



Hierarchical models for the spatial-temporal carbon nanotube height variations

Jialing Tao, Kaibo Wang, Bo Li, Liang Liu & Qi Cai

To cite this article: Jialing Tao, Kaibo Wang, Bo Li, Liang Liu & Qi Cai (2016) Hierarchical models for the spatial-temporal carbon nanotube height variations, International Journal of Production Research, 54:21, 6613-6632, DOI: [10.1080/00207543.2016.1181809](https://doi.org/10.1080/00207543.2016.1181809)

To link to this article: <http://dx.doi.org/10.1080/00207543.2016.1181809>



Published online: 05 May 2016.



Submit your article to this journal [↗](#)



Article views: 47



View related articles [↗](#)



View Crossmark data [↗](#)

Hierarchical models for the spatial–temporal carbon nanotube height variations

Jialing Tao^a, Kaibo Wang^{a*}, Bo Li^b, Liang Liu^c and Qi Cai^c

^aDepartment of Industrial Engineering, Tsinghua University, Beijing, China; ^bSchool of Economics and Management, Tsinghua University, Beijing, China; ^cTianjin FuNaYuanChuang Technology Co. Ltd, Tianjin, China

(Received 22 May 2015; accepted 15 April 2016)

Carbon nanotubes (CNTs) are allotropes of carbon with a cylindrical nanostructure. Due to their low production cost and potentially high demand, the large-scale production of CNTs is urgently needed and will be highly profitable. However, quality control will be a great challenge to a large-scale production due to the delicate nature of the production process. Among the problems involved in the quality control of CNT array production, height variation is one of the primary concerns. The objective of this study is to model the height along both the spatial and temporal dimensions, so that the height variations can be controlled during the production process, thus improving the quality and stability of CNT arrays. Specifically, the height variation of the CNT arrays is decomposed into macro-scale and micro-scale variations. The macro-scale variation is modelled by state-space and regression models, and the micro-scale variation is modelled as a spatial process. The models successfully capture both the macro-trends and the micro-patterns. A practical case study shows the effectiveness of the proposed models in terms of goodness of fit and prediction accuracy, and the distinction of the models is summarised to aid in choosing a model to apply to other spatial–temporal data modelling problems.

Keywords: quality control; statistical methods; semiconductor manufacturing; spatial representation; spatial–temporal model

1. Introduction

Carbon Nanotube (CNT) arrays are a type of nanostructure material that grows vertically on silicon wafer substrates with a dense alignment. Due to the van der Waals force among the molecules, the arrays can be drawn into continuous yarns or films. With their extraordinary properties, such as their light weight, good electrical and thermal conductivity, and high mechanical strength, the applications of CNT yarns and films are countless, including flexible loudspeakers, touch screens, incandescent displays, etc. (Jiang et al. 2011). Scaling up the CNT production while maintaining stable quality has great business potential, and thus has generated increasing interest in recent years. The nanotechnology is gradually extending from the laboratory and entering a massive production stage. There is a great need for techniques to help improve quality, productivity and reduce cost. We intend to apply statistical methods to address the needs of the nano manufacturing process and demonstrate the effectiveness of production techniques in the new manufacturing environment.

The quality control of CNT arrays is motivated by industrial demands of metal content, single-walled carbon purity, diameter distribution, height of nanotubes, crystalline quality, etc. (Kumar and Ando 2010). Among others, the height of CNT arrays is of great importance as this parameter is influential to other quality characteristics, and the uniformity of height also affects material utility. In practice, reaction time, catalyst film thickness, temperature and vapour pressure are the dominating factors. In the literature, researchers have focused more on engineering experiments of various reaction ingredients and environment conditions to pursue a desirable height of CNT array. However, in a scaled-up nanomanufacturing scenario where real data are available, we believe a data-driven approach will be effective. The joint consideration of growth mechanisms and statistical techniques can enable production engineers and quality experts to better understand and control the height variation and thus improve quality.

To have a better understanding of the variation in height, we first introduce the production procedures of CNT arrays briefly. Currently, chemical vapour deposition (CVD) is one of the most popular production methods for CNT production (Kumar and Ando 2010). A schematic illustration of the CVD equipment is shown in Figure 1. First, wafer substrates that have been coated with a thin catalyst film are placed into a quartz tray, and the tray is put into a tube furnace. Second, certain chemical gas is fed into the tube furnace. The chamber is heated and thus leads to a series of

*Corresponding author. Email: kbwang@tsinghua.edu.cn

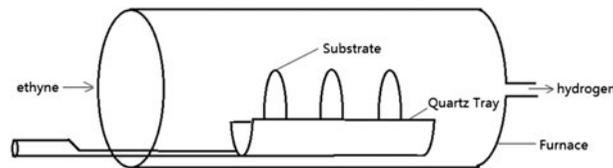


Figure 1. Production system via the CVD method.

chemical reactions; hydrogen is liberated, and carbon atoms are deposited onto the substrate and grow into the form of nanotubes. Variations within a sample may be caused by the competition of reaction resources in neighbourhoods of each nanotube or the diffusion of the reaction gas. In addition, the variation across samples may be caused by the process or external factors, such as gas source or temperature.

In this study, we aim to construct a spatial–temporal model to quantitatively characterise the height variation, with the temporal trend describing the variation among the samples and the spatial trend embodying variations within a sample. Such a model is important for many quality control applications. For instance, the model can be used to develop control charts for process monitoring and to reduce process variation via run-to-run control.

This paper is organised as follows. Section 2 presents a review of the existing spatial–temporal models and estimation methods. In Section 3, the variation is presented and analysed on both the macro- and micro-scales. In Section 4, three spatial–temporal models are proposed, and the corresponding estimation methods are described. Section 5 presents a case study and performance comparison results of proposed models. Finally, conclusions and practical implications are discussed in Section 6.

2. State-of-the-art modelling of spatial–temporal variations

In recent years, spatial–temporal models have been widely applied in various fields, especially those involving remote sensing and monitoring. For example, Smith, Kolenikov, and Cox (2003) modelled a $PM_{2.5}$ network data covering three states in the USA, and Nguyen, Cressie, and Braverman (2012) addressed the massive global aerosol measurements collected by NASA. Theoretical systems of spatial models and time series models are already well developed. However, modelling spatial and temporal data remain a topic of interest.

Almost all of the spatial–temporal models are composed of a macro-scale variation component and a micro-scale variation component. Macro-scale variation includes a general mean, a spatial trend, a temporal trend and an interactive trend (e.g. a site-specific temporal trend), while micro-scale variation contains the spatial autocorrelation or the spatio-temporal autocorrelation of a process. These two components are characterised by different model formats. The macro-scale variations are usually structured with regression and state-space models, and micro-scale variations are often modelled as a zero-mean Gaussian process. In addition, a locally weighted regression model can also be used to interpolate and extrapolate spatial–temporal data. Details of such type of models are unfolded in the following.

Statistical approaches to address the spatial–temporal data focus on two paradigms (Kang, Cressie, and Shi 2010). First, *time* is viewed as an additional dimension in spatial. That is, a d -dimensional spatial process with temporal changes becomes a $(d + 1)$ -dimensional process, and thus can be model as a usually spatial process.

The best-known geostatistics modelling technique, Kriging, is frequently modified and applied to spatial data. The Kriging model yields estimates for a particular site by taking observations of nearby sites as predictors. Cameletti, Ignaccolo, and Bande (2011) used a linear regression term to model the macro-scale variation. Sharples and Hutchinson (2005) adopted an additive regression spline term in their model. In addition, a generalised linear regression model or additive model (Hastie and Tibshirani 1990) is also flexible enough to handle such type of data. Smoothing-based methods are alternatives for modelling spatial–temporal data, although they have a disadvantage of lacking a clear model presentation. Locally weighted regression, a popular model proposed by Cleveland (1979), has been extended to multivariate predictors in Cleveland and Devlin (1988).

The idea of the second paradigm is to address the temporal dimension using dynamic statistical models, thus as ARIMA time-series models, state-space models, etc. In the models proposed by Cameletti, Ignaccolo, and Bande (2011), a latent process AR(1) term was used to embody the time-variant general mean. Fasso and Cameletti (2010) and Cressie, Shi, and Kang (2010) proposed state-space models of two levels: (1) a latent temporal process to model the state transition along time, and then (2) a process loaded by a spatial observation matrix to obtain the output. The model proposed by Cressie, Shi, and Kang (2010) is more widely applicable because the observation matrix can be time-variant. The noticeable advantages of their model are the following: (1) spatial dimension reduction from the

number of observation sites to the number of state-vector dimensions, and (2) rapid iterations using a Kalman smoother and filter.

Comparing the two paradigms above, dynamic statistical models are preferred to spatial models for the following two reasons. First, dynamic statistical models can be derived from knowledge of the time-variant pattern, if known previously, and they are readily interpreted based on how observation changes with time. Second, dynamic statistical models allow for rapid iterations for smoothing and forecasting, which results in computation efficiency. According to Cameletti, Ignaccolo, and Bande (2011), modelled using the same micro-scale variation component and applied to PM_{10} data, dynamic statistical models perform better than linear regression models in their predictability.

To model the micro-scale variation component, assumptions must be primarily built and verified. Excluding the macro-scale trend and studying the innovations, it must first be identified whether purely spatial autocorrelation or spatial-temporal autocorrelation exists, and then the assumptions for modelling the micro-scale variation can be defined. If only the spatial autocorrelation is strong, then the data can be modelled as a purely spatial process; otherwise, the data should be modelled as a spatial-temporal process.

For a purely spatial process, the Gaussian Markov random field (GMRF) is a powerful tool for modelling such information. It is a random vector with finite dimensions following a multivariate Gaussian distribution and conditional independence property. The critical part of GMRF is an inverse covariance matrix, or a precision matrix, that defines the density of the GMRF. The precision matrix controls the influence of nearby points on the target point. There exist many alternatives for modelling covariances: Matern, exponential, Gaussian, etc. Families of spatial covariance functions can be found in Koehler and Owen (1996).

For a spatial-temporal process, Bilonick (1985) proposed an additive variogram as the sum of a spatial variogram and a temporal variogram, which was proven unsatisfactory (Rouhani and Myers 1990). Rodríguez-Iturbe and Mejía (1974) proposed a separable spatial-temporal covariance, which can be written as the product of a purely spatial covariance and a purely temporal covariance. However, the lack of modelling space-time interaction leads to a corresponding limited class of random processes. Cressie and Huang (1999) proposed a non-separable spatial-temporal covariance, which is a generic approach for developing parametric models for spatial-temporal processes. Although the non-separable stationary covariance is more widely applicable than separable covariance, the computation is much more time-consuming.

Aside from the models above, the Bayesian Hierarchical Space-Time model is another alternative to solve spatio-temporal modelling problems (Wikle, Berliner, and Cressie 1998). The model consists of five stages: (1) measurement error process for the observational data, (2) site-specific time series model for the state variable, (3) spatial structures and dynamics, (4) priors of the parameters and (5) hyper-priors. In the model above, the second stage is of the most interest, and it could be regarded as the macro-scale variation construction mentioned in this section, and the third stage corresponds to the micro-scale variation construction.

3. Variation pattern analysis of a real example

The data used in this study are collected from a real CNT production line, including a series of 109 CNT samples. All of the data are taken from the same CVD furnace and at the same position on the quartz tray in the furnace. To construct an appropriate model for the wafers, we first introduce the sampling scheme, analyse the whole data-set and then try to have a good understanding of the variation patterns embedded in the data.

3.1 The sampling scheme

The height values of the CNT arrays are measured at sample. As shown in Figure 2(a), the measuring device takes samples along the x -axis, and then traverses 10 mm along the y -axis to the next line and keep sampling along the x -axis again. In this way, for each sample, the data sampling can be regarded as 19 lines along the x -axis. Figure 2(b) shows a 3-D scatterplot of a typical sample. The bow shape on the projection plane implies that a convexity variation features within the sample. To simplify the problem, the measurement line obtained at $y = 0$, which is the centre one of the 19 lines, is chosen as the study object of each sample, as it captures the convexity and also the general height trend information of the whole wafer. On each centre line, height values are measured at 17 evenly distributed points, as shown by the dots in Figure 2(a).

3.2 Macro-scale variation pattern

Figure 3 shows the main variation pattern of the temporal trend among the samples and the spatial trend within a sample. Figure 3(a) randomly picks 5 out of the 109 samples; the plot reveals a quadric height variation trend in the spatial

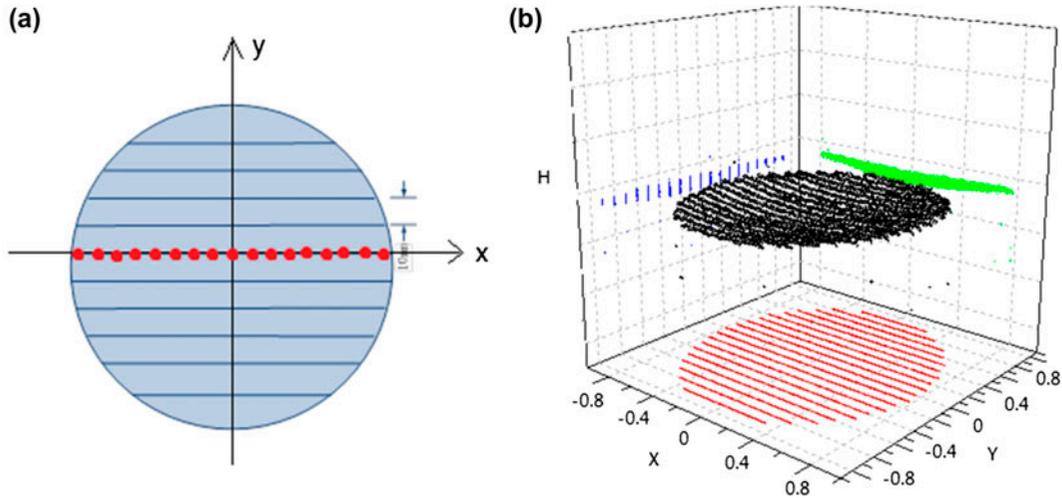


Figure 2. (a) Positions of height sampling points; (b) The 3-D scatterplot of a real sample.

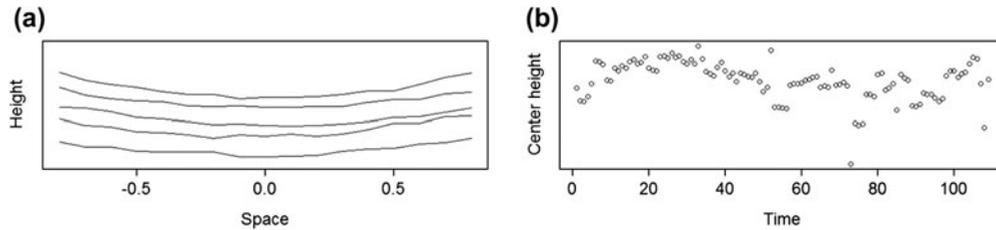


Figure 3. (a) A preview of the spatial pattern; (b) A preview of the temporal pattern.

domain. Figure 3(b) shows the centre height of each line (that is, the height value at the centre point of each line). The centre point can be regarded as the general position of each line. There is no obvious cycling pattern in the plot. Therefore, we concluded from Figure 3 that the macro-scale height variation is mainly dominated by a quadratic spatial variation and an empirical temporal variation.

3.3 Micro-scale variation pattern

To study the micro-scale variation, samples at three time steps are randomly selected. By regression and removing the quadratic spatial trend from the original data, the remaining residuals are used to (1) test the significance of spatial autocorrelation of residuals with Moran’s I statistic, (2) test the significance of serial autocorrelation of residuals with Durbin–Watson statistic, (3) choose a proper spatial correlation function structure (SCF).

First, Moran’s I is used to test the significance of spatial autocorrelation of the residuals. The test was developed by (Moran 1950), and is defined as

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (Z_i - \bar{Z})(Z_j - \bar{Z})}{\sum_i (Z_i - \bar{Z})^2} \tag{1}$$

where N denotes the number of sites (i, j) , w_{ij} denotes the weight of each site, Z denotes the observation variable, and \bar{Z} denotes the average of observations. Moran’s I ranges from -1 to 1 , and larger positive values imply stronger positive correlation, i.e. the closer in space, the more similar in observations. In addition, the p -value can help examine the significance of autocorrelation. Among the 109 samples, 3 samples are randomly chosen for Moran’s I test, and it is implemented with the *ape* package in R software, the results of which are shown in Table 1. Under the confidence level of 95%, a significant spatial positive autocorrelation exists in all the three samples. Hence, strong spatial autocorrelation exists and must be modelled.

Table 1. Statistical test for spatial autocorrelation.

	Sample1	Sample2	Sample3
Moran's I	0.30	0.29	0.28
p-value	9.14e-6	1.71e-5	2.33e-5
$\alpha = 0.05$	Significant	Significant	Significant

Second, the Durbin–Watson test is used to detect the presence of serial autocorrelation in the residuals from a regression analysis (Durbin and Watson 1971). For panel data, they are generalised by (Bhargava, Franzini, and Narendranathan 1982), which is defined as

$$d = \frac{\sum_{i=1}^N \sum_{t=2}^T (e_{i,t} - e_{i,t-1})^2}{\sum_{i=1}^N \sum_{t=1}^T e_{i,t}^2} \tag{2}$$

where $e_{i,t}$ denotes the residual after regression with fixed effects of K regressors, N denotes number of individuals in the panel and T denotes time length. In our case, there is only one regressor, the quadratic univariate spatial term, and thus $K = 1$. From the number of batches, $T = 109$. The DW statistic for our data is 1.75 computed from the formula above. According to significance table, at level 0.05, for $K = 1$ and $T = 109$, the upper and lower critical values are $(d_u, d_l) = (1.70, 1.67)$. Because $d > d_u$ and $(4 - d) > d_u$, there is no statistical evidence that the errors are positively or negatively autocorrelated. Hence, no serial autocorrelation exists, and the micro-scale variation can be safely modelled as a purely spatial process. Thus, the structure of SCF ought to be decided.

Third, to choose a proper SCF structure, we start with explaining the mathematical relationship between the variogram and the SCF, and then compute the empirical variogram and fit the candidate variogram function family to the empirical variogram; finally, we decide the SCF family.

Let $Z(x)$ denote the residual at site x , let h denote the distance between two sites, let σ_Z^2 denote $Var(Z(x))$, let $r(h)$ denote the SCF and let $2\gamma(h)$ denote the variogram.

$$r(h) = Corr(Z(x+h), Z(x)) = \frac{Cov(Z(x+h), Z(x))}{\sigma_Z^2} = \frac{E(Z(x+h)Z(x))}{\sigma_Z^2} \tag{3}$$

$$2\gamma(h) = E[(Z(x+h) - Z(x))^2] = \sigma_Z^2(2 - 2r(h)) \tag{4}$$

From Equation (4), variogram and SCF are structured by the same function family. Empirical variograms were introduced by Cressie and Cassie (1993), and in our case, for each time step, they are computed as

$$2\gamma(h) = \frac{1}{|N(h)|} \sum_{(i,j) \in N(h)} |z_i - z_j|^2, \tag{5}$$

where x_i denotes the i th site ($i = 1, 2, \dots, 17$), $N(h)$ denotes the set of pairs of observations (i, j) such that $|x_i - x_j| = h$, $|N(h)|$ denotes the number of pairs in the set and z_i denotes the residual at site x_i .

The variograms of the three samples are presented in Figure 4. In each plot, the distance ranges from 0 to 1.6 because the sampling sites range from -0.8 to 0.8 in steps of 0.1 . The variograms increases with the distance, which

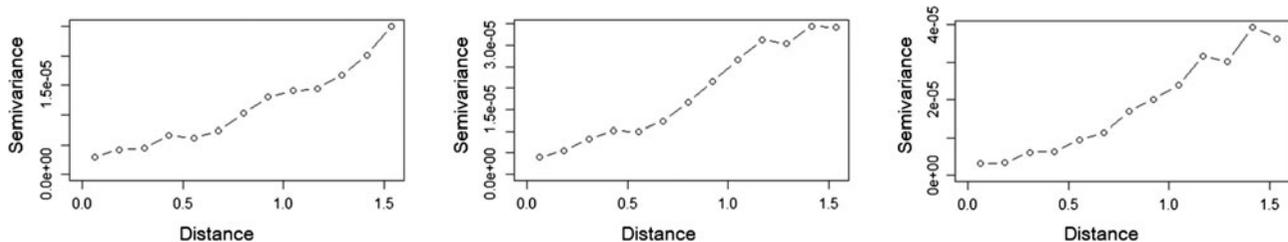


Figure 4. Empirical variograms of residuals at three time steps.

implies that the correlation decreases when the neighbours become further apart. The SCF family of spatial process is chosen as follows.

There are many families of SCF: Matern, exponential, Gaussian, spherical, etc. The exponential and Gaussian families are two commonly used examples of the Matern family. Here, we utilise the geoR package in R software to compute an empirical variogram and fit it against one of the above families. The analysis of real data shows that the Gaussian family is an appropriate choice for the variogram family as well as the SCF family in our case.

From the data analysis, three conclusions are reached. (1) The features of the macro-scale variation include a quadratic trend along the space dimension and an empirical trend along the time dimension. (2) The micro-scale variation can be structured as a spatial process because strong spatial autocorrelation exists but no significant serial autocorrelation exists. (3) The SCF family is assumed as Gaussian, for simplicity, it is structured as follows

$$r(h) = e^{-\theta h^2} \quad (6)$$

4. Modelling spatial–temporal variations

As mentioned previously, both regression models and dynamic statistical models are used for modelling macro-scale variations, and the micro-scale variation component is modelled as a purely spatial process, as analysed in Section 3. In all, three spatial–temporal models are proposed in this section, and the corresponding estimation methods will also be given in detail.

The mind map of modelling is presented in Figure 5, and Table 2 gives a brief summary of the three proposed models. Each model is composed of a macro-scale variation and a micro-scale variation component. The micro-scale variation components in the three models are the same, modelled using a zero-mean Gaussian spatial process. The macro-scale variation components are different. In Model 1, the macro-scale variation component is the sum of the spatial and temporal trends, assuming the effects are separable and additive. From the conclusions of the data analysis, the quadratic spatial trend is modelled using a polynomial regression, and the empirical temporal trend is modelled using regression splines in Model 1A and by a locally weighted regression in Model 1B. In Model 2, the spatial–temporal trend is integrated in a state-space model. The details of the modelling are introduced in this section.

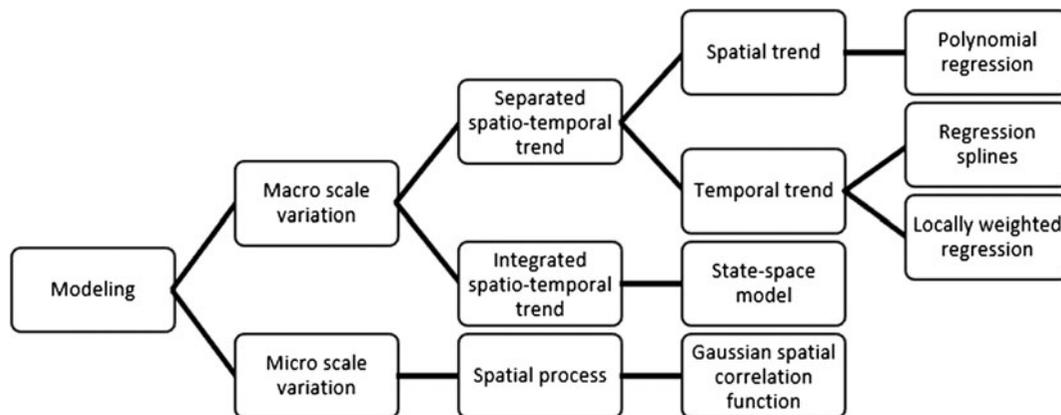


Figure 5. The mind map of modelling strategy.

Table 2. A brief outline of the proposed models.

Contents of each model
Model 1A: Spatial trend (polynomial regression term) + Temporal trend (regression splines term) + Error in spatial process
Model 1B: Spatial trend (polynomial regression term) + Temporal trend (locally weighted regression) + Error in spatial process
Model 2: Spatial–temporal trend (state-space model) + Error in spatial process

4.1 Model 1

4.1.1 Modelling

Model 1 has two varieties, termed Model 1A and Model 1B. Both models are based on the assumption that the spatial trend and the temporal trend are separable and additive. The difference in the models is the modelling method of the empirical temporal trend. Regression splines are used in Model 1A, and locally weighted regression is used in Model 1B. Their commonality is introduced first, and the estimation method of each model is presented.

Model 1 is a two-level hierarchical model

$$Y(s; t) = \mu_s(s) + \mu_t(t) + Z(s) + \varepsilon \tag{7}$$

$$\mu_s(s) = \sum_{m=1}^M \beta_{s_m} s^m \tag{8}$$

Let s denote the sampling sites, let N_s denote the sampling size, let t denote the time steps, let T denote the number of all the time steps and let Y denote the response variable height following a normal distribution. At the first level (Equation (7)), the macro-variation is decomposed into a spatial trend term $\mu_s(s)$, a temporal trend term $\mu_t(t)$, a spatial process term $Z(s)$ and a white noise $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. $Z(s)$ satisfies the following condition: for any s , $Z(s)$ follows a normal distribution, and its expectation and covariance are as follows

$$E(Z(s)) = 0$$

$$Cov(Z(s), Z(s+h)) = \sigma_z^2 r(h), \tag{9}$$

assuming $r(h)$ is a Gaussian family correlation function, as in Equation (6). Thus, the covariance of Y is

$$Cov(Y) = Cov(Z + \varepsilon) = Cov(Z) + Var(\varepsilon) = \sigma_z^2 e^{-\theta h^2} + \sigma_\varepsilon^2 I \tag{10}$$

For the second level, the spatial trend $\mu_s(s)$ is modelled via a polynomial regression of degree M , as in Equation (8), as the global nature of polynomial regression suits the spatial trend observed from the data analysis.

4.1.1.1 Model 1A. At the second level of Model 1, the temporal trend $\mu_t(t)$ is modelled by a cubic B-spline regression term with the intercept as in Equation (11), due to the empirical and local nature observed.

$$\mu_t(t) = \beta_{t_0} + \sum_{i=1}^{i=K+3} \beta_{t_i} f_i(t) \tag{11}$$

The cubic B-spline term defines a special set of piecewise cubic basis functions centred at pre-chosen knots and is a linear combination of the $(K + 3)$ basis functions and the intercept. K denotes the number of knots, $f_i(t)$ denotes the already known cubic basis functions and β_{t_i} denotes the coefficients before the basis functions. The basis functions are generated as follows:

$$B_{j,0}(t) = \begin{cases} 1 & t_j \leq t < t_{j+1} \\ 0 & \text{otherwise} \end{cases}, \tag{12}$$

$$B_{j,p}(t) = \frac{t-t_j}{t_{j+p}-t_j} B_{j,p-1}(t) + \frac{t_{j+p+1}-t}{t_{j+p+1}-t_{j+1}} B_{j+1,p-1}(t)$$

where j denotes the j th span, and p denotes the degree of the function $B(\cdot)$. For example, given knots $(-1, 0, 1)$ (K knots) and boundary knots $(-2, 2)$, the knot vector will be $(-2, -2, -2, -2, -1, 0, 1, 2, 2, 2, 2)$, forming 10 spans and $B_{j,0}$ where $j = 1, 2, \dots, 10$. In the above Equation (12), also known as the Cox-de Boor recursion formula, $B_{j,3}$ can be computed, where $j = 1, 2, \dots, 7$, and there are $(K+3)$ cubic basis functions and the intercept. In Equation (11), $f_i(t) = B_{i,3}(t)$, where $i = j = 1, 2, \dots, K + 3$.

In summary, the number of parameters in Model 1A depends on the degree of the polynomial regression chosen and the number of knots chosen, and there are $(M + K + 7)$ parameters:

$$\psi = \{\beta_{s_1}, \dots, \beta_{s_M} \beta_{t_0}, \beta_{t_1}, \dots, \beta_{t_{(K+3)}}, \sigma_z^2, \theta, \sigma_\varepsilon^2\}$$

4.1.1.2 *Model 1B.* The modelling method of the temporal trend $\mu_t(t)$ is a locally weighted regression in Model 1B. The purpose of the locally weighted regression is to smooth and extrapolate from the historical data. First, the temporal trend history data must be obtained, which, for example, could be the mean of the observations over all spatial space at each time step. In our case, the centre points of the symmetric line that we study are used to investigate the temporal trend. Let p_t denote the temporal trend observation at step t . The smoothing and extrapolation procedure details are described as follows (Cleveland 1979).

Step 1: Define a span (in percentage) as neighbourhood size ratio of the obtained data. Set t_r as the target point. For each t_q in the neighbourhood of $t_r, q = 1, 2, \dots, N_q$, and denote the furthest distance of points in neighbourhood from the target point as Δ_r . Define the weight function as

$$w_q(t_r) = W\left(\frac{|t_q - t_r|}{\Delta_r}\right) = \left[1 - \left(\frac{|t_q - t_r|}{\Delta_r}\right)^3\right]^3 \quad \left(\frac{|t_q - t_r|}{\Delta_r} < 1\right)$$

Step 2: For each t_r , compute the coefficients of the polynomial regression of degree D of p_{t_r} on t_r , which is fitted by the weighted least squares with weight $w_q(t_r)$. Thus, $\hat{\beta}_{qd}$ are the values that minimise $\sum_{q=1}^{N_q} w_q(t_r) \left(p_{t_r} - \sum_{d=0}^D \beta_{qd} t_q^d\right)$. Thus, at a target point $t_r, \hat{p}_{t_r} = \sum_{d=0}^D \hat{\beta}_{qd} t_q^d$.

Step 3: Let $e_r = p_{t_r} - \hat{p}_{t_r}$, and let m_{e_r} denote the median of $|e_r|$. Define the robustness weights as $v_r = \left[1 - \left(\frac{e_r}{m_{e_r}}\right)^2\right]^2$.

Step 4: Compute a new \hat{p}_r for each t_r by fitting a polynomial regression of degree D using weighted least squares with weight $v_q w_q(t_q)$.

Step 5: Repeat Steps 3 and 4 until convergence or the maximum iterations are reached. The final \hat{p}_r is a robust locally weighted regression fitted value.

Although the above procedure is complicated for the robustness of smoothing, the key idea is simply to weight observations in the neighbourhood, with decreasing weights assigned to the further neighbours. For smoothing, the observation at the target points is included in the neighbours and has the highest weight. For extrapolation, the observation at the target point cannot be obtained, and the observation at the point next to it has the highest weight.

Therefore, the smoothed temporal trend history data is regarded as $\mu_t(t)$. By removing it from the spatial-temporal data, the problem remains in estimating the parameters of $(\mu_s(s) + \varepsilon)$ with the new data, which only account for the spatial trend and the spatial process. In addition, the parameters for Model 1B are $\psi = \{\beta_{s_1}, \dots, \beta_{s_M}, \sigma_z^2, \theta, \sigma_\varepsilon^2\}$. Therefore, the estimation solution of Model 1B is included in that of Model 1A, and thus only the estimation method of Model 1A is discussed in the next section.

4.1.2 Parameter estimation

The first level of the model motivates us to split the parameter vector into two vectors

$$\psi_1 = \{\beta_{s_1}, \dots, \beta_{s_M} \beta_{t_0}, \beta_{t_1}, \dots, \beta_{t_{(K+3)}}\}$$

$$\psi_2 = \{\sigma_z^2, \theta, \sigma_\varepsilon^2\}$$

Let

$$B = (s \quad s^2 \quad \dots \quad s^M \quad 1 \quad f_1(t) \quad f_2(t) \quad \dots \quad f_{K+3}(t)).$$

Because the number of sites is N_s , and the number of time steps is T , the dimension of matrix B is $(N_s T) \times (M + K + 4)$.

Let

$$\Sigma(\psi_2) = \text{Cov}(Y),$$

then

$$Y \sim N(B\psi_1, \Sigma(\psi_2)).$$

The log likelihood function is

$$\log L(Y|\psi) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(\|\Sigma(\psi_2)\|) - \frac{1}{2} (Y - B\psi_1)' \Sigma(\psi_2)^{-1} (Y - B\psi_1).$$

The algorithm combines the general procedure introduced by Fang, Li, and Sudjianto (2005) and some modifications recommended by Mardia and Marshall (1984), and the primary idea is to iteratively update the values of $\hat{\psi}_1$ and $\hat{\psi}_2$.

Step 1: Give the initial parameter $\hat{\psi}_2$.

Step 2: Update $\hat{\psi}_1$.

According to the generalised least square estimator,

$$\psi_1 = (B' \Sigma^{-1}(\psi_2) B)^{-1} B' \Sigma^{-1}(\psi_2) Y$$

Step 3: Update $\hat{\psi}_2$.

Mardia and Marshall (1984) suggested using the one-step Fisher scoring algorithm to update the value of ψ_2 .

$$\hat{\psi}'_2 = \hat{\psi}_2 + I_2^{-1}(\hat{\psi}) V_2(Y : \hat{\psi})$$

$V_2(Y; \hat{\psi})$ is the score function with respect to $\hat{\psi}_2$:

$$\begin{aligned} V_2(Y : \hat{\psi}) &= -\frac{1}{2} \frac{\partial}{\partial \hat{\psi}_2} \log \|\Sigma(\hat{\psi}_2)\| - \frac{1}{2} (Y - B\hat{\psi}_1)' \frac{\partial}{\partial \hat{\psi}_2} \Sigma^{-1}(\hat{\psi}_2) (Y - B\hat{\psi}_1) \\ \frac{\partial}{\partial \hat{\psi}_2^{(k)}} \log \|\Sigma(\hat{\psi}_2)\| &= \sum_{j=1}^m \sum_{i=1}^m (\Sigma^{-1}(\hat{\psi}_2))_{i,j} \left(\frac{\partial}{\partial \hat{\psi}_2^{(k)}} \Sigma(\hat{\psi}_2) \right)_{i,j} \\ \frac{\partial}{\partial \hat{\psi}_2^{(k)}} \Sigma^{-1}(\hat{\psi}_2) &= \Sigma^{-1}(\hat{\psi}_2) \left(\frac{\partial}{\partial \hat{\psi}_2^{(k)}} \Sigma(\hat{\psi}_2) \right) \Sigma^{-1}(\hat{\psi}_2) \end{aligned}$$

where $\hat{\psi}_2^{(k)}$ is the k th parameter in $\psi_2 = \{\sigma_z^2, \theta, \sigma_\varepsilon^2\}$.

$I_2(\psi)$ is the Fisher information matrix with respect to ψ_2 . It has been proven that

$$I(\psi) = \begin{pmatrix} I_1(\psi) & 0 \\ 0 & I_2(\psi) \end{pmatrix},$$

where

$$I_1(\psi) = Y' \Sigma^{-1}(\psi_2) Y,$$

and the (i, j) th element of $I_2(\psi)$ is

$$[I_2(\psi)]_{i,j} = \frac{1}{2} \text{tr} \left(\Sigma(\psi_2)^{-1} \frac{\partial}{\partial \psi_2^{(i)}} \Sigma(\psi_2) \Sigma^{-1}(\psi_2) \frac{\partial}{\partial \psi_2^{(j)}} \Sigma(\psi_2) \right)$$

Step 4: Run Step 2 and Step 3 until convergence.

4.2 Model 2

4.2.1 Modelling

Model 2 is a three-level hierarchical model; the main part is a state-space equation for the macro-scale variations.

$$Y(s; t) = \mu(s; t) + Z(s) + \varepsilon \tag{13}$$

$$\mu(s; t) = S(s)\eta_t \tag{14}$$

$$\eta_{t+1} = G\eta_t + \tau + \delta \tag{15}$$

The role of the first level of Model 2 (Equation (13)) is equivalent to that in Model 1. For the second level (Equation (14)), the macro-scale variation component is modelled as the product of a time-variant state vector η_t and a loading matrix $S(s)$. The loading matrix $S(s)$ is a combination of the spatial basis functions. Moreover, the mean and variance of η_0 are denoted as m_{η_0} and Σ_{η_0} , respectively. The third level of Model 2 (Equation (15)) is a latent temporal process of state η_t : AR(1) with a transition matrix G and an intercept τ . δ is white noise that follows the distribution $N(0, \Sigma_\delta)$.

The error $e(s) = Z(s) + \varepsilon$ has a zero-mean Gaussian distribution with covariance matrix as given by Equation (10). For the convenience of the estimation, the covariance matrix is represented as $\Sigma_e = \sigma_z^2 \Gamma(h)$. Γ is the following scaled spatial covariance function

$$\Gamma_{\zeta, \theta}(h) = \begin{cases} 1 + \zeta & h = 0 \\ r(h) & h > 0 \end{cases} \tag{16}$$

where $\zeta = \frac{\sigma_\eta^2}{\sigma_\varepsilon^2}$ and $r(h)$ follow Equation (6). For positive definiteness reasons, it is preferable to estimate $\log(\zeta)$ instead of σ_ε^2 .

In summary, the parameters in Model 2 are

$$\psi = \{m_{\eta_0}, \Sigma_{\eta_0}, G, \tau, \Sigma_\delta, \sigma_z^2, \theta, \log(\zeta)\}$$

4.2.2 Parameter estimation

We here propose to estimate Model 2 using an EM algorithm that is also embedded with an NR algorithm. The model is modified based on the estimation algorithms proposed by Fasso and Cameletti (2010), and some iteration equations are developed by utilising the work of Deng and Shen (1997) in our model. The estimation algorithm is implemented by modifying the Stem package developed by Michela Cameletti in R software.

We assume having the complete data of both the latent process η_t and the response variable Y_t , where $i = 1, 2, \dots, T$. T denotes the number of time steps. Let $D = \{\eta, Y\}$ denote the complete data, and then the complete log-likelihood is given by

$$\begin{aligned} \log L(D) \propto & -\frac{T}{2} \log |\Sigma_e| - \frac{1}{2} \sum_{t=1}^T (Y_t - \mu_t) \Sigma_e^{-1} (Y_t - \mu_t)' \\ & - \frac{1}{2} \log |\Sigma_{\eta_0}| - \frac{1}{2} (\eta_0 - m_{\eta_0}) \Sigma_{\eta_0}^{-1} (\eta_0 - m_{\eta_0})' \\ & - \frac{T}{2} \log |\Sigma_\delta| - \frac{1}{2} \sum_{t=1}^T (\eta_t - G\eta_{t-1} - \tau) \Sigma_\eta^{-1} (\eta_t - G\eta_{t-1} - \tau)' \end{aligned} \tag{17}$$

The EM algorithm consists of an E and an M step. The E step computes the conditional expectation of the complete log-likelihood function $\log L(D)$ on the observation matrix Y , and the updated ψ is given by maximising $E[\log L(D)|Y]$ at the M step.

4.2.2.1 E step. The conditional expectation of the log-likelihood function is computed as follows:

$$\begin{aligned} -2Q(D) &= -2E[\log L(D)|Y] \\ &= T \log |\Sigma_e| + \text{tr}\{\Sigma_e^{-1} \sum_{t=1}^T [(Y_t - S\eta_t^T)(Y_t - S\eta_t^T)' + SP_t^T S']\} \\ &\quad + \log |\Sigma_{\eta_0}| + \text{tr}\{\Sigma_{\eta_0}^{-1} [(\eta_0 - m_{\eta_0})(\eta_0 - m_{\eta_0})' + P_0^T]\} \\ &\quad + T \log |\Sigma_\delta| + \text{tr}\{\Sigma_\delta^{-1} [S_{11}^* - S_{10}^* G' - GS_{10}^* + GS_{00}^* G']\} \end{aligned} \tag{18}$$

where

$$\begin{aligned} S_{00}^* &= \frac{\sum_{t=1}^T (\eta_{t-1}^T \eta_{t-1}^{T'} + P_{t-1}^T)}{T} \\ S_{10}^* &= \frac{\sum_{t=1}^T [(\eta_t^T - \tau) \eta_{t-1}^{T'} + P_{t,t-1}^T]}{T}, \\ S_{11}^* &= \frac{\sum_{t=1}^T [(\eta_t^T - \tau)(\eta_t^T - \tau)' + P_t^T]}{T} \end{aligned} \tag{19}$$

and $\eta_t' = E(\eta_t | Y_t)$, $P_{t_1, t_2}' = E\{(\eta_{t_1} - \eta_{t_1}')(\eta_{t_2} - \eta_{t_2}')'\}$, $P_t' = P_{t,t}'$. Plus, $\eta_1^0 = m_{\eta_0}$, $P_1^0 = \Sigma_{\eta_0}$. And $P_{t,t-1}^T$, η_t^T , P_t^T are computed using the Kalman smoother with the k th iteration of parameters $\psi^{(k)}$ as the true value. The computation details, except the Kalman smoother, are described in Appendix 1. For details of the Kalman smoother, refer to Shumway and Stoffer (2010).

4.2.2.2 *M step*. The parameters are updated as follows, and the computation details are presented in Appendix 2.

$$\begin{aligned} (G \quad \tau) &= \left(S_{10} \quad \frac{1}{T} \sum_{t=1}^T \eta_t^T \right) \left(\begin{array}{cc} S_{00} & \frac{1}{T} \sum_{t=1}^T \eta_{t-1}^T \\ \frac{1}{T} \sum_{t=1}^T \eta_{t-1}^{T'} & 1 \end{array} \right)^{-1}, \\ \Sigma_\delta &= S_{11} - (G \quad \tau) \left(S_{10} \quad \sum_{t=1}^T \eta_t^T \right)' \end{aligned} \tag{20}$$

where

$$\begin{aligned} S_{00} &= \frac{\sum_{t=1}^T (\eta_{t-1}^T \eta_{t-1}^{T'} + P_{t-1}^T)}{T} \\ S_{10} &= \frac{\sum_{t=1}^T (\eta_t^T \eta_{t-1}^{T'} + P_{t,t-1}^T)}{T} \\ S_{11} &= \frac{\sum_{t=1}^T (\eta_t^T \eta_t^{T'} + P_t^T)}{T} \end{aligned} \tag{21}$$

According to the conditional maximisation approach (McLachlan and Krishnan 2007), the solution of $\frac{\partial Q(\psi; \psi^{(k)})}{\partial \psi} = 0$ is approximated by partitioning $\psi = \{\psi_1, \psi_2\}$. $\psi_1 = \{m_{\eta_0}, G, \Sigma_\delta, \sigma_z^2\}$. Holding $\psi_2 = \{\Sigma_{\eta_0}, \theta, \log(\zeta)\}$ constant, the $(k + 1)$ th iterations of m_{η_0} and σ_z^2 are given by

$$\begin{aligned} m_{\eta_0}^{(k+1)} &= \eta_0^T \\ (\sigma_z^2)^{(k+1)} &= \frac{(\sigma_z^2)^{(k)}}{Tn} \text{tr}(\Sigma_e^{-1} W) \end{aligned}$$

where

$$W = \sum_{t=1}^T [(Y_t - S\eta_t^T)(Y_t - S\eta_t^T)' + SP_t^T S'] \tag{22}$$

At the k th iteration of the EM algorithm, ψ_1 is updated to $\psi_1^{(k)}$ according to the algorithm above. Because there are no closed forms for the remaining parameters, $\psi_2 = \{\Sigma_{\eta_0}, \theta, \log(\zeta)\}$, the NR algorithm is used for minimising

$$Q_1(\psi_2) = T \log |\Sigma_e| + \text{tr}\{\Sigma_e^{-1} W\}$$

The updating formula for the i th iteration of the inner NR algorithm is given by

$$\psi_2^{(i+1)} = \psi_2^{(i)} - H_{\psi_2 = \psi_2^{(i)}}^{-1} \times \Delta_{\psi_2 = \psi_2^{(i)}}$$

where H and Δ are the Hessian matrix and the gradient vector of $Q_1(\psi_2)$, respectively. The update step repeats until the NR algorithm converges, and then the $\psi_2^{(k)}$ and the complete $\psi^{(k)}$ are obtained. The details of the NR algorithm are reported in Appendix 3.

The E and M steps are repeated until the EM algorithm converges, and final parameter estimates of $\hat{\psi}$ are obtained.

4.2.3 Prediction

The prediction of the height at the sampling sites at next time step is based on Kalman-filtered data, η_t^t , P_t^t ($t = 1, 2, \dots, T$), which was previously obtained during the estimation.

$$\begin{aligned} \eta_t^t &= E(\eta_t | Y_t) \\ P_t^t &= E\{(\eta_t - \eta_t^t)(\eta_t - \eta_t^t)'\} \end{aligned}$$

The prediction of the latent process can be computed as follows:

$$\begin{aligned} \eta_{T+1}^T &= \hat{G}\eta_T^T + \hat{\tau} \\ P_{T+1}^T &= \hat{G}P_T^T\hat{G}' + \hat{\Sigma}_\delta \end{aligned}$$

Thus, the prediction of the response variable is $Y_{T+1} = S\eta_{T+1}^T$. For details of the Kalman filter, refer to Shumway and Stoffer (2010).

5. Case study

In this section, the three models introduced above, with their specific formats that are applicable to the real data-set, are also described here. The performances of these models are evaluated based on both goodness of fit and prediction accuracy.

5.1 Model specifications

Based on the generalised models presented in Section 4 and the conclusions of the data analysis in Section 3, Model 1A, Model 1B and Model 2 are considered in this section. The specifications correspond to the macro-scale variation component of each model.

5.1.1 Model 1

The spatial trend is quadratic and symmetric, as observed in data preview, and thus, the spatial trend term in both Model 1A and Model 1B is

$$\mu_s(s) = s^2$$

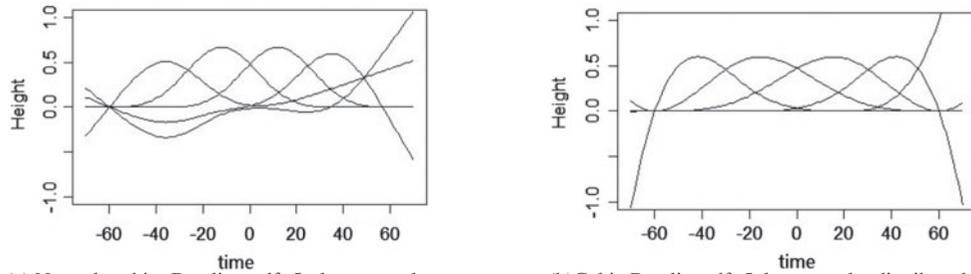
To decide the specifications of the trend component in Model 1A, the following three components are critical: the spline function (i.e. cubic B-spline or natural cubic B-spline), the degree of freedom and the position of knots chosen. Here we explain how these components are determined.

First, consider the difference between natural and cubic splines. The Natural cubic B-spline differs from the cubic B-spline by a linear limit near boundaries and leads to greater stability near or beyond the boundaries. Holding their degrees of freedom and knot positions the same, Figure 6(a) and (b) shows that the greatest difference appears around the boundaries ($-60, 60$); the natural cubic B-spline function is linear out of boundaries while cubic B-spline function can deviate far away. Thus, when predicting with cubic B-spline, the prediction site must be within the boundaries, and natural cubic B-spline is more conservative as for the boundary concern.

Second, consider the degrees of freedom (df). In our case, with the natural cubic B-spline and evenly distributed knot positions, Figure 6(a) and (c) plot the basis functions when df is set as three and five. Figure 6(e) plots the fitted temporal trend of centre height, and the comparison shows that the fitting performance is better when the degree of freedom is 5 than that when it is 3. When the degree of freedom becomes higher, the fitted line will generally be curlier.

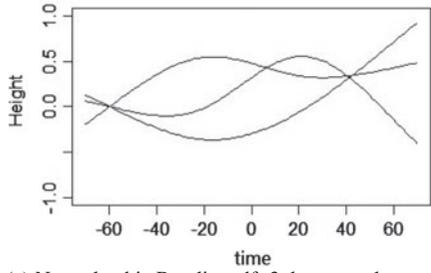
Third, consider the knot positions. Appropriately chosen knot position will help to reach a better fitness. The default choice is evenly distributed between the boundaries, while in our case the knot positions are chosen as those where observation fluctuates greatly (i.e. (5, 50, 70, 80, 100)). Figure 6 shows that the latter choice leads to better fitness where the fluctuations are (but please note that it may not lead to improved prediction).

Therefore, the temporal term in Model 1A is modelled using the natural cubic B-spline, with the degree of freedom set to five. When computing the natural cubic B-spline basis function evaluated at each time step (e.g. $t = 1, 2, \dots, 109$), the predictor input should better be symmetric (e.g. $t = -54, \dots, 54$). The boundary knots are set as ($-60, 60$), with the knots evenly distributed for better prediction performance.

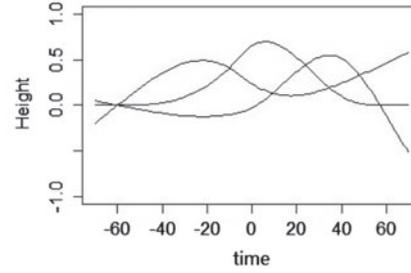


(a) Natural cubic B-spline, $df=5$, knot evenly distributed

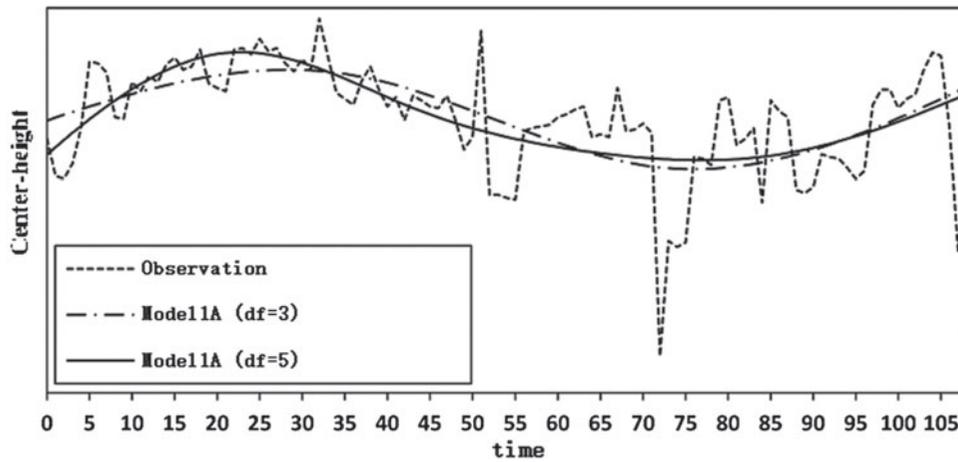
(b) Cubic B-spline, $df=5$, knot evenly distributed



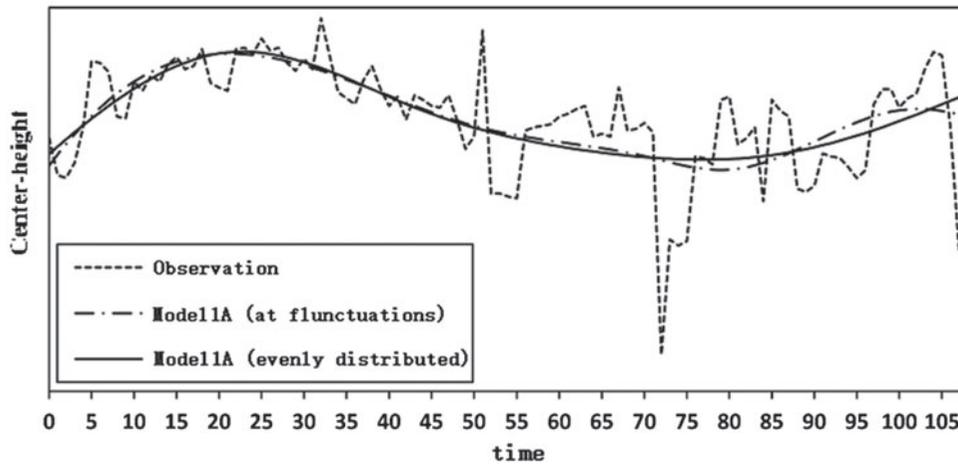
(c) Natural cubic B-spline, $df=3$, knot evenly distributed



(d) Cubic B-spline, $df=3$, knot at fluctuations



(e) A comparison of model fitness under different degrees of freedom



(f) A comparison of model fitness under different knot positions

Figure 6. Comparison of choices of spline functions, degrees of freedom and knot positions.

The temporal term in Model 1B is modelled based on a locally weighed regression involving a degree of inner polynomial regression and a span percentage, as mentioned in Section 4.1.1 Model 1B. Here, the degree is set as 1, and the span is set as 0.6 via trial and error.

For Model 2, the spatial and temporal trends are integrated. The loading matrix and the state vector are defined as

$$\eta_t = (a_t \quad c_t)'$$

$$S(s) = \begin{pmatrix} s^2 & 1 \end{pmatrix},$$

where the convexity a_t and the position c_t of a quadratic symmetric curve are exacted as features in the state vector η_t . Assuming no independent changing pattern of a_t and c_t , the transition matrix G is a time-invariant diagonal matrix $diag(G_a, G_c)$ and the covariance matrix of states is $\Sigma_\delta = diag(\sigma_{\delta a}^2, \sigma_{\delta c}^2)$. The independence assumption is feasible for our problem. From the chemical and physical knowledge of CNT growth, the degree of convexity (a_t) depends on the internal distribution of the catalyst that costs the film on the substrates, whereas the general position (c_t) relies on external environment variables, such as temperature and gas volume. Loaded with $S(s)$, the macro-scale variation component $\mu(s; t) = S(s)\eta_t$ embodies a time-variant quadric symmetric curve.

5.2 Estimation result and performance comparison

All of the estimation algorithms are implemented using R software. The parameters of the three models and a base model are shown in Table 3. In addition, the performances are compared based on the goodness of fit and the prediction accuracy.

To show the effectiveness of our models, simple exponential smoothing is chosen as a base model without consideration of the hierarchical nature of the spatial-temporal variation. At each location is a time series of 109 steps. All of the time series are smoothed with the same parameter, α , which can also be viewed as smoothing a vector, $Y_t = (Y_{s1}, Y_{s2}, \dots, Y_{sN_s})$, with parameter α , as follows:

$$\hat{Y}_1 = Y_1$$

$$\hat{Y}_{t+1} = \alpha Y_t + (1 - \alpha) \hat{Y}_t$$

The smoothed data of the 109 steps are used to evaluate the goodness of fit, and the smoothed data of the last 59 steps are used to evaluation the accuracy of the prediction. In our case, $\alpha = 0.9$ for the best goodness of fit.

The performance evaluation is based on two factors: the goodness of fit and the prediction accuracy. For both factors, two popular measures are used: the residual sum of squares (RSS) and the mean absolute percentage error (MAPE).

$$RSS = \sum_{t=1}^T \sum_{s=1}^{N_s} (Y - \hat{Y})^2$$

$$MAPE = \frac{1}{N_s T} \sum_{t=1}^T \sum_{s=1}^{N_s} \frac{|Y - \hat{Y}|}{Y}$$

Table 3. Parameter estimates.

Model 1A						
β_0	β_s	β_t	σ_ε^2	σ_z^2	θ	
0.225	0.037	(0.052, 0.028, -0.019, 0.171, 0.128)	7.75e-4	2.48e-6	0.062	
Model 1B						
β_s	σ_ε^2	σ_z^2		θ		
0.038	4.45e-4	3.23e-6		0.7		
Model 2						
G	τ	Σ_δ	σ_ε^2	σ_z^2	θ	
diag(0.579, 0.512)	diag(0.017, 0.139)	diag(4.77e-6, 3.82e-4)	2.3e-4	4.5e-6	6.59e-3	
m_{η_0}	Σ_{η_0}					
(-0.012, 0.163)	diag(1e-6, 1e-6)					

Table 4 summarises the performance of the proposed models and the base model. The prediction accuracy is based on one-step-ahead predictions and a total of 59 predictions. Model 2 performs the best both for goodness of fit and prediction, and Model 1B provides a good fit but poor prediction. The performance of Model 1A does not exceed the base model regarding either factor. The analyses of their performances are as follows.

Table 4. Goodness of fit and prediction accuracy of the models.

	Fit (109 steps)		Prediction (59 steps: one step ahead)	
	RSS	MAPE (%)	RSS	MAPE (%)
Exp smoothing	1.635	7.08	1.428	9.35
Model 1A	1.439	7.81	1.533	12.07
Model 1B	0.829	5.78	1.503	11.40
Model 2	0.021	0.97	1.244	9.30

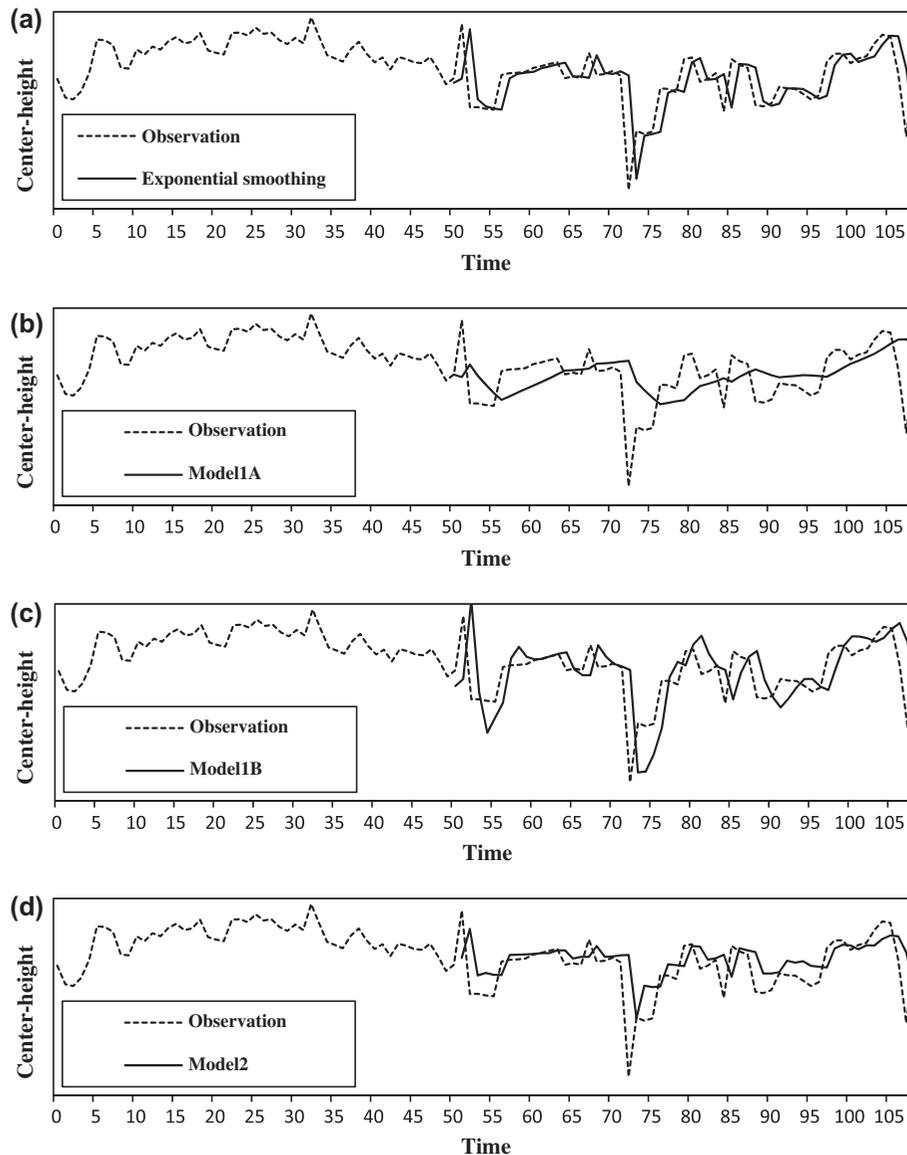


Figure 7. Predictions of the temporal trend.

Table 5. Advantages and disadvantages of the proposed models.

	Advantage	Disadvantage
Model 1A	Smooth in prediction; can provide multi-step-ahead prediction	Ignorant of local features and time-variant spatial patterns; unsatisfactory fitness and prediction performance
Model 1B	Good fitness performance, flexible in capturing local features	Ignorant of time-variant spatial patterns; can only provide one-step-ahead prediction
Model 2	Great fitness and prediction performance; consideration of time-variant spatial patterns; can provide multi-step-ahead prediction	Number of parameters to be estimated is large

It is seen from Table 4 that Model 2 fits best because the temporal trend fits the latent process very well, and there is an error adjustment scheme in each iteration of the Kalman filter and the smoother. Model 1B fits better than Model 1A due to the local nature resulting from the locally weighted regression, and thus the trend fitted by Model 1A is smoother than Model 1B. If the degree of freedom of the regression splines chosen in Model 1A is higher, the curve fluctuates more globally.

Next, we compare the three proposed models in terms of their prediction accuracy. It should be noted that Model 1B only performs a one-step-ahead prediction, whereas the other two models are capable of multiple-step-ahead prediction. The prediction accuracy primarily depends on the prediction of the temporal trend (i.e. the prediction of height at the centre point). Figure 7 compares the temporal prediction trends of each model. The more global nature of Model 1A results in a smooth prediction and thus, unsatisfactory prediction accuracy. The base model, Model 1B and Model 2 predictions all exhibit a lag effect, and there are two primary differences, as follows. (1) Model 1B cannot quickly recover from an outlier and return to normal levels due to the regression nature, such as the predictions at the steps 70–80. (2) Due to the transition scheme of the latent process in Model 2, the general position the curve quickly recovers to normal, and the fitted value on which the prediction depends is very accurate in Model 2.

6. Conclusions

CNTs possess great physical and thermal properties and are produced using low-cost materials but via a delicate procedure. Hence, the large-scale production of CNT arrays with stable quality holds great commercial value. Current problems are that (1) the height of a CNT array is concave within the piece, which interrupts the continuousness when drawing CNT film or yarns, and (2) the height lacks uniformity among samples of different batches. To describe the height variations within and among samples, three spatial–temporal models are proposed, and their performances are evaluated based on CNTs.

The models proposed in Section 4 are generalised and can be applied to other spatial–temporal modelling problems. As summarised in Table 2, the three models can all be decomposed into macro-scale variation and micro-scale variation components. The micro-variation component is modelled as a spatial process. For the macro-scale variation, Models 1A and 1B assume separable and additive spatial and temporal trends, with the spatial trend modelled as polynomial regression term in both models, and the temporal trend modelled as a regression splines term in Model 1A and as a locally weighted regression term in Model 1B. The macro-variation in Model 2 is modelled as a state-space model, integrating both spatial and temporal trends. The estimation algorithms for each model are also presented in detail, which are primarily the combination of EM and NR algorithms.

Based on the specifications of the model components, the models are applied and evaluated for goodness of fit and prediction accuracy. A base model of exponential smoothing is added to the comparison to provide a more complete evaluation. Table 5 summarises the advantages and disadvantages of each model to provide some clues for choosing spatial–temporal models. Models 1A and 1B are regression models and are thus smoother than Model 2. Regarding the local nature, Model 1B responds more rapidly and greatly to changes compared to Model 1A. Model 2 describes the trend more delicately than Model 1B due to the error fix step in the Kalman filter and the smoother, and the transition scheme allows for a quicker recovery from outliers compared to Model 1B. The degree of freedom of the three models (the number of parameters required to estimate) is generally ordered as Model 2 > Model 1A > Model 1B, although the degree of freedom also depends on the degree of regression or model structure that is chosen. Model 2 is more widely applicable than the other two models because the modelling of the macro-scale variation allows for a time-variant spatial pattern. That is, in our case, the convexity of the curve can change with time. Regarding the computation, unlike other two models, the estimation of Model 1B is not integrated due to the limitation of the model structure of the temporal

trend, and the temporal trend must be estimated first and then the spatial part. Regarding functionality of prediction, Model 1A and Model 2 can make multi-step-ahead predictions, whereas Model 1B is only capable of one-step-ahead prediction.

In conclusion, Model 2 (state-space model) performs the best for CNT spatial–temporal height modelling, and can also be applied to many other spatial–temporal data. Moreover, though only the centre line of the height surface is used in modelling for simplicity, as mentioned in Section 3.1, the fits or prediction can be extended to location-dependent heights other than centre-line heights by spinning the centre line height around and obtaining the whole surface. In addition, this paper modified the existing state-space model for spatial–temporal data by adding a intercept in the AR(1) process expressed by Equation (15), and correspondingly modified the existing algorithm of parameter estimation with detailed steps and proof in Appendices 1 and 2. The modification extends the original model proposed by Fasso and Cameletti (2010), and allows more flexibility and wider application of the state-space model. Finally, the paper applied some common existing models (spline regression, locally weighted regression) to compare the fitting and prediction performance in the CNT case, and hopefully it will provided some hints for readers facing similar spatial–temporal problems.

The main focus of this paper is modelling variation patterns of CNT arrays, thus providing researchers and practitioners a better understanding of the process. In practice, more important questions that exist in this process are still left untouched. For example, given a process that is characterized by the spatial-temporal models studied above, how to adjust process inputs so that its output has minimised variation. As another example, after a good understanding of the natural variation the profiles, how to detect abnormal process changes through an implementation of control charts. All these questions are interesting topics that deserve future research efforts.

Acknowledgements

We would like to thank the editor and three anonymous reviewers for their constructive comments, which have helped us improve this work greatly.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work is supported by the National Natural Science Foundation of China (NSFC) [grant number 71471096].

References

- Bhargava, Alok, Luisa Franzini, and Wiji Narendranathan. 1982. “Serial Correlation and the Fixed Effects Model.” *The Review of Economic Studies* 49 (4): 533–549.
- Bilonick, Richard A. 1985. “The Space-time Distribution of Sulfate Deposition in the Northeastern United States.” *Atmospheric Environment (1967)* 19 (11): 1829–1845.
- Cameletti, Michela, Rosaria Ignaccolo, and Stefano Bande. 2011. “Comparing Spatio-temporal Models for Particulate Matter in Piemonte.” *Environmetrics* 22 (8): 985–996.
- Cleveland, William S. 1979. “Robust Locally Weighted Regression and Smoothing Scatterplots.” *Journal of the American Statistical Association* 74 (368): 829–836.
- Cleveland, William S., and Susan J. Devlin. 1988. “Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting.” *Journal of the American Statistical Association* 83 (403): 596–610.
- Cressie, Noel A. C., and Noel A. Cassie. 1993. *Statistics for Spatial Data*. Vol. 900. New York: Wiley.
- Cressie, N., and H. C. Huang. 1999. “Classes of Nonseparable, Spatio-temporal Stationary Covariance Functions”. *Journal of the American Statistical Association* 94: 1330–1339
- Cressie, Noel, Tao Shi, and Emily L. Kang. 2010. “Fixed Rank Filtering for Spatio-temporal Data.” *Journal of Computational and Graphical Statistics* 19 (3): 724–745.
- Deng, Li, and Xuemin Shen. 1997. “Maximum Likelihood in Statistical Estimation of Dynamic Systems: Decomposition Algorithm and Simulation Results.” *Signal Processing* 57 (1): 65–79.
- Durbin, James, and G. S. Watson. 1971. “Testing for Serial Correlation in Least Squares Regression. III.” *Biometrika* 58 (1): 1–19.
- Fang, Kai-Tai, Runze Li, and Agus Sudjianto. 2005. *Design and Modeling for Computer Experiments*. Boca Raton: Chapman & Hall/CRC.

- Fasso, Alessandro, and Michela Cameletti. 2010. "A Unified Statistical Approach for Simulation, Modeling, Analysis and Mapping of Environmental Data." *Simulation* 86 (3): 139–153.
- Harville, David A. 1997. *Matrix Algebra from a Statistician's Perspective*. Vol. 157. New York: Springer.
- Hastie, Trevor J., and Robert J. Tibshirani. 1990. *Generalized Additive Models*, Vol. 157. Boca Raton: Chapman & Hall/CRC.
- Jiang, Kaili, Jiaping Wang, Qunqing Li, Liang Liu, Changhong Liu, and Shoushan Fan. 2011. "Superaligned Carbon Nanotube Arrays, Films, and Yarns: A Road to Applications." *Advanced Materials* 23 (9): 1154–1161.
- Kang, Emily L., Noel Cressie, and Tao Shi. 2010. "Using Temporal Variability to Improve Spatial Mapping with Application to Satellite Data." *Canadian Journal of Statistics* 38 (2): 271–289.
- Koehler, J. R., and A. B. Owen. 1996. "Computer Experiments." *Handbook of Statistics* 13 (13): 261–308.
- Kumar, Mukul, and Yoshinori Ando. 2010. "Chemical Vapor Deposition of Carbon Nanotubes: A Review on Growth Mechanism and Mass Production." *Journal of Nanoscience and Nanotechnology* 10 (6): 3739–3758.
- Mardia, Kanti V., and R. J. Marshall. 1984. "Maximum Likelihood Estimation of Models for Residual Covariance in Spatial Regression." *Biometrika* 71 (1): 135–146.
- McLachlan, Geoffrey, and Thriyambakam Krishnan. 2007. *The EM Algorithm and Extensions*. Vol. 382. Hoboken, NJ: John Wiley & Sons.
- Moran, Patrick A. P. 1950. "Notes on Continuous Stochastic Phenomena." *Biometrika* 37: 17–23.
- Nguyen, Hai, Noel Cressie, and Amy Braverman. 2012. "Spatial Statistical Data Fusion for Remote Sensing Applications." *Journal of the American Statistical Association* 107 (499): 1004–1018.
- Rodríguez-Iturbe, Ignacio, and José M. Mejía. 1974. "The Design of Rainfall Networks in Time and Space." *Water Resources Research* 10 (4): 713–728.
- Rouhani, Shahrokh, and Donald E. Myers. 1990. "Problems in Space-time Kriging of Geohydrological Data." *Mathematical Geology* 22 (5): 611–623.
- Sharples, J. J., and M. F. Hutchinson. 2005. "Spatio-temporal Analysis of Climatic Data Using Additive Regression Splines." In *MODSIM 2005 International Congress on Modelling and Simulation*, edited by Z. Erger, A. and R.M. Argent. 1695–1701. Queensland: Modelling and Simulation Society of Australia and New Zealand.
- Shumway, Robert H., and David S. Stoffer. 2010. *Time Series Analysis and Its Applications: With R Examples*. New York: Springer Science & Business Media.
- Smith, Richard L., Stanislav Kolenikov, and Lawrence H. Cox. 2003. "Spatiotemporal Modeling of PM_{2.5} Data with Missing Values." *Journal of Geophysical Research: Atmospheres (1984–2012)* 108: (D24).
- Wand, M. P. 2002. "Vector Differential Calculus in Statistics." *The American Statistician* 56 (1): 55–62.
- Wikle, Christopher K., L. Mark Berliner, and Noel Cressie. 1998. "Hierarchical Bayesian Space-time Models." *Environmental and Ecological Statistics* 5 (2): 117–154.

Appendix 1. E step computation in the estimation method of Model 2

The computation details of the E step of the EM algorithm for estimating Model 2 are outlined below, including the derivation of Equation (18) from Equation (17). In the study of Fasso and Cameletti (2010), there exists a similar conditional expectation of the log-likelihood function. However, our model is extended by adding an intercept term in the temporal process of the state (Equation (15)), and the computation must be modified accordingly.

$$\begin{aligned}
 -2Q(D) &= -2E[\log L(D)|Y] = Q_1(D) + Q_2(D) + Q_3(D) \\
 Q_1 &= T \log |\Sigma_e| + E\left(\sum_{t=1}^T (Y_t - \mu_t) \Sigma_e^{-1} (Y_t - \mu_t)' | Y\right) \\
 &= T \log |\Sigma_e| + tr\left\{\Sigma_e^{-1} \sum_{t=1}^T E[(Y_t - \mu_t)(Y_t - \mu_t)' | Y]\right\} \\
 &= T \log |\Sigma_e| + tr\left\{\Sigma_e^{-1} \sum_{t=1}^T [(Y_t - S\eta_t^T)(Y_t - S\eta_t^T)' + SP_t^T S']\right\}
 \end{aligned} \tag{23}$$

The first equivalence in Equation (23) is based on $aVa' = tr(Vaa')$, and the second is based on $Var(X) = E[(X - E(X))(X - E(X))'] = E[(XX') - (E(X))E(X)']$. Here we provide additional details of the second equivalence. By replacing X with $Y_t - \mu_t$, we obtain

$$\begin{aligned}
 &E[(Y_t - \mu_t)(Y_t - \mu_t)' | Y] \\
 &= E[(Y_t - \mu_t) | Y]E[(Y_t - \mu_t)' | Y] + E[(Y_t - \mu_t) - E(Y_t - \mu_t)](Y_t - \mu_t - E(Y_t - \mu_t))' | Y] \\
 &= (Y_t - S\eta_t^T)(Y_t - S\eta_t^T)' + Var(Z_t | Y) \\
 &= (Y_t - S\eta_t^T)(Y_t - S\eta_t^T)' + SP_t^T S'
 \end{aligned}$$

Similar to Q_1 ,

$$\begin{aligned}
 Q_2 &= \log |\Sigma_{\eta_0}| + E[(\eta_0 - m_{\eta_0}) \Sigma_{\eta_0}^{-1} (\eta_0 - m_{\eta_0})' | Y] \\
 &= \log |\Sigma_{\eta_0}| + tr\{\Sigma_{\eta_0}^{-1} [(\eta_0 - m_{\eta_0})(\eta_0 - m_{\eta_0})' + P_0^T]\}
 \end{aligned}$$

$$\begin{aligned}
 Q_3 &= T \log |\Sigma_\delta| + E \left[\sum_{t=1}^T (\eta_t - G\eta_{t-1} - \tau) \Sigma_\delta^{-1} (\eta_t - G\eta_{t-1} - \tau)' | Y \right] \\
 &= T \log |\Sigma_\delta| + tr \left\{ \Sigma_\delta^{-1} \sum_{t=1}^T E[(\eta_t - G\eta_{t-1} - \tau)(\eta_t - G\eta_{t-1} - \tau)' | Y] \right\}
 \end{aligned}$$

In addition,

$$\begin{aligned}
 &E[(\eta_t - G\eta_{t-1} - \tau)(\eta_t - G\eta_{t-1} - \tau)' | Y] \\
 &= (\eta_t^T - G\eta_{t-1}^T - \tau)(\eta_t^T - G\eta_{t-1}^T - \tau)' + E[(\eta_t - G\eta_{t-1} - \eta_t^T + G\eta_{t-1}^T)(\eta_t - G\eta_{t-1} - \eta_t^T + G\eta_{t-1}^T)' | Y] \\
 &= \eta_t^T \eta_t^{T'} - G\eta_{t-1}^T \eta_{t-1}^{T'} - \tau \eta_t^{T'} - \eta_t^T \eta_{t-1}^{T'} G' + G\eta_{t-1}^T \eta_{t-1}^{T'} G' + \tau \eta_{t-1}^{T'} G' - \eta_t^T \tau' + G\eta_{t-1}^T \tau' + \tau \tau' \\
 &\quad + P_t^T - GP_{t,t-1}^{T'} - P_{t,t-1}^T G' + GP_{t-1}^T G' \\
 &= S_{11}^* - S_{10}^* G' - GS_{10}^* + GS_{00}^* G'
 \end{aligned}$$

where $S_{11}^*, S_{10}^*, S_{00}^*$ are defined as Equation (19).

Therefore, Equation (18) is obtained.

Appendix 2. M step computation in the estimation method of Model 2

The development of the updating Equations (20) is based on Deng and Shen (1997)’s study. They proposed a generalised EM algorithm to estimate a state-space model, which is equivalent to the macro-scale variation component of our Model 2.

$$\begin{aligned}
 (G \ \tau) &= \left(\sum_{t=1}^T E(\eta_t \eta_{t-1}' | Y) \quad \sum_{t=1}^T E(\eta_t | Y) \right) \begin{pmatrix} \sum_{t=1}^T E(\eta_{t-1} \eta_{t-1}' | Y) & \sum_{t=1}^T E(\eta_{t-1} | Y) \\ \sum_{t=1}^T E(\eta_{t-1}' | Y) & T \end{pmatrix}^{-1} \\
 &= \left(\sum_{t=1}^T (\eta_t^T \eta_{t-1}^{T'} + P_{t,t-1}^T) \quad \sum_{t=1}^T \eta_t^T \right) \begin{pmatrix} \sum_{t=1}^T (\eta_{t-1}^T \eta_{t-1}^{T'} + P_{t-1}^T) & \sum_{t=1}^T \eta_{t-1}^T \\ \sum_{t=1}^T \eta_{t-1}^{T'} & T \end{pmatrix}^{-1} \\
 &= \left(S_{10} \quad \frac{1}{T} \sum_{t=1}^T \eta_t^T \right) \begin{pmatrix} S_{00} & \frac{1}{T} \sum_{t=1}^T \eta_{t-1}^T \\ \frac{1}{T} \sum_{t=1}^T \eta_{t-1}^{T'} & 1 \end{pmatrix}^{-1} \\
 \Sigma_\delta &= \frac{1}{T} \sum_{t=1}^T E(\eta_t \eta_t' | Y) - (G \ \tau) \begin{pmatrix} \frac{1}{T} \sum_{t=1}^T E(\eta_t \eta_{t-1}' | Y) & \frac{1}{T} \sum_{t=1}^T E(\eta_t | Y) \\ \frac{1}{T} \sum_{t=1}^T E(\eta_{t-1}' | Y) & 1 \end{pmatrix}' \\
 &= S_{11} - (G \ \tau) \begin{pmatrix} S_{10} & \sum_{t=1}^T \eta_t^T \\ \sum_{t=1}^T \eta_{t-1}^{T'} & T \end{pmatrix}'
 \end{aligned}$$

where S_{11}, S_{10}, S_{00} are defined as Equation (21).

Appendix 3. NR iteration in the estimation method of Model 2

The details of the NR algorithm are equivalent to those reported in Fasso and Cameletti (2010)’s study, except that the correlation function $r(h)$ in the micro-scale variation component in our Model 2 is Gaussian instead of exponential.

Let $\psi_{2(j)}$ for Σ_{η_0} ($j = 1$), θ ($j = 2$), and $\log(\zeta)$ ($j = 3$). From the standard matrix differential rules (Harville 1997; Wand 2002),

$$\frac{\partial Q_1}{\partial \psi_2} = T tr \left(\Gamma^{-1} \frac{\partial \Gamma}{\partial \psi_2} \right) - \frac{1}{\sigma_\varepsilon^2} tr \left(\Gamma^{-1} \frac{\partial \Gamma}{\partial \psi_2} \Gamma^{-1} W \right),$$

where W follows Equation (22). In addition,

$$\begin{aligned}
 \frac{\partial \Gamma}{\partial \theta} &= [1 - I_0(h)] \times \frac{\partial r(h)}{\partial \theta} \\
 \frac{\partial \Gamma}{\partial \log(\zeta)} &= I_0(h) \times \exp[\log(\zeta)]
 \end{aligned}$$

where $I_0(h)$ is an indicator function returning 1 for $h = 0$ and 0 elsewhere. For the second-order derivative,

$$\begin{aligned} \frac{\partial^2 Q_1}{\partial \psi_{2(i)} \partial \psi'_{2(i)}} &= \text{Tr} \left(\Gamma^{-1} \frac{\partial^2 \Gamma}{\partial \psi_{2(i)} \partial \psi'_{2(i)}} \right) - \text{Tr} \left(\Gamma^{-1} \frac{\partial \Gamma}{\partial \psi_{2(i)}} \Gamma^{-1} \frac{\partial \Gamma}{\partial \psi'_{2(i)}} \right) \\ &\quad - \frac{1}{\sigma_z^2} \text{tr} \left(\Gamma^{-1} \frac{\partial^2 \Gamma}{\partial \psi_{2(i)} \partial \psi'_{2(i)}} \Gamma^{-1} W \right) - \frac{2}{\sigma_z^2} \text{tr} \left(\Gamma^{-1} \frac{\partial \Gamma}{\partial \psi_{2(i)}} \Gamma^{-1} \frac{\partial \Gamma}{\partial \psi'_{2(i)}} \Gamma^{-1} W \right). \end{aligned}$$

Furthermore,

$$\frac{\partial^2 \Gamma}{\partial \theta \partial \theta'} = [1 - I_0(h)] \times \frac{\partial^2 r(h)}{\partial \theta \partial \theta'};$$

For $\log(\varsigma)$, it holds that $\frac{\partial^2 \Gamma}{\partial^2 \log(\varsigma)} = \frac{\partial \Gamma}{\partial \log(\varsigma)}$. Finally, considering that $\frac{\partial^2 \Gamma}{\partial \log(\varsigma) \partial \theta}$ is a null matrix, the second-order mixed derivative is

$$\frac{\partial^2 Q_1}{\partial \theta \partial \log(\varsigma)} = -\text{Tr} \left(\Gamma^{-1} \frac{\partial \Gamma}{\partial \theta} \Gamma^{-1} \frac{\partial \Gamma}{\partial \log(\varsigma)} \right) + \frac{2}{\sigma_z^2} \text{tr} \left(\Gamma^{-1} \frac{\partial \Gamma}{\partial \theta} \Gamma^{-1} \frac{\partial \Gamma}{\partial \log(\varsigma)} \Gamma^{-1} W \right)$$

The first-order and second-order derivative of Γ respect to θ depends on the form of $r(h)$ (exponential, Gaussian, etc.)