ENVIRONMETRICS
Environmetrics, **9**, 283–301 (1998)

SPATIAL MODELLING OF EXTREME SEA-LEVELS

MARK J. DIXON,[1]* JONATHAN A. TAWN[2] AND JOHN M. VASSIE[3]

[1]*Department of Statistics, University of Newcastle upon Tyne, Newcastle NE1 7RU, UK*
[2]*Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, UK*
[3]*The Proudman Oceanographic Laboratory, Bidston Observatory, Birkenhead, UK*

SUMMARY

The problem of estimating the probability of extreme sea-levels along a coastline has received little attention. Most of the existing analyses are univariate approaches that are applied independently to data from individual sites. We present a spatial extension of the methods that integrates information from the data sites and exploits knowledge of the spatial variation of the tidal and surge constituents of the sea-level along a coastline, to produce estimates at any coastal location. We illustrate the method by application to the UK east coast providing a set of design level estimates along the entire coastline. © 1998 John Wiley & Sons, Ltd.

KEY WORDS coastal flooding; extreme sea-levels; extreme value theory; joint probabilities method; kernel regression smoothing; spatial estimation

1. INTRODUCTION

Given data from a network of sites on the coast, our aim is to estimate extreme sea-level probabilities at all points along a coastline for use in designing coastal flood defence schemes. Treating the coastline as one-dimensional, we define G_d to be the distribution function of the annual maximum sea-level at a point on the coast at a coastal distance d from an origin. If the probability of flooding in a year is to be limited to a specified level p, then the flood defences need to be designed to the $[-\log(1-p)]^{-1}(\approx 1/p$ for small $p)$ year return level $z_p(d)$, which is defined by

$$z_p(d) = G_d^{-1}(1-p),$$

where G_d^{-1} is the inverse distribution function. The design question is: how do we estimate $z_p(d)$, for small p, at any given distance d, using data from a network of m sites at distances $d_1, \ldots, d_m$? For the UK east coast, the data are various spans of hourly sea-levels from 14 sites shown in Figure 1, and d is taken to be coastal distance (in km) from Wick. Before considering possible

* Correspondence to: M. J. Dixon, Department of Statistics, University of Newcastle upon Tyne, Newcastle NE1 7RU, UK.

Contract grant sponsor: Ministry of Agriculture, Fisheries and Food.
Contract grant sponsor: Natural Environment Research Council.

CCC 1180–4009/98/030283–19$17·50

Received 21 May 1997
Revised 5 December 1997

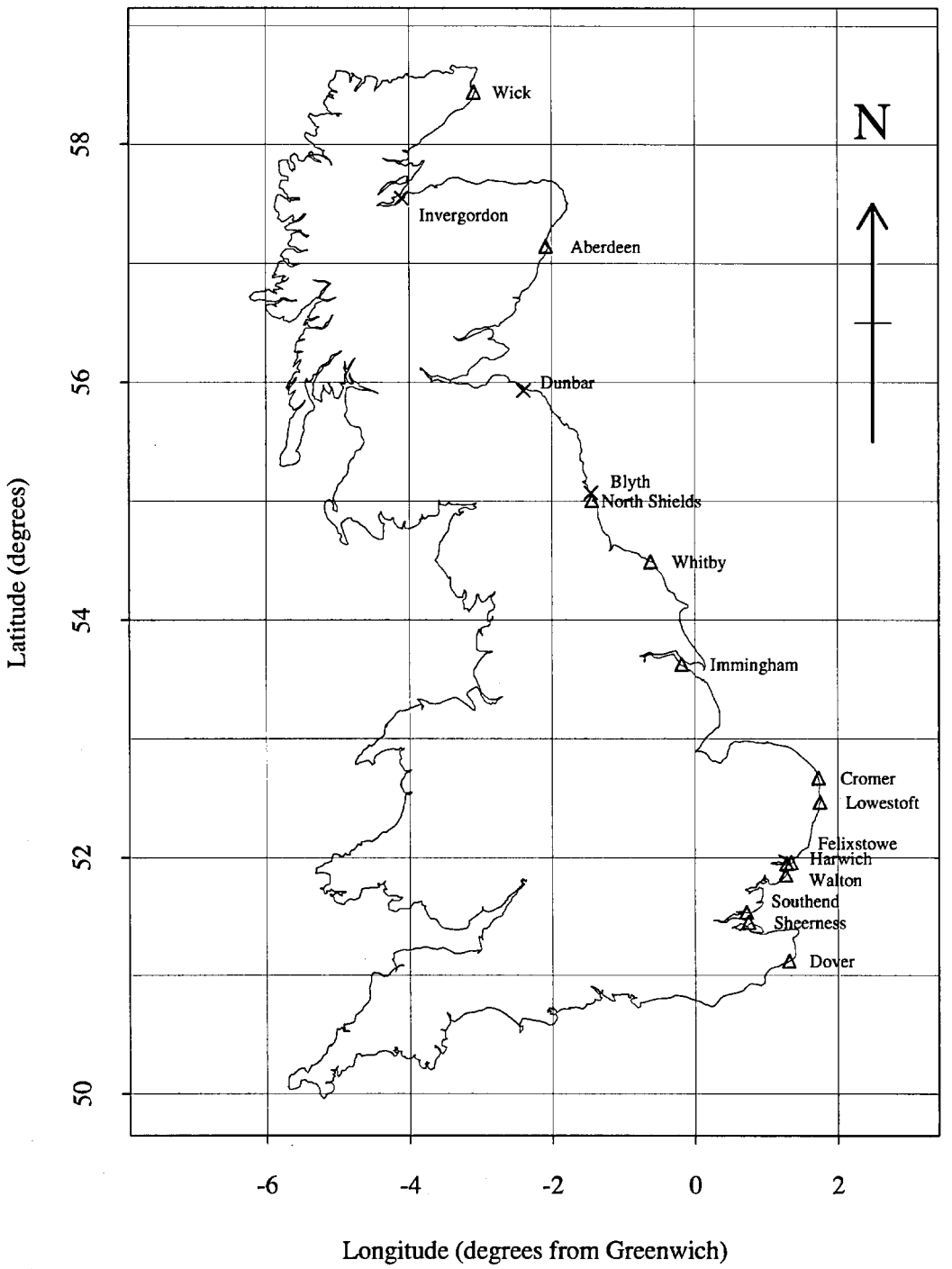

Figure 1. Map showing positions of the east coast A-class data sites and the extra tidal test sites of Invergordon, Blyth and Dunbar as crosses

solutions to this spatial problem, we describe the sea-level process at a site, and summarize the existing univariate approaches.

First consider the sea-level process at a site. The observed hourly sea-level Z_t at time t, after averaging out the high frequency surface waves, consists of the astronomically induced tidal level X_t and a meteorologically induced surge level Y_t, so that

$$Z_t = X_t + Y_t. \tag{1}$$

Although the tide and surge components X_t and Y_t are not observable directly, they can be estimated precisely from a series of hourly sea-level data by exploiting the deterministic, periodic behaviour of the tide. At each site with data, the tide can be expressed as the sum of harmonic sinusoidal constituents and has an 18·61 year cycle, called the nodal cycle (Pugh 1987). Tides vary substantially from point to point along the coastline, and the data from the network of sites is insufficient to estimate the tide at intermediate sites. In contrast, surges change slowly along a coastline so data from a network of sites is adequate for predicting the distribution of surges at intermediate sites.

Now consider the existing univariate methods for estimating extreme sea-levels. The traditional approach is the annual maximum method (AMM; Gumbel 1958). This involves fitting a distribution to observations of the annual maximum hourly sea-level. The most commonly used distribution is the generalized extreme value distribution (GEV). This distribution arises as the limiting distribution of the linearly normalized maximum of a weakly stationary sequence of random variables (Leadbetter *et al.* 1983), and has a distribution function, evaluated at z, of

$$\exp\{-[1 + \xi(z - \mu)/\sigma]_+^{-1/\xi}\} \quad \text{for } z \in \mathcal{R}, \tag{2}$$

where μ, $\sigma > 0$ and ξ are location, scale and shape parameters respectively and $[a]_+ = \max\{a, 0\}$. High quantiles of the fitted GEV (μ, σ, ξ) distribution are used to provide return level estimates and this procedure is applied independently to each data site.

There have been several developments to improve extreme sea-level estimation by incorporating more data than just the annual maximum (Smith 1986; 1989; Tawn 1988), or by exploiting knowledge of the constituent tide and surge processes (Pugh and Vassie 1979; 1980; Walden *et al.* 1982; Middleton and Thompson 1986; Tawn and Vassie 1989; Tawn 1992). Although these methods give more precise estimates and improved estimation at short data sites, their restriction to site-by-site application means that both the spatial coherence of estimates from neighbouring data sites and the tidal variation between the data sites are ignored.

Now consider spatial extensions of the univariate methods. The simplest way of obtaining estimates on a spatial scale, i.e. at any distance along the coastline, is to interpolate return level estimates obtained at each site by one of the site-by-site methods. Application of a particular univariate method, the revised joint probability method (RJPM), to the UK east coast data sites gives estimates of the return level plotted against distance in Figure 2. The spatial estimate obtained by a kernel regression smoother (see the Appendix for details) of the site-by-site return level estimates is also shown on this figure. The general pattern in Figure 2 is that return levels increase gradually with distance from north to south with a dip south of the Wash region (at a distance of about 900 km). However, we will see later that this spatial estimate gives poor return level estimates between data sites as it ignores the spatial variation in the tides.

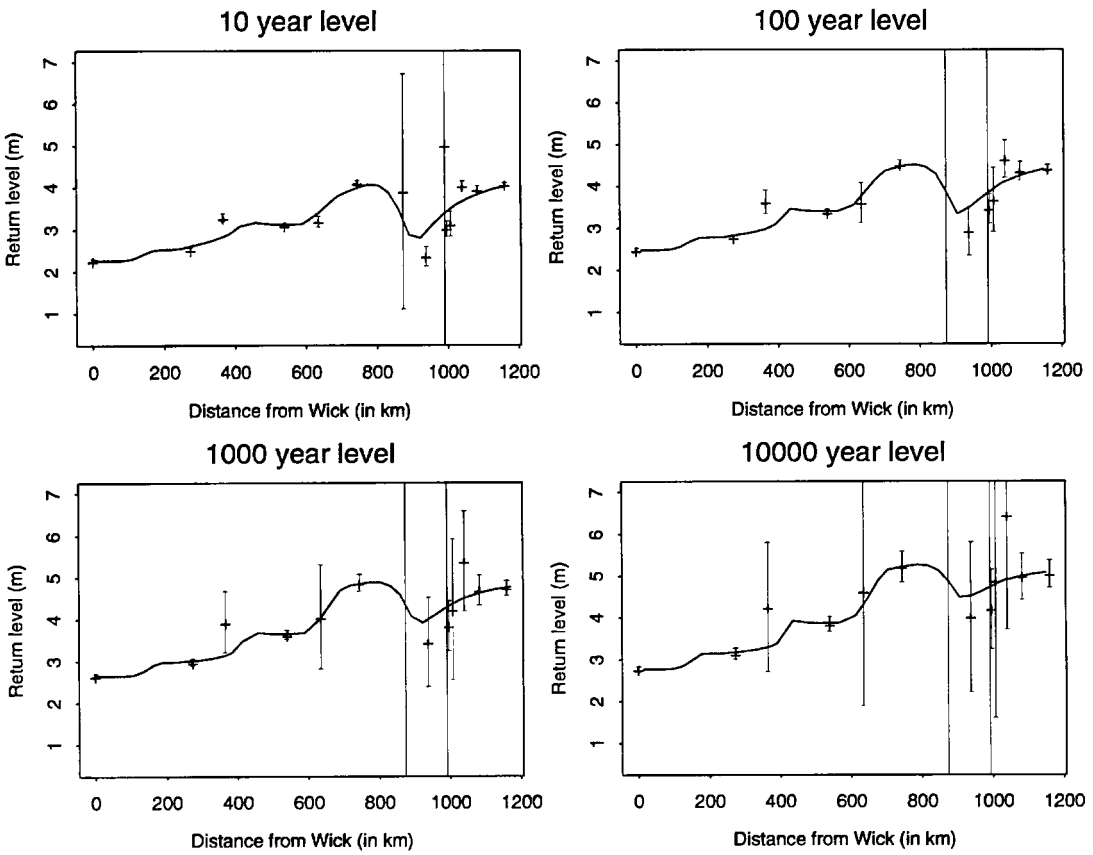

Figure 2. Return level estimates plotted against distance from Wick (in km) for 1990 obtained using the RJPM with 95 per cent confidence intervals given as bars. The kernel regression estimate is shown as a solid line

Simple methods that enable more sophisticated spatial interpolation are considered by Lennon (1963) and Coles and Tawn (1990) who develop a spatial extension of the AMM. However, the methods do not exploit knowledge of either the constituent tide and surge processes, which can lead to substantial bias (Dixon and Tawn 1998), or the variation in the tides between the main data sites. Coles and Tawn (1990) were aware of the importance of the tide but did not have any between-site information on tides. They attempted to explain the variations in the location parameter of the GEV, fitted to annual maximum sea-level data, by constructing a covariate based on the four major harmonics of the tide, listed in the Admiralty Tide Tables (1989). However, their approach failed to overcome problems due to the lack of information about tides and tide–surge interaction between sites.

In this paper we present a method which overcomes these deficiencies. Our approach is to spatially model each of the constituent parts of the RJPM along the coastline, namely the tidal series, the tide–surge interaction, and the distribution of extreme surges. This enables us to incorporate the fine structure of the tide between data sites, using tidal data from a numerical model for the north-west European continental shelf (Flather 1987), and to exploit the smooth spatial coherence of the surge process. These spatial estimates are then combined to give return

level estimates along the whole coast. In Section 2, we review the existing RJPM and present modifications of certain aspects which make the method amenable to spatial extension. In Section 3 the method is applied to the UK east coast data, and in Section 4 the method is compared with the more naive method shown in Figure 2.

Examination of the series of extreme surge levels shows that for many sites there are non-stationary features such as linear trends. Although modelling the spatial behaviour of trends is important for determining return levels for practical use (Dixon and Tawn 1992; Tawn *et al.* 1994), for simplicity of presentation we describe the methods assuming that trends are known at each point along the coastline and have been removed from the data. Extension of our methods to incorporate trends is straightforward and is described by Dixon and Tawn (1995).

2. THE REVISED JOINT PROBABILITY METHOD (RJPM)

First we summarize an extended version of the RJPM as described by Tawn and Vassie (1989) and Tawn (1992) before modifying the structure of the parameters so that they have smooth spatial variation. The RJPM is applicable to data from a single site, and exploits the compositional form (1) of the sea-level by modelling the surge extremes and then combining the resulting estimate with the tide. The RJPM uses the following approximation from extreme value theory. If $W_1, \ldots, W_n$ is a stationary sequence of random variables that satisfy a condition that ensures extremes far apart in the sequence are independent (see Leadbetter *et al.* 1983), then for large n

$$\Pr\left(\max_{1 \leqslant i \leqslant n} W_i \leqslant w\right) \approx \left[\prod_{i=1}^{n} \Pr(W_i \leqslant w)\right]^{\theta} \quad \text{for } w > u, \tag{3}$$

where $0 < \theta \leqslant 1$ is termed the extremal index for the process, and u is a high threshold. The extremal index is a measure of the degree of clustering in the extremes of the process, with θ^{-1} being the limit of the mean number of exceedances of a threshold in a cluster as the threshold is increased to the endpoint of the distribution of W. This result continues to hold for certain non-stationary sequences (see Husler 1986).

If the sea-level series $\{Z_t\}$ satisfies the conditions for (3), $N = 8766$ is the number of hours in a year and $T = 18{\cdot}61 \times N$ is the number of hourly observations in the nodal tidal cycle of 18·61 years, then from (1) and (3) the distribution of the maximum sea-level over one nodal cycle is

$$H(z) = \Pr\left(\max_{1 \leqslant t \leqslant T} Z_t \leqslant z\right) \approx \left\{\prod_{t=1}^{T} F_{Y|X_t}(z - X_t)\right\}^{\theta_Z} \quad \text{for large } z, \tag{4}$$

where $F_{Y|X}$ is the distribution function of the hourly surge Y conditional on the tidal level X, and θ_Z is the extremal index for the hourly sea-level process. If the distribution of the tidal series is approximately the same every year, then the annual maximum sea-level distribution $G(z)$ is given by

$$G(z) \approx [H(z)]^{N/T}. \tag{5}$$

The dependence of the surge distribution on the tidal level in (4), termed tide–surge interaction, occurs mainly in shallow water areas (see Pugh 1987 for oceanographic details). Interaction takes

the form of amplifying and dampening the surge at mid-range and high tidal levels respectively. Consequently the largest surges generally do not occur at high tides.

We first describe estimation of $G(z)$ when the tide and surge are independent, i.e. where $F_{Y|X} = F_Y$, before extending the modelling and estimation to the tide–surge interaction case. In order to exploit equation (5) to obtain an estimate of $G(z)$, a model for F_Y is required. Below a high surge threshold u_Y, F_Y is estimated non-parametrically using an empirical estimate $\tilde{F}_Y$. Above u_Y, an asymptotically justified parametric model is used to smooth and extrapolate the surge distribution tail. Assuming that the hourly surge values from a year, $Y_1, Y_2, \ldots, Y_N$, satisfy the conditions for (3), we have

$$\Pr[\max(Y_1, Y_2, \ldots, Y_N) \leqslant y] = [\Pr(Y_1 \leqslant y)]^{N\theta_Y} = F_Y(y)^{N\theta_Y} \quad \text{for } y > u_Y,$$

where θ_Y is the surge extremal index. As the surge is approximately stationary over the winter storm season, it is reasonable to assume that the annual maximum surge distribution is GEV(μ_Y, σ_Y, ξ_Y), and so

$$F_Y(y) = \begin{cases} \exp\{-(N\theta_Y)^{-1}[1 + \xi_Y(y - \mu_Y)/\sigma_Y]_+^{-1/\xi_Y}\} & \text{for } y > u_Y, \\ \tilde{F}_Y(y) & \text{for } y \leqslant u_Y. \end{cases} \tag{6}$$

Although in general there is no simple relationship between the distribution of extreme surges and sea-levels, if $\xi_Y = 0$ then Tawn (1992) showed that, when the surge is independent of the tide, the distribution of the annual maximum is approximately GEV($\mu_Y^*, \sigma_Y, 0$), where

$$\mu_Y^* = \mu_Y + \sigma_Y \log\left[\theta_Z(T\theta_Y)^{-1} \sum_{t=1}^{T} \exp(X_t/\sigma_Y)\right].$$

More generally, G has to be evaluated numerically using (4).

We now fit model (6) to surge data. The parameters μ_Y, σ_Y and ξ_Y are estimated using a point process approach described by Smith (1989). First the (dependent) hourly surge process is declustered using a simple declustering technique such as that described by Dixon and Tawn (1994), which assumes that two local maxima of the process are independent if they are greater than a predetermined time separation apart. Then, if the hourly surge series satisfies the conditions of (3), and $\boldsymbol{y} = (y_1, \ldots, y_n)$ is a realization of the surge series, then, for a suitably high threshold u_Y, the censored likelihood of each hourly declustered surge above u_Y is approximately the likelihood of a non-homogeneous Poisson process with integrated intensity on $[y, \infty]$ of $N^{-1}[1 + \xi_Y(y - \mu_Y)/\sigma_Y]_+^{-1/\xi_Y}$, i.e.

$$L(\boldsymbol{y}; \mu_Y, \sigma_Y, \xi_Y) \propto \prod_{t=1}^{n} \exp\left\{-\frac{1}{N}\left[1 + \xi_Y\left(\frac{u_Y - \mu_Y}{\sigma_Y}\right)\right]_+^{-1/\xi_Y}\right\}\left[1 + \xi_Y\left(\frac{y_t - \mu_Y}{\sigma_Y}\right)\right]_+^{(-1/\xi_Y - 1)\delta_t}, \tag{7}$$

where δ_t is an indicator function which is 1 when $y_t > u_Y$ and y_t is a cluster maximum and is 0 otherwise; see Robinson and Tawn (1997) for a justification of this approach. This approach is less direct than fitting the GEV(μ_Y, σ_Y, ξ_Y) to the annual maximum surge data; however, it has the advantages of using all extreme hourly declustered surges and extends naturally to the tide–surge interaction case.

Maximization of likelihood (7) gives estimates of the parameters μ_Y, σ_Y and ξ_Y. The surge and sea-level extremal indices are estimated using the reciprocal of the mean cluster size above a high threshold (see Tawn 1992). The return level is estimated by inserting (6) into (4), replacing parameters by their estimated values, setting (5) to be equal to $1 - p$ and inverting.

When tides and surges interact, the surge series is non-stationary and has a distribution that depends on the concurrent tidal level. Thus the distribution of the surges which exceed a high threshold also depends on the concurrent tidal level. This suggests replacing the parameters μ_Y, σ_Y and ξ_Y by suitably modelled functions $\mu_Y(X)$, $\sigma_Y(X)$ and $\xi_Y(X)$ of the tide. Tawn (1988) applied this approach to data from some UK east coast sites using low-order polynomial regression functions for each parameter, and found that the shape of the tail was homogeneous across tides, i.e. that $\xi_Y(X) = \xi$. The disadvantage of using simple parametric polynomial regression functions is that they are not flexible enough to model the complex tide–surge interaction process observed over the coastal network. Following Dixon and Tawn (1994) we use more flexible covariate forms. We assume that there exist functions of the tide $a(X)$, $b(X)$, such that the location-scale normalization of the surge series Y_t,

$$S_t = [Y_t - a(X_t)]/b(X_t), \tag{8}$$

is approximately stationary over time for $Y_t > u_Y(X_t)$. Here $u_Y(X)$ is a high threshold which depends on the tide X. Experience suggested that suitable functions are given by $a(X) = a_1(X)$ and $b(X) = a_2(X) - a_1(X)$, where $a_i(X)$ is the $1 - p_i$ quantile of the surge distribution conditional on the tide level X for $i = 1, 2$. The p_i values are taken to be small enough so that the quantiles contain sufficient information about the extremal tail, but low enough so that they can be estimated empirically with sufficient accuracy.

If we assume that $a(X)$ and $b(X)$ are known, then we can apply the models for Y developed in the tide–surge independent case to data on S. Specifically, assuming that annual maximum transformed surges are GEV(μ_S, σ_S, ξ_S), we can estimate μ_S, σ_S, ξ_S by maximizing the likelihood $L(\boldsymbol{s}; \mu_S, \sigma_S, \xi_S)$, where L is as in (7) and $\boldsymbol{s} = (s_1, \ldots, s_n)$ are hourly observations of the transformed surge. A simple procedure is to estimate $a_i(X)$ using empirical quantiles of the surge conditional on the tide, and proceed as if $a(X)$ and $b(X)$ were known.

More generally this formulation is equivalent to using (7) in a regression setting with

$$\mu_Y(X_t) = \mu_S b(X_t) + a(X_t), \quad \sigma_Y(X_t) = \sigma_S b(X_t), \quad \text{and } \xi_Y(X_t) = \xi_S, \tag{9}$$

i.e. $a(X)$ and $b(X)$ are treated as covariates and the shape parameter is constant, as in Tawn (1988). Our approach is to parametrically specify $a(X_t)$ and jointly estimate these and the extreme parameters using likelihood (7) with forms (9). Specifically, we take the annual p_1 quantile of the surge conditional on the tide X_t to be independent and distributed normally with mean $a_1(X_t)$, so that $a_1(X_t)$ are taken as unknown parameters which are then estimated by maximum likelihood. This involves numerical maximization of the likelihood given by the product of (7) with μ_Y, σ_Y and ξ_Y replaced by (9), and the likelihood for the $a_1(X)$ parameters based on the associated annual empirical estimates. This is computationally intensive since each likelihood evaluation requires the product over the n terms, where n, the number of hours at the site, is typically very large. In practice we group tides into bands so that $a_1(X_t)$ is the parameter corresponding to the p_1 quantile in the tidal band that X_t falls in. The function $b(X)$ is estimated empirically for each tidal band.

Although the covariate choice in (9) is to some extent *ad hoc*, the actual covariate form is natural in an extreme value setting and gives rise to invariance of the extremal parameters μ_S, σ_S and ξ_S with respect to location and scale changes of the surge series. This leads to a reduction in the spatial variability of the transformed surge series S_t, which is important for simplifying the later spatial analysis. A similar approach was adopted by Coles and Pan (1996) in the context of seasonal non-stationarity for pollution levels.

There are a few practical details associated with application of the method. Firstly, it is convenient to work with a transformed uniformly distributed tide $X^* = F_{tide}(X)$, where F_{tide} is the distribution function of the tide. This is helpful for the later spatial analysis where tidal ranges can differ substantially across sites. Secondly, experience with data suggests that 10 tidal bands are adequate to describe the interaction, and so the a function is specified by 10 parameters, at tides $X^*(j) = (j - 0{\cdot}5)/10$, for $j = 1, \ldots, 10$. Intermediate values are obtained using weighted kernel regression smoothing of the points $\{(X^*(j), \hat{a}_1(X^*(j))); j = 1, \ldots, 10\}$, where $\hat{a}_1$ are the estimated values of a_1 and the weights are inversely proportional to the standard error of $\hat{a}_1$. See the Appendix for details of the smoothing method.

3. SPATIAL ESTIMATION

3.1. Tides

In order to estimate return levels spatially, we require a spatial estimate of the tidal series. In this section we describe the estimation of the hourly tidal series at a given distance d, $X_t(d)$: $t = 1, \ldots, T$. For a given location, the tide can be represented as

$$X_t = Z_0 + \sum_{i=1}^{m} h_i \cos(\omega_i t - g_i) \tag{10}$$

(Pugh 1987). Here Z_0, h_i and g_i are (unknown) parameters which determine the mean level signal and the amplitude and phase of the harmonic constituents respectively. The number of constituents m is typically taken to be at least 60, and ω_i are known constants.

Each constituent is labelled by a letter and a number, for example M and S denote lunar and solar constituents respectively. The associated number defines the tidal species which are grouped around 1, 2, 3, 4 and 6 cycles per day. At any given location the constituents give a varying contribution to the tide, depending on their amplitude and phase. In UK waters, and in many other parts of the world, the tide is dominated by the lunar constituent M_2 and the solar constituent S_2, which makes the tides predominantly semi-diurnal and causes the density function of the tide to be bimodal. Other important constituents are the declinational terms O_1 and K_1, which are caused by the moon/sun moving away from the equator; N_2 and K_2, which are a function of the elliptical orbits of the moon and sun; the non-linear terms M_4, MS_4 and M_6, which appear in shallow water; and the fortnightly tide MS_f.

Now consider estimation of the tide at any given coastal location. For purposes of spatially estimating the tide, the position of a coastal point is defined by its distance along the shoreline distance metric (SDM) as opposed to geographical distance. The SDM is a fine resolution representation of the UK coastline, obtained from the World Vector Shoreline compiled by the US Defence Mapping Agency.

If we assume that the mean sea-level $Z_0(d)$ is known at every distance d, then estimation of the tidal series reduces to estimation of the amplitude and phase of the ith constituent at distance d, denoted by $h_i(d)$ and $g_i(d)$ respectively. Since the tide is fully specified, at any time, by the constituents, with $\mathbf{h} = (h_1, \ldots, h_m)$ and $\boldsymbol{g} = (g_1, \ldots, g_m)$, our aim is to obtain estimates $\hat{\boldsymbol{h}}(d)$ and $\hat{\boldsymbol{g}}(d)$ at every distance d. This is achieved by interpolating the values of $\boldsymbol{h}$ and $\boldsymbol{g}$ from *reference points*. Reference points are defined as sites which have enough hourly data to obtain accurate estimates of the tidal series. They include the 14 data sites in Figure 1, and 24 additional sites of the National Tide Gauge Network (Rae 1988) which have long enough records for accurate tidal analysis but which are too short for extreme sea-level analyses. Harmonic analysis of the hourly tidal data at the reference points gives $\boldsymbol{h}(d_j)$ and $\boldsymbol{g}(d_j)$.

Even with these extra sites, the spatial resolution of the reference points is too low to capture the rapid spatial variation in tidal features. Thus output from a hydrodynamical model is used to improve the mapping of the tides. This output consists of hourly tidal predictions at a discrete 12 km grid of 83 grid points down the UK east coast. Since the grid points lie some way off the coast, usually between 0 and 50 km, we use an additional distance metric termed the model distance metric (MDM). Distances d on this metric are translated into distances d along the coastline (SDM) using a transformation which gives the SDM for the nearest point on the coast. The hourly tidal values at the grid points are denoted by $V_t(d_k^*)$ for $k = 1, \ldots, n_g$, where $n_g = 83$, $t = 1, \ldots, N$ and d_k^* is the distance of the kth grid point as measured on the MDM. Harmonic analysis of the synthetic tidal data then gives $\boldsymbol{h}^*(d_k^*)$ and $\boldsymbol{g}^*(d_k^*)$.

Estimated hourly tides at distance d are now derived by first obtaining the numerical model series at distance d, and then correcting this for observed differences at the reference sites and nearby numerical model points. Thus initially we calculate a value of $\boldsymbol{h}^*$ and $\boldsymbol{g}^*$ at distance d^*, the distance on the MDM corresponding to distance d on the SDM. First let the neighbouring grid points to distance d^* be the kth and $(k+1)$th, and define functions $f_{i,1}, f_{i,2}$ of the ith tidal constituent at distance a by

$$f_{i,1}(a) = h_i(a) \cos g_i(a), \quad f_{i,2}(a) = h_i(a) \sin g_i(a), \text{ for } a \in \mathcal{R}.$$

Then for each constituent i, we obtain $h_i(d^*)$ and $g_i(d^*)$ by first linearly interpolating $f_{i,j}(d_k^*)$ and $f_{i,j}(d_{k+1}^*)$ to distance d^* for $j = 1, 2$. Then $f_{i,j}(d^*)$ and $f_{i,j}(d^*)$ are resolved to give $h_i(d^*)$ and $g_i(d^*)$. A year of hourly tidal levels $\hat{V}_t(d)$, termed the model tidal series at distance d (on the SDM), is then obtained from $\boldsymbol{h}^*(d^*)$ and $\boldsymbol{g}^*(d^*)$.

Now define the error in the modelled tides at the reference points, at time t, by

$$\varepsilon_t(d_j) = X_t(d_j) - \hat{V}_t(d_j) \quad \text{for } j = 1, \ldots, 38,$$

and approximate the derivative in this error by

$$\varepsilon_t'(d_j) = [\varepsilon_t(d_{j+1}) - \varepsilon_t(d_j)]/(d_{j+1} - d_j).$$

The estimated hourly tides in this year at distance d are then obtained by

$$\hat{X}_t(d) = \hat{V}_t(d) + \varepsilon_t(d_j) + (d - d_j)\varepsilon_t'(d_j) \quad \text{for } t = 1, \ldots, N.$$

Harmonic analysis of the year-long $\hat{X}_t(d)$ series then gives the required estimates $\hat{\boldsymbol{h}}(d)$ and $\hat{\boldsymbol{g}}(d)$.

Table I. Constituents M_2, S_2, K_1 and O_1 in centimetres and degrees. Rows denoted A and B refer to values from the Admiralty Tide Tales and from the interpolation scheme respectively. The terms h and g denote amplitude and phase respectively. The sites are shown in Figure 1 as crosses

Site	Grid reference	Source	M_2		S_2		K_1		O_1	
			h	g	h	g	h	g	h	g
Invergordon	57°41N, 4°10W	A	136	335	48	012	11	180	12	034
		B	133	336	47	013	12	179	12	034
Blyth	55°07N, 1°29W	A	160	087	55	126	13	234	14	081
		B	158	087	53	129	11	242	14	080
Dunbar	56°00N, 2°31W	A	161	055	56	096	11	219	13	067
		B	161	056	55	097	11	221	14	065
Walton	51°51N, 1°16E	A	143	330	40	024	11	358	13	177
		B	141	331	40	024	10	357	13	177

Two checks for the adequacy of this interpolation model are made. Firstly, a cross-validation test was performed for four sites, which are selected to be close to and distant from neighbouring sites. Table I gives the derived values from the interpolation scheme and actual values for the harmonic constituents obtained using the tide model with the specified site removed. The interpolated amplitude and phases of the constituents M_2, S_2, K_1 and O_1 agree well with actual values. Secondly, from the harmonic constituents, we obtain kernel density estimates of the tidal distribution by generating a tidal series over the full nodal cycle. Figure 3 shows the 90, 95, 99 and 99·9 per cent tidal quantiles for each data site and the continuous spatial tidal quantile curves for the whole coastline, obtained from the tidal interpolation scheme. Note that in Figure 3 the quantiles are based on a zero mean sea-level at each site, and are independent of the tidal datum used to obtain the measured data for each site. Out of all the data sites only the quantiles for Harwich and Walton (at a distance of approximately 1000 km) are not near perfectly described by the spatial estimate, and even in these cases the spatial estimate provides only a slight underestimation.

3.2. Marginal surge and interaction

Now consider estimating the conditional surge distribution at each distance d. In principle a multivariate parametric model could be fitted to surge data from all sites simultaneously, incorporating distance as a covariate. As our main interest is in extremes of the process, such a model requires high dimensional multivariate extreme value models that are yet to be developed for practical application. Coles and Tawn (1991) illustrate that, even in relatively simple settings, the applications become restrictively complex. In view of the high dimensionality and complexity of the sea-level process, the approach we take is more simplistic, but retains the flexibility to model complex forms of spatial variation.

Rather than attempting to fit a spatial model to the surge data from all sites simultaneously, we develop a spatial model in two stages. First we estimate the components of the RJPM at each site, ignoring possible dependence and spatial coherence in the surge data from different sites. We then use these site-by-site estimates to obtain spatial estimates of each RJPM parameter along the whole coastline and these are combined, via equation (6), to give an estimate of the conditional surge distribution at each distance d. Concentrating on aspects of the surge distribution which have dominant effect on high return level estimates, we define the 10th tidal band interaction

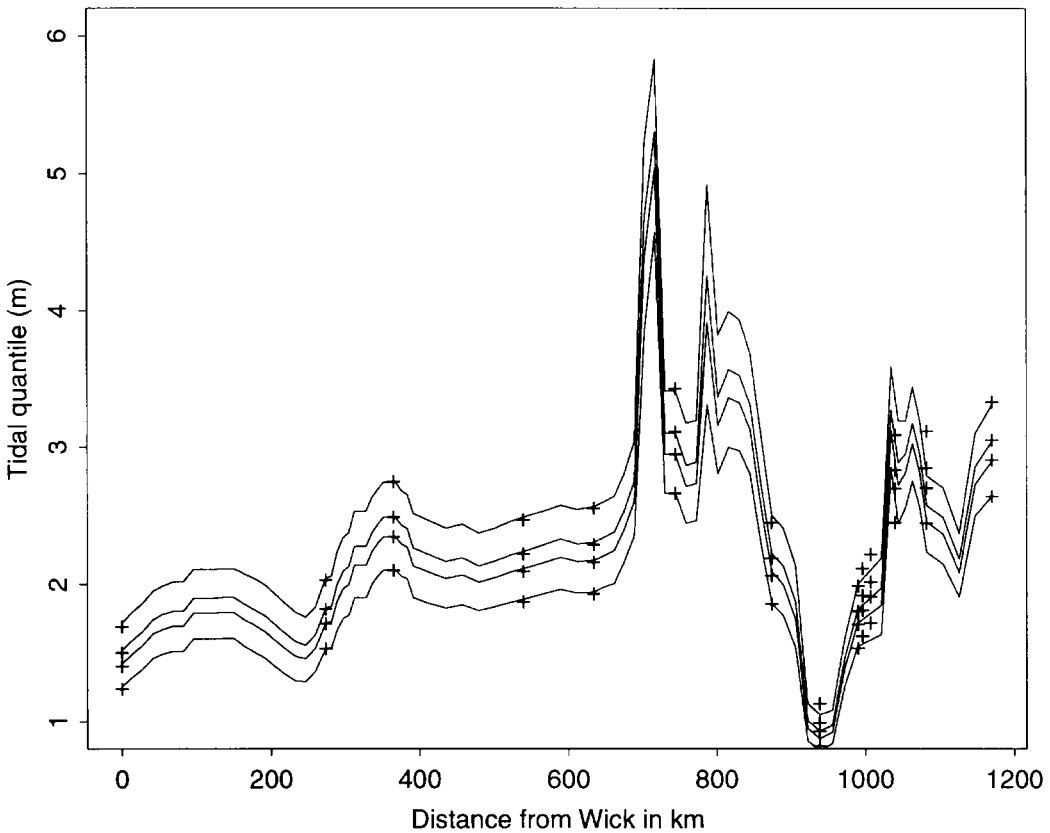

Figure 3. Quantiles of the tidal distribution at the sites, and at all distances obtained using the tide interpolation model. The quantiles shown are the 90, 95, 99 and 99·9 per cent levels. The + symbols indicate the corresponding quantiles, based on tidal predictions from actual observations, for the data sites

functions at distance d by $\upsilon_1(d)$ and $\upsilon_2(d)$ (corresponding to $a(X^*(10))$ and $b(X^*(10))$ at distance d respectively), and extreme parameters at distance d by $\mu_S(d)$, $\sigma_S(d)$ and $\xi_S(d)$. Also define $\boldsymbol{\theta}(d) = \{\upsilon_1(d), \upsilon_2(d), \mu_S(d), \sigma_S(d), \xi_S(d)\}$ and denote estimates of the parameters at site j, with distance d_j by $\hat{\boldsymbol{\theta}}_j = \hat{\boldsymbol{\theta}}(d_j)$, for $j = 1, \ldots, 14$.

Now we assume that each parameter changes slowly with distance along the coast, and that all parameter estimates both within a site and across sites are independent. In other words, the components of $\hat{\boldsymbol{\theta}}_j$ are mutually independent for all j, as are $\hat{\boldsymbol{\theta}}_i$ and $\hat{\boldsymbol{\theta}}_j$ for all $i \neq j$. Under these assumptions, we separately estimate each parameter at distance d by univariate weighted kernel regression estimation. More explicitly, using the shape parameter ξ_S as an example, a spatial estimate $\tilde{\xi}_S(d)$ is given by kernel regression estimation of $\{(d_j, \hat{\xi}_S(d_j)), j = 1, \ldots, 14\}$ with weights inversely proportional to the marginal standard errors of $\hat{\xi}_S(d_j)$. This procedure is repeated for each component of $\boldsymbol{\theta}(d)$: details are given in the Appendix.

Figure 4 shows the resulting kernel regression estimates $\tilde{\upsilon}_1(d)$, $\tilde{\mu}_S(d)$, $\tilde{\sigma}_S(d)$ and $\tilde{\xi}_S(d)$ plotted against d. Each parameter exhibits spatial smoothness and the kernel regression estimate reflects the pattern in the site estimates and their relative uncertainty. The general lack of sensitivity to the choice of bandwidth in the kernel regression smoothing is illustrated in the Appendix. The $\tilde{\upsilon}_1(d)$

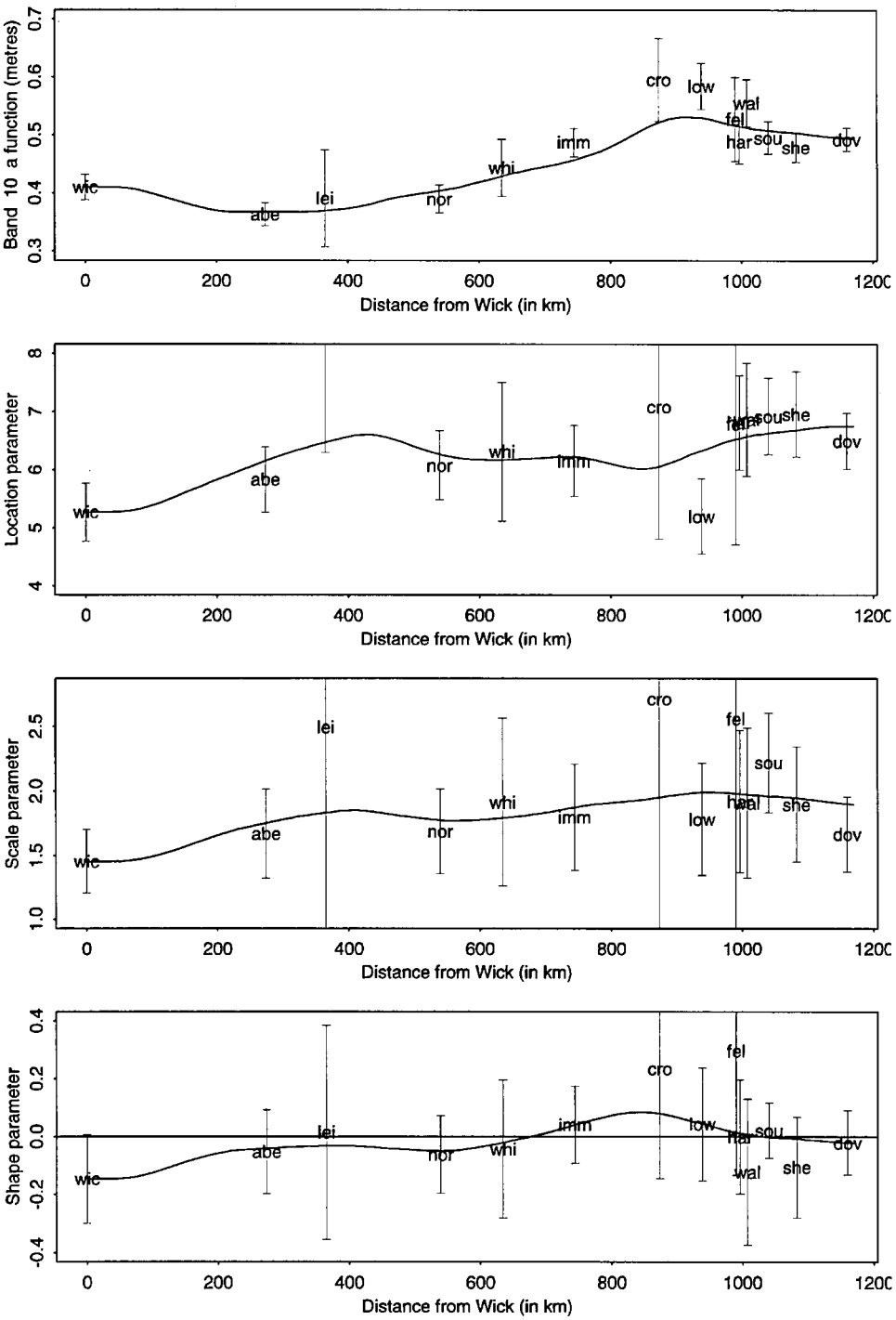

Figure 4. The highest tidal level *a*-function and point process extreme value parameters for the surge against distance; the site-by-site estimates and the spatial estimate with approximate 95 per cent confidence intervals

parameter estimate generally increases from Wick, peaking around Lowestoft and levelling off for the south-eastern coast. There is a gradual increase in the location, scale and shape parameters with distance, although with respect to the uncertainty in the estimates it would not be unreasonable to take the scale and shape parameters as constant along the coastline; in particular, the shape parameter could be taken to be zero for the whole coastline, corresponding to exponential tails. We have not pursued this approach, despite the mathematical simplicity it produces. This is mainly because the estimated extrapolations, and corresponding uncertainties, are sensitive to the shape parameter, and fixing it somewhat arbitrarily at zero would result in substantially underestimated precision.

To simplify the analysis thus far, we have made independence assumptions which are not supported by the data. For example, there may be a slight negative dependence between $\tilde{\upsilon}_1$ and $\tilde{\mu}_S$ evident in Figure 4 as seen in the marginal estimates at Lowestoft which have an atypically high (and low) estimate of $\tilde{\upsilon}_1$ and ($\tilde{\mu}_S$) respectively relative to the spatial pattern. This feature feeds into the spatial estimates of the two parameters.

For other parameter combinations, however, the independence assumption appears to have greater validity, and is supported by theoretical considerations. For instance, estimates of high quantiles of a distribution are asymptotically independent (David 1981) so that $\tilde{\upsilon}_1$ and $\tilde{\upsilon}_2$ should be approximately independent of $\tilde{\sigma}_S$ and $\tilde{\xi}_S$. Also, examination of the variance–covariance matrix of maximum likelihood estimates, given by Prescott and Walden (1983), shows that generally there is low dependence between $\hat{\mu}_S(d_j)$ and $\hat{\sigma}_S(d_j)$ and between $\hat{\mu}_S(d_j)$ and $\hat{\xi}_S(d_j)$. The features of dependence observed in other parameter combinations could be reduced by reparameterizating the model to have greater orthogonality. However, since the main spatial variation is driven by the form of the tide, improved smoothing techniques here will have a minimal impact on return level estimates, and we retain our more naive approach for simplicity. Furthermore, even though the independence assumptions are violated, the proposed approach gives an unbiased spatial estimate of the return level with too small a standard error. A general approach which modifies the standard error evaluation to account for the neglected dependence is given by Liang and Self (1996), but this was not implemented here.

3.3. Return levels

Combining the parameter and tidal estimates at each distance d, using equations (4) and (5), leads to a spatial estimate of the return level. Figure 5 shows the corresponding estimated 100 and 1000 year return levels. The large spatial variation between data sites (due to the tide) is clearly evident, for example around Immingham. The spatial estimate is generally close to the site-by-site estimates that have the smallest standard errors, and picks up the variation of the spatial tide estimate between data sites. Three of the data sites, Leith, Cromer and Felixstowe, have records of less than 5 years of data, and have correspondingly poor site-by-site estimates. The spatial model provides improved estimates at these sites by transferring information about the surge process from neighbouring sites with longer records.

4. COMPARING ESTIMATES

Using the naive model in Section 1 as a benchmark, we assess the performance of our model using a cross-validation technique. Specifically, we refit the model described in Sections 2 and 3 and the naive model of Section 1 (shown in Figure 2), but sequentially leave out each site

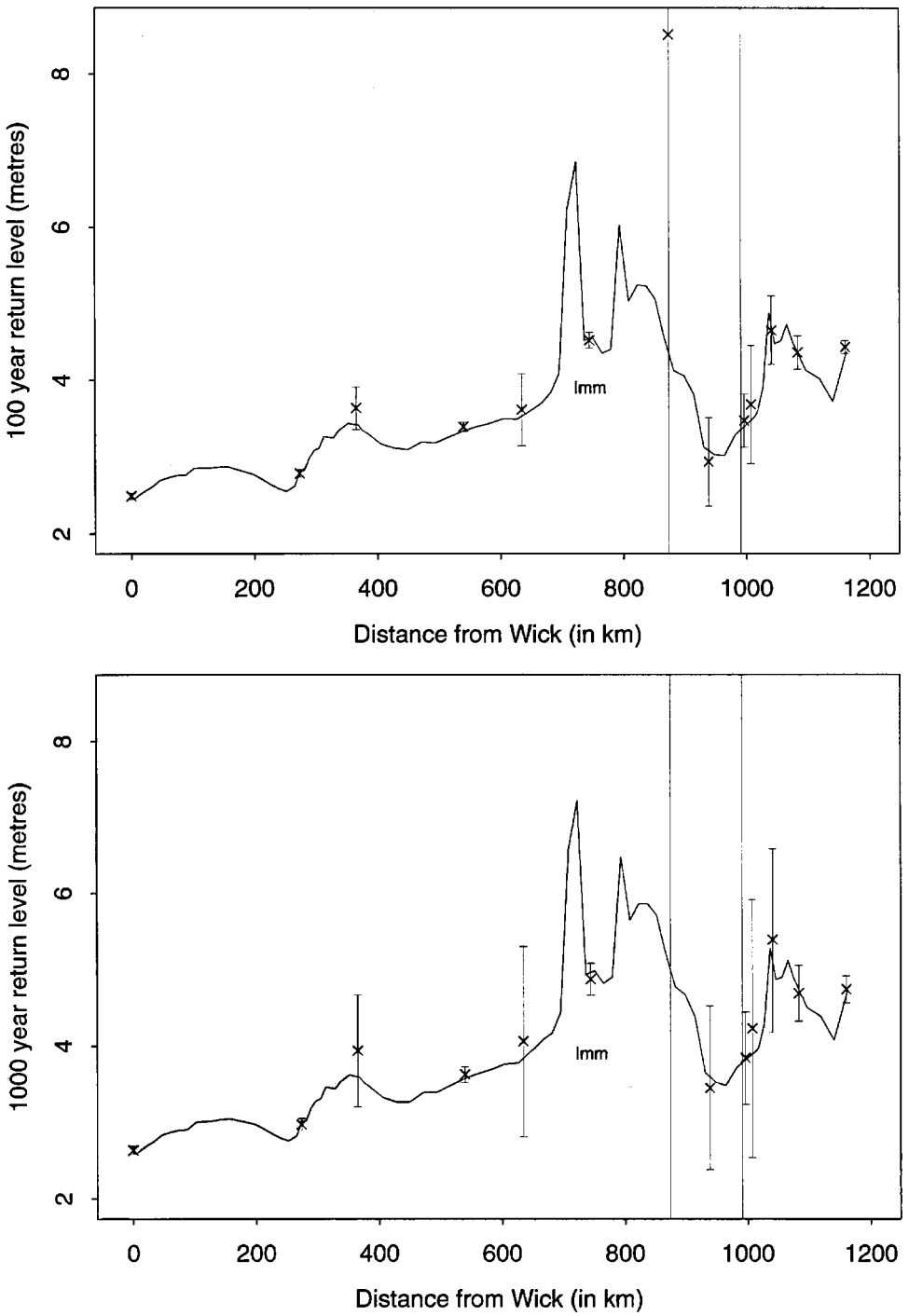

Figure 5. Spatial estimate of the 100 and 1000 year return level. The site-by-site estimates, with 95 per cent confidence intervals, are also shown. Immingham is marked as site abbreviation Imm

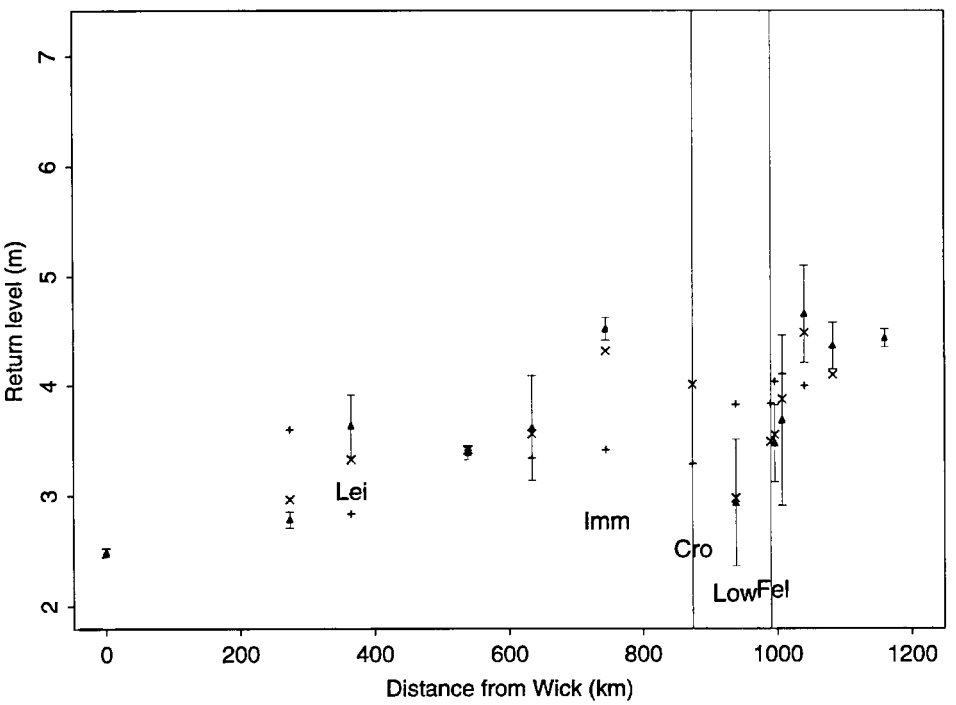

Figure 6. Cross-validation estimates at each site for our spatial method (× symbol) and the crude method (+ symbol). The site estimates and standard errors are shown as triangles and error bars respectively. Site abbreviations are Leith, Immingham, Cromer, Lowestoft and Felixstowe

(excluding the end sites, Wick and Dover) before applying the spatial methods. The estimated return level at the site that has been removed is then compared with the corresponding estimate at each of the 12 sites for each method. Figure 6 displays the estimates of the 100 year return level for each of the 12 sites plotted against distance. The picture is similar for other return levels. Our spatial method estimates the return level at the removed sites well, and the estimate is contained within most of the 95 per cent confidence intervals of the site-by-site estimates. In contrast, the naive method fails to estimate some sites well, especially around complex shaped coastline areas. In particular, the estimates at Immingham and Lowestoft are very poor. This suggests that our model gives an improvement for return level estimates *between* data sites.

5. CONCLUSIONS

We have presented an approach for obtaining extreme sea-level probabilities along the whole UK east coast. From a purely statistical viewpoint, there are obvious criticisms of our approach. Principally, these concern the poor handling of the dependence in the surge and tide–surge interaction components when spatially smoothing the surge distribution. However, a multivariate extreme value model on this large scale is intractable, and even less ambitious methods quickly become complex within this setting. Despite this, the spatial model does capture the important features of the spatial sea-level process and substantially improves on a more naive approach as it exploits the separate spatial variation in tides and surge levels. In particular, our

model gives improved estimates between data sites, and at data sites with short records, and can easily be extended to other coastlines which have a high enough spatial density of sites relative to the complexity of the coastline. The main drawback of falsely assuming independence is that the spatial estimate standard error would be too small. We have not calculated standard errors for the spatial estimate as the main source of uncertainty is in the spatial tidal estimate between data sites. This uncertainty is difficult to quantify as it arises from the numerical model limitations on the 12 km grid which depend critically (and unpredictably) on the local bathymetry and on the geometry of the nearby coastline, with complex shaped regions having greater uncertainty than simple linear coastal stretches.

We have provided estimates of the tide and the sea-level distribution at 20 km intervals along the east coast. For some regions, for example the Humber and the Wash, the local bathymetry induces rapid non-linear spatial changes in the tidal characteristics, so that 20 km intervals are too coarse to provide accurate estimates. These complex coastal regions are the subject of future study by incorporating data from finer scale tide–surge numerical models.

APPENDIX: KERNEL REGRESSION ESTIMATION

This section summarizes our procedure for the kernel regression estimation. There are many possible implementations of kernel regression and descriptions are given in various books including Härdle (1990) and Wand and Jones (1995). Following advice in these texts, our procedure is as follows.

Using the shape parameter as an example, and following the notation of Section 3·2, we wish to find a kernel regression estimate of the points

$$\{(d_j, \hat{\xi}_S(d_j)), j = 1, \ldots, n\}$$

with weights $w_j = [\mathrm{SE}(\hat{\xi}_S(d_j))]^{-1}$, $j = 1, \ldots, n$.

Assuming that $\hat{\xi}_S(d_j)$ are independent, we apply a version of the Nadaraya–Watson (for example see Wand and Jones 1995) estimator which has kernel regression estimate at distance d given by

$$\hat{\xi}_S(d) = \frac{\sum_{j=1}^{n} K_h(d - d_j)\hat{\xi}_S(d_j)w_j}{\sum_{j=1}^{n} K_h(d - d_j)w_j},$$

where K_h is the kernel function with bandwidth h.

There is no all-round optimal kernel function. We use the Epanechnikov kernel which has some optimality properties in kernel density estimation (for example see Wand and Jones 1995) and is given by

$$K_h(x) = \begin{cases} \frac{3}{4}\{1 - x^2/(5h^2)\}/(h\sqrt{5}) & \text{for } |x| < h\sqrt{5}, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

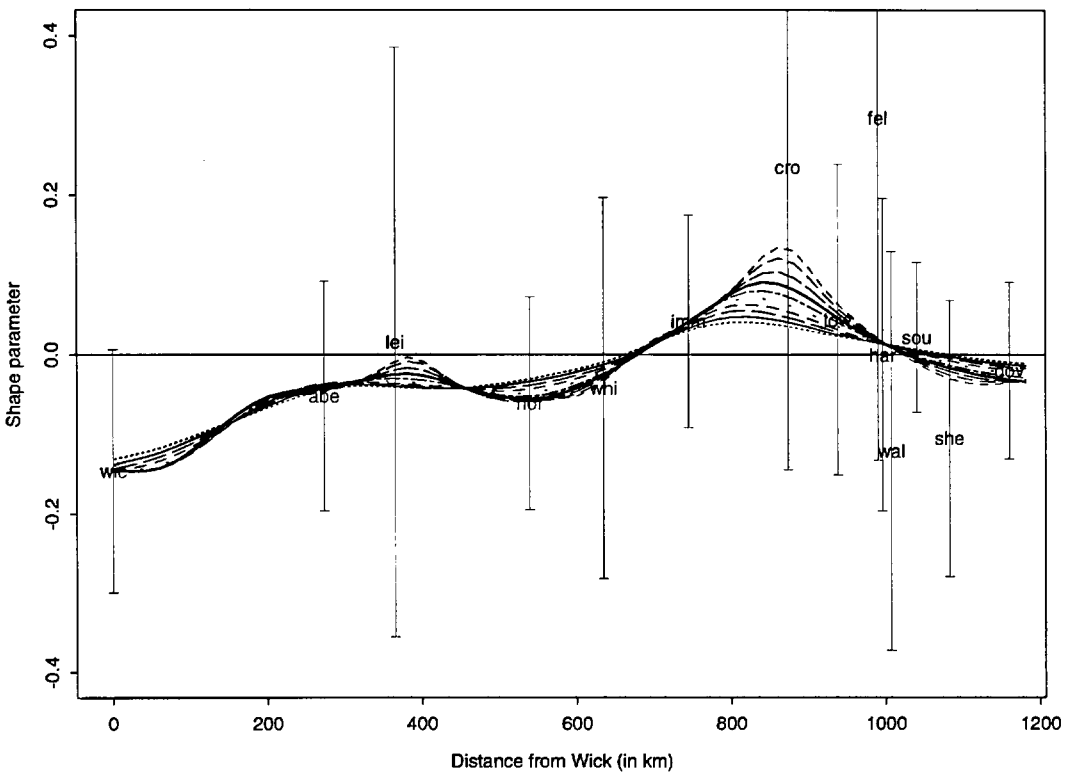

Figure 7. Choice of bandwidth for the kernel regression estimate of the shape parameter. The smoothing ranges from $h = 120$ ($\cdots$) to $h = 30$ ($- - -$)

There have been various suggestions as to how to choose the bandwidth. For example Härdle and Marron (1985) describe a procedure based on least squares for the Nadaraya–Watson estimator. Our procedure is to choose a bandwidth subjectively by looking at a range of plausible values, and making a selection based on oceanographic knowledge of the underlying sea-level processes. Figure 7 illustrates this procedure for the shape parameter. The smoothed regression curves, using kernel (11), are for h ranging between 30 and 120 with kilometres as the units of distance. Based on this figure we chose to use $h = 60$, shown by the thick solid curve. Clearly other choices of h would have led to slightly different spatial estimates of the final return level. However, the difference is small relative to the errors in the marginal estimates, and for much of the coastline is consistently estimated.

Finally note that this aspect complicates any attempt to provide uncertainty estimates of the spatial estimate.

ACKNOWLEDGEMENTS

We thank the Proudman Oceanographic Laboratory and the British Oceanographic Data Centre for supplying the data. M.J.D. and J.A.T. were partly supported by a Ministry of Agriculture, Fisheries and Food grant. J.M.V. was supported by NERC.

REFERENCES

Admiralty Tide Tables (1989). *Volume 1: European Waters including the Mediterranean Sea*, Hydrographer of the Navy.

Coles, S. G. and Pan, F. (1996). 'The analysis of extreme pollution levels: a case study', *Jnl. of Appl. Statist.*, **23**, 333–348.

Coles, S. G. and Tawn, J. A. (1990). 'Statistics of coastal flood prevention', *Phil. Trans. R. Soc. Lond., A*, **332**, 457–476.

Coles, S. G. and Tawn, J. A. (1991). 'Modelling extreme multivariate events', *J. R. Statist. Soc., B*, **53**, 377–392.

David, H. A. (1981). *Order Statistics*, 2nd edn., Wiley, New York.

Dixon, M. J. and Tawn, J. A. (1992). 'Trends in U.K. extreme sea levels: a spatial approach', *Geophys. J. Int.*, **111**, 607–616.

Dixon, M. J. and Tawn, J. A. (1994). *Estimates of Extreme Sea Conditions: Extreme Sea-Levels at the UK A-Class Sites: Site-By-Site Analyses*, Proudman Oceanographic Laboratory internal document no. 65.

Dixon, M. J. and Tawn, J. A. (1995). *Estimates of Extreme Sea Conditions: Extreme Sea-Levels at the UK A-Class Sites: Optimal Site-by-Site Analyses and Spatial Analysis for the East Coast*, Proudman Oceanographic Laboratory internal document no. 72.

Dixon, M. J. and Tawn, J. A. (1998). 'The impact of non-stationarity in extreme sea-level estimation', *Appl. Stat.*, submitted.

Flather, R. A. (1987). 'Estimates of extreme conditions of tide and surge using a numerical model of the north-west European continental shelf', *Estuarine Coastal Shelf Sci.*, **24**, 69–93.

Gumbel, E. J. (1958). *Statistics of Extremes*, Columbia University Press, New York.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press, Cambridge.

Härdle, W. and Marron, J. S. (1985). 'Optimal bandwidth selection in non-parametric regression function estimation', *Ann. Statist.*, **13**, 1365–1381.

Husler, J. (1986). 'Extreme values of non-stationary sequences', *J. Appl. Probab.*, **23**, 937–950.

Leadbetter, M. R., Lindgren, G. and Rootzen, H. (1983). *Extremes and Related Properties of Random Sequences and Series*, Springer, New York.

Lennon, G. W. (1963). 'A frequency investigation of abnormally high tidal levels at certain west coast ports', *Proc. Inst. Civ. Engrs.*, **25**, 451–484.

Liang, K. Y. and Self, S. G. (1996). 'On the asymptotic behaviour of the pseudolikelihood ratio test statistic', *J. R. Statist. Soc. Series B*, **58**, 785–797.

Middleton, J. F. and Thompson, K. R. (1986). 'Return periods of extreme sea-levels from short records', *J. Geophys. Res.*, **91**, 11,707–11,716.

Prescott, P. and Walden, A. T. (1983). 'Maximum likelihood estimation of the parameters of the three-parameter generalised extreme-value distribution from censored samples', *J. Statist. Comput. Simul.*, **16**, 241–250.

Pugh, D. T. (1987). *Tides, Surges and Mean Sea-Level*, Wiley, Chichester.

Pugh, D. T. and Vassie, J. M. (1979). 'Extreme sea-levels from tide and surge probability', in *Proceedings 16th Coastal Engineering Conference, 1978, Hamburg*, American Society of Civil Engineers, New York, vol. 1, pp. 911–930.

Pugh, D. T. and Vassie, J. M. (1980). 'Applications of the joint probability method for extreme sea-level computations', *Proc. Instn. Civ. Engrs., Part 2*, **69**, 959–975.

Rae, J. B. (1988). 'Centralised data collection and monitoring systems for coastal tide gauge measurements', in Kitching, J. A. (ed.), *Tidal Measurement and Instrumentation*, Hydrographic Society Special Publication no. 19, London, pp. 19–25.

Robinson, M. E. and Tawn, J. A. (1997). 'Statistics for extreme sea currents', *Appl. Statist.*, **46**, 183–205.

Smith, R. L. (1986). 'Extreme value theory based on the *r* largest annual events', *J. Hydrol.*, **86**, 27–43.

Smith, R. L. (1989). 'Extreme value analysis of environmental time series: an application to trend detection in ground level ozone', *Statist. Sci.*, **4**, 367–393.

Tawn, J. A. (1988). 'An extreme value theory model for dependent observations', *J. Hydrol.*, **101**, 227–250.

Tawn, J. A. (1992). 'Estimating probabilities of extreme sea-levels'. *Appl. Statist.*, **41**, 77–93.

Tawn, J. A., Dixon, M. J. and Woodworth, P. L. (1994). 'Trends in sea-levels', in Barnett, V. and Turkman, F. K. (eds.), *Statistics for the Environment 2: Water Related Issues*, Wiley, Chichester, pp. 147–181.
Tawn, J. A. and Vassie, J. M. (1989). 'Extreme sea-levels: the joint probabilities method revisited and revised', *Proc. Instn. Civ. Engrs. Part 2*, **87**, 429–442.
Walden, A. T., Prescott, P. and Webber, N. B. (1982). 'An alternative approach to the joint probability method for extreme sea level computations', *Coastal Engineering*, **6**, 71–82.
Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*, Chapman and Hall, London.

