facilitates the estimation at a location $\mathbf{s}$ with no direct observations (i.e. $\mathbf{s} \neq \mathbf{s}_i$ for $1 \leqslant i \leqslant N$).

2.2.1. Time series estimation

For a fixed $\mathbf{s}$, we estimate the direction $\boldsymbol{\alpha}(\mathbf{s})$ and the coefficient functions $a(\mathbf{s}, \cdot)$ and $\mathbf{b}(\mathbf{s}, \cdot)$. Fan *et al.* (2003) proposed a back-fitting algorithm to solve this estimation problem. We argue that, with modern computer power, the problem can be dealt with in a more direct manner even for moderately large d such as d between 10 and 20. Once the direction $\boldsymbol{\alpha}(\mathbf{s})$ has been fixed, the estimators for a and $\mathbf{b}$ may be obtained by one-dimensional smoothing over $\boldsymbol{\alpha}(\mathbf{s})^{\mathrm{T}} \mathbf{X}_t(\mathbf{s})$ using, for example, standard kernel regression methods.

We consider the estimation for $\boldsymbol{\alpha}(\mathbf{s})$ first. If the coefficient functions a and $\mathbf{b}$ were known, we may estimate $\boldsymbol{\alpha}$ by solving a non-linear least squares problem; see equation (8) below. Since a and $\mathbf{b}$ are unknown, we may plug in their appropriate estimators instead. By doing so, we shall have used the same observations twice. Those estimators themselves are functions of $\boldsymbol{\alpha}$. Hence a cross-validation approach will be employed to mitigate the effect of the double use of the same data set. For any $\boldsymbol{\alpha}(\mathbf{s})$, define the leave-one-out estimators $\breve{a}_{-t}(\mathbf{s}, z) = \breve{a}$ and $\breve{\mathbf{b}}_{-t}(\mathbf{s}, z) = \breve{\mathbf{b}}$, where $(\breve{a}, \breve{\mathbf{b}})$ minimizes

$$\frac{1}{T-1}\sum_{\substack{1\leqslant j\leqslant T\\ j\neq t}}\left\{Y_j(\mathbf{s})-a-\mathbf{b}^{\mathrm{T}}\check{\mathbf{X}}_j(\mathbf{s})\right\}^2 K_h\{\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{s})\,\mathbf{X}_j(\mathbf{s})-z\}, \tag{7}$$

where $K_h(\cdot)=h^{-1}K(\cdot/h)$, $K(\cdot)$ is a kernel function and $h>0$ is a bandwidth. Then we choose $\hat{\boldsymbol{\alpha}}(\mathbf{s})$ and h which minimize

$$R(\boldsymbol{\alpha},h)=\frac{1}{T}\sum_{t=1}^{T}[Y_t(\mathbf{s})-\check{a}_{-t}\{\mathbf{s},\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}_t(\mathbf{s})\}-\check{\mathbf{b}}_{-t}\{\mathbf{s},\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}_t(\mathbf{s})\}^{\mathrm{T}}\check{\mathbf{X}}_t(\mathbf{s})]^2\,w\{\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}_t(\mathbf{s})\}, \tag{8}$$

where $w(\cdot)$ is a weight function which controls the boundary effect. In the above definition for the estimator $\hat{\boldsymbol{\alpha}}(\mathbf{s})$, we used the Nadaraya–Watson estimators $\check{a}$ and $\check{\mathbf{b}}$ in expression (7) to keep the function $R(\boldsymbol{\alpha},h)$ simple. The asymptotic property of the estimator $\hat{\boldsymbol{\alpha}}(\mathbf{s})$ has been derived on the basis of the results of Lu *et al.* (2007b); see equation (25) in Section 4 below. The minimization of $R(\cdot)$ may be carried out using, for example, the downhill simplex method (section 10.4 of Press *et al.* (1992)).

Once $\boldsymbol{\alpha}(\mathbf{s})=\hat{\boldsymbol{\alpha}}(\mathbf{s})$ is known, we may estimate a and $\mathbf{b}$ by using univariate local linear regression estimation. For this, let $(\hat{a},\hat{\mathbf{b}},\hat{c},\hat{\mathbf{d}})$ be the minimizers of

$$\frac{1}{T}\sum_{j=1}^{T}\{Y_j(\mathbf{s})-a-c(Z_j-z)-(\mathbf{b}-\mathbf{d}(Z_j-z))^{\mathrm{T}}\check{\mathbf{X}}_j(\mathbf{s})\}^2 K_{\bar{h}}(Z_j-z), \tag{9}$$

where $Z_t=\hat{\boldsymbol{\alpha}}(\mathbf{s})^{\mathrm{T}}\mathbf{X}_t(\mathbf{s})$, and a different bandwidth $\bar{h}$ from the h in expression (7) will be used. Note that both h and $\bar{h}$ depend on T. The estimators for the functions $a(\mathbf{s},\cdot)$ and $\mathbf{b}(\mathbf{s},\cdot)$ at z are defined as $\hat{a}(\mathbf{s},z)=\hat{a}$ and $\hat{\mathbf{b}}(\mathbf{s},z)=\hat{\mathbf{b}}$.

2.2.2. Spatial smoothing

Although we do not assume stationarity over space, improvement in estimation of a, $\mathbf{b}$ and $\boldsymbol{\alpha}$ over $\mathbf{s}$ may be gained by extracting information from neighbourhood locations, owing to the continuity of the functions concerned. Zhang *et al.* (2003) showed for the model that they considered that it was indeed the case in the presence of the *nugget effect* (Cressie (1993), section 2.3.1). See also Lu *et al.* (2008). Furthermore, the spatial smoothing provides a way to extrapolate our estimators to the locations at which observations are not available.

Let $W(\cdot)$ be a kernel function defined on $\mathbb{R}^2$. Put $W_{\tilde{h}}(\mathbf{s})=\tilde{h}^{-2}W(\mathbf{s}/\tilde{h})$, where $\tilde{h}>0$ is a bandwidth depending on the size N of spatial samples. The bandwidth $\tilde{h}$ is different from both h and $\bar{h}$ in the last subsection. On the basis of the estimators that were obtained above for each of the locations $\mathbf{s}_1,\ldots,\mathbf{s}_N$, the spatially smoothed estimators at the location $\mathbf{s}_0$ are defined as

$$\tilde{\boldsymbol{\alpha}}(\mathbf{s}_0)=\sum_{j=1}^{N}\hat{\boldsymbol{\alpha}}(\mathbf{s}_j)\,W_{\tilde{h}}(\mathbf{s}_j-\mathbf{s}_0)\Big/\sum_{j=1}^{N}W_{\tilde{h}}(\mathbf{s}_j-\mathbf{s}_0), \tag{10}$$

$$\tilde{a}(\mathbf{s}_0,z)=\sum_{j=1}^{N}\hat{a}(\mathbf{s}_j,z)\,W_{\tilde{h}}(\mathbf{s}_j-\mathbf{s}_0)\Big/\sum_{j=1}^{N}W_{\tilde{h}}(\mathbf{s}_j-\mathbf{s}_0), \tag{11}$$

$$\tilde{\mathbf{b}}(\mathbf{s}_0,z)=\sum_{j=1}^{N}\hat{\mathbf{b}}(\mathbf{s}_j,z)\,W_{\tilde{h}}(\mathbf{s}_j-\mathbf{s}_0)\Big/\sum_{j=1}^{N}W_{\tilde{h}}(\mathbf{s}_j-\mathbf{s}_0). \tag{12}$$

2.3. Bandwidth selection

Fan *et al.* (2003) outlined an empirical rule for selecting a bandwidth in fitting non-spatial

varying-coefficient regression models. Below we adapt that idea to determining bandwidths $\bar{h}$ in expression (9) and $\tilde{h}$ in equations (10)–(12). Note that the bandwidth h in equation (8) is selected, together with $\hat{\boldsymbol{\alpha}}(\mathbf{s})$, by cross-validation (cf. Stone (1974)).

We first deal with the selection of $\bar{h}$ in expression (9). Let

$$\mathrm{CV}_1(\bar{h}) = \frac{1}{T}\sum_{t=1}^{T}\{Y_t(\mathbf{s}) - \breve{a}_{-t,\bar{h}}(\mathbf{s}, \hat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{X}_t(\mathbf{s})) - \breve{\mathbf{b}}_{-t,\bar{h}}(\mathbf{s}, \hat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{X}_t(\mathbf{s}))^{\mathrm{T}}\breve{\mathbf{X}}_t(\mathbf{s})\}^2 w\{\hat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{X}_t(\mathbf{s})\}, \tag{13}$$

where $\breve{a}_{-t,\bar{h}}(\mathbf{s}, z)$ and $\breve{\mathbf{b}}_{-t,\bar{h}}(\mathbf{s}, z)$ are the leave-one-out estimators that were defined as in expression (9) but with the term with $j = t$ removed from the sum. Under appropriate regularity conditions (see Lu *et al.* (2007b)), it holds that

$$\mathrm{CV}_1(\bar{h}) = c_0 + c_1\bar{h}^4 + \frac{c_2}{T\bar{h}} + o_P(\bar{h}^4 + T^{-1}\bar{h}^{-1}).$$

Thus, up to first-order asymptotics, the optimal bandwidth is $\bar{h}_{\mathrm{opt}} = (c_2/4Tc_1)^{1/5}$. In practice, the coefficients c_0, c_1 and c_2 will be estimated from solving the least squares problem

$$\min_{c_0,c_1,c_2}\left\{\sum_{k=1}^{q}(\mathrm{CV}_{1k} - c_0 - c_1\bar{h}_k^4 - c_2/T\bar{h}_k)^2\right\}, \tag{14}$$

where $\mathrm{CV}_{1k} = \mathrm{CV}_1(\bar{h}_k)$ is obtained from equation (13), and $\bar{h}_1, \ldots, \bar{h}_q$ are q prescribed bandwidth values. We let $\bar{h} = (\hat{c}_2/4T\hat{c}_1)^{1/5}$ when both $\hat{c}_1$ and $\hat{c}_2$ are positive. In the event that one of them is non-positive, we let $\bar{h} = \arg\min_{1\leqslant k\leqslant q}\{\mathrm{CV}_1(\bar{h}_k)\}$. This bandwidth selection procedure is computationally efficient as q is moderately small, i.e. we need to compute only q CV-values; see remark 2(c) in Fan *et al.* (2003). Furthermore the least squares estimation (14) also serves as a smoother for the CV bandwidth estimates. Also see Ruppert (1997).

The bandwidth $\tilde{h}$ in equations (10)–(12) may be determined in the same manner. For example, for the estimator (10), the CV-function is defined as

$$\mathrm{CV}_2(\tilde{h}) = \frac{1}{N}\sum_{i=1}^{N}\left\{\hat{\boldsymbol{\alpha}}(\mathbf{s}_i) - \tilde{\boldsymbol{\alpha}}_{-\mathbf{s}_i,\tilde{h}}(\mathbf{s}_i)\right\}^2,$$

which admits the asymptotic expansion

$$\mathrm{CV}_2(\tilde{h}) = d_0 + d_1\tilde{h}^4 + \frac{d_2}{N\tilde{h}^2} + o_P(\tilde{h}^4 + N^{-1}\tilde{h}^{-2}).$$

The resulting bandwidth is of the form $\tilde{h} = (\hat{d}_2/2N\hat{d}_1)^{1/6}$.

The CV bandwidth selection has been extensively studied in the literature. In the sense of mean integrated squared error the relative error of a CV-bandwidth is higher than that of, for example, a plug-in method of Ruppert *et al.* (1995). However, there is a growing body of opinion (e.g. Mammen (1990), Jones (1991), Härdle and Vieu (1992) and Loader (1999)) maintaining that the selection of bandwidth should be targeted at estimating the unknown function instead of the ideal bandwidth itself. Therefore, one should seek a bandwidth minimizing the integrated squared error rather than the mean integrated squared error, i.e. focusing on loss rather than risk. From this point of view, the CV-bandwidth performs reasonably well (Hall and Johnstone (1992), page 479). In the time series context, the argument for why cross-validation is an appropriate selection method for the bandwidth can be found in Kim and Cox (1995), Quintela-del-Río (1996) and Xia and Li (2002), among others.

3. A simulated example

Consider a spatiotemporal process

$$Y_{t+1}(\mathbf{s}_{ij}) = a\{\mathbf{s}_{ij}, Z_t(\mathbf{s}_{ij})\} + b_1\{\mathbf{s}_{ij}, Z_t(\mathbf{s}_{ij})\}\, Y_t(\mathbf{s}_{ij}) + b_2\{\mathbf{s}_{ij}, Z_t(\mathbf{s}_{ij})\}\, X_t(\mathbf{s}_{ij}) + 0.1\, \varepsilon_t(\mathbf{s}_{ij})$$

defined on the grid points $\mathbf{s}_{ij} = (u_i, v_j)$, where

$$\begin{aligned}
a(\mathbf{s}_{ij}, z) &= 3 \exp\left\{ \frac{-2z^2}{1 + 7(u_i + v_j)} \right\}, \\
b_1(\mathbf{s}_{ij}, z) &= 0.7 \sin\{7\pi(u_i + v_j)\}, \\
b_2(\mathbf{s}_{ij}, z) &= 0.8z, \\
Z_t(\mathbf{s}_{ij}) &= \tfrac{1}{3}\{2\, Y_t(\mathbf{s}_{ij}) + 2\, X_t(\mathbf{s}_{ij}) + X_t(\mathbf{s}_{i,j+1})\}, \\
X_t(\mathbf{s}_{ij}) &= \frac{2}{9} \sum_{k=-1}^{1} \sum_{l=-1}^{1} e_t(\mathbf{s}_{i+k,j+l}),
\end{aligned}$$

and all $\varepsilon_t(\mathbf{s}_{ij})$ and $e_t(\mathbf{s}_{ij})$ are independent $N(0, 1)$ random variables. Observations were taken over $N = 64$ grid points $\{(u_i, v_j), 1 \leqslant i, j \leqslant 8\}$, where $u_i = v_i = (i - 1)/8$ with $T = 60$ or $T = 100$. For each given $\mathbf{s} = \mathbf{s}_{ij}$, we estimated the curves $a(\mathbf{s}, \cdot)$, $b_1(\mathbf{s}, \cdot)$ and $b_2(\mathbf{s}, \cdot)$ on the 11 grid points $z_l = -0.5 + 0.1(l - 1)$ $(l = 1, \ldots, 11)$, as well as the index $\boldsymbol{\alpha}(\mathbf{s}) = (\alpha_1(\mathbf{s}), \alpha_2(\mathbf{s}), \alpha_3(\mathbf{s}))^{\mathrm{T}} \equiv (\frac{2}{3}, \frac{2}{3}, \frac{1}{3})^{\mathrm{T}}$, which is independent of $\mathbf{s}$. The accuracy of the estimation is measured by the squared estimation errors:

$$\begin{aligned}
\mathrm{SEE}\{\hat{\alpha}_j(\mathbf{s})\} &= \{\hat{\alpha}_j(\mathbf{s}) - \alpha_j\}^2, \qquad j = 1, 2, 3, \\
\mathrm{SEE}\{\hat{a}(\mathbf{s})\} &= \frac{1}{11} \sum_{l=1}^{11} \{\hat{a}(\mathbf{s}, z_l) - a(\mathbf{s}, z_l)\}^2, \\
\mathrm{SEE}\{\hat{b}_k(\mathbf{s})\} &= \frac{1}{11} \sum_{l=1}^{11} \{\hat{b}_k(\mathbf{s}, z_l) - b_k(\mathbf{s}, z_l)\}^2, \qquad k = 1, 2.
\end{aligned}$$

For each setting, we replicated the experiments 10 times. The SEEs over the 64 grid points are collectively displayed as boxplots in Fig. 1 for $T = 60$, and in Fig. 2 for $T = 100$ (i.e. each of the boxplots is based on $10 \times 64 = 640$ SEE-values). It is clear that the estimates that were defined in Section 2.2.1 are less accurate than those defined in Section 2.2.2. In fact the gain from the spatial smoothing is substantial. The plots also indicate that the estimation becomes more accurate when T increases.

The improvement from the spatial smoothing is due to the fact that the functions to be estimated are continuous in $\mathbf{s}$. If we view the observations $\{Y_t(\mathbf{s}_{ij}), X_t(\mathbf{s}_{ij})\}$ as a sample from a spatiotemporal process with $\mathbf{s}$ varying continuously over space, the sample paths of such an underlying process are discontinuous over space, as both $\varepsilon_t(\mathbf{s})$ and $e_t(\mathbf{s})$ are independent at different locations no matter how close they are. The spatial smoothing pools together the information on the mean function from neighbour locations. It is intuitively clear that there would not be any gains from such a 'local pooling' if the sample realizations of both $X_t(\cdot)$ and $\varepsilon_t(\cdot)$ are continuous in $\mathbf{s}$. The improvement may be obtained when the observations that are pooled together bring in different information, which is only possible when the sample realizations are discontinuous. The theory that is developed in the next section indicates that the spatial smoothing may indeed improve the estimation when the nugget effect is present, which results in the discontinuity in sample realizations as a function of $\mathbf{s}$. See also Lu *et al.* (2008) for more detailed discussions regarding the asymptotic theory with or without the nugget effect.

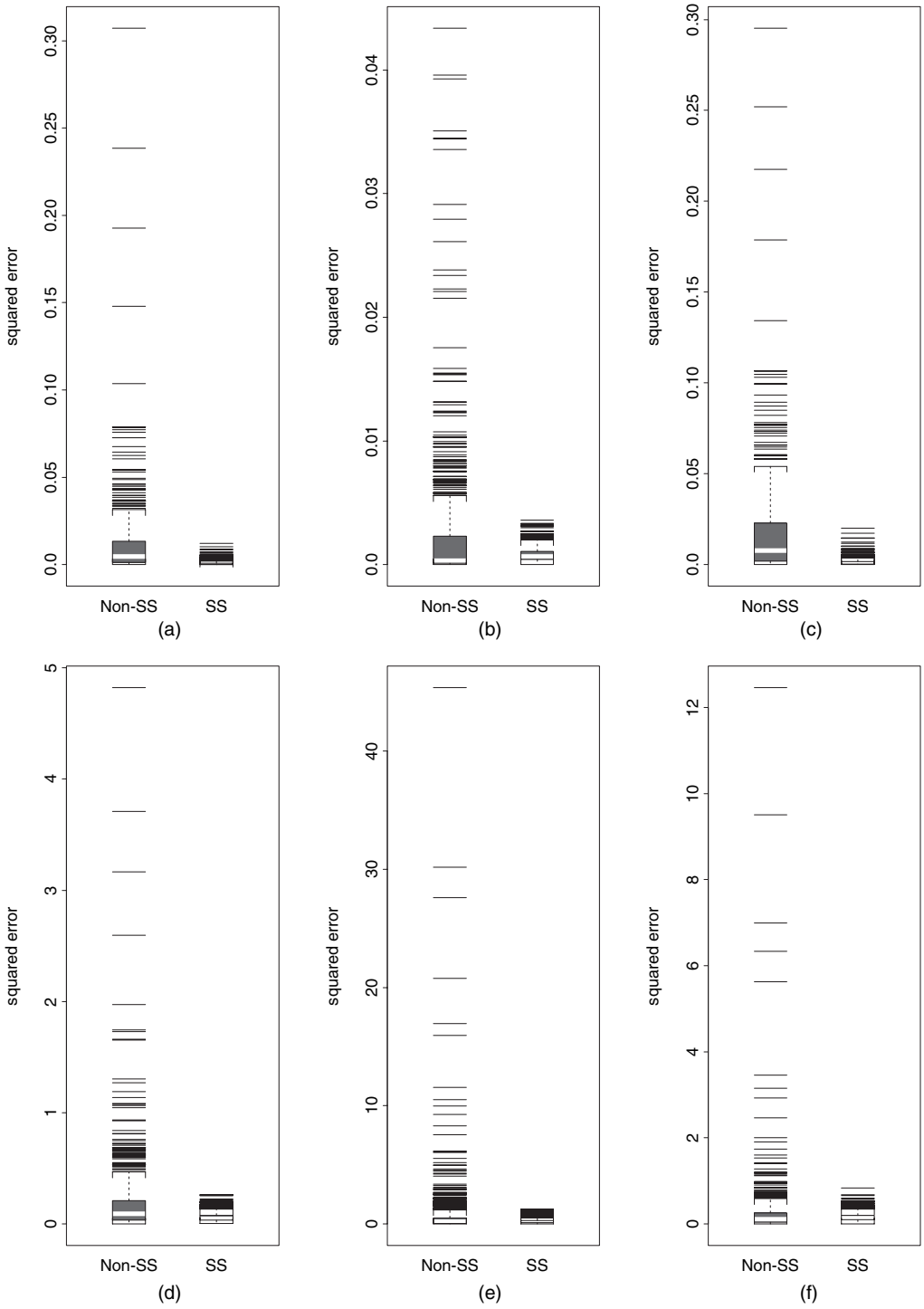

Fig. 1. Simulation with $T = 60$—boxplots for (a)–(c) the three components of α and (d)–(f) the varying-coefficient functions (the plots for the estimates that are defined in Section 2.2.1 are labelled as 'Non-SS', and the plots for the spatial smoothing estimates are labelled as 'SS'): (a) α_1; (b) α_2; (c) α_3; (d) $a(\cdot)$; (e) $b_1(\cdot)$; (f) $b_2(\cdot)$

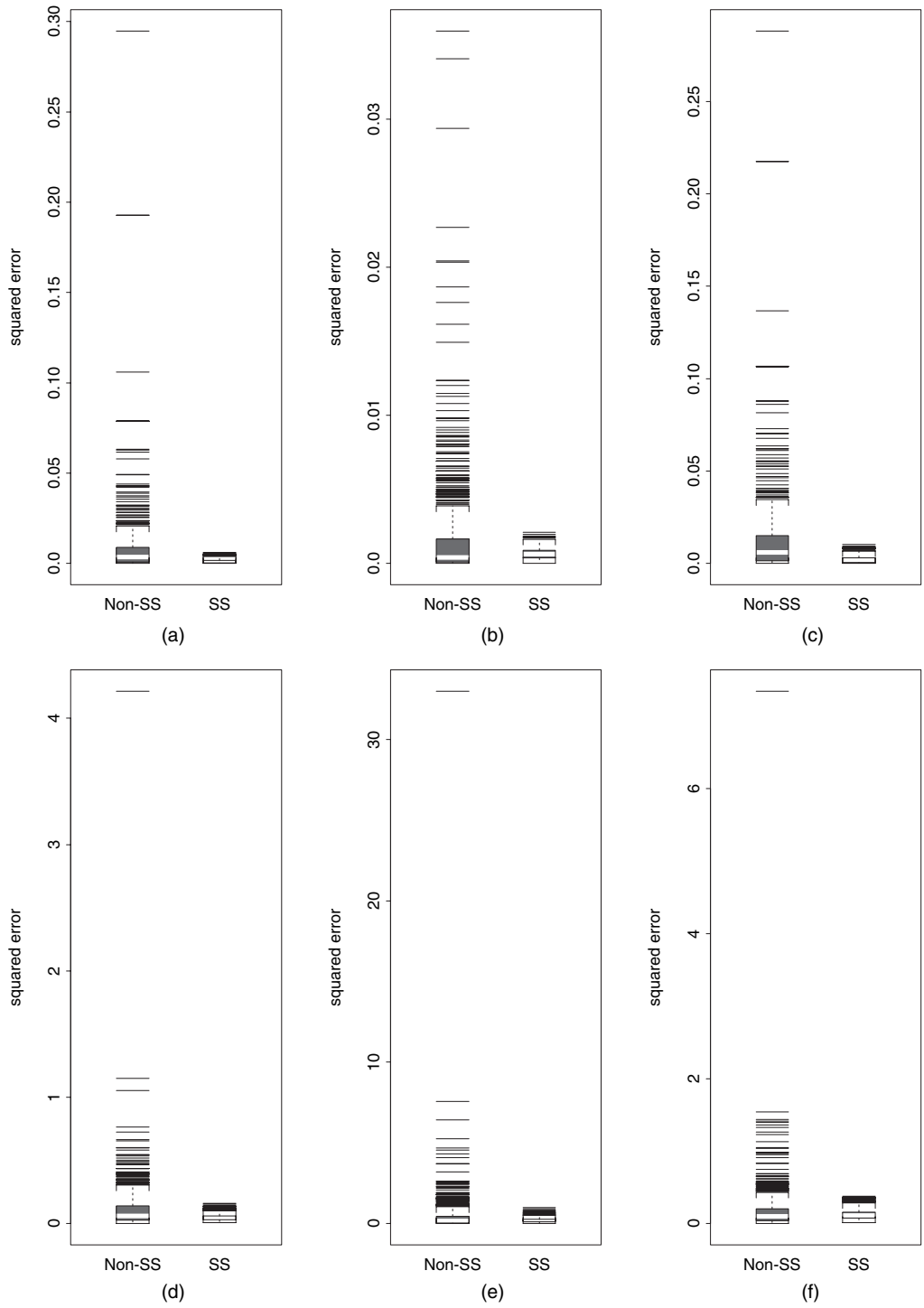

Fig. 2. Simulation with $T = 100$—boxplots for (a)–(c) the three components of α and (d)–(f) the varying-coefficient functions (the plots for the estimates that are defined in Section 2.2.1 are labelled as 'Non-SS', and the plots for the spatial smoothing estimates are labelled as 'SS'): (a) α_1; (b) α_2; (c) α_3; (d) $a(\cdot)$; (e) $b_1(\cdot)$; (f) $b_2(\cdot)$

4. Asymptotic properties

The improvements due to the spatial smoothing have been illustrated via simulation in Section 3. In this section, we derive the asymptotic bias and variance of the estimators (10)–(12), which also show the benefits from spatial smoothing in the presence of the so-called nugget effect. Note that those asymptotic approximations are derived without stationarity over space. However, the asymptotic normality for the time series estimators that were defined in Section 2.2.1 may be derived from theorems 3.1 and 3.2 of Lu *et al.* (2007b). We introduce some notation first.

We always assume the true index $\boldsymbol{\alpha}(\mathbf{s}) \in \mathbb{B}$ for all $\mathbf{s} \in \mathcal{S}$, where

$$\mathbb{B} = \{\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^{\mathrm{T}} \in \mathbb{R}^d : \|\boldsymbol{\beta}\| = 1, \text{the first non-zero element is positive, and the last element is non-zero with} |\beta_d| \geqslant \varepsilon_0\}, \tag{15}$$

and $\varepsilon_0 > 0$ is a fixed constant. Put $\mathbb{X}_t(\mathbf{s}) = (1, \breve{\mathbf{X}}_t(\mathbf{s})^{\mathrm{T}})^{\mathrm{T}} = (1, X_{t1}(\mathbf{s}), \ldots, X_{t,d-1}(\mathbf{s}))^{\mathrm{T}}$, and

$$\mathbf{g}(\mathbf{s}, z, \boldsymbol{\beta}) = E\{\mathbb{X}_t(\mathbf{s})\, \mathbb{X}_t^{\mathrm{T}}(\mathbf{s}) | \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_t(\mathbf{s}) = z\}^{-1} E\{\mathbb{X}_t(\mathbf{s})\, Y_t(\mathbf{s}) | \boldsymbol{\beta}^{\mathrm{T}}\, \mathbf{X}_t(\mathbf{s}) = z\},$$

where the matrix inverse is ensured by assumption 6 in Appendix A. We denote the derivatives of $\mathbf{g}$ as follows:

$$\begin{aligned}
\dot{\mathbf{g}}_z(\mathbf{s}, z, \boldsymbol{\beta}) &= \partial \mathbf{g}(\mathbf{s}, z, \boldsymbol{\beta}) / \partial z, \\
\dot{\mathbf{g}}_\beta(\mathbf{s}, z, \boldsymbol{\beta}) &= \partial \mathbf{g}(\mathbf{s}, z, \boldsymbol{\beta}) / \partial \boldsymbol{\beta}^{\mathrm{T}}, \\
\ddot{\mathbf{g}}_z(\mathbf{s}, z, \boldsymbol{\beta}) &= \partial^2 \mathbf{g}(\mathbf{s}, z, \boldsymbol{\beta}) / \partial z^2.
\end{aligned}$$

When $\boldsymbol{\beta} = \boldsymbol{\alpha}(\mathbf{s})$, we shall often drop $\boldsymbol{\alpha}(\mathbf{s})$ in $\mathbf{g}\{\mathbf{s}, z, \boldsymbol{\alpha}(\mathbf{s})\}$ and its derivatives if no confusion arises. It therefore follows from model (3) that $\mathbf{g}(\mathbf{s}, z)^{\mathrm{T}} = \mathbf{g}(\mathbf{s}, z, \boldsymbol{\alpha}(\mathbf{s}))^{\mathrm{T}} = \{a(\mathbf{s}, z), \mathbf{b}(\mathbf{s}, z)^{\mathrm{T}}\}$. Further, we let $Z_t(\mathbf{s}) = \boldsymbol{\alpha}(\mathbf{s})^{\mathrm{T}}\, \mathbf{X}_t(\mathbf{s})$ and define

$$\Gamma(\mathbf{s}) = E[\tilde{g}(\mathbf{s})^{\mathrm{T}}\, \mathbb{X}_t(\mathbf{s})\, \mathbb{X}_t(\mathbf{s})^{\mathrm{T}} \tilde{g}(\mathbf{s})\, w\{Z_t(\mathbf{s})\}], \tag{16}$$

where $\tilde{g}(\mathbf{s}) = \dot{\mathbf{g}}_z\{\mathbf{s}, Z_t(\mathbf{s}), \alpha(\mathbf{s})\}\, \mathbf{X}_t(\mathbf{s})^{\mathrm{T}} + \dot{\mathbf{g}}_\beta\{\mathbf{s}, Z_t(\mathbf{s}), \alpha(\mathbf{s})\}$.

Let $\alpha_j(\mathbf{s})$ and $\ddot{\alpha}_j(\mathbf{s})$ be the jth component of $\boldsymbol{\alpha}(\mathbf{s})$ and its second-order derivative matrix with respect to $\mathbf{s}$ respectively. Recall that $h = h_T$ and $\bar{h} = \bar{h}_T$ are the two bandwidths that are used for time series smoothing in Section 2.2.1, and $\tilde{h} = \tilde{h}_N$ is the bandwidth for spatial smoothing in Section 2.2.2. Put $\mu_{2,K} = \int u^2 K(u)\, \mathrm{d}u$, $\boldsymbol{\mu}_{2,W} = \int_{\mathcal{S}} \mathbf{u}\mathbf{u}^{\mathrm{T}}\, W(\mathbf{u})\, \mathrm{d}\mathbf{u}$ and $\nu_{0,K} = \int K^2(u)\, \mathrm{d}u$, where K and W are the two kernel functions that are used in our estimation.

Now we are ready to present the asymptotic biases and variances for the spatial smoothing estimators $\tilde{\boldsymbol{\alpha}}(\mathbf{s}_0)$, $\tilde{a}(\mathbf{s}_0)$ and $\tilde{b}(\mathbf{s}_0)$, which are derived when $T \to \infty$. The key is to show that the time series estimators $\hat{\boldsymbol{\alpha}}(\mathbf{s})$, $\hat{a}(\mathbf{s})$ and $\hat{b}(\mathbf{s})$ converge uniformly in $\mathbf{s}$ over a small neighbourhood of $\mathbf{s}_0$, which is established by using the results in Lu *et al.* (2007b). Naturally the derived asymptotic approximations for the biases and variances of $\tilde{\boldsymbol{\alpha}}(\mathbf{s}_0)$, $\tilde{a}(\mathbf{s}_0)$ and $\tilde{b}(\mathbf{s}_0)$ depend on N, and those approximations admit more explicit expressions when $N \to \infty$. Note that we write $a_N \simeq b_N$ if $\lim_{N\to\infty}(a_N / b_N) = 1$.

Theorem 1. Let conditions 1–10 listed in Appendix A hold. Assume that $\mathbf{s}_0 \in \mathcal{S}$ and $f(\mathbf{s}_0) > 0$, where $f(\cdot)$ is given in condition 8. Then, as $T \to \infty$, it holds that

$$\begin{aligned}
\tilde{\boldsymbol{\alpha}}(\mathbf{s}_0) - \boldsymbol{\alpha}(\mathbf{s}_0) = \mathcal{B}_{1,N}(\mathbf{s}_0) h^2 \{1 + o_P(1)\} + \mathcal{B}_{2,N}(\mathbf{s}_0) + T^{-1/2} \{\mathcal{V}_{1,N}(\mathbf{s}_0) \\
+ \mathcal{V}_{2,N}(\mathbf{s}_0)\}\, \eta(\mathbf{s}_0) \{1 + o_P(1)\},
\end{aligned}$$

where $\eta(\mathbf{s}_0)$ is a $d \times 1$ random vector with zero mean and identity covariance matrix,

$$\mathcal{B}_{1,N}(\mathbf{s}) \simeq -\tfrac{1}{2}\,\Gamma^{-1}(\mathbf{s})\; E[\tilde{g}(\mathbf{s})^{\mathrm{T}}\mathbb{X}_t(\mathbf{s})\mathbb{X}_t(\mathbf{s})^{\mathrm{T}}\ddot{\mathbf{g}}_z\{\mathbf{s}, Z_t(\mathbf{s}), \boldsymbol{\alpha}(\mathbf{s})\}\; w\{Z_t(\mathbf{s})\}]\,\mu_{2,K}, \tag{17}$$

$$\mathcal{B}_{2,N}(\mathbf{s}) \simeq \tfrac{1}{2}\,\tilde{h}^2(\mathrm{tr}\{\ddot{\alpha}_1(\mathbf{s})\mu_{2,W}\}, \ldots, \mathrm{tr}\{\ddot{\alpha}_d(\mathbf{s})\mu_{2,W}\})^{\mathrm{T}}, \tag{18}$$

and $\mathcal{V}_{1,N}(\mathbf{s}_0)$ and $\mathcal{V}_{2,N}(\mathbf{s}_0)$ are two $d \times d$ matrices, satisfying

$$\mathcal{V}_{1,N}(\mathbf{s})\,\mathcal{V}_{1,N}^{\mathrm{T}}(\mathbf{s}) \simeq \{\sigma_1^2(\mathbf{s}) + \sigma_2^2(\mathbf{s})\}\,\Gamma^{-1}(\mathbf{s})\,\mathbf{V}(\mathbf{s},\mathbf{s})\,\Gamma^{-1}(\mathbf{s})\left\{\frac{1}{N\tilde{h}^2} f(\mathbf{s})^{-1}\int W^2(\mathbf{u})\,\mathrm{d}\mathbf{u}\right\}, \tag{19}$$

$$\mathcal{V}_{2,N}(\mathbf{s})\,\mathcal{V}_{2,N}^{\mathrm{T}}(\mathbf{s}) \simeq \sigma_1^2(\mathbf{s})\,\Gamma^{-1}(\mathbf{s})\,\mathbf{V}_1(\mathbf{s},\mathbf{s})\,\Gamma^{-1}(\mathbf{s}). \tag{20}$$

In the above expressions, $\Gamma(\mathbf{s})$ is defined in equation (16), and $\sigma_1^2(\mathbf{s})$ and $\sigma_2^2(\mathbf{s})$ are defined in condition 2 and $\mathbf{V}$ and $\mathbf{V}_1$ in condition 3 in Appendix A.

Let $\tilde{\vartheta}(\mathbf{s}_0, z) \equiv (\tilde{a}(\mathbf{s}_0, z), \tilde{\mathbf{b}}(\mathbf{s}_0, z)^{\mathrm{T}})^{\mathrm{T}}$ and $\vartheta(\mathbf{s}, z) \equiv (a(\mathbf{s}, z), \mathbf{b}(\mathbf{s}, z)^{\mathrm{T}})^{\mathrm{T}}$. Denote by $\vartheta_j(\mathbf{s}, z)$ the jth component of $\vartheta(\mathbf{s}, z)$, and $\ddot{\vartheta}_{\mathbf{s},j}(\mathbf{s}, z)$ its second-derivative matrix with respect to $\mathbf{s}$. Denote by $\ddot{\vartheta}_z(\mathbf{s}, z) = (\ddot{a}_z(\mathbf{s}, z), \ddot{\mathbf{b}}_z(\mathbf{s}, z)^{\mathrm{T}})^{\mathrm{T}}$ the second-order derivatives with respect to z.

Theorem 2. Let the conditions of theorem 1 hold, and $\bar{h} = O(T^{-1/5})$. Let the density function of $\boldsymbol{\alpha}(\mathbf{s}_0)^{\mathrm{T}}\,\mathbf{X}_t(\mathbf{s}_0)$ be positive at z. Then, as $T \to \infty$, it holds that

$$\begin{aligned}\tilde{\vartheta}(\mathbf{s}_0, z) - \vartheta(\mathbf{s}_0, z) = {} & \mathcal{B}_{3,N}(\mathbf{s}_0, z)\bar{h}^2\{1 + o_P(1)\} + \mathcal{B}_{4,N}(\mathbf{s}_0, z) + \{(T\bar{h})^{-1/2}\mathcal{V}_{3,N} \\ & + T^{-1/2}\mathcal{V}_{4,N}\}\,\xi(\mathbf{s}_0)\{1 + o_P(1)\},\end{aligned}$$

where $\xi(\mathbf{s}_0)$ is a $d \times 1$ random vector with zero mean and identity covariance matrix,

$$\mathcal{B}_{3,N}(\mathbf{s}_0, z) \simeq \tfrac{1}{2}\mu_{2K}\,\ddot{\vartheta}_z(\mathbf{s}_0, z), \tag{21}$$

$$\mathcal{B}_{4,N}(\mathbf{s}_0, z) \simeq \tfrac{1}{2}\,\tilde{h}^2(\mathrm{tr}\{\ddot{\vartheta}_{\mathbf{s},1}(\mathbf{s}_0, z)\mu_{2,W}\}, \ldots, \mathrm{tr}\{\ddot{\vartheta}_{\mathbf{s},d}(\mathbf{s}_0, z)\mu_{2,W}\})^{\mathrm{T}}, \tag{22}$$

and $\mathcal{V}_{3,N}$ and $\mathcal{V}_{4,N}$ are two $d \times d$ matrices, satisfying

$$\mathcal{V}_{3,N}\mathcal{V}_{3,N}^{\mathrm{T}} \simeq \{\sigma_1^2(\mathbf{s}_0) + \sigma_2^2(\mathbf{s}_0)\}\nu_{0,K}\,\mathbf{A}^{-1}(\mathbf{s}_0, z)\,f_Z^{-1}(\mathbf{s}_0, z)\left\{\frac{1}{N\tilde{h}^2} f^{-1}(\mathbf{s}_0)\int W^2(\mathbf{s})\,\mathrm{d}\mathbf{s}\right\}, \tag{23}$$

$$\mathcal{V}_{4,N}\mathcal{V}_{4,N}^{\mathrm{T}} \simeq \sigma_1^2(\mathbf{s}_0)\{\mathbf{A}^{-1}(\mathbf{s}_0, z)\,\mathbf{A}_1(\mathbf{s}_0, \mathbf{s}_0; z, z)\,\mathbf{A}^{-1}(\mathbf{s}_0, z)\}\,f_Z^{-2}(\mathbf{s}_0, z)\,f^*(\mathbf{s}_0, z), \tag{24}$$

where $\sigma_1^2(\mathbf{s})$ and $\sigma_2^2(\mathbf{s})$ are defined in condition 2, $\mathbf{A}(\mathbf{s}, z) = \mathbf{A}(\mathbf{s}, \mathbf{s}, z, z)$ and $\mathbf{A}_1(\mathbf{s}, \mathbf{s}, z, z)$ in condition 3 and $f^*(\mathbf{s}_0, z)$ in condition 4 in Appendix A.

Remark 1.

(a) In theorems 1 and 2, expressions (17), (18), (21) and (22) are approximate biases whereas expressions (19), (20), (23) and (24) are approximate variances.

(b) In the presence of the nugget effect that is specified in conditions 2 and 3 in Appendix A, the spatial smoothing estimators $\tilde{\boldsymbol{\alpha}}(\mathbf{s}_0)$, $\tilde{a}(\mathbf{s}_0, z)$ and $\tilde{\mathbf{b}}(\mathbf{s}_0, z)$ have smaller asymptotic variances than the corresponding time series estimators $\hat{\boldsymbol{\alpha}}(\mathbf{s}_0)$, $\hat{a}(\mathbf{s}_0, z)$ and $\hat{\mathbf{b}}(\mathbf{s}_0, z)$. In fact, using the results in Lu *et al.* (2007b), we may deduce that

$$\hat{\boldsymbol{\alpha}}(\mathbf{s}) - \boldsymbol{\alpha}(\mathbf{s}) = \mathcal{B}_1(\mathbf{s})h^2\{1 + o_P(1)\} + T^{-1/2}\,\mathcal{A}_1(\mathbf{s})\,\eta(\mathbf{s})\{1 + o_P(1)\}, \tag{25}$$

where $\mathcal{B}_1(\mathbf{s})$ is as defined in theorem 1, and

$$\mathcal{A}_1(\mathbf{s})\,\mathcal{A}_1^{\mathrm{T}}(\mathbf{s}) = \{\sigma_1^2(\mathbf{s}) + \sigma_2^2(\mathbf{s})\}\,\Gamma^{-1}(\mathbf{s})\,\mathbf{V}(\mathbf{s},\mathbf{s})\,\Gamma^{-1}(\mathbf{s}),$$

and that

$$\hat{\vartheta}(\mathbf{s},z)-\vartheta(\mathbf{s},z)=\mathcal{B}_3(\mathbf{s},z)\bar{h}^2\{1+o_P(1)\}+(T\bar{h})^{-1/2}\,\mathcal{A}_3(\mathbf{s},z)\,\xi(\mathbf{s})\{1+o_P(1)\}, \tag{26}$$

where $\mathcal{B}_3(\mathbf{s},z)$ is as defined in theorem 2, and

$$\mathcal{A}_3(\mathbf{s},z)\,\mathcal{A}_3^{\mathrm{T}}(\mathbf{s},z)=\{\sigma_1^2(\mathbf{s})+\sigma_2^2(\mathbf{s})\}\nu_{0,K}\,\mathbf{A}^{-1}(\mathbf{s},z)\,f_Z(\mathbf{s},z)^{-1}.$$

Comparing equations (25) and (26) with theorems 1 and 2 respectively, we note that both expression (19) and expression (23) tend to 0, and the asymptotic variances of the spatial smoothing estimators are therefore much smaller.

(c) In the case of no nugget effect (i.e. $\sigma_2^2(\mathbf{s}_0)=0$ in condition 2, $\mathbf{A}_0(\mathbf{s}_0,z)=\mathbf{A}_2(\mathbf{s}_0,\mathbf{s}_0,z,z)\equiv 0$ and $\mathbf{V}_2(\mathbf{s}_0,\mathbf{s}_0)=0$ in condition 3), spatial smoothing cannot reduce the asymptotic variance of the time series estimators $\hat{\boldsymbol{\alpha}}(\mathbf{s}_0)$, $\hat{a}(\mathbf{s}_0)$ and $\hat{\mathbf{b}}(\mathbf{s}_0)$. See also Lu *et al.* (2008). This is due to the fact that the spatial smoothing uses effectively the data at locations that are within a distance $\tilde{h}$ from $\mathbf{s}_0$. Owing to the continuity of the function $\gamma_1(\cdot,\cdot)$ that is stated in condition 2, all the $\varepsilon_t(\mathbf{s})$s from those locations are asymptotically identical. We argue that asymptotic theory under this setting presents an excessively gloomy picture. Adding a nugget effect in the model brings the theory closer to reality since in practice the data that are used in local spatial smoothing usually contain some noise components which are not identical even within a very small neighbourhood. Note that the nugget effect is not detectable in practice since we can never estimate $\gamma(\mathbf{s}+\Delta,\mathbf{s})$, which is defined in expression (29) in Appendix A for $\|\Delta\|$ less than the minimum pairwise distance between observed locations. See also remark 3 of Zhang *et al.* (2003).

The proofs of theorems 1 and 2 are included in an extended version of this paper that is available from `http://stats.lse.ac.uk/q.yao/qyao.links/paper/spatioVLM.pdf`.

5. A real data example

We illustrate the proposed methodology with a meteorological data set.

5.1. Data

We analyse the daily MSLP readings measured in units of pascals in an area of the North Sea with longitudes from 20° E to 20° W and latitudes from 50° to 60° N. This area is heavily influenced by travelling low pressure systems. The grid points are of size $2.5^\circ\times 2.5^\circ$ with a total number of $17\times 5=85$ spatial locations. The time period is winter 2001–2002 with 100 daily observations at each location starting from December 1st, 2001. The data were provided by the National Centers for Environmental Prediction–National Center for Atmospheric Research. This type of data is commonly used in climate analysis, and more detailed information about the data can be found in Kalnay *et al.* (1996). Trends are not removed, and therefore the original time series of MSLP at each location is generally non-stationary, and we have chosen to work with the differenced series (daily change series). Fig. 3 displays some plots for MSLP data at five randomly selected locations. We denote the daily changes of MSLP as $Y_t(\mathbf{s}_{ij})=Y_t(u_i,v_j)$, $t=1,\ldots,99$, $u_i=60-(i-1)\times 2.5$ with $i=1,\ldots,5$, and $v_j=-17.5+(j-1)\times 2.5$ with $j=1,2,\ldots,17$. From Fig. 3(b), the daily change series of MSLP, $Y_t(\mathbf{s})$, looks approximately stationary at each site.

The contour plots of the daily changes are given in Fig. 4 for the time period from $t=1$ to $t=20$. These plots show the difference from one day to another as a function of spatial co-ordinate. For the high pressure period (approximately from $t=8$ to $t=16$) it is seen that for most spatial locations there are small changes, corresponding to the fact that a high pressure

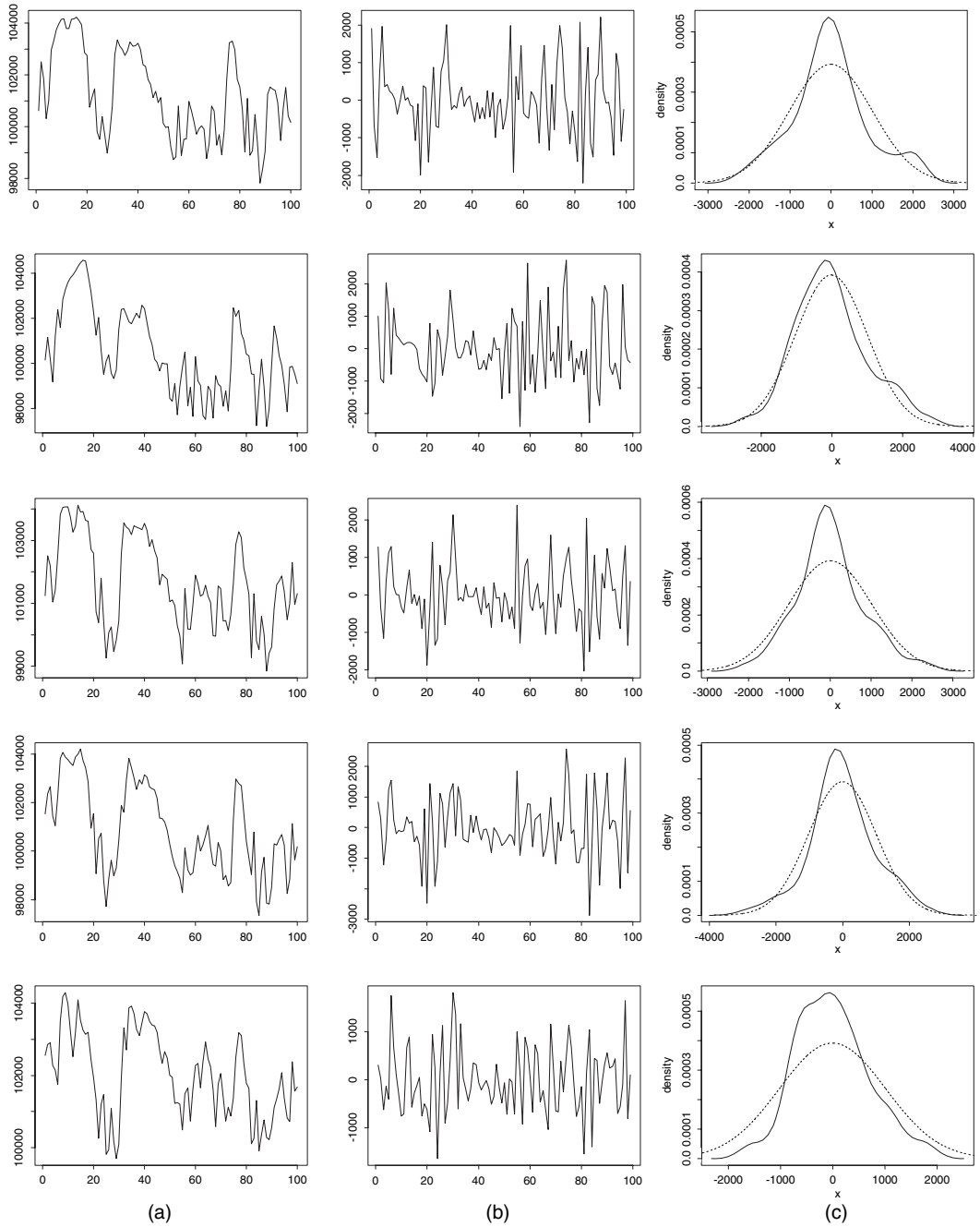

Fig. 3. Modelling of MSLP data: (a) time series plots of the daily MSLP series from December 1st, 2001, to March 10th, 2002; (b) daily changes in MSLP; (c) estimated density function (———) of the daily changes together with the normal density with the same mean and variance (· · · · · · ·)

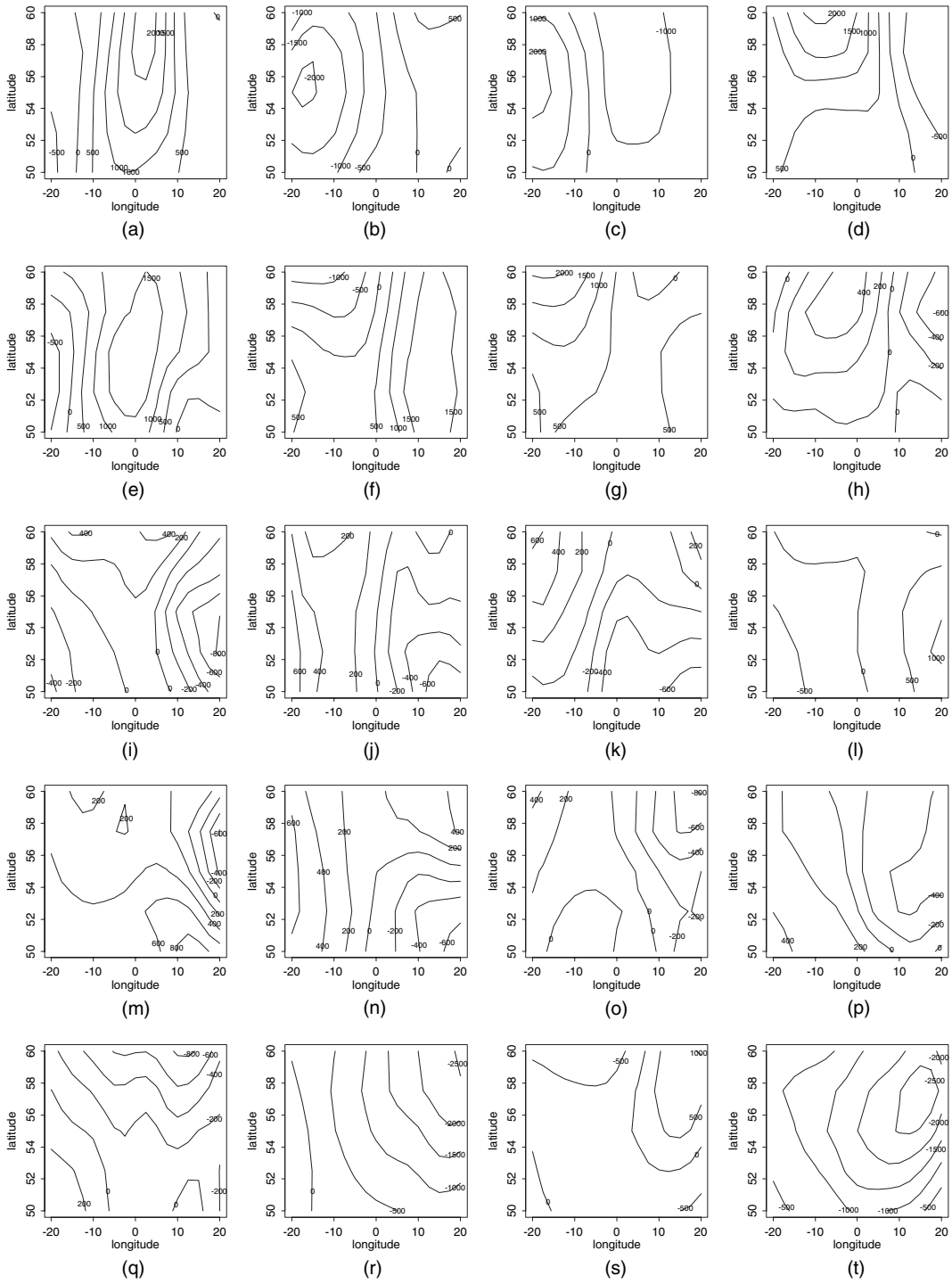

Fig. 4. Modelling of MSLP data—contour plots of the daily MSLP changes $\{Y_t(\mathbf{s})\}$ for the first 20 days t: (a) $t = 1$; (b) $t = 2$; (c) $t = 3$; (d) $t = 4$; (e) $t = 5$; (f) $t = 6$; (g) $t = 7$; (h) $t = 8$; (i) $t = 9$; (j) $t = 10$; (k) $t = 11$; (l) $t = 12$; (m) $t = 13$; (n) $t = 14$; (o) $t = 15$; (p) $t = 16$; (q) $t = 17$; (r) $t = 18$; (s) $t = 19$; (t) $t = 20$

system is fairly stable in time and in space. The zero contour of no change is also seen to be fairly stable running in the north–south direction roughly in the middle of this area. Just before $t = 8$, positive pressure gradients are dominating, corresponding to the build-up of the high pressure system, whereas negative gradients are dominating after $t = 16$, when the high pressure system is deteriorating.

5.2. Varying-coefficient modelling

Now we model the daily changes $Y_t(\mathbf{s})$ by the varying-coefficient model (1) with

$$\mathbf{X}_t(\mathbf{s}_{ij})^{\mathrm{T}} = (Y_{t-1}(\mathbf{s}_{i-1,j}), Y_{t-1}(\mathbf{s}_{i+1,j}), Y_{t-1}(\mathbf{s}_{i,j-1}), Y_{t-1}(\mathbf{s}_{i,j+1}), Y_{t-1}(\mathbf{s}_{ij})), \tag{27}$$

i.e. $\mathbf{X}(\mathbf{s}_{ij})$ consists of the lagged values at the four nearest neighbours of the location $\mathbf{s}_{ij}$, and the lagged value at $\mathbf{s}_{ij}$ itself. We use only time lag 1 since the auto-correlation of the daily change data is weak. It is clear that this specification does not require that the data are recorded regularly over the space. With $\mathbf{X}(\mathbf{s}_{ij})$ specified above, model (1) is now of the form

$$\begin{aligned} Y_t(\mathbf{s}_{ij}) = {} & a\{\mathbf{s}_{ij}, Z_t(\mathbf{s}_{ij})\} + b_1\{\mathbf{s}_{ij}, Z_t(\mathbf{s}_{ij})\} Y_{t-1}(\mathbf{s}_{i-1,j}) + b_2\{\mathbf{s}_{ij}, Z_t(\mathbf{s}_{ij})\}\, Y_{t-1}(\mathbf{s}_{i+1,j}) \\ & + b_3\{\mathbf{s}_{ij}, Z_t(\mathbf{s}_{ij})\}\, Y_{t-1}(\mathbf{s}_{i,j-1}) + b_4\{\mathbf{s}_{ij}, Z_t(\mathbf{s}_{ij})\}\, Y_{t-1}(\mathbf{s}_{i,j+1}) + b_5\{\mathbf{s}_{ij}, Z_t(\mathbf{s}_{ij})\} \\ & \times Y_{t-1}(\mathbf{s}_{i,j}) + \varepsilon_t(\mathbf{s}_{ij}) \end{aligned} \tag{28}$$

with $Z_t(\mathbf{s}_{ij}) = \boldsymbol{\alpha}(\mathbf{s}_{ij})^{\mathrm{T}}\, \mathbf{X}_t(\mathbf{s}_{ij})$, $\boldsymbol{\alpha}(\mathbf{s}_{ij})^{\mathrm{T}} = (\alpha_1(\mathbf{s}_{ij}), \alpha_2(\mathbf{s}_{ij}), \alpha_3(\mathbf{s}_{ij}), \alpha_4(\mathbf{s}_{ij}), \alpha_5(\mathbf{s}_{ij}))$ and $\mathbf{b}^{\mathrm{T}} = (b_1, b_2, b_3, b_4, b_5)$. The estimation was based on data from 45 locations: $Y_t(\mathbf{s}_{ij}) = Y_t(u_i, v_j)$, $t = 1, \ldots, 99$, $i = 2, 3, 4$, and $j = 2, \ldots, 16$, owing to the boundary effect in space. The bandwidths were selected by using the methods that were specified in Section 2.3, in which we let $q = 10$ and take $\bar{h}_k = C_T \bar{h}_{ck}$ with $C_T = \mathrm{std}(Z_t)(99)^{-1/5}$ and $\mathrm{std}(Z_t)$ being the sample standard deviation of $\{Z_t = \hat{\boldsymbol{\alpha}}^{\mathrm{T}}\, \mathbf{X}_t(\mathbf{s})\}_{t=1}^{99}$ at location $\mathbf{s}$, and $\bar{h}_{ck} = 0.1k$ for $k = 1, 2, \ldots, q$. The estimated $a(z)$, $b_1(z)$, $b_2(z)$, $b_3(z)$, $b_4(z)$ and $b_5(z)$ at 45 locations, together with their averages (over the 45 locations), are plotted in Fig. 5 without spatial smoothing, and Fig. 6 with the spatial smoothing.

If the meteorological data could be described linearly, the functions that are displayed in Figs 5 and 6 would be constant. But clearly they are not. If these plots are continued for higher and lower values of z, the asymptote turns out to be roughly horizontal. This means that in a low pressure situation the system behaves roughly as a linear system, but not in the high pressure zones ($z \approx 0$) and the transition zones. This indicates that altogether the weather system is non-linear. In another context it would be of interest to try to give a detailed meteorological interpretation of the curves. But, since this is a general paper, we satisfy ourselves by noting the following.

(a) $z \approx 0$: this corresponds to a high pressure situation primarily. The curves for b_i, $i = 1, \ldots, 5$, have their maximum or minimum at zero, the negative and positive peak values roughly balancing out, and, given that the $\mathbf{X}_t$-variables stay around 0 for a high pressure situation, the effect is that Y_t stays around 0, and the high pressure continues.

(b) $z < 0$: this is the situation when a high pressure area deteriorates, or when new low pressure areas are coming in. It is seen that there are only small contributions from the north (b_1), east (b_4) and the site itself (b_5), but a main positive contribution from west (b_3) having the effect of maintaining a negative difference in X_t. It is tempting to interpret this as having to do with low pressure systems coming primarily from the west in the North Sea.

The contour plots (which are not shown) of the estimated index $\boldsymbol{\alpha}^{\mathrm{T}}\, \mathbf{X}(\mathbf{s})$ resemble the patterns of Fig. 4, indicating that it catches the major spatial variations of $Y(\mathbf{s})$. Furthermore

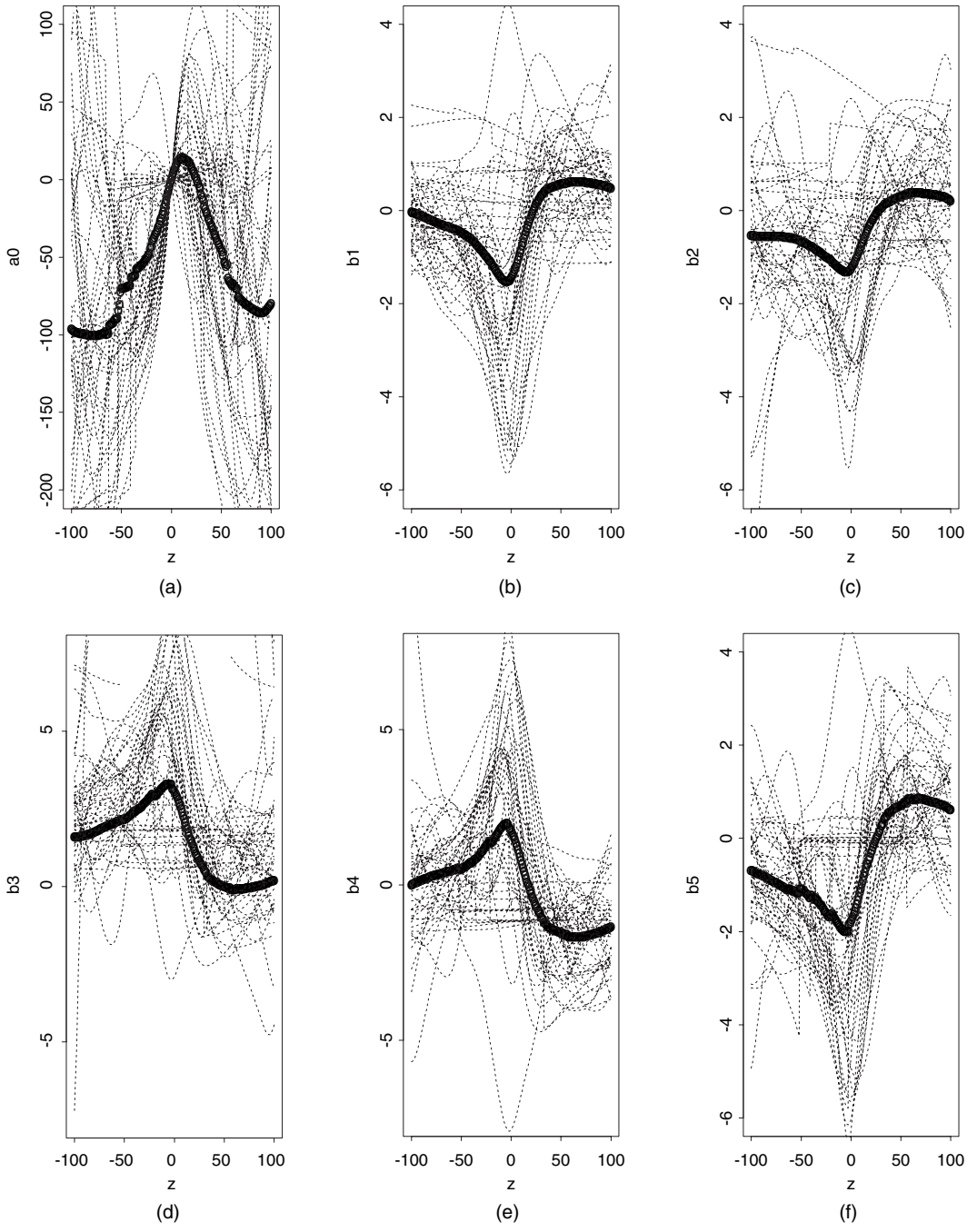

Fig. 5. Modelling of MSLP data—estimated curves (·····) of 45 locations together with the averaged curve (———) for (a) $a(\cdot)$, (b) $b_1(\cdot)$, (c) $b_2(\cdot)$, (d) $b_3(\cdot)$, (e) $b_4(\cdot)$ and (f) $b_5(\cdot)$: no spatial smoothing applied in estimation

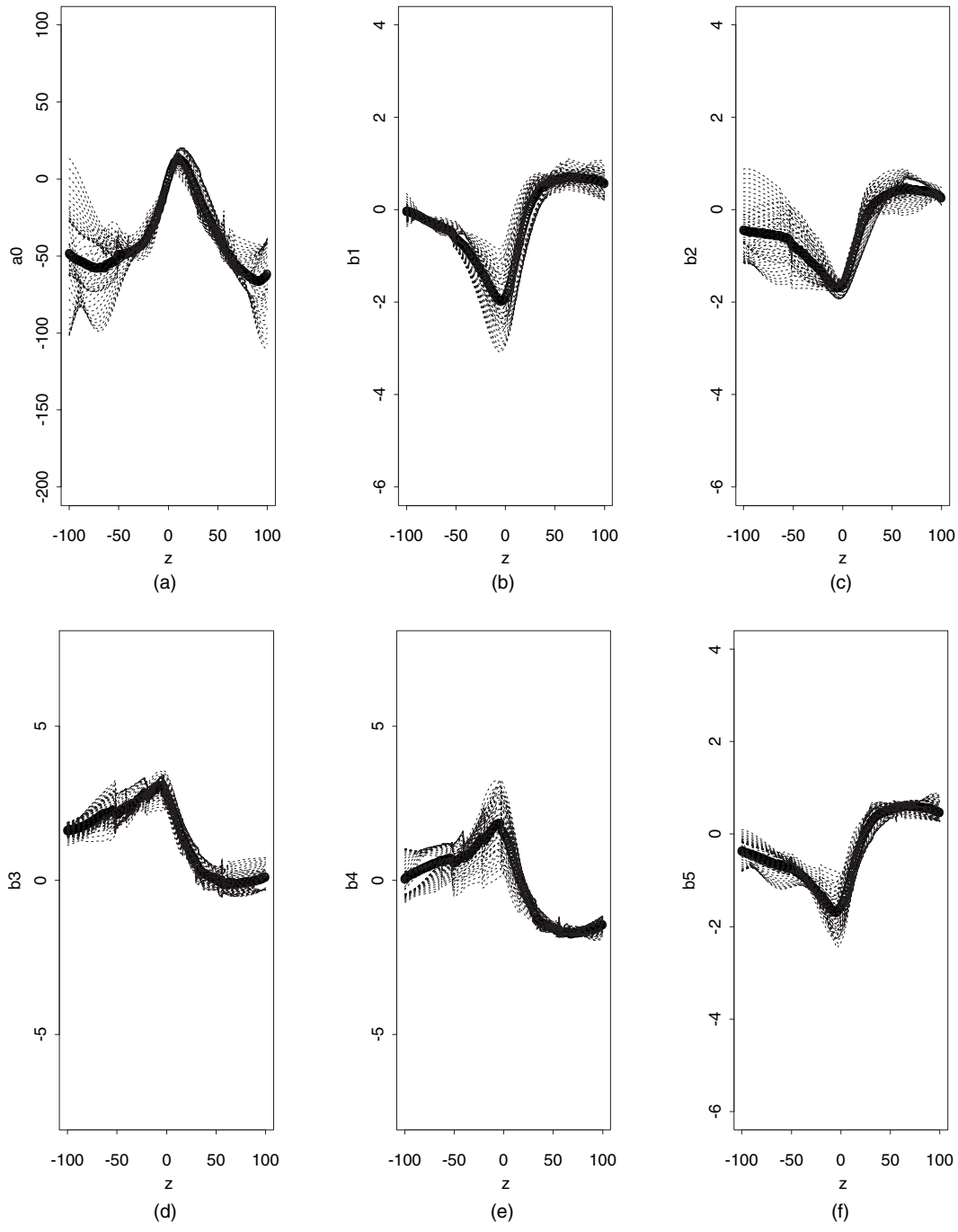

Fig. 6. Modelling of MSLP data—spatial-smoothed estimated curves (· · · · · · ·) of 45 locations together with the averaged curve (———) for (a) $a(\cdot)$, (b) $b_1(\cdot)$, (c) $b_2(\cdot)$, (d) $b_3(\cdot)$, (e) $b_4(\cdot)$ and (f) $b_5(\cdot)$

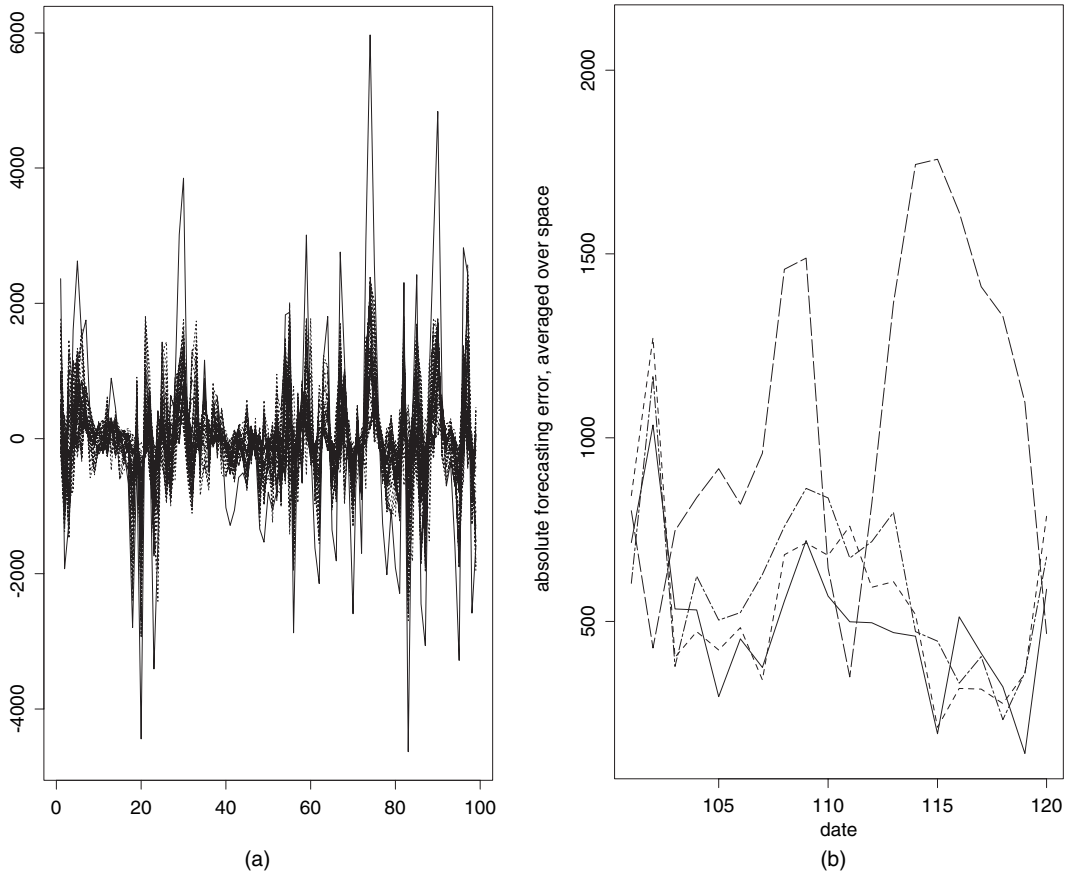

Fig. 7. MSLP data: (a) time series plots of the 45 estimated index series (⋮) and the global spatial index series (i.e. the first principal component series (|), scaled down by 3) and (b) absolute one-step-ahead forecasting errors, averaged over 45 locations, over the period of March 11th–30th, 2002 (———, adaptive forecast; -------, linear forecast; — —, mean forecast; – · – ·, last day forecast)

there were little variations in the index vector $\boldsymbol{\alpha}(\mathbf{s})$ over the space. The average value is $\bar{\boldsymbol{\alpha}}^{\mathrm{T}} = (-0.367, -0.252, 0.174, 0.219, 0.851)$. Note that the last component of the index vector is substantially greater than all the others, indicating the dominating role of the lagged value from the same site; see expression (27). We plot the 45 estimated index time series $Z_t(\mathbf{s})$ in Fig. 7(a), together with the first principal component series $\gamma(t) = \boldsymbol{\lambda}_1^{\mathrm{T}} \mathbf{Y}_t$ where $\boldsymbol{\lambda}_1$ is the first principal component of the data (explaining about 46% of the total variation). Note that the oscillatory patterns of $\gamma(t)$ and the estimated index series $Z_t(\mathbf{s})$ are similar. The high pressure areas can again be identified as regions where the $Z_t(\mathbf{s})$ series are close to zero, building up a high pressure area corresponds to positive $Z_t(\mathbf{s})$ (but not conversely), and deterioration to negative $Z_t(\mathbf{s})$ (but not conversely). There does not seem to be a corresponding regularity for the $\gamma(t)$ series, which is sometimes taken as a representative index for the north Atlantic oscillation on another timescale.

5.3. Forecasting comparisons

To assess further the capability of model (28), we compared its post-sample forecasting performance with those of other models including the linear model. We used the adaptive

Table 1. Mean absolute predictive errors MAPE of the four different models

Model	AVCM	Linear	Mean	Last day
MAPE	494.1	553.0	1052.8	599.9

Table 2. Mean absolute predictive errors MAPE of the AVCM with various bandwidths

(a) $\bar{h}_c = \bar{h}_{0c}$; $\tilde{h} = \tilde{h}_0$				
h	$h_0 - 4$	$h_0 - 1$	$h_0 + 1$	$h_0 + 4$
MAPE	499.9	496.8	498.9	508.7
(b) $h = h_0$; $\tilde{h} = \tilde{h}_0$				
$\bar{h}_c$	$\bar{h}_{0c} - 0.4$	$\bar{h}_{0c} - 0.1$	$\bar{h}_{0c} + 0.1$	$\bar{h}_{0c} + 0.4$
MAPE	523.3	494.1	496.2	495.0
(c) $h = h_0$; $\bar{h}_c = \bar{h}_{0c}$				
$\tilde{h}$	$\tilde{h}_0 - 0.4$	$\tilde{h}_0 - 0.1$	$\tilde{h}_0 + 0.1$	$\tilde{h}_0 + 0.4$
MAPE	505.2	495.3	493.9	493.3

varying-coefficient model (AVCM) fitted from the above for one-step ahead prediction for the 20 MSLP values in the period of March 11th–30th, 2002, over the 45 locations. Also included in the comparison are the linear model which may be viewed as a special case of the AVCM with constant coefficient functions, the mean forecast, which forecasts the future on the basis of the mean of the past values from December 1st, 2001, onwards, and the last day forecast, which forecasts tomorrow's value by today's value. The absolute forecasting errors, averaged over the 45 space locations, are plotted in Fig. 7(b). The mean absolute predictive errors MAPE, over the 45 space locations and the 20 days of March 11th–30th, 2002, are listed in Table 1. Overall our AVCM outperforms the other three simple models. This lends further support to the assertion that the dynamic structure of the MSLP is non-linear and it cannot be accommodated, for example, in a linear dynamic model.

In the above comparison, the bandwidths that were used in fitting the AVCM were selected by using the methods specified in Section 2.3. Over the 45 locations, the selected values for h, $\bar{h}_c$ in $\bar{h} = C_T \bar{h}_c$ and $\tilde{h}$ are around respectively 15, 0.6 and 0.7. To examine the sensitivity of the AVCM forecasting on the bandwidths, we repeat the above exercises with the bandwidths shifted from the selected values. We denote by $h_0, \bar{h}_{0c}$ and $\tilde{h}_0$ the selected values of h, $\bar{h}_c$ and $\tilde{h}$ by using the methods that were specified in Section 2.3. Table 2 lists the MAPE of the AVCM with various bandwidths. Comparing with Table 1, the MAPE of the AVCM varies with the bandwidths used in the estimation, but the variation is not excessive. Further the AVCM always offers better prediction than the three other models.

Acknowledgements

The authors thank the Joint Editor, an Associate Editor and two referees for very helpful comments and suggestions.

This work was partially supported by a Leverhulme Trust research grant and Engineering and Physical Sciences Research Council research grant. Lu's work was also partially supported by

an internal research grant from Curtin University of Technology and a 'Discovery' grant from the Australian Research Council.

Appendix A: Regularity conditions

We list below the regularity conditions, among which conditions 4, parts (a) and (b), 5–7, 9 and 10 are basically inherited from Lu *et al.* (2007b) in the time series context.

Condition 1. For each $\mathbf{s}$, $\{\varepsilon_t(\mathbf{s}), t \geqslant 1\}$ is a sequence of independent and identically distributed random variables. The distribution of $\varepsilon_t(\mathbf{s})$ is the same for all $\mathbf{s} \in \mathcal{S}$. Further, for each $t > 1$, $\{\varepsilon_t(\mathbf{s}), \mathbf{s} \in \mathcal{S}\}$ is independent of $\{(Y_{t-j}(\mathbf{s}), \mathbf{X}_{t+1-j}(\mathbf{s})), \mathbf{s} \in \mathcal{S} \text{ and } j \geqslant 1\}$. The spatial covariance function

$$\gamma(\mathbf{s}_1, \mathbf{s}_2) \equiv \operatorname{cov}\{\varepsilon_t(\mathbf{s}_1), \varepsilon_t(\mathbf{s}_2)\} \tag{29}$$

is bounded over $\mathcal{S}^2$.

Condition 2. For any $t \geqslant 1$ and $\mathbf{s} \in \mathcal{S}$,

$$\varepsilon_t(\mathbf{s}) = \varepsilon_{1,t}(\mathbf{s}) + \varepsilon_{2,t}(\mathbf{s}) \tag{30}$$

where $\{\varepsilon_{1,t}(\mathbf{s}), t \geqslant 1, \mathbf{s} \in \mathcal{S}\}$ and $\{\varepsilon_{2,t}(\mathbf{s}), t \geqslant 1, \mathbf{s} \in \mathcal{S}\}$ are two independent processes, and both fulfil the conditions that are imposed on $\varepsilon_t(\mathbf{s})$ in condition 1 above. Further, $\gamma_1(\mathbf{s}_1, \mathbf{s}_2) \equiv \operatorname{cov}\{\varepsilon_{1,t}(\mathbf{s}_1), \varepsilon_{1,t}(\mathbf{s}_2)\}$ is continuous in $(\mathbf{s}_1, \mathbf{s}_2)$ (we shall denote $\sigma_1^2(\mathbf{s}_1) = \gamma_1(\mathbf{s}_1, \mathbf{s}_1)$), and $\gamma_2(\mathbf{s}_1, \mathbf{s}_2) \equiv \operatorname{cov}\{\varepsilon_{2,t}(\mathbf{s}_1), \varepsilon_{2,t}(\mathbf{s}_2)\} = 0$ if $\mathbf{s}_1 \neq \mathbf{s}_2$, and $\sigma_2^2(\mathbf{s}) = \gamma_2(\mathbf{s}, \mathbf{s}) > 0$ is continuous.

Condition 3.

(a) $\mathbf{A}(\mathbf{s}, \mathbf{s}'; z_1, z_2) \equiv E\{\mathbb{X}_t(\mathbf{s})\, \mathbb{X}_t(\mathbf{s}')^{\mathrm{T}} | Z_t(\mathbf{s}) = z_1, Z_t(\mathbf{s}') = z_2\} = \mathbf{A}_1(\mathbf{s}, \mathbf{s}'; z_1, z_2) + \mathbf{A}_2(\mathbf{s}, \mathbf{s}'; z_1, z_2)$, where $\mathbf{A}_1(\mathbf{s}, \mathbf{s}'; z_1, z_2)$ is continuous, and $\mathbf{A}_2(\mathbf{s}, \mathbf{s}; z_1, z_2)$ is a positive definite matrix. Further, $\mathbf{A}_2(\mathbf{s}, \mathbf{s}'; z_1, z_2) = 0$ if $\mathbf{s} \neq \mathbf{s}'$, and $\mathbf{A}_0(\mathbf{s}, z) \equiv \mathbf{A}_2(\mathbf{s}, \mathbf{s}; z, z)$ is continuous.

(b) Set $\mathbb{V}_t(\mathbf{s}) = \mathbb{W}_t(\mathbf{s}) - \mathbf{B}_t(\mathbf{s})\, \mathbb{X}_t(\mathbf{s})$, where $\mathbf{B}_t(\mathbf{s}) = E\{\mathbb{W}_t(\mathbf{s})\, \mathbb{X}_t(\mathbf{s})^{\mathrm{T}} | Z_t(\mathbf{s})\}\, E\{\mathbb{X}_t(\mathbf{s})\, \mathbb{X}_t(\mathbf{s})^{\mathrm{T}} | Z_t(\mathbf{s})\}^{-1}$ and $\mathbb{W}_t(\mathbf{s}) = \mathbf{X}_t(\mathbf{s})\, \dot{\mathbf{g}}_z\{\mathbf{s}, Z_t(\mathbf{s})\}^{\mathrm{T}}\, \mathbb{X}_t(\mathbf{s})$. Then $\mathbf{V}(\mathbf{s}_1, \mathbf{s}_2) \equiv E[\mathbb{V}_t(\mathbf{s}_1) \mathbb{V}_t(\mathbf{s}_2)^{\mathrm{T}} w\{Z_t(\mathbf{s}_1)\}\, w\{Z_t(\mathbf{s}_2)\}] = \mathbf{V}_1(\mathbf{s}_1, \mathbf{s}_2) + \mathbf{V}_2(\mathbf{s}_1, \mathbf{s}_2)$, where $\mathbf{V}_1(\mathbf{s}_1, \mathbf{s}_2)$ is continuous, $\mathbf{V}_2(\mathbf{s}_1, \mathbf{s}_2) = 0$ if $\mathbf{s}_1 \neq \mathbf{s}_2$ and $\mathbf{V}_2(\mathbf{s}, \mathbf{s})$ is positive definite and continuous.

Condition 4.

(a) $\mathbf{g}(\mathbf{s}, z, \boldsymbol{\beta})$ is twice continuously differentiable with respect to $\mathbf{s}$, and $\boldsymbol{\alpha}(\mathbf{s})$ is twice continuously differentiable. Also in the expression

$$R\{a(\cdot), \mathbf{b}(\cdot), \boldsymbol{\beta}\} = E[\{Y_t - a(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_t) - \mathbf{b}(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_t)^{\mathrm{T}} \check{\mathbf{X}}_t\}^2\, w(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_t)],$$

differentiation with respect to $\boldsymbol{\beta}$ and the expectation are exchangeable.

(b) The density function $f_{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_t(\mathbf{s})}(\mathbf{s}, z)$ of $\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_t(\mathbf{s})$ is continuous. For any fixed $\mathbf{s} \in \mathcal{S}$, it is uniformly bounded away from 0 for $z \in [-L, L]$ and $\boldsymbol{\beta} \in \mathbb{B}$, where $L > 0$ is a constant. Furthermore, the joint probability density function of $(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{t_0}(\mathbf{s}), \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{t_1}(\mathbf{s}), \ldots, \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_{t_k}(\mathbf{s}))$ exists and is bounded uniformly for any $t_0 < t_1 < \ldots < t_k$ and $0 \leqslant k \leqslant 2(r-1)$ and $\boldsymbol{\beta} \in \mathbb{B}$, where $r > 3d$ is a positive integer.

(c) Denote by $f_{Z,Z}(\mathbf{s}_1, \mathbf{s}_2; z_1, z_2)$ the joint density function of $Z_t(\mathbf{s}_1)$ and $Z_t(\mathbf{s}_2)$. Then $f_{Z,Z}(\mathbf{s}_1, \mathbf{s}_2; z, z)$ is continuous in z, and it converges to $f^*(\mathbf{s}, z)$ as $\|\mathbf{s}_i - \mathbf{s}\| \to 0$ for $i = 1, 2$.

Condition 5. $E|Y_t(\mathbf{s})|^{\rho r} < \infty$ and $E\|\mathbf{X}_t(\mathbf{s})\|^{\rho r} < \infty$. Furthermore, it holds for some $\rho > 6$ and r given in condition 4 that $\sup_{\boldsymbol{\beta} \in \mathbb{B}}[E\{|Y_t(\mathbf{s}) - \mathbf{g}(\mathbf{s}, \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_t(\mathbf{s}), \boldsymbol{\beta})^{\mathrm{T}} \mathbb{X}_t|^{\rho r}\}] < \infty$.

Condition 6. The matrix $E(\mathbb{X}_t \mathbb{X}_t^{\mathrm{T}} | \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}_t = z)$ is positive definite for $z \in [-L, L]$ and $\boldsymbol{\beta} \in \mathbb{B}$.

Condition 7. For each fixed $\mathbf{s} \in \mathcal{S}$, the time series $\{(Y_t(\mathbf{s}), \mathbf{X}_t(\mathbf{s})), t \geqslant 1\}$ is strictly stationary and β mixing with the mixing coefficients satisfying the condition $\beta(t) = O(t^{-b})$ for some $b > \max\{2(\rho r + 1)/(\rho r - 2), (r + a)/(1 - 2/\rho)\}$, where r and ρ are specified in conditions 4 and 5, and $a \geqslant (r\rho - 2)r/(2 + r\rho - 4r)$.

Condition 8. There is a continuous sampling intensity function (i.e. density function) f defined on $\mathcal{S}$ for which $N^{-1} \sum_{i=1}^{N} I(\mathbf{s}_i \in A) \to \int_A f(\mathbf{s})\, \mathrm{d}\mathbf{s}$ for any measurable set $A \subset \mathcal{S}$.

Condition 9. $W(\cdot)$ is a symmetric density function on $\mathbb{R}^2$ with bounded support. $K(\cdot)$ is a bounded and symmetric density function on $\mathbb{R}^1$ with bounded support S_K. Furthermore, $|K(x) - K(y)| \leqslant C|x - y|$ for $x, y \in S_K$ and some $0 < C < \infty$.

Condition 10. As $N \to \infty$, $\tilde{h} \to 0$ and $N\tilde{h}^2 \to \infty$. As $T \to \infty$ $Th^4 = O(1)$, $Th^{3+3d/r} \to \infty$ and $\liminf_{T\to\infty}(Th^{\{2(r-1)a+(\rho r-2)\}/(a+1)\rho}) > 0$ for r given in condition 4, ρ given in condition 5 and a and b given in condition 7. Furthermore, there is a sequence of positive integers $m_T \to \infty$ such that $m_T = o\{(Th)^{1/2}\}$, $Tm_T^{-b} \to 0$ and $2m_T h^{(\rho r-2)/\{2+b(\rho r-2)\}} > 1$.

Conditions 1 and 2 were introduced in Zhang *et al.* (2003). Condition 2 manifests the so-called nugget effect in the noise process when $\sigma_2^2(\mathbf{s}_1) > 0$. The concept of the nugget effect was introduced by G. Matheron in the early 1960s. It implies that the variogram $E\{\varepsilon_t(\mathbf{s}_1) - \varepsilon_t(\mathbf{s}_2)\}^2$ does not converge to 0 as $\|\mathbf{s}_1 - \mathbf{s}_2\| \to 0$ or, equivalently, the function $\tilde{\gamma}(\mathbf{s}) \equiv \gamma(\mathbf{s}_1 + \mathbf{s}, \mathbf{s}_1)$ is not continuous at $\mathbf{s} = 0$ for any given $\mathbf{s}_1 \in \mathcal{S}$. Note that $\varepsilon_{1,t}(\mathbf{s})$ in equation (30) represents so-called system noise which typically has continuous sample realization (in $\mathbf{s}$), whereas $\varepsilon_{2,t}(\mathbf{s})$ stands for microscale variation and/or measurement noise; see Cressie (1993), section 2.3.1. Condition 3 represents the possible nugget effect in the regressor process $\mathbf{X}_t(\mathbf{s})$. Note that, when $\mathbf{X}_t(\mathbf{s})$ contains some lagged values of $Y_t(\mathbf{s})$, condition 3 may be implied by condition 2.

The function $f^*(\mathbf{s}, \cdot)$ in condition 4, part (c), may not be the density function of $Z_t(\mathbf{s})$ owing to the nugget effect; see the discussion below expression (3.1) in Lu *et al.* (2008). Other smooth conditions in condition 4 and the moment conditions in condition 5 are standard, though some of them may be reduced further at the cost of an increase in the already lengthy technical arguments. Condition 6 is not essential and is introduced for technical convenience.

Condition 7 imposes the β-mixing condition on the time series. The complex restriction on the mixing coefficients is required to ensure that the time series estimators $\hat{\alpha}(\mathbf{s})$, $\hat{a}(\mathbf{s})$ and $\hat{\mathbf{b}}(\mathbf{s})$ converge uniformly in $\mathbf{s}$. Further discussion on mixing time series may be found in, for example, Section 2.6 of Fan and Yao (2003).

Condition 8 specifies that the asymptotic approximations in the spatial domain are derived from the fixed domain or infill asymptotics (Cressie (1993), section 3.3).

The conditions on W and K in condition 9 and $\tilde{h}$ in condition 10 are standard for non-parametric regression. More sophisticated conditions on the bandwidth h in condition 10 are required to ensure the uniform convergence of the time series estimators.

References

Anselin, L. (1988) *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer.

Biau, G. and Cadre, B. (2004) Nonparametric spatial prediction. *Statist. Inf. Stochast. Process.*, **7**, 327–349.

Brunsdon, C., McClatchey, J. and Unwin, D. J. (2001) Spatial variations in the average rainfall-altitude relationship in Great Britain: an approach using geographically weighted regression. *Int. J. Clim.*, **21**, 455–466.

Chilès J.-P. and Delfiner, P. (1999) *Geostatistics: Modeling Spatial Uncertainty*. New York: Wiley.

Cressie, N. A. C. (1993) *Statistics for Spatial Data*. New York: Wiley.

Diggle, P. J. (2003) *Statistical Analysis of Spatial Point Patterns*. London: Arnold.

Fan, J. and Yao, Q. (2003) *Nonlinear Time Series: Nonparametric and Parametric Methods*. New York: Springer.

Fan, J., Yao, Q. and Cai, Z. (2003) Adaptive varying-coefficient linear models. *J. R. Statist. Soc.* B, **65**, 57–80.

Fotheringham, A. S., Charlton, M. E. and Brunsdon, C. (1998) Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. *Environ. Planng* A, **30**, 1905–1927.

Fuentes, M. (2001) A high frequency kriging approach for non-stationary environmental processes. *Environmetrics*, **12**, 469–483.

Gao, J., Lu, Z. and Tjøstheim, D. (2006) Estimation in semiparametric spatial regression. *Ann. Statist.*, **34**, 1395–1435.

Guyon, X. (1995) *Random Fields on a Network: Modelling, Statistics, and Application*. New York: Springer.

Hall, P. and Johnstone, I. (1992) Empirical functionals and efficient smoothing parameter selection (with discussion). *J. R. Statist. Soc.* B, **54**, 475–530.

Hallin, M., Lu, Z. and Tran, L. T. (2004a) Density estimation for spatial processes: the L_1 theory. *J. Multiv. Anal.*, **88**, 61–75.

Hallin, M., Lu, Z. and Tran, L. T. (2004b) Local linear spatial regression. *Ann. Statist.*, **32**, 2469–2500.

Härdle, W. and Vieu, P. (1992) Kernel regression smoothing of time series. *J. Time Ser. Anal.*, **13**, 209–224.

Jones, M. C. (1991) The roles of ISE and MISE in density estimation. *Statist. Probab. Lett.*, **12**, 51–56.

Kalnay, E., Kanamitsu, M., Kitsler, R., Collins, W., Deaven, D., Candin, L., Iredell, M., Saha, S., White, G., Wollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R. and Joseph, D. (1996) The NCEP/NCAR 40-year Reanalysis Project. *Bull. Am. Meteorol. Soc.*, **77**, 437–471.

Kim, T. Y. and Cox, D. D. (1995) Asymptotic behaviors of some measures of accuracy in non-parametric curve estimation with dependent observations. *J. Multiv. Anal.*, **53**, 67–93.

Knorr-Held, L. (2000) Bayesian modelling of inseparable space time variation in disease risk. *Statist. Med.*, **19**, 2555–2567.

Lagazio, C., Dreassi, E. and Biggeri, A. (2001) A hierarchical Bayesian model for space time variation of disease risk. *Statist. Modllng*, **1**, 17–19.
Loader, C. R. (1999) Bandwidth selection: classical or plug-in? *Ann. Statist.*, **27**, 415–438.
Lu, Z., Lundervold, A., Tjøstheim, D. and Yao, Q. (2007) Exploring spatial nonlinearity using additive approximation. *Bernoulli*, **13**, 447–472.
Lu, Z., Tjøstheim, D. and Yao, Q. (2007) Adaptive varying-coefficient linear models for stochastic processes: asymptotic theory. *Statist. Sin.*, **17**, 177–197.
Lu, Z., Tjøstheim, D. and Yao, Q. (2008) Spatial smoothing, Nugget effect and infill asymptotics. *Statist. Probab. Lett.*, **78**, 3145–3151.
Mammenn, E. (1990) A short note on optimal bandwidth selection for kernel estimators. *Statist. Probab. Lett.*, **9**, 23–25.
Matheron, G. (1976) A simple substitute for conditional expectation: the disjunctive kriging. In *Proc. Advanced Geostatistics in the Mining Industry*, *Rome, Oct. 1975*, pp. 221–236. Dordrecht: Reidel.
Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (1992) *Numerical Recipes in C++: the Art of Scientific Computing*. Cambridge: Cambridge University Press.
Quintela-del-Río, A. (1996) Comparison of bandwidth selectors in nonparametric regression under dependence. *Comput. Statist. Data Anal.*, **21**, 563–580.
Ripley, B. D. (1981) *Spatial Statistics*. New York: Wiley.
Rivoirard, J. (1994) *Introduction to Disjunctive Kriging and Non-linear Geostatistics*. Oxford: Clarendon.
Ruppert, D. (1997) Empirical-bias bandwidths for local polynomial regression and density estimation. *J. Am. Statist. Ass.*, **92**, 1049–1062.
Ruppert, D., Sheather, S. J. and Wand, M. P. (1995) An effective bandwidth selector for local least squares regression. *J. Am. Statist. Ass.*, **90**, 1257–1270.
Stein, M. L. (1999) *Interpolation of Spatial Data*. New York: Springer.
Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions (with discussion). *J. R. Statist. Soc.* B, **36**, 111–147; correction, **38** (1976), 102.
Xia, Y. and Li, W. K. (1999) On single-index coefficient regression models. *J. Am. Stastist. Ass.*, **94**, 1275–1285.
Xia, Y. and Li, W. K. (2002) Asymptotic behavior of bandwidth selected by the cross-validation method for local polynomial fitting. *J. Multiv. Anal.*, **83**, 265–287.
Yao, Q. and Brockwell, P. J. (2006) Gaussian maximum likelihood estimation for ARMA models: II, spatial processes. *Bernoulli*, **12**, 403–429.
Zhang, W., Yao, Q., Tong, H. and Stenseth, N. C. (2003) Smoothing for spatial-temporal models and its application in modelling muskrat-mink interaction. *Biometrics*, **59**, 813–821.