

## RESEARCH ARTICLE

# Missing data imputation for paired stream and air temperature sensor data

Han Li | Xinwei Deng | Eric Smith

Department of Statistics, Virginia Tech,  
Blacksburg, 24061, VA, U.S.A.

**Correspondence**

Eric Smith, Department of Statistics, Virginia  
Tech, 406A Hutcheson Hall, Blacksburg, VA  
24061, U.S.A.

Email: epsmith@vt.edu

Stream water temperature is an important factor in determining the impact of climate change on hydrologic systems. Near continuous monitoring of air and stream temperatures over large spatial scales is possible due to inexpensive temperature recorders. However, missing water temperature data commonly occur due to the failure or loss of equipment. Missing data creates difficulties in modeling relationships between air and stream water temperatures. It also imposes challenges if the objective is an analysis, for example, clustering streams in terms of the effect of changes in water temperature. In this work, we propose to use a novel spatial–temporal varying coefficient model to impute missing water temperatures. Modeling the relationship between air and water temperature over time and space increases the effectiveness of imputing the missing water temperatures. A parameter estimation method is developed, which utilizes the temporal covariation in the relationship, borrows strength from neighboring stream sites, and is useful for imputing sequences of missing data. A simulation study is conducted to examine the performance of the proposed method in comparison with several existing imputation methods. The proposed method is applied to cluster streams with missing water temperatures into groups from 156 streams with meaningful interpretations.

**KEYWORDS**

missing water temperature, spatiotemporal, streaming clustering, varying coefficient

## 1 | INTRODUCTION

Water temperature is a determining factor in water quality and may be one of the most important inputs in modeling the impact of climate change on hydrologic systems (Beitinger, Bennett & McCauley, 2000; Caissie, 2006; Chadwick, Moore & Green, 1995; Flebbe, Roghair & Bruggink, 2006; Keleher & Rahel, 1996; Meisner, 1990; Minns, Randall, Chadwick, Moore & Green, 1995; Sinokrot & Stefan, 1993). An efficient way to study water temperature is through a real-time monitoring system with modern sensors for inexpensive acquisition of temperature data (Dunham, Gwynne, Reiman & Martin, 2005; Huff, Hubler & Borisenko, 2005; Wang et al., 2013). The low cost of sensor equipment allows for intensive temporal monitoring of water and air temperatures over a large spatial region (Hudy, Thieling, Gillespie & Smith, 2008; O'Donnell, Rushworth, Bowman, Scott & Hallard, 2014; Trumbo et al., 2014). Such data are valuable for relating

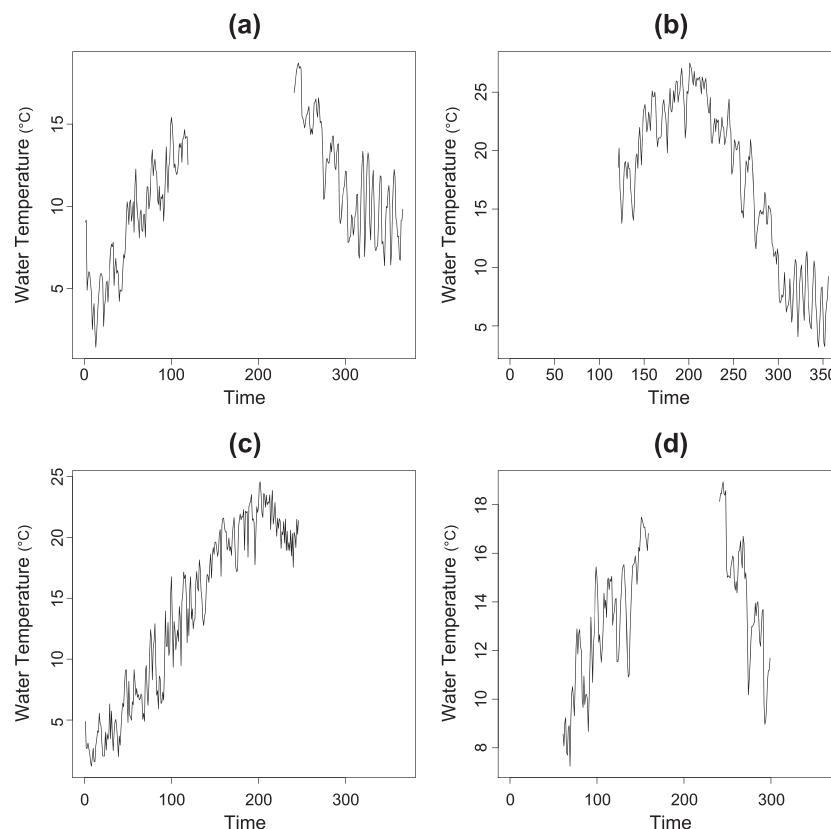
changes in water temperature to changes in the health and status of aquatic species. Brook trout, for example, prefers cooler water found in higher elevation streams, and temperatures greater than 21°C are viewed as highly stressful to the health of trout (Meisner, 1990; Beitinger et al., 2000). As part of a study on water temperature and brook trout, over 150 paired (air and water) thermographs (HOBO Watertemp Pro v2; accuracy 0.2°C; drift < 0.1 annually (Onset Computer Corporation, 2009)) were placed at the pour point of randomly selected stream catchments in southeast USA. A detailed explanation of the data collection procedure can be found in Li, Deng, Kim & Smith (2014).

Here, we focus on the daily maximum water and air temperatures collected at 156 sites in 2011. Daily maximum water temperature is used to summarize the daily effect of increased water temperature as increased maximum temperature is likely to result in increased stress for Brook trout Trumbo et al. (2014). A problem with these sensors,

especially those in the water is that there is sensor failure. Although efforts are made to monitor the sensors, it is difficult and due to the failure of equipment, there are only around 25% of the streams having a complete record of stream water temperatures. Missing values for water temperatures, especially large sequences of missing values, are not uncommon and may create issues with analysis. If data on water temperatures are missing in the summer period, it may be difficult to make proper inference about the survival of trout. Missing values during the spring season might affect modeling of survival of young. While missing data imputation is thus important, imputing missing values in water temperature is challenging, especially when there are missing values over extended time. Figure 1 shows four time series of daily maximum water temperatures collected from the sensors at different stream sites for a year (Time Point 1 is January 1). Clearly, there are large segments of missing values for the water temperature that will complicate multistream analysis. It is however challenging to impute missing temperatures from individual time series because there is not sufficient information from the neighboring time points.

In this work, we propose to impute missing values in water temperature by considering a novel spatiotemporal varying coefficient model (STVCM). The varying coefficient model (VCM) was first proposed by Hastie and Tibshirani (1993). The methodology has been advanced theoretically by several authors (Fan & Zhang, 2008; Hoover, Rich, Wu & Yang,

1998; Huang, Wu & Zhou, 2002; Wu, Chiang & Hoover, 1998), and it has been used in many applications (Cheng, Zhang & Chen, 2009; Ferguson, Bowman, Scott & Carvalho, 2007; Ferguson, Bowman, Scott & Carvalho, 2009; Li et al., 2014). The VCM usually is based on a linear relationship between the response and covariates in which the coefficients vary as a function of other variables. The missing values in the response can be imputed by the estimated VCM equation. Although there are many recent studies on VCM with missing values in the response (Huang et al. 2015; Xu & Zhu, 2013; Zhao & Xue, 2011), few of them consider environmental data or use location in the varying coefficients thus do not take advantage of the spatial correlation of the response. To take the spatial correlation of the response into consideration, the VCM was extended to the STVCM using various estimation methods (Lu, Steinskog, Tjøstheim, & Yao, 2009; Serban, 2011). However, those methods are not adequate for the situation where there are large sequences of missing values in the response curves. To address the challenge, the proposed method (denoted as STVCM) focuses on estimating the spatially temporal varying coefficients when there are missing values in the response. The proposed STVCM uses a linear model form that relates water and air temperatures, although the model coefficients vary with space and time. An important feature of the environmental data in this work is that the correlation over time is considerably higher than the correlation over space. To accommodate this property,



**FIGURE 1** Data with large sequences of missing water temperatures during the monitoring period: (a) missing at the middle, (b) missing at the beginning, (c) missing at the end, (d) missing at multiple segments

we adopt a polynomial spline method to model the temporal effect and use a local kernel method to fit the spatial effect. Polynomial splines with parsimonious expression can easily capture the temporal correlation in water temperature (Li et al., 2014). The spatially based kernel method borrows strength from neighboring sites to impute missing values. For the bandwidth in the kernel method, an adaptive nearest neighbor bandwidth selection is used to include the site that is most relevant to the target site (Fan & Gijbels, 1996). It leads to an adaptive, self-learning fitting algorithm with accurate missing data imputation.

There are two commonly used methods for imputing missing water temperature: using spatiotemporal correlation of water temperatures at multiple sites and using covariates to impute water temperature. For the first method, the water temperatures are measured at multiple locations over time thus are spatially and temporally correlated. Water temperatures from neighboring sites or nearby time points can be used to impute missing values. For the second method, covariates such as air temperature are useful for imputing water temperature. Colocated air sensors or air temperature measurements from weather networks may have a complete data record and would be expected to be strongly correlated with water temperature. Therefore, those measurements can be used to impute water temperature, potentially with reasonable accuracy (Mohseni et al., 1998, Webb, Clack & Walling, 2003). The proposed method takes advantage of both types of approaches thus can effectively model the paired air and stream water temperatures when there are groups of missing values in the response.

Conventional methods for modeling water temperature with missing data include the linear regression model (Neumann, Rajagopalan, & Zagona, 2003) and nonlinear logistic regression model (Mohseni et al., 1998; Mohseni & Stefan, 1999) using air temperature as a covariate. These methods have meaningful interpretation. However, such approaches overlook the spatial and temporal effects and may not be appropriate for modeling water temperature over a large spatial region. Advanced models are also used in missing data imputation. For example, the Gaussian process based on spatial correlation is widely used in the literature (Cressie, 1993; Cressie & Wikle, 2011). The Gaussian process model can capture both the water–air relationship in the mean component and the spatiotemporal correlation of water temperature in the covariance component. However, it is often challenging to deal with large data sets due to the computational difficulty, especially when the Gaussian process model is applied to bivariate or multivariate time series data (Kaufman, Schervish & Nychka, 2008). The neural network (NN) is another possible method for analysis of spatiotemporal environmental data (Coulibaly & Evora, 2007). NN are flexible in modeling the water–air relationship. As a black box for the inputs and outputs, it does not provide a clear explanation of the relationship or help with interpretation of the spatiotemporal correlation.

The imputed data are useful for a number of applications. For example, risk metrics that are based on the entire or part of the time series may be used to rank streams and predict streams that might be at risk for losing trout (Trumbo et al., 2014). Rankings might be used to identify streams that would be candidates for restoration or preservation. Clustering of streams based on air–water temperature relationships is also useful for identifying groups of streams. An approach to clustering was proposed in Li, Deng, Dolloff and Smith (2016); however, the method was limited by the lack of complete data. Here, the imputed data will be used with the clustering approach of Li et al. (2016) to identify groups of streams related to the air–water temperature relationship.

The remainder of this work is organized as follows. Section 2 details the proposed method. Simulation studies are conducted in Section 3 to examine the performance of the proposed method. Section 4 applies the proposed method to impute the missing values in the real data for 156 trout streams in the eastern United States (Trumbo et al., 2014). The clustering of streams using the imputed data is conducted to illustrate the effectiveness of investigating the characteristics of streams. We conclude this work with some discussion in Section 5.

## 2 | THE PROPOSED METHOD

### 2.1 | Spatiotemporal varying coefficient model

Denote by  $T$  the number of time points and  $S$  the number of sites. Let  $W_{s,t}$  be the maximum daily water temperature and  $A_{s,t}$  be the maximum daily air temperature for site  $s$  at time  $t$ ,  $t = 1, 2, \dots, T$ ,  $s = 1, 2, \dots, S$ . We propose a STVCM for the air–water temperature relationship as

$$W_{s,t} = \theta_0(s, t) + A_{s,t}\theta_1(s, t) + \epsilon_{s,t}, \quad (1)$$

where  $\theta_0(s, t)$  and  $\theta_1(s, t)$  are varying intercept and slope coefficients and  $\epsilon_{s,t}$  is the error term in the model. We assume  $E(\epsilon_{s,t}) = 0$  and  $\text{var}(\epsilon_{s,t}) = \sigma^2$ . For the inference in Section 2.3, we assume that the error term  $\epsilon_{s,t}$  is independently normally distributed. Here,  $\theta_0(s, t)$  can be viewed as an intercept function in terms of  $s$  and  $t$ . Similar interpretation can be applied for  $\theta_1(s, t)$ .

Note that there is strong correlation over time for water temperatures in the data we use, but relatively weak correlation over space due to the effects of other landscape characteristics of streams and the degree of spatial separation. To capture most of the variation in maximum water temperature, the temporal effect needs to be emphasized. Note that this is in contrast with recent stream network methodologies where there is a strong spatial component (Ver Hoef, Peterson, Clifford, & Shah, 2014; Rushworth, Peterson, Ver Hoef, & Bowman, 2015). Here, the stream sites are not part of a stream network but rather dispersed over a broad geographic region.

We thus model the varying coefficients as

$$\theta_0(s, t) = \sum_{j=1}^K \alpha_j(s) b_j(t), \quad \theta_1(s, t) = \sum_{j=1}^K \beta_j(s) b_j(t), \quad (2)$$

where  $\{b_1(t), \dots, b_K(t)\}$  are a set of  $K$  time dependent basis functions chosen for the temporal effect. It means that the model in (1) can be expressed as

$$W_{s,t} = \sum_{j=1}^K \alpha_j(s) b_j(t) + \sum_{j=1}^K \beta_j(s) b_j(t) A_{s,t} + \epsilon_{s,t}, \quad (3)$$

Here,  $\alpha_j(s)$ ,  $\beta_j(s)$ ,  $j = 1, 2, \dots, K$  are viewed as coefficients of the basis functions  $b_j(t)$ , and they vary with the spatial index,  $s$ . In this sense, we consider the spatial effect as a varying coefficient of the time effect. Hence, the proposed model in (6) has a linear model form for water temperature with respect to the covariates

$$\mathbf{x}_s(t) \equiv [b_1(t), \dots, b_K(t), A_{s,t} b_1(t), \dots, A_{s,t} b_K(t)]', \quad (4)$$

and the corresponding space coefficients as

$$\boldsymbol{\gamma}(s) \equiv [\alpha_1(s), \dots, \alpha_K(s), \beta_1(s), \dots, \beta_K(s)]'. \quad (5)$$

Clearly, the coefficients  $\boldsymbol{\gamma}(s)$  in (5) can be viewed as varying coefficients of the coefficients  $\theta_0(s, t)$  and  $\theta_1(s, t)$ . For a fixed  $s$ , define a response vector  $\mathbf{W}_s = (W_{s,1}, \dots, W_{s,T})'$ , a regression matrix  $\boldsymbol{\Gamma}_s = (\mathbf{x}_s(1), \dots, \mathbf{x}_s(T))'$ , and an error vector  $\boldsymbol{\epsilon}_s = (\epsilon_{s,1}, \dots, \epsilon_{s,T})'$ . Then the model in (6) can be written in a matrix form as

$$\mathbf{W}_s = \boldsymbol{\Gamma}_s \boldsymbol{\gamma}(s) + \boldsymbol{\epsilon}_s. \quad (6)$$

For the temporal effects  $\mathbf{x}_s(t)$  in (4), we adopt the quadratic spline method in Li et al. (2014) with the following basis functions:

$$\{b_1(t), \dots, b_K(t)\} = \{1, t, t^2, (t - \xi_1)_+^2, \dots, (t - \xi_N)_+^2\}, \quad (7)$$

where,  $\xi_1, \xi_2, \dots, \xi_N$  are  $N$  knots and  $(t - \xi_n)_+$ ,  $n = 1, 2, \dots, N$  are the splines with  $(t - \xi_n)_+ = t - \xi_n$  if  $t \geq \xi_n$  and  $(t - \xi_n)_+ = 0$  if  $t < \xi_n$ . The splines can easily capture the strong correlation over time associated with the change in water temperature. The rationale for using a quadratic spline ( $m = 2$ ) is from a combination of good cross-validation and model interpretation (smoothness of the fitted curve). The Mallows's  $C_p$  (Mallows, 1973) was used as a selection statistic. We compared the results in one site to the cases with a linear spline ( $m = 1$ ) and a cubic spline ( $m = 3$ ). A cubic spline ( $m = 3$ ) results in a larger  $C_p$  value compared to the quadratic spline ( $m = 2$ ). The linear spline ( $m = 1$ ) gives a smaller  $C_p$  value but results in a curve that is less smooth. Therefore, the choice of a quadratic spline is based on a combination of curve fitting, model parsimony and selection statistic. To select the number of knots  $N$ , we follows the strategy in Li et al. (2014) and choose  $N = 4$  and set the location of the knots evenly distributed over the time range. We compared the fit using  $N = 3$ ,  $N = 5$ , and  $N = 12$  knots with different degrees of

polynomials for the spline models. Using  $N = 4$  knots with quadratic splines results in a smooth fitted curve, good interpretation in terms of seasonal effect, and a good  $C_p$  statistic.

For the spatial coefficients in (5), we consider the local kernel method for estimation (Fan & Zhang, 2008) rather than the spline approach for the spatial coefficients in (5). Because the spatial correlation of water temperature may not be strong (i.e., two sites located close to each other may not necessarily have a higher correlation than two distant sites), a spline basis for the spatial effect (based on latitude and longitude) would make the shape of the fitted varying coefficients unreliable and highly dependent on the chosen bases. The use of a local kernel method results in more flexibility for modeling the spatial correlation between water temperature at different stream sensor sites. Specifically, we use the local constant method and express, for a fixed site  $u$ :

$$\alpha_j(u) = \alpha_{j,u}, \beta_j(u) = \beta_{j,u}, u = 1, 2, \dots, S. \quad (8)$$

For a given  $u$ , we define  $B_k(u)$  as a set containing  $k$  neighboring sites of  $u$ . To estimate the parameters  $\boldsymbol{\alpha}_u = (\alpha_{1,u}, \dots, \alpha_{K,u})'$  and  $\boldsymbol{\beta}_u = (\beta_{1,u}, \dots, \beta_{K,u})'$  in the proposed STVCM, we propose to minimize the following objective function,

$$L(\boldsymbol{\alpha}_u, \boldsymbol{\beta}_u) = \sum_t \sum_{s \in B_k(u)} \left[ W_{s,t} - \sum_{j=1}^K \alpha_{j,u} b_j(t) - \sum_{j=1}^K \beta_{j,u} b_{j,u}(t) A_{s,t} \right]^2 \times G_h(\|u - s\|) + \lambda (\|\boldsymbol{\alpha}_u\|_2^2 + \|\boldsymbol{\beta}_u\|_2^2), \quad (9)$$

where  $\lambda \geq 0$  is a smoothing parameter,  $\|\cdot\|_2$  is the vector  $L_2$  norm, and  $\|u - s\|$  is the distance between site  $u$  and site  $s$ .  $G_h(x)$  is the kernel function with bandwidth  $h$ . In this study, we use the Epanechnikov kernel  $G_h(x) = \frac{3}{4}(1 - x^2)_+ I_{|x| < h}$  (Ruppert, Wand, & Carroll, 2003). Clearly, the estimation of  $\boldsymbol{\alpha}_u$  and  $\boldsymbol{\beta}_u$  can be obtained through standard quadratic programming. One can also consider using weighted least squares for parameter estimation. However, because of the penalty terms  $\lambda (\|\boldsymbol{\alpha}_u\|_2^2 + \|\boldsymbol{\beta}_u\|_2^2)$  in (9), the weighted least squares estimation approach needs to be solved under the ridge regression setting. Note that water temperatures from different streams typically contain missing values over different time ranges. For the proposed model to achieve accurate imputation, it is crucial to properly choose  $B_k(u)$  and the bandwidth  $h$ . Here, we adopt the adaptive nearest neighbor selection in a spirit similar to that in Fan and Gijbels (1996). In choosing  $B_k(u)$ , we include neighboring sites that are similar to the target site and also have complete data for the period when data are missing in the target site. Denote by  $n_i$  the number of observed records at time  $i$  for water temperatures from a set of  $k$  sites. Then we find  $B_k(u)$  to be the set of  $k$  sites closest to the target site  $u$  satisfying  $\min n_i > 0$ . Thus, for a bandwidth  $h$ , it can be selected adaptively as

$$h = \max_{s \in B_k} \|u - s\|, \quad (10)$$

where  $\|u - s\|$  is the distance between site  $s$  and site  $u$ .

It is worth pointing out that the choice of the local constant method in (9) instead of a local linear or other complex expression results from the pursuit of a parsimonious model. The data from other sites may not directly improve the imputation of water temperature but may help to impute the missing information over the time. In other words, it would be preferable to use a small bandwidth  $h$  such that we select only a few sites from the neighborhood to borrow information.

## 2.2 | Tuning parameter selection

For the proposed method, there are two tuning parameters  $k$  and  $\lambda$ . Because of a relatively weak spatial correlation pattern in the water temperatures, we fix  $k = 1$  resulting in a “1-nearest-neighborhood” method. A large  $k$  may result in inaccurate imputation because a large number of sites with different landscape characteristics may not be helpful for fitting the model. To select  $\lambda$ , we use the generalized cross-validation (GCV) method (Wahba, 1990). The  $GCV(\lambda)$  is defined as

$$GCV(\lambda) = \sum_{s=1}^S \left( \hat{\mathbf{W}}_s - \mathbf{W}_s \right)' \left( \hat{\mathbf{W}}_s - \mathbf{W}_s \right) / (1 - \text{tr}(\mathbf{S}_\lambda)/T), \quad (11)$$

where  $\mathbf{W}_s = (W_{s,1}, \dots, W_{s,T})'$  is the observed water temperature and  $\hat{\mathbf{W}}_s$  is the imputation of  $\mathbf{W}_s$ . Here,  $\mathbf{S}_\lambda$  is the so-called *smoother matrix* in the spline context (Wahba, 1990). One can refer to Li et al. (2014) for more details. The optimal tuning parameter  $\lambda_{GCV}$  is then selected as the one minimizing  $GCV(\lambda)$ . Note that there are other criteria for tuning parameter selection, such as leave-one-out cross-validation and Mallows's  $C_p$  statistic (Ruppert et al., 2003). The leave-one-out cross-validation is computationally expensive and can be approximated by GCV. The GCV is approximately equal to  $C_p$  and does not require a prior estimate of the variance of the error term (Ruppert et al., 2003). Therefore, we choose GCV here as the criterion for smoothing parameter selection.

## 2.3 | Inference on prediction

Because the proposed STVCM can be viewed as a regression-based spline model, inference on the prediction of stream temperature can be easily conducted for each stream site. The prediction interval for a future observation is similar to the interval for the regression model under the assumption of normally distributed error terms (Rencher & Schaalje, 2008). Specifically, the  $100(1 - \alpha)\%$  prediction interval for a predicted water temperature  $\hat{w}_0$  at a future observation  $x_0$  is

$$\hat{w}_0 \pm t_{\alpha/2, N_c} \hat{\sigma} \sqrt{1 + \mathbf{x}'_0 (\mathbf{X}' \mathbf{D} \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{x}_0}, \quad (12)$$

where  $t_{\alpha/2, N_c}$  is the  $100(1 - \alpha)\%$  percentile of  $t$  distribution with degree of freedom  $N_c$ . Here,  $N_c$  is the number of non-missing observations,  $\hat{\sigma} = \sum_{s=1}^S (\hat{\mathbf{W}}_s - \mathbf{W}_s)' (\hat{\mathbf{W}}_s - \mathbf{W}_s) / N_c$

the estimate for the standard deviation  $\sigma$ , and  $\mathbf{I}$  is the identity matrix.  $\mathbf{X}$  is a model matrix defined as  $\mathbf{X} = (\mathbf{\Gamma}'_1, \mathbf{\Gamma}'_2, \dots, \mathbf{\Gamma}'_S)'$ .  $\mathbf{D}$  is a diagonal matrix with the nonzero diagonal elements being the kernel distance for the selected neighbor sites.

## 3 | SIMULATION STUDY

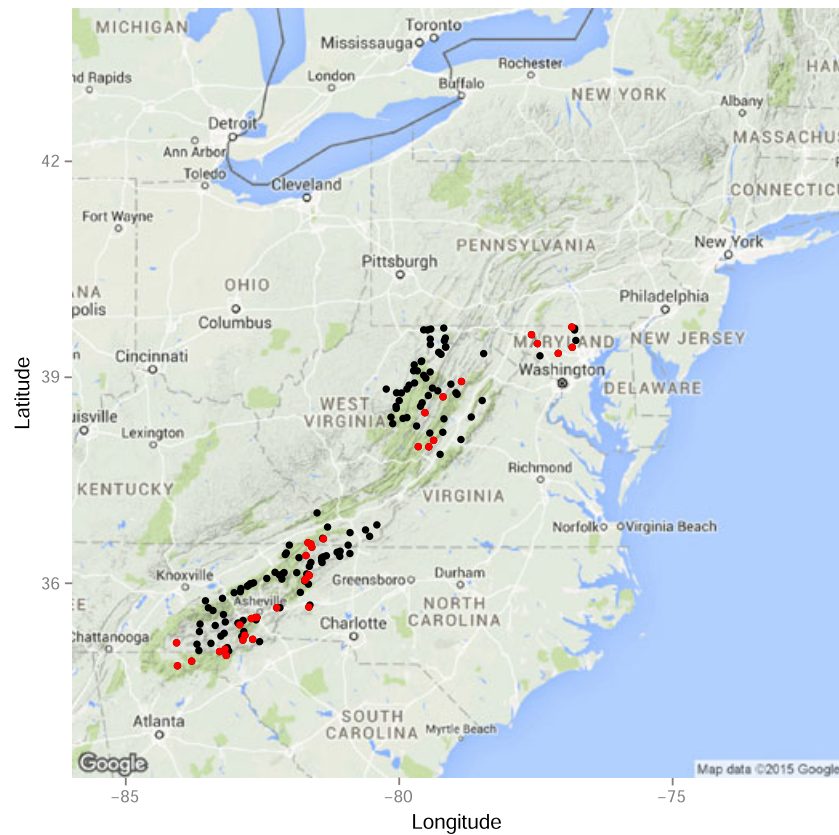
In this section, we evaluate the proposed method through a simulation study based on a subset of the data collected from thermograph sensors. For this simulation study, we use the sensor data available on 35 stream sites with complete records of both daily maximum water and maximum air temperatures. For each site, there is a full year of data with the same starting date (January 1, 2011) and ending date (December 31, 2011). The sensor locations are shown in Figure 2. The longitude and latitude of the streams are used to provide spatial information.

For these 35 sites with complete records, we deliberately remove a part of the water temperature data and use the removed data as test sets to evaluate imputation. The remaining data are used for model estimation. The root mean squared errors (RMSE) statistic is used to evaluate the imputation performance. The RMSE is defined as

$$RMSE = \sqrt{\frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} (W_i - \hat{W}_i)^2}, \quad (13)$$

where  $\mathcal{H}$  is the test set of missing values and  $|\mathcal{H}|$  is the number of the observations in set  $\mathcal{H}$  (i.e., the number of missing water values). Here,  $W'_i$ s are the water temperatures removed from the original data set (i.e., not used to estimate the model), and  $\hat{W}'_i$ s are the imputed values.

The models used in the comparison with the proposed STVCM include the linear regression model (Neumann et al., 2003), the nonlinear logistic model (Mohseni et al., 1998), the SAS MI procedure (Allison, 2005), the Gaussian process model (Cressie, 1993), and the NN (Hastie, Tibshirani, & Friedman, 2009). For the linear regression model and the nonlinear logistic model, we impute the missing water temperature data from the fitted model based on air temperature. For the Gaussian process model, the linear form of air temperature is used for the mean component. The empirical correlation matrix is used for the spatial correlation, and the autoregressive structure of order one is used for the temporal correlation. For the NN, we used the Matlab Neural Network Toolbox (The MathWorks, 2014) using the default settings, where time, location (latitude and longitude), and air temperature are inputs and water temperature is the output. A two-layer network with 10 neurons was used. The network was trained with Levenberg–Marquardt backpropagation algorithm. For each method compared, we consider three scenarios for missing values for each site. Scenario 1 (S1) considers missing values at random time points. Scenario 2 (S2) considers one sequence of missing values occurring at a random time point. Scenario 3 (S3) considers one sequence of missing values occurring at the beginning or end of the



**FIGURE 2** Locations of 35 stream sites with complete records in the simulation (red dots) and locations of 156 streams in the real data study (black dots)

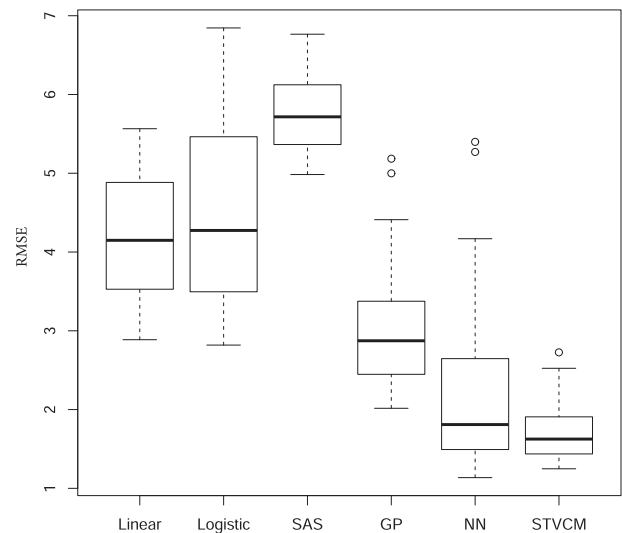
period. For each scenario, we treat 10%, 20%, and 30% of the data as missing, respectively. Thus, there are nine simulation scenarios, denoted as S1-10%, S1-20%, ..., S3-30%.

### 3.1 | Overview of results

Results from the different simulation settings were similar in pattern and results from the setting of S3-30% are summarized below. For this scenario, for each site, we randomly remove 30% of the water temperature values from either the beginning or the end of the water series and treat them as missing values. Then the RMSE is calculated with respect to the missing values for each site, and thus, we can obtain 35 RMSEs for each method in comparison. The boxplots of the RMSEs are shown in Figure 3 to examine the imputation performance for different models. From Figure 3, it is clear that the proposed STVCM has the lowest RMSEs and the NN is the only method comparable to STVCM. Therefore, we will compare the performance of STVCM and NN comprehensively as follows.

### 3.2 | Comprehensive study of STVCM versus NN

In this section, we focus on evaluating the performance of the STVCM and NN methods for the nine scenarios. The number of missing values are the same for all the sites in the same scenario. Then we impute those missing values and cal-



**FIGURE 3** Boxplots of RMSEs (y-axis) for six different methods of imputing missing data in 35 streams

culate the RMSEs based on the STVCM and NN methods, respectively. For each setting, we repeat the procedure of generating and imputing missing values 1,000 times and obtain 1,000 RMSEs. Then, for each setting, we graph boxplots of the RMSEs for the STVCM and NN methods, respectively. Boxplots based on 1,000 RMSEs are shown in Figure 4.

There are a few findings based on the results from Figure 4. First, imputation using the STVCM has greater accuracy than

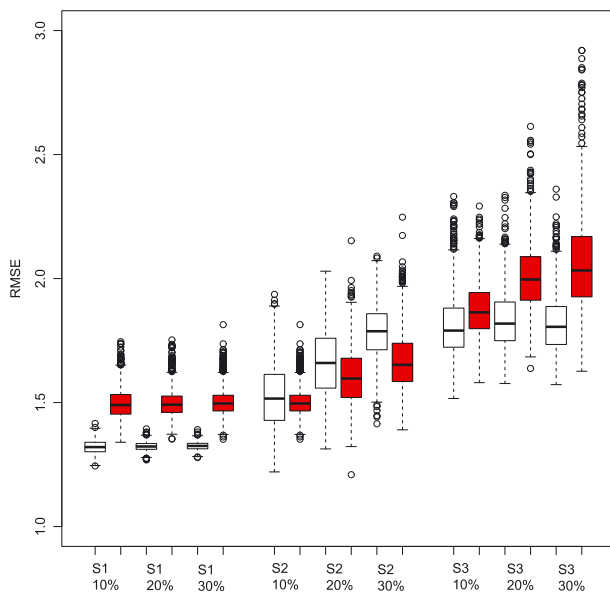
the NN method for S1. Note that the STVCM uses a smooth method for  $\theta_0(s, t)$  and  $\theta_1(s, t)$ . Missing values at the random locations thus have little impact on the proposed method on modeling (smoothing) the trend of water temperature. Therefore, the STVCM can easily borrow information from the neighboring time points to accurately impute the missing values. Second, the STVCM is slightly worse than that of the NN for S2. One possible explanation is that one large sequence of missing values might not be well predicted by the series at another location. Also we can see that the performance of the proposed method (as well as that of the NN method) degenerates as the percent of missing values increases. Third, the STVCM achieves better performance than the NN for S3. In this scenario, the time range for the missing water temperatures is at the end of the time range of the nonmissing data. As the proposed STVCM considers the spatial correlation between sites and searches for the best neighboring site to

borrow information, it works more efficiently and precisely to impute the missing values. In contrast, the NN uses all available information without regard to location and may result in less accurate imputation of the missing values.

We also compare the computational time for the proposed STVCM and the NN. The computational time (for the 1,000 runs) for both methods is shown in Table 1. Clearly, the proposed STVCM is much faster than the NN in terms of computational time; in some cases, the computational time of the STVCM is about 80–90% faster than that of the NN method.

In addition, we also calculate the coverage probability of the 95% prediction interval based on (12). Specifically, for each simulation, we calculate the prediction interval for every missing water temperature. Then one can count the number of times that the true value of missing water temperature is covered by the prediction interval. For each simulation, we can thus estimate the coverage probability by

$$\text{Coverage Probability} = \frac{\#\{\text{imputed missing water temperature is in the prediction interval}\}}{\#\{\text{missing water temperatures}\}}$$



**FIGURE 4** Boxplots for RMSEs (y-axis) for STVCM (in white color) and NN (in red color) under different scenarios

**TABLE 1** Computation time (for 1000 runs) for STVCM and NN methods (in seconds)

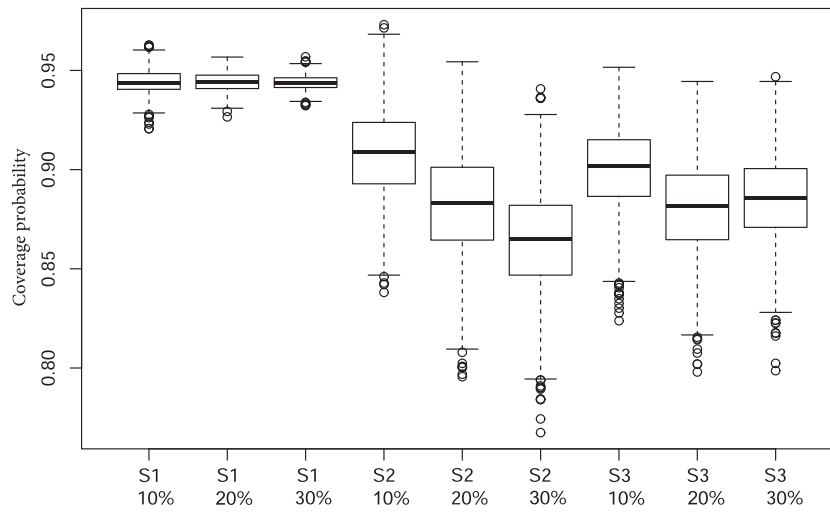
	10%		20%		30%	
	STVCM	NN	STVCM	NN	STVCM	NN
Randomly missing (S1)	6,579	39,461	6,519	35,536	6,439	30,700
One sequence of missing at random location (S2)	4,525	46,276	4,521	45,282	4,483	30,306
One sequence of missing at beginning or end (S3)	5,242	39,899	5,183	37,254	5,137	31,666

NN = neural network; STVCM = spatiotemporal varying coefficient model.

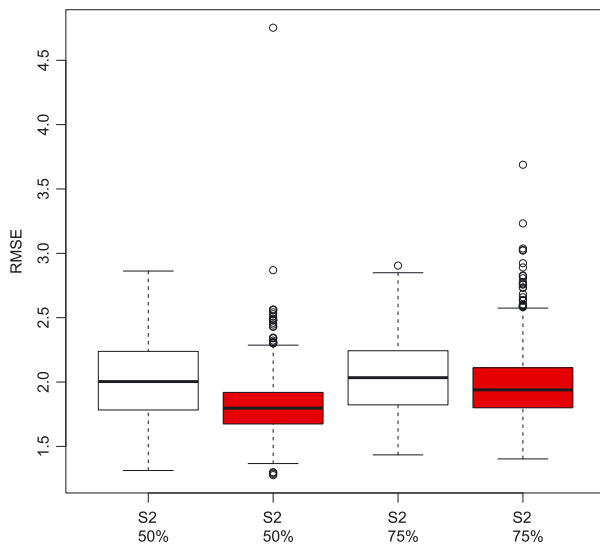
Then, for each simulation setting, we can obtain 1,000 coverage probabilities from the simulations for each setting. Figure 5 shows the boxplots of coverage probability of the 95% prediction interval for the proposed STVCM method under the nine settings. Clearly, one can see that for in the case of S1 with missing at random time points, the coverage probability the 95% prediction interval from the proposed method is very close to 95%. In the case of S2 and S3 when involving missing over time, the coverage probability of the 95% prediction interval appears to be lower, at around 90%. A possible explanation is that the proposed method may lose estimation efficiency slightly when imputing the missing values over time.

### 3.3 | Comparison of STVCM and NN with large sequences of missing values

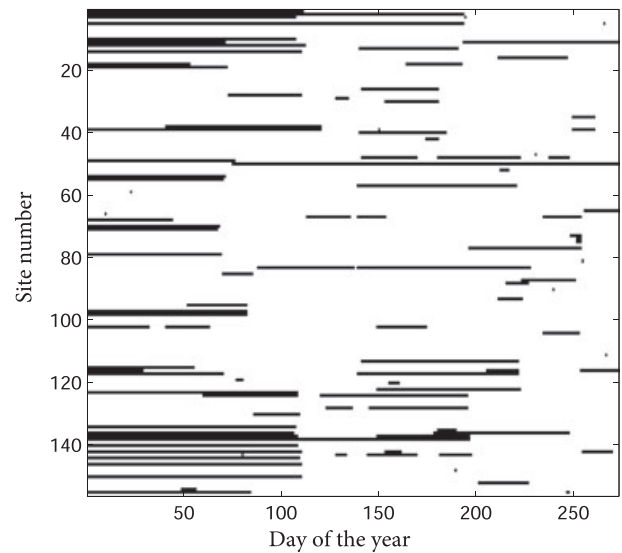
Note that the real data in this work contains several streams with more than 50% missing on water temperatures and as suggested by the referees, we consider a further simulations under S2 for 50% and 75% of missing data, respectively. In this simulation study, 15 sites with a complete full-year data are randomly selected, and one sequence of the water temperatures in those sites was set to missing. We repeat this procedure for 1,000 iterations for both STVCM and NN. The RMSEs are reported in Figure 6 and the computational time is shown in Table 2. From the comparison results, one can see that the RMSEs from STVCM and NN are comparable, but the NN method is not very stable with more outliers. In terms of computational time in Table 2, the proposed method is much faster than the NN method.



**FIGURE 5** Boxplots of coverage probability of prediction intervals in the simulation



**FIGURE 6** Boxplots for RMSEs (y-axis) for STVCM (in white color) and NN (in red color) under S2 with large sequence of missing values



**FIGURE 7** Binary description for missing water temperature for the 156 sites. Black line segments indicate the date and duration of missing values

#### 4 | CASE STUDY: CLUSTERING TROUT STREAMS

In this section, we apply the proposed STVCM to the water temperature measured in 156 streams primarily in the southeastern United State (the locations are shown in Figure 2). The data used here were measured for 9 months starting from January 1 to September 30, 2011 (272 days) and do not contain missing values for the air temperatures. Imputation of potential missing values in the air temperature will be discussed in

Section 5. Figure 7 shows the missing pattern for daily maximum water temperatures for the 156 streams. Among the 156 streams, there are only around 25% of streams having a complete record of stream water temperatures.

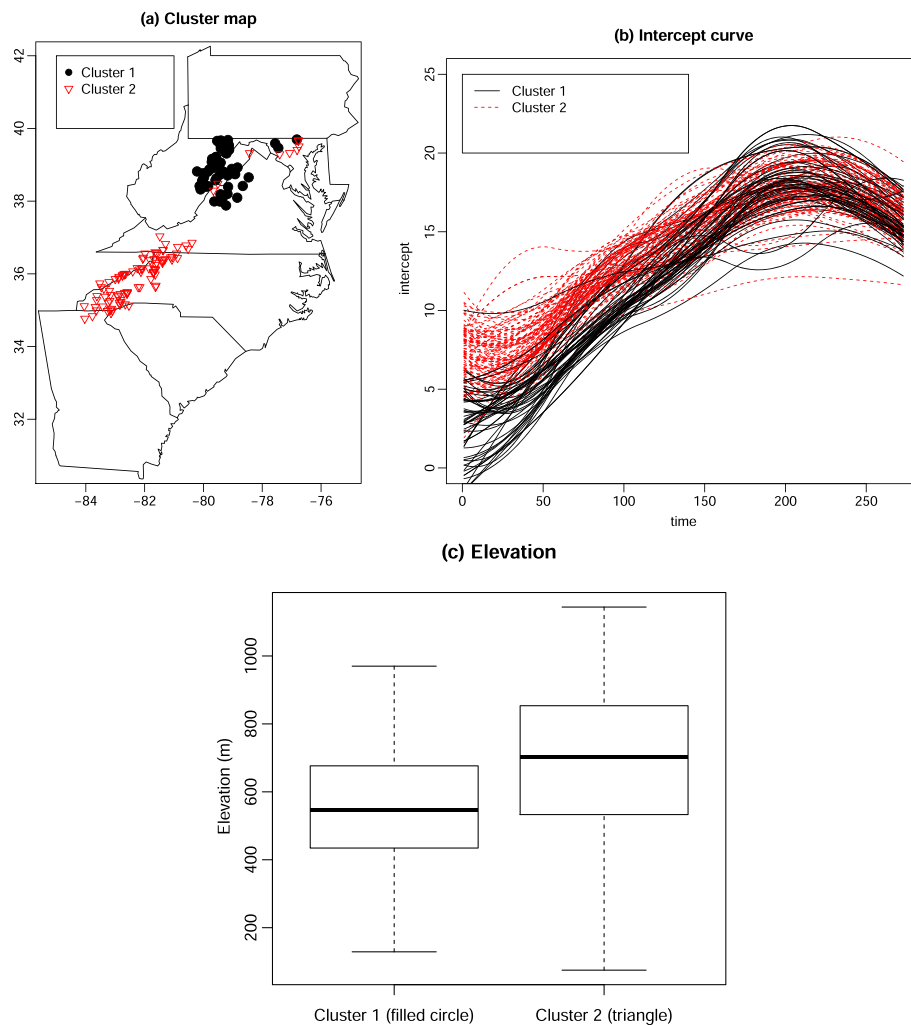
After imputing the water temperature data using the proposed method, the clustering method in Li et al. (2016) is used for grouping the 156 streams. For the clustering, we used a K-medoids approach with the Canberra distance metrics (details are in Li et al. (2016)). For the management of trout streams (or sites) in terms of fish habitat and temper-

**TABLE 2** Computational time (for 1,000 runs) for STVCM and NN methods (in seconds) under S2 with large sequence of missing values

	50%		75%	
	STVCM	NN	STVCM	NN
One sequence of missing at random location (S2)	3,737	13,375	3,642	16,654

NN = neural network; STVCM = spatiotemporal varying coefficient model.





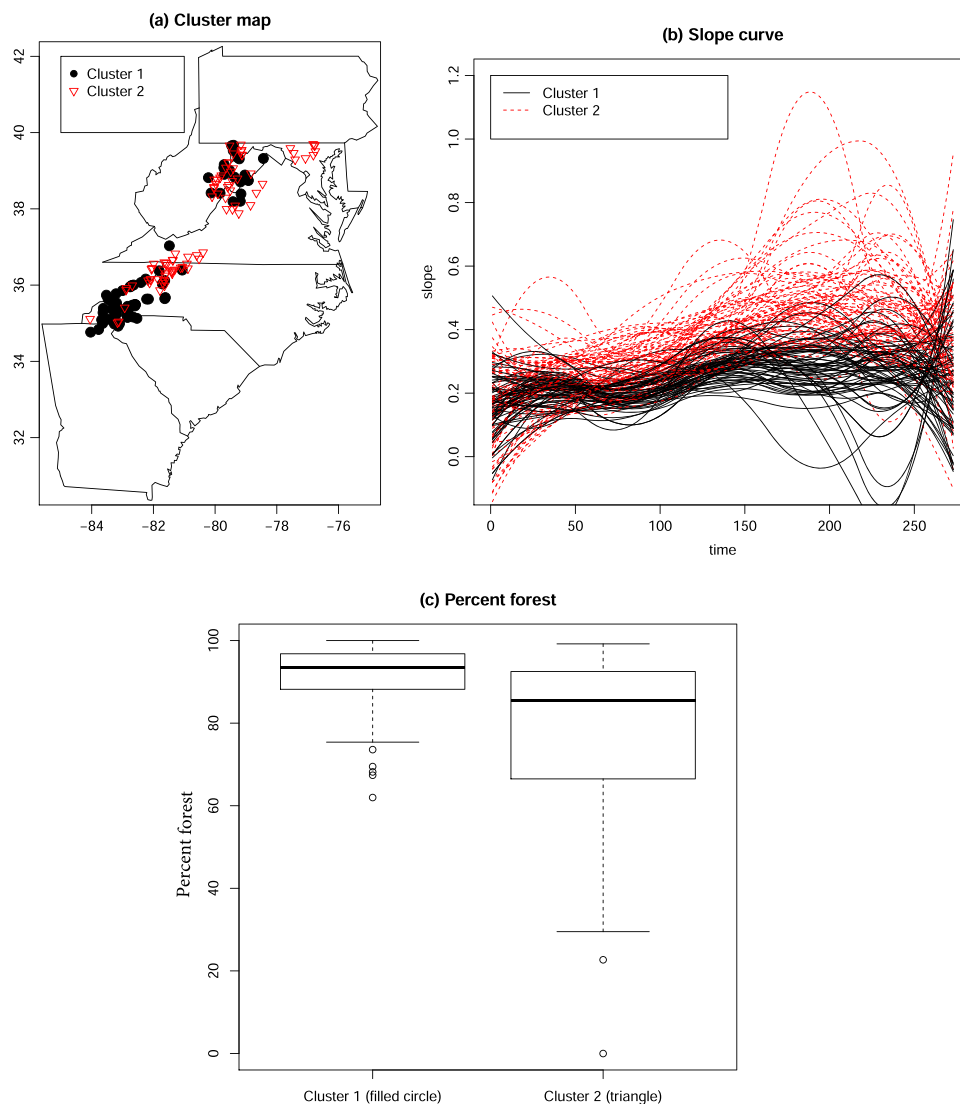
**FIGURE 8** Clustering results based on intercept curves: (a) cluster map, (b) intercept curves, (c) boxplots of elevation in the two clusters

ature risk, it is helpful to identify clusters of streams with similar air–water temperature relationships. If streams within the same cluster have similar profiles of risk, then agencies can better manage streams and watersheds based on the water–air temperature relationship (Mayer, 2012). Clustering streams is also valuable in other aspects. For example, many climate and landscape factors affect the water and air temperature relationship in streams (Chen, Carsel, McCutcheon, & Nutter, 1998). Those factors tend to be distinct for each stream but may show certain similarity when compared within the same cluster. In addition to latitude and longitude of stream sites, several other stream characteristics are also measured: elevation and the percentage of forest in the area around the stream. The variables elevation and forest percentage are used as descriptive information to evaluate clustering results.

Cluster results are summarized graphically for intercept and slope profiles in Figures 8 and 9. There are two findings in the clustering results supporting the reliability of the imputed data by STVCM. Consider first, Figure 8 that summarizes the clustering results based on the intercept curves (i.e.,  $\theta_0(s, t)$  in STVCM). It results in two clusters, which are

spatially well separated with respect to the stream locations. Moreover, the clustered streams also have a close connection with the elevation variable (Figure 8c). The two clusters differ significantly in elevation (two sample  $t$  test,  $p$  value =  $3.6e-4$ ). Note that the intercept curves are related to the local average daily maximum water temperature (Li et al., 2014). It shows that the completed water temperature data using imputed missing values from the STVCM are consistent with the physical interpretation that high elevation usually results in low water temperature, especially in the cooler part of the year.

Second, consider the clustering results based on the slope curves (i.e.,  $\theta_1(s, t)$  in STVCM) in Figure 9. It is seen that there are two clusters, which do not have clear separation in terms of spatial location. However, the resultant clusters have a meaningful connection with the percentage of forest variable. Figure 9 shows the boxplots for the percent forest of streams for the two clusters, based on the slope curve. The mean percent forest is significantly different for the two clusters ( $p$  value =  $1.4e-6$  from two sample  $t$  test). The differences suggests that the slope curves are associated with stream percent forest and that altering percent forest may affect stream sensitivity to changes in air temperature



**FIGURE 9** Clustering results based on slope curves: (a) cluster map, (b) slope curves, (c) boxplots of percent forest in the two clusters

(Li et al. 2014). The resultant clustering thus gives evidence that the use of the STVCM is reasonable because high-percent forest usually makes water temperature less sensitive to changes in air temperature due to shading. Elevation is associated with average air temperature, especially in the winter and spring.

## 5 | CONCLUSION

In this work, we propose a novel missing data imputation method based on a STVCM for the stream water temperature. The proposed method considers both the temporal and the spatial variation in water temperature and provides an accurate imputation of the missing water temperatures. The simulation study shows that the performance of the proposed method for missing data imputation is better than several existing methods such as the NN. By imputing the missing water temperature data in 156 streams, the complete set of sensor data can be used to successfully cluster the streams and results in clusters with meaningful interpretation.

Note that in this study of maximum water temperature from multiple stream sensors, the sensor data have strong temporal correlation and weak spatial correlation. The adoption of polynomial splines for the temporal effect and local kernel method for the spatial effect provides a flexible structure to quantify the varying coefficients. We also conducted some experiments using polynomial splines for both time and spatial effects. Although the imputation results remain plausible, the fitted intercept and slope curves do not provide useful information for the cluster analysis. One future research direction can be how to use more flexible basis functions for both temporal and spatial effects. It is also worth pointing out that the proposed method has not taken advantage of other potential covariates such as precipitation, slope, and aspect that may be available.

Because the proposed method can provide valid inference on estimation and imputation, one could incorporate the variance of the estimated intercept and slope curves into the clustering. In this case, the clustering algorithm adopted from Li et al. (2016) needs to be modified to accommodate the

unequal variance with different amounts of “shrinkage” to the cluster medoid. The K-medoids algorithm used in Li et al. (2016) is effective in detecting compact spherical-shaped clusters and is easy to use in practice (Aggarwal & Reddy, 2013). Alternatively, one may consider other clustering algorithms such as hierarchical clustering, support vector machines, classification tree, and random forest (Hastie et al. 2009). Finally, the current work assumes that the data on air temperatures are complete without missing values. In practice, missing value are also occurred in the air temperatures. How to conduct an effective method for imputation of both missing air and water temperatures would be an interesting topic for future research.

## REFERENCES

- Aggarwal, C. C., & Reddy, C. K. (2013). *Data clustering: Algorithms and applications*. Boca Raton, FL: CRC Press.
- Allison, P. D. (2005). Imputation of categorical variables with PROC MI. Philadelphia, PA.: SAS Users Group International, 30th Meeting (SUGI 30).
- Beitinger, T. L., Bennett, W. A., & McCauley, R. W. (2000). Temperature tolerances of North American freshwater fishes exposed to dynamic changes in temperature. *Environmental Biology of Fishes*, *58*, 237–275.
- Caissie, D. (2006). The thermal regime of rivers: A review. *Freshwater Biology*, *51*, 1389–1406.
- Chen, Y. D., Carsel, R. F., McCutcheon, S. C., & Nutter, W. L. (1998). Stream temperature simulation of forested riparian areas: 1. Watershed-scale model development. *Journal of Environmental Engineering*, *124*, 304–315.
- Cheng, M. Y., Zhang, W., & Chen, L. H. (2009). Statistical estimation in generalized multiparameter likelihood models. *Journal of the American Statistical Association*, *104*, 1179–1191.
- Coulibaly, P., & Evora, N. D. (2007). Comparison of neural network methods for infilling missing daily weather records. *Journal of Hydrology*, *341*, 27–41.
- Cressie, N. (1993). *Statistics for Spatial Data*. New York, NJ: John Wiley & Sons.
- Cressie, N., & Wikle, C. K. (2011). *Statistics for spatio-temporal data*. Hoboken, NJ: Wiley.
- Dunham, J., Chandler, G., Reiman, B., & Martin, D. (2005). Technical report. Report RMRS-GTR-150WWW. Measuring stream temperature with digital data loggers: A user’s guide. Fort Collins, CO: USDA Forest Service.
- Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications*. New York, NY: Chapman and Hall.
- Fan, J., & Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and Its Interface*, *1*, 179–195.
- Ferguson, C. A., Bowman, A. W., Scott, E. M., & Carvalho, L. (2007). Model comparison for a complex ecological system. *Journal of the Royal Statistical Society, Series A*, *170*, 691–711.
- Ferguson, C. A., Bowman, A. W., Scott, E. M., & Carvalho, L. (2009). Multivariate varying-coefficient models for an ecological system. *Environmetrics*, *20*, 460–476.
- Flebbe, P. A., Roghair, L. D., & Bruggink, J. L. (2006). Spatial modeling to project southern Appalachian trout distribution in a warmer climate. *Transactions of the American Fisheries Society*, *135*, 1371–1382.
- Hastie, T., & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, *55*, 757–796.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer.
- Hoover, D. R., Rich, J. A., Wu, C. O., & Yang, L. P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, *85*, 809–822.
- Huang, J. Z., Wu, C. O., & Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, *89*, 111–128.
- Huang, Z., Li, J., Nott, D., Feng, L., Ng, T. P., & Wong, T. Y. (2015). Bayesian estimation of varying-coefficient models with missing data, with application to the Singapore longitudinal aging study. *Journal of Statistical Computation and Simulation*, *85*(12), 2364–2377.
- Hudy, M., Thieling, T. M., Gillespie, N., & Smith, E. P. (2008). Distribution, status, and land use characteristics of subwatersheds within the native range of brook trout in the eastern United States. *North American Journal of Fisheries Management*, *28*(4), 1069–1085.
- Huff, D. D., Hubler, S. L., & Borisenko, A. N. (2005). Using field data to estimate the realized thermal niche of aquatic vertebrates. *North American Journal of Fisheries Management*, *25*, 346–360.
- Kaufman, C. G., Schervish, M. J., & Nychka, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, *103*, 1545–1555.
- Keleher, C. J., & Rahel, F. J. (1996). Thermal limits to salmonid distributions in the Rocky Mountain region and potential habitat loss due to global warming: A geographic information systems (GIS) approach. *Transactions of the American Fisheries Society*, *125*, 1–13.
- Li, H., Deng, X., Dolloff, C. A., & Smith, E. P. (2016). Bivariate functional data clustering: Grouping streams based on a varying coefficient model of the stream water and air temperature relationship. *Environmetrics*, *27*(1), 15–26.
- Li, H., Deng, X., Kim, D. Y., & Smith, E. P. (2014). Modeling maximum daily temperature using a varying coefficient regression model. *Water Resources Research*, *50*, 3073–3087.
- Lu, Z., Steinskog, D. J., Tjøstheim, D., & Yao, Q. (2009). Adaptively varying-coefficient spatiotemporal models. *Journal of the Royal Statistical Society. Series B (Methodological)*, *71*, 859–880.
- Mallows, C. L. (1973). Some comments on c p. *Technometrics*, *15*(4), 661–675.
- Mayer, T. D. (2012). Controls of summer stream temperature in the Pacific Northwest. *Journal of Hydrology*, *475*, 323–335.
- Meisner, J. D. (1990). Effect of climate warming on the southern margins of the native range of brook trout. *Salvelinus fontinalis*. *Canadian Journal of Fisheries and Aquatic Science*, *47*, 1065–1070.
- Minns, C. K., Randall, R. G., Chadwick, E. M. P., Moore, J. E., & Green, R. (1995). Potential impact of climate change on the habitat and production dynamics of juvenile Atlantic salmon (*Salmo salar*) in eastern Canada. In Beamish, R. J. (Ed.), *Climate Change and Northern Fish Population* (pp. 699–708). NRC Research Press, Ottawa.
- Mohseni, O., & Stefan, H. G. (1999). Stream temperature/air temperature relationship: A physical interpretation. *Journal of Hydrology*, *218*, 128–141.
- Mohseni, O., Stefan, H. G., & Erickson, T. R. (1998). A nonlinear regression model for weekday stream temperatures. *Water Resources Research*, *34*, 2685–2692.
- Neumann, D. W., Rajagopalan, B., & Zagana, E. A. (2003). Regression model for daily maximum stream temperature. *Journal of Environmental Engineering*, *129*, 667–674.
- O’Donnell, D., Rushworth, A., Bowman, A. W., Scott, E. M., & Hallard, M. (2014). Flexible regression models over river networks. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *63*(1), 47–63.
- Onset Computer Corporation (2009). HOBO U22 water temp pro v2 users manual. Document number 10366-c.
- Rencher, A. C., & Schaalje, G. B. (2008). *Linear models in statistics*. Hoboken, NJ: John Wiley & Sons, Inc.
- Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. New York, NY: Cambridge University Press.
- Rushworth, A. M., Peterson, E. E., Ver Hoef, J. M., & Bowman, A. W. (2015). Validation and comparison of geostatistical and spline models for spatial stream networks. *Environmetrics*, *26*, 327–338.
- Serban, N. (2011). A space–time varying coefficient model: The equity of service accessibility. *Annals of Applied Statistics*, *5*, 2024–2051.
- Sinokrot, B. A., & Stefan, H. G. (1993). Stream temperature dynamics: Measurements and modeling. *Water Resources Research*, *29*, 2299–2312.
- The MathWorks, Inc. (2014). *Neural network toolbox user’s guide*.

- Trumbo, B. A., Nislow, K. H., Stallings, J., Hudy, M., Smith, E. P., Kim, D. Y., Dolloff, C. A. (2014). Ranking site vulnerability to increasing temperatures in southern appalachian brook trout streams in virginia: An exposure-sensitivity approach. *Transactions of the American Fisheries Society*, *143*(1), 173–187.
- Ver Hoef, J. M., Peterson, E. E., Clifford, D., & Shah, R. (2014). Ssn: An R package for spatial statistical modeling on stream networks. *Journal of Statistical Software*, *56*(3), 1–45.
- Wahba, G. (1990). *Spline models for observational data*. Philadelphia, PA: SIAM.
- Wang, Y., Zheng, T., Zhao, Y., Jiang, J., Wang, Y., Guo, L., & Wang, P. (2013). Monthly water quality forecasting and uncertainty assessment via bootstrapped wavelet neural networks under missing data for Harbin, China. *Environment Science Pollution Research*, *20*, 8909–8923.
- Webb, B. W., Clack, P. D., & Walling, D. E. (2003). Water air temperature relationships in a Devon River system and the role of flow. *Hydrological Processes*, *17*, 3069–3084.
- Wu, C. O., Chiang, C. T., & Hoover, D. R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American Statistical Association*, *93*, 1388–1402.
- Xu, W., & Zhu, L. (2013). Testing the adequacy of varying coefficient models with missing responses at random. *Metrika*, *76*, 53–69.
- Zhao, P. X., & Xue, L. G. (2011). Variable selection for semiparametric varying-coefficient partially linear models with missing response at random. *Acta Mathematica Sinica*, *27*(11), 2205–2216.

**How to cite this article:** Li, H., Deng, X., and Smith, E. Missing data imputation for paired stream and air temperature sensor data. *Environmetrics*. 2017;28:e2426. doi: 10.1002/env.2426.