Original Research Article

Prediction and assessment of stream water quality by considering multiple water quality parameters

ABSTRACT

Geostatistical studies entail identifying the most appropriate model to describe the observed data so that it can be used to accurately predict responses across a range of possible locations. The purpose of such a model is to depict the link between the response variables and the predictors while taking into account uncertainties in space and time. We propose a novel approach to model such data via a multivariate spatio-temporal additive model derived through considering a multivariate normal approximation. To demonstrate how the proposed approach works, we use numerous water quality parameters to model and predict the water quality of a stream network. To reflect the spatial variability of the stream network, we employed hydrologic distances in the model, which allowed certain properties of streams and rivers, such as stream flow connectivity, to be effectively described. It was observed that the proposed multivariate model produces accurate predictions at un-sampled locations compared to its univariate counterparts. Accordingly, this study reveals that the proposed multivariate modelling approach is a viable alternative for modelling complicated data such as the data found in water quality monitoring.

Keywords: Geostatistical studies, Hydrologic distance, Spatial-temporal additive model, Spatial variability, Stream network.

1 Introduction

Streams and rivers are among the most precious environmental resources that belong to any nation or country. They have numerous advantages for humans, animals, and plants. Stream and river water is used for drinking, agriculture, fishing, and industrial purposes including manufacturing and electricity. Additionally, minerals such as quartz and granite are extracted using stream and river water. It also aids fish, insects, plants, and birds in finding suitable habitats. Unfortunately, various factors such as industrial pollutants, chemicals used in agriculture, and human behaviour have put this important resource in danger. Thus, appropriate monitoring and management of water quality in streams and rivers is vital to minimize such dangers and maximize the use of this valuable resource.

Due to the cost and scarcity of other resources, many stream and river water quality monitoring programs assess a limited number of water quality parameters and water samples. It is vital but

difficult to identify the key parameters that explain changes in water quality. As such, taking measurements for insufficient number of parameters reduces the accuracy of these monitoring programs. As a result, complex statistical models for predicting water quality at unobserved places utilizing various water quality parameters are gaining popularity (Peterson et al., 2007; Isaak et al., 2014).

An additive model is a type of nonparametric regression model that describes nonlinear trends using a one-dimensional smoother (Friedman and Stuetzle, 1981). Additive models provide greater flexibility than parametric models (Seya et al., 2011) and are easier to interpret than general regression surfaces (Qiu, 1998). Thus, additive models have recently attracted interest for modelling ecological time series data (Vercelloni et al., 2014, 2017). Furthermore, spatial additive models have been presented in other disciplines for capturing the true underlying spatial and time series variabilities within the data, and they have been proved to be effective for prediction under different conditions (Nandy et al., 2017). Accordingly, such modelling approaches should aid in developing more appropriate models for predicting water quality at unobserved locations.

In spatial additive models, the assumption of independence of the errors is relaxed to allow spatial autocorrelation, which is modelled as a function of the distance separating any two locations. However, given the branching network structure, stream flow connectivity, direction and volume (Peterson et al., 2013, 2007), spatial auto-correlation may exist in streams data that is not well described using Euclidean distance (Davis and Curriero, 2019; Curriero, 2006; Murphy et al., 2015). Thus, standard spatial statistical modelling approaches such as Kriging, which requires inter-point distances to be Euclidean distances, may not be sufficient to describe the unique spatial relationships found in stream networks. Therefore, developing new approaches for modelling specific covariance structures observed in streams data is a key concern in this area of research (Peterson et al., 2013; Cressie et al., 2006; Ganio et al., 2005).

In general, various physical, chemical and biological properties of water are measured to determine the water quality of a stream network, including temperature, turbidity, pH level, alkalinity, total hardness, dissolved phosphates, dissolves sulphate and algae concentration (Korashey, 2009). However, due to the complex nature of stream networks, a vast majority of previous studies were limited to only consider a single response variable (i.e., a single physical, chemical or biological property of water) when developing statistical models for stream data (Peterson and Ver Hoef, 2010; Peterson et al., 2013). Such an approach requires developing multiple models to describe the water quality of a stream network. This increases the computational cost and time for such methods and also limits the applicability in describing water quality of large-scale stream networks (Smith et al., 1997).

In this paper, we propose a multivariate spatio-temporal additive model (STAM) to describe changes in the water quality of a stream network over time. To improve the accuracy of water quality predictions at unobserved locations, the proposed approach takes into account spatial variability and nonlinear time series trends observed in multiple water quality parameters. Hydrologic distances are utilised to depict spatial auto-correlation in this way, allowing specific aspects of streams and rivers, such as stream flow connectedness, to be effectively defined. Using water quality monitoring data from the River Thames in the United Kingdom as a case study, we evaluate the proposed modelling approach in terms of predicting water quality at unsampled locations. Mean square prediction errors are used to evaluate the performance of the univariate and multivariate spatial models. As the methodologies developed throughout this study are generic in nature, this research has the potential to improve water quality modelling

as well as other ecological modelling in the future.

The rest of the paper is organised as follows. In Section 2, the formulation of multivariate STAM and the method of estimating the model parameters are presented. We show how the proposed model can be extended to measure the spatial variability in the stream network. The water quality monitoring application considered in this study is described in Section 3. Section 4 presents the results obtained from the analysis. It includes a comparison of analyses based on univariate and multivariate STAM that are performed with an emphasis on predictive accuracy. The important findings and recommendations for further research are presented in Section 5.

2 Spatio-temporal additive models

Given spatial dependency in the data, one can model the mean μ_i of a univariate response y_i as follows:

$$\mu_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + s_i, \quad E(y_i|s_i) = g^{-1}(\mu_i), \text{ for } i = 1, 2, \ldots, n, \tag{1}$$

where s_i represents spatially dependent random effects related to the ith sampling location (e.g. a monitoring station), $\boldsymbol{X}_i = (1, X_{i1}, X_{i2}, \cdots, X_{ip-1})^T$ is the covariate vector (excluding time covariate) associated with the ith sampling location, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \cdots, \beta_{p-1})$ is the vector of regression coefficients (fixed effects). In addition, $g(\cdot)$ represents a link function which defines the relationship between the linear predictor μ_i and the expected outcome given s_i.

Given the spatially dependent nature of the collected data, random effects in the spatial regression models allow for spatial autocorrelation in the errors. Such random effects s_i in the above model are assumed to be from a Gaussian random field with zero-mean and a covariance $\boldsymbol{\Sigma}$. Given the separation distance, the covariance measures the level of spatial autocorrelation between two sampling locations (Olea, 1991). Here, the separation distance is simply the distance traveled from one sampling location to another, and it may be measured using a suitable distance measure (Ver Hoef et al., 2006; Peterson et al., 2007; Lyon et al., 2008).

Hydrologic distances are used to characterise spatial autocorrelation in biological, chemical, and physical stream data (Ganio et al., 2005; Gardner et al., 2003). In particular, the symmetric hydrologic distance can be used to fit autocovariance functions in order to obtain the appropriate covariances for modelling spatial data. Simply, the shortest distance between two points along a stream is known as hydrologic distance or stream distance. In general, it is the expected value of all conceivable distances between the two points. The hydrologic distance differs from the Euclidean distance since it is determined by the flow connection between the two locations. Here, it is worth noting that the exponential covariance model is the only known autocovariance function that is valid when using covariance matrices based on symmetric hydrologic distances to make predictions at unobserved sampling locations (Ver Hoef et al., 2001).

The exponential covariance function can be defined to determine the covariance matrix $\boldsymbol{\Sigma}$ as follows:

$$\text{cov}(h, \boldsymbol{\gamma}) = \begin{cases} \gamma_0 + \gamma_1 & ; \text{if } h = 0 \\ \gamma_1 \exp\left(\dfrac{-h}{\gamma_2}\right) & ; \text{otherwise} \end{cases}$$

where h is the hydrological distance between two locations and the parameters $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)$

are the nugget effect, partial sill and the spatial range, respectively. Here, the nugget effect indicates the variation across sampling locations as the separation distance approaches zero. The sill is the autocovariance asymptote that represents variance among uncorrelated data. The range parameter determines how quickly autocovariance decays as distance increases.

In order to capture nonlinear time series trends, the spatial model defined in Equation (1) can be extended to incorporate an additive smooth component in terms of time as follows:

$$\mu_{ij} = \boldsymbol{X}_{ij}^T\boldsymbol{\beta} + m(t_{ij}, \boldsymbol{\lambda}) + s_i, \;\; E(y_{ij}|s_{ij}, t_{ij}) = g^{-1}(\mu_{ij}) \text{ for } i = 1, 2, \ldots, n \text{ and } j = 1, \ldots, l, \quad (2)$$

where $m(t_{ij}, \boldsymbol{\lambda})$ is a smooth function and in which t_{ij} represents the jth time point where data have been collected from ith sampling location and $\boldsymbol{\lambda}$ represents all the parameters related to the corresponding smooth function. Of note, such a smooth function can be employed to capture nonlinear patterns not only in time but also in any other covariate. The smooth function can be modelled in a number of ways, including cubic splines, B-splines, truncated polynomials, radial splines and etc (Crainiceanu et al., 2005). Low-rank thin-plate splines are appealing because they require fewer parameters, unaffected by knot selection, and inherent prevention against overfitting (Wood, 2003). Accordingly, the smooth function in Equation (2) can be expressed as follows:

$$m(t_{ij}, \boldsymbol{\lambda}) = \alpha t_{ij} + \sum_{k=1}^{K} \zeta_k |t_{ij} - \tau_k|^3, \quad (3)$$

where $\boldsymbol{\lambda}$ includes the regression coefficients for time α, and random coefficients $\boldsymbol{\zeta} = (\zeta_1, \cdots, \zeta_K)$, and τ_k are knots, and K is the total number of knots. In the equation, $|t_{ij} - \tau_k|$ are calculated based on the sample quantile of t_{ij}'s, and the reader is referred to Crainiceanu et al. (2005) for more details on modelling a smooth function using thin plate splines.

2.1 Spatio-temporal additive model for multiple responses

Here, we focus on extending the approach proposed in the previous section for modelling more than one response. In the case of considering two responses Y_1 and Y_2 being normally distributed with means (μ_1, μ_2) and variances (σ_1^2, σ_2^2), the corresponding bivariate response $\boldsymbol{Y} = (Y_1, Y_2)$ can be modelled through combining the individual univariate responses. In doing so, a bivariate normal distribution is assumed where the corresponding mean $\boldsymbol{\mu_{ij}} = (\mu_{1ij}, \mu_{2ij})'$ for $i = 1, 2, \ldots, n$ and $j = 1, \ldots, l$ is obtained as follows:

$$\mu_{1ij} = \boldsymbol{X}_{ij}^T\boldsymbol{\beta}_1 + m_1(t_{ij}, \boldsymbol{\lambda}_1) + s_{1i} \text{ and } \mu_{2ij} = \boldsymbol{X}_{ij}^T\boldsymbol{\beta}_2 + m_2(t_{ij}, \boldsymbol{\lambda}_2) + s_{2i}. \quad (4)$$

Next, the corresponding covariance matrix can be obtained as $\boldsymbol{\Omega} = \begin{pmatrix} \sigma_1^2\boldsymbol{I} + \boldsymbol{\Sigma_1} & \sigma_{12}\boldsymbol{I} \\ \sigma_{21}\boldsymbol{I} & \sigma_2^2\boldsymbol{I} + \boldsymbol{\Sigma_2} \end{pmatrix}$, where $\sigma_{12} = \sigma_{21}$ is the covariance between Y_1 and Y_2, and $\boldsymbol{\Sigma}_1 = cov(h, \boldsymbol{\gamma}_1)$ and $\boldsymbol{\Sigma}_2 = cov(h, \boldsymbol{\gamma}_2)$ are covariance matrices based on spatial distances related to Y_1 and Y_2, respectively. Furthermore, it is straight forward to extend such a model to incorporate more than two responses. For example, in the case of having three responses, the mean and covariance matrix of the spatial model can be defined as $\boldsymbol{\mu_{ij}} = (\mu_{1ij}, \mu_{2ij}, \mu_{3ij})'$ for $i = 1, 2, \ldots, n$ and $j = 1, \ldots, l$, and $\boldsymbol{\Omega} = \begin{pmatrix} \sigma_1^2\boldsymbol{I} + \boldsymbol{\Sigma_1} & \sigma_{12}\boldsymbol{I} & \sigma_{13}\boldsymbol{I} \\ \sigma_{21}\boldsymbol{I} & \sigma_2^2\boldsymbol{I} + \boldsymbol{\Sigma_2} & \sigma_{23}\boldsymbol{I} \\ \sigma_{31}\boldsymbol{I} & \sigma_{32}\boldsymbol{I} & \sigma_3^2\boldsymbol{I} + \boldsymbol{\Sigma_3} \end{pmatrix}$, respectively. Here, $\sigma_{13} = \sigma_{31}$ and $\sigma_{23} = \sigma_{32}$ are the covariances between the corresponding responses.

2.2 Estimating the model parameters

Given the probability distributions of data and the model parameters, maximum likelihood estimation entails establishing a likelihood function to evaluate the conditional probability of observing a data sample. Typically, the log likelihood is considered as it makes further mathematical analysis easier, but also it helps numerically. Assuming the univariate response variable follows a normal distribution with mean μ_{ij} and variance σ^2, the log likelihood function can be defined for the model in Equation (2) as follows:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{l} \log p(y_{ij}|\alpha, \boldsymbol{\beta}, \boldsymbol{\zeta}, s_i, \sigma^2, \boldsymbol{X}, \boldsymbol{Z}) + \sum_{i=1}^{n} \log p(s_i|\boldsymbol{\gamma}) + \sum_{k=1}^{K} p(\zeta_k|\sigma_\zeta, \boldsymbol{Z}), \tag{5}$$

where $p(y_{ij}|\alpha, \boldsymbol{\beta}, \boldsymbol{\zeta}, s_i, \sigma^2, \boldsymbol{X}, \boldsymbol{Z})$ is the likelihood conditional on the random effects and random coefficients, $p(s_i|\boldsymbol{\gamma})$ is the distribution of random effects, and $p(\zeta_k|\sigma_\zeta, \boldsymbol{Z})$ is the distribution of random coefficients, and $\boldsymbol{\theta}$ represents all model parameters including the regression coefficients $(\alpha, \boldsymbol{\beta})$, random coefficients related to smooth function $\boldsymbol{\zeta}$, covariance parameters $\boldsymbol{\gamma}$, and the standard deviation related to random coefficient in the smooth function $\sigma_{\boldsymbol{\zeta}}$, and $\boldsymbol{Z}$ represents the data matrix related to the sample quantiles.

The likelihood function defined in Equation (5) can be extended for a bivariate response as follows:

$$\begin{aligned} l(\boldsymbol{\theta^*}) = &\sum_{i=1}^{n}\sum_{j=1}^{l} \log p(y_{1ij}, y_{2ij}|\alpha_1, \alpha_2, \boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{\zeta_1}, \boldsymbol{\zeta_2}, s_{1i}, s_{2i}, \sigma_1^2, \sigma_2^2, \sigma_{12}, \boldsymbol{X}, \boldsymbol{Z}) + \\ &\sum_{i=1}^{n} \log p(s_{1i}|\boldsymbol{\gamma_1}) + \sum_{i=1}^{n} \log p(s_{2i}|\boldsymbol{\gamma_2}) + \sum_{k=1}^{K} p(\zeta_{1k}|\sigma_{\zeta_1}, \boldsymbol{Z}) + \sum_{k=1}^{K} p(\zeta_{2k}|\sigma_{\zeta_2}, \boldsymbol{Z}), \end{aligned} \tag{6}$$

where $\boldsymbol{\theta^*} = (\alpha_1, \alpha_2, \boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{\gamma_1}, \boldsymbol{\gamma_2}, \boldsymbol{\zeta_1}, \boldsymbol{\zeta_2}, \sigma_{\zeta_1}, \sigma_{\zeta_2}, \sigma_1^2, \sigma_2^2, \sigma_{12})$ and the rest of the distributions have the same meaning as described in relation to Equation (5). When comparing this $\boldsymbol{\theta^*}$ to $\boldsymbol{\theta}$, it is clear that the bivariate model involves the estimation of a covariance parameter σ_{12} in addition to the univariate counterparts. Similarly, a multivariate model with three responses necessarily requires the estimate of three more covariance parameters as compared to the respective univariate models $(\sigma_{12}, \sigma_{13}, \sigma_{23})$.

3 Application

This work was motivated by a water quality monitoring program in the River Thames in the United Kingdom. The case study data were collected from the Environmental Information Data Centre of the Centre for Ecology and Hydrology in the United Kingdom (Bowes et al., 2017). This data set includes weekly water quality monitoring data from seven sampling locations along the River Thames, as well as fifteen of its major tributaries, from March 2009 to February 2013, see Table 1 and Figure 1. There are 4300 observations in total, with 45 columns of water quality parameters including water temperature, pH value, nitrogen species, dissolved sodium and chloride concentrations. Here, some observations were excluded from the data set as they included missing values. The remaining weekly data were then converted to monthly data by taking monthly averages, due to their flexibility and ease of representation.

Table 1: Water quality monitoring stations along the River Thames in the United Kingdom, as well as fifteen of its major tributaries. Station numbers have been assigned arbitrary for the identification purposes only.

Station number	Monitoring station	Station number	Monitoring station
1	Jubilee River at Pocock's Bridge	12	River Ray at Islip
2	River Cherwell at Hampton Poyle	13	River Thame at Wheatley
3	River Cole at Lynt Bridge	14	River Thames at Hannington Wick
4	River Coln at Whelford	15	River Thames at Newbridge
5	River Enborne at Brimpton	16	River Thames at Runnymede
6	River Evenlode at Cassington Mill	17	River Thames at Sonning
7	River Kennet at Woolhampton	18	River Thames at Swinford
8	River Leach at Mill Lane,Lechlade	19	River Thames at Wallingford
9	River Lodden at Charvil	20	River Windrush at Newbridge
10	River Ock at Abingdon	21	River Wye at Bourne End
11	River Pang at Tidmarsh	22	The Cut at Paley Street

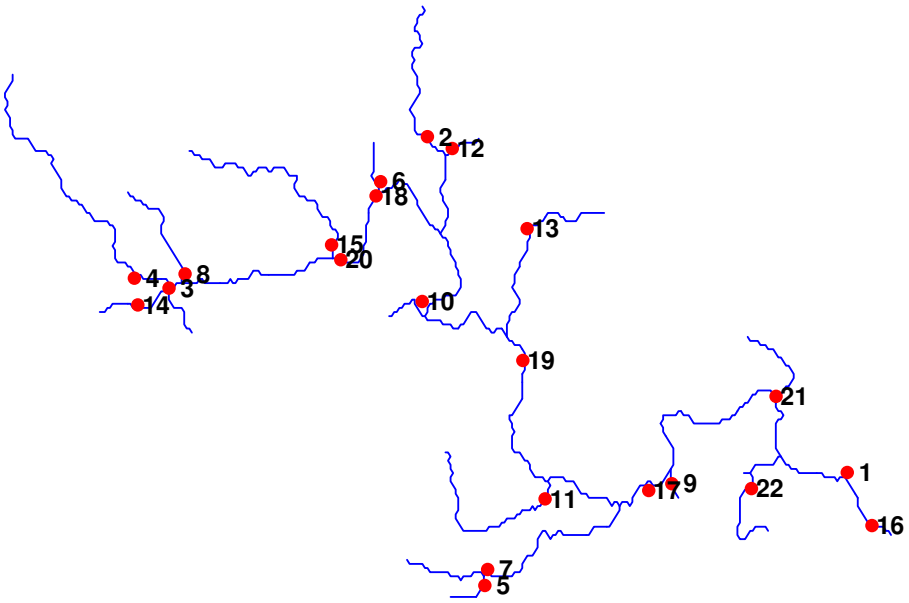

Figure 1: The River Thames water quality monitoring network. The names of the monitoring stations that correspond to the numbers indicated here are listed in Table 1.

To investigate the performance of the proposed approach for modeling a multivariate response, we considered three water quality parameters; dissolved sodium (Y_1), dissolved chloride (Y_2) and dissolved nitrite (Y_3) concentrations, as response variables. These responses were chosen for their known adverse effects on human health (Kumar and Puri, 2012; Mueller et al., 1997; Hallenbeck et al., 1981) and their usefulness in evaluating water quality. In addition, we observed a significant correlation between each pair of responses, which is another motivation for employing a multivariate STAM in this study. Water quality characteristics such as water temperature and pH value were employed as predictor variables in the models as they are straightforward to measure. To evaluate the effectiveness of the proposed modelling approach, two scenarios

were considered, which differ in terms of the number of responses utilised when formulating the multivariate model, see Table 2. Furthermore, under each scenario, two cases were considered where each case is different in terms of the number of prediction locations included in the analysis. These two cases were considered to demonstrate the predictive performance of the proposed multivariate models, under different conditions.

Table 2: The number of response variables and the number of prediction locations utilised under the two scenarios adopted in this study.

Scenario	Dimension of the multivariate model	Number of prediction locations
1	bivariate (Y_1,Y_2)	5
		10
2	trivariate (Y_1,Y_2,Y_3)	5
		10

All simulations were run using RStudio 1.4.1106, and R code to reproduce the results in this paper is available via the following GitHub repository, https://github.com/SenarathneSGJ/Water_Quality_Modelling.

4 Results

Scenario 1: Here, we first investigated the predictive performance of the bivariate STAM and its univariate counterparts based on a randomly selected five and ten prediction locations, see Figures 2 and 3, respectively. For this comparison, we used station wise prediction means and prediction intervals, where the width of a prediction interval is equal to the prediction mean plus or minus standard deviation. As it can be seen, the proposed bivariate model produces narrow prediction intervals than univariate models. This implies that the proposed bivariate spatial model has higher prediction accuracy compared to its univariate counterparts.

Next, to confirm the results obtained for a random set of locations shown in Figures 2 and 3, we re-evaluated the prediction performance of the bivariate model based on 500 independent simulations. In each simulation, a set of prediction locations were selected randomly, and then the multivariate and the corresponding univariate models were used for predictions for those locations. Those results are summarised in Figures 4 and 5. Accordingly, for the majority of prediction sites, the bivariate model yields lower prediction intervals and predicted mean values that are closer to actual values than the univariate models. Of note, the actual values for the monitoring stations 21 and 22 were not included in the prediction intervals of both bivariate and univariate models. It may be due to the water quality parameters that we predict for these locations are significantly lower than the actual values.

To statistically compare the prediction accuracies of the bivariate and univariate STAMs, the absolute prediction errors (obtained using 500 runs) were assessed using two samples Hotelling's T-square test. The Hotelling's T-square is a statistic for a multivariate test of discrepancies between the mean values of two groups, where the null hypothesis states that centroids of the groups are identical. As shown in Table 3, the p-values of the hypothesis tests are less than 0.05 significance level, which indicates that there is a significant difference between the mean absolute prediction error (MAPE) of bivariate model and the univariate models. Following

the results of this test, post-hoc comparisons were made using 95% confidence intervals for the difference of MAPE of bivariate and univariate models, see Table A.1 in Appendix. According to the post-hoc test findings, the MAPE of the bivariate model is less than that of the univariate model for the majority of prediction locations.

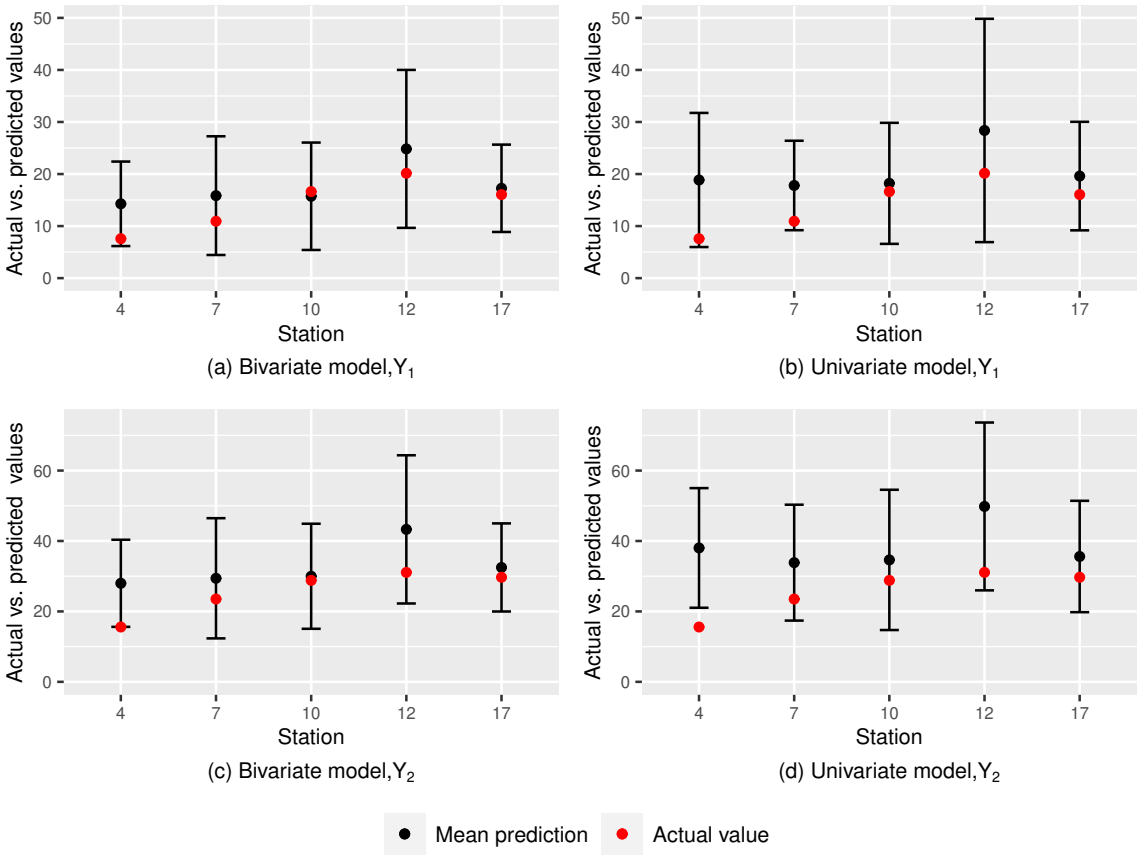

Figure 2: Predictions (mean±standard deviation) from the bivariate and univariate STAMs for the randomly selected five locations.

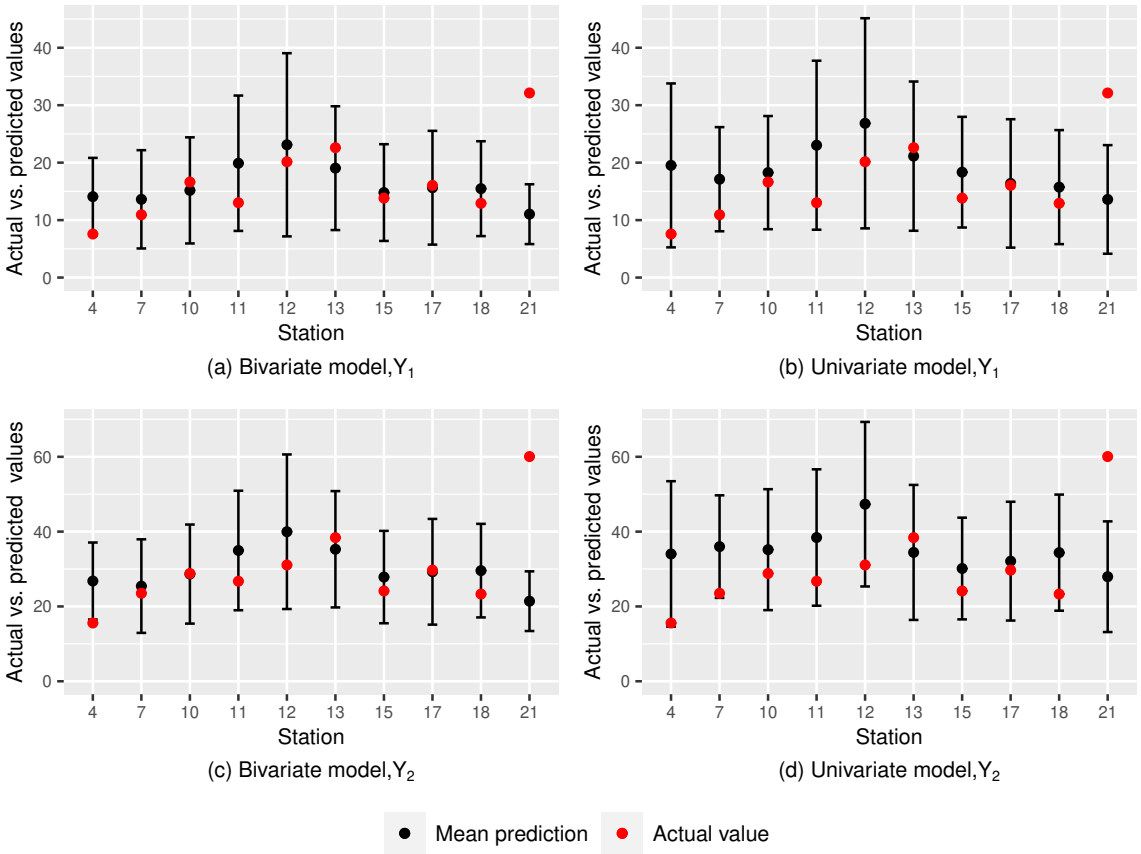

Figure 3: Predictions (mean±standard deviation) from the bivariate and univariate STAMs for the randomly selected ten locations.

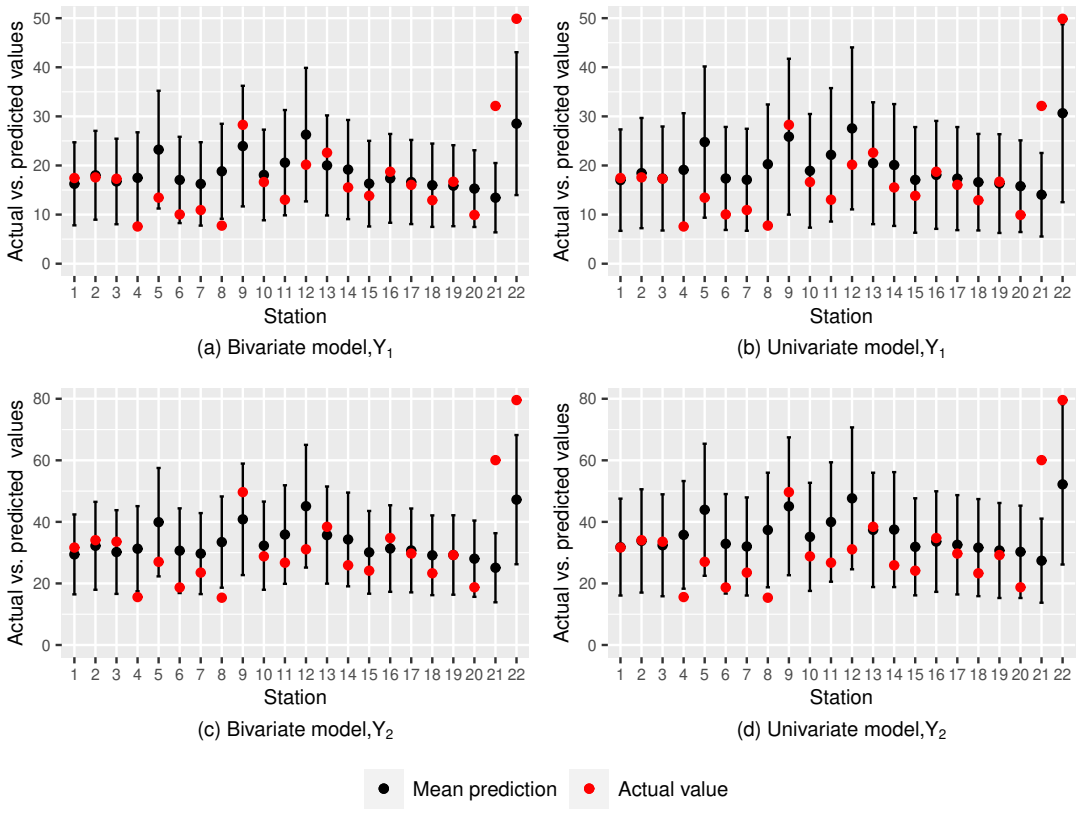

Figure 4: Prediction summary (mean±standard deviation) from 500 simulations in each with bivariate and univariate STAMs were used to predict the randomly selected five locations.

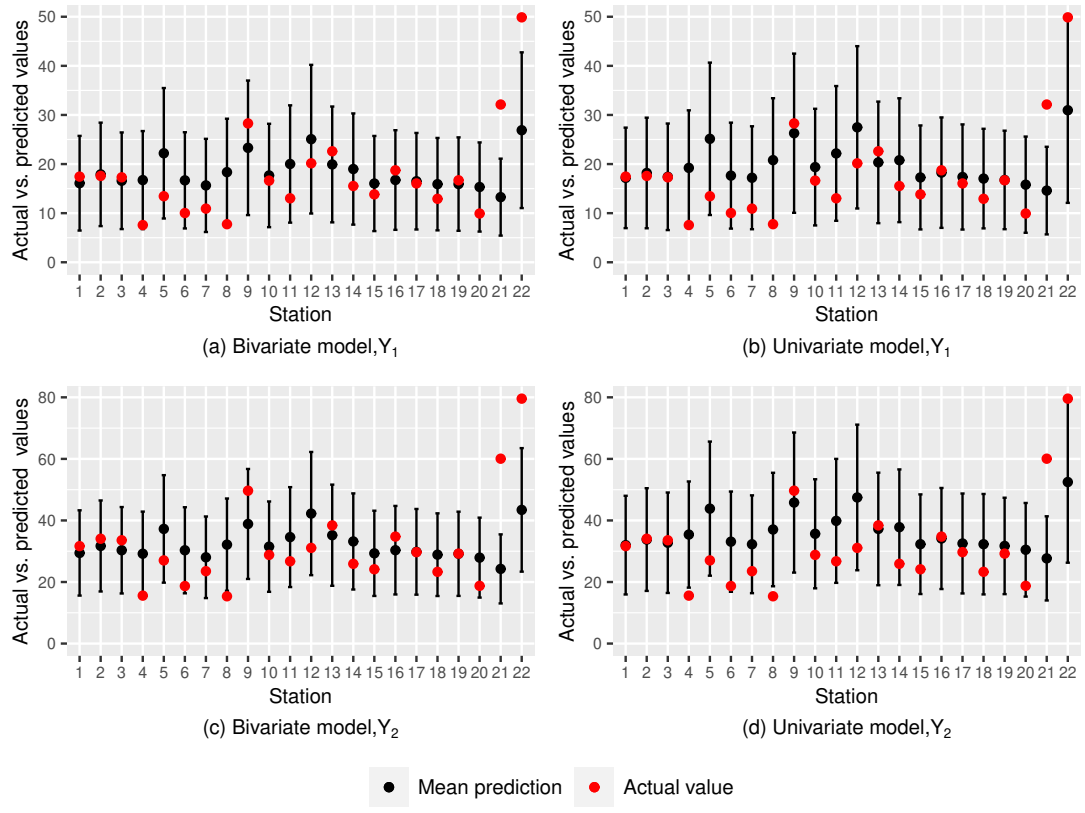

Figure 5: Prediction summary (mean±standard deviation) from 500 simulations in each with bivariate and univariate STAMs were used to predict the randomly selected ten locations.

Table 3: Comparison of bivariate and univariate models using two samples Hotelling's T-square test.

Number of prediction locations	Response	Test statistic	p-value
n=5	Y_1	173.8810	0.0000
	Y_2	318.5971	0.0000
n=10	Y_1	721.7687	0.0000
	Y_2	1249.1760	0.0000

Scenario 2: Similar to Scenario 1, we first assessed the prediction performance of the proposed multivariate model with three responses by utilising a randomly selected five and ten prediction locations, see Figures 6 and 7, respectively. For some prediction locations, it can be observed that the prediction intervals produced from multivariate models are smaller than those derived from univariate models. In summary, these results show that the multivariate model slightly outperformed univariate models in terms of prediction accuracy.

Next, we examined the prediction performance of the multivariate model based on 500 independent simulations, with Figures 8 and 9 depicting the results for five and ten randomly selected prediction locations, respectively. For the majority of prediction locations, it can be seen that the multivariate STAM has lower prediction intervals and prediction mean values that are closer to actual values than univariate models. Furthermore, as previously stated, the maximum values of the prediction intervals of the responses Y_1 and Y_2 for both multivariate and univariate models were less than the actual values for the monitoring stations 21 and 22.

The two sample Hotelling's T-square test results obtained for this scenario is shown in Table 4, which shows that there is a significant difference between the predictions of multivariate and univariate models in both cases (i.e., for $n = 5$ and $n = 10$). Furthermore, using the post-hoc comparison findings acquired for this scenario, it was enable to identify the places where the multivariate model outperforms the comparable univariate models in terms of MAPE. A summary of these results are shown in Table A.2 in Appendix.

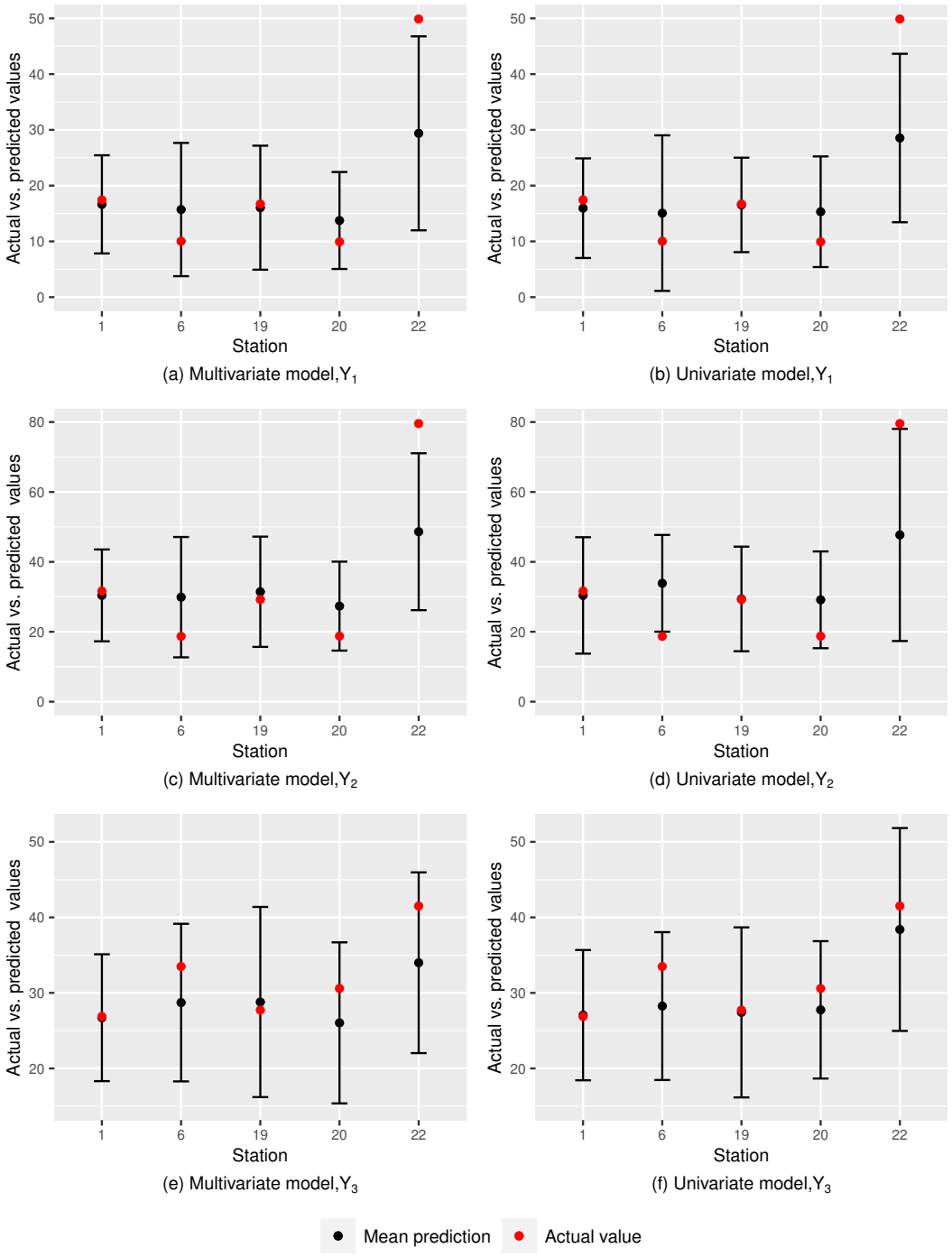

Figure 6: Predictions (mean±standard deviation) from the multivariate and univariate STAMs for the randomly selected five locations.

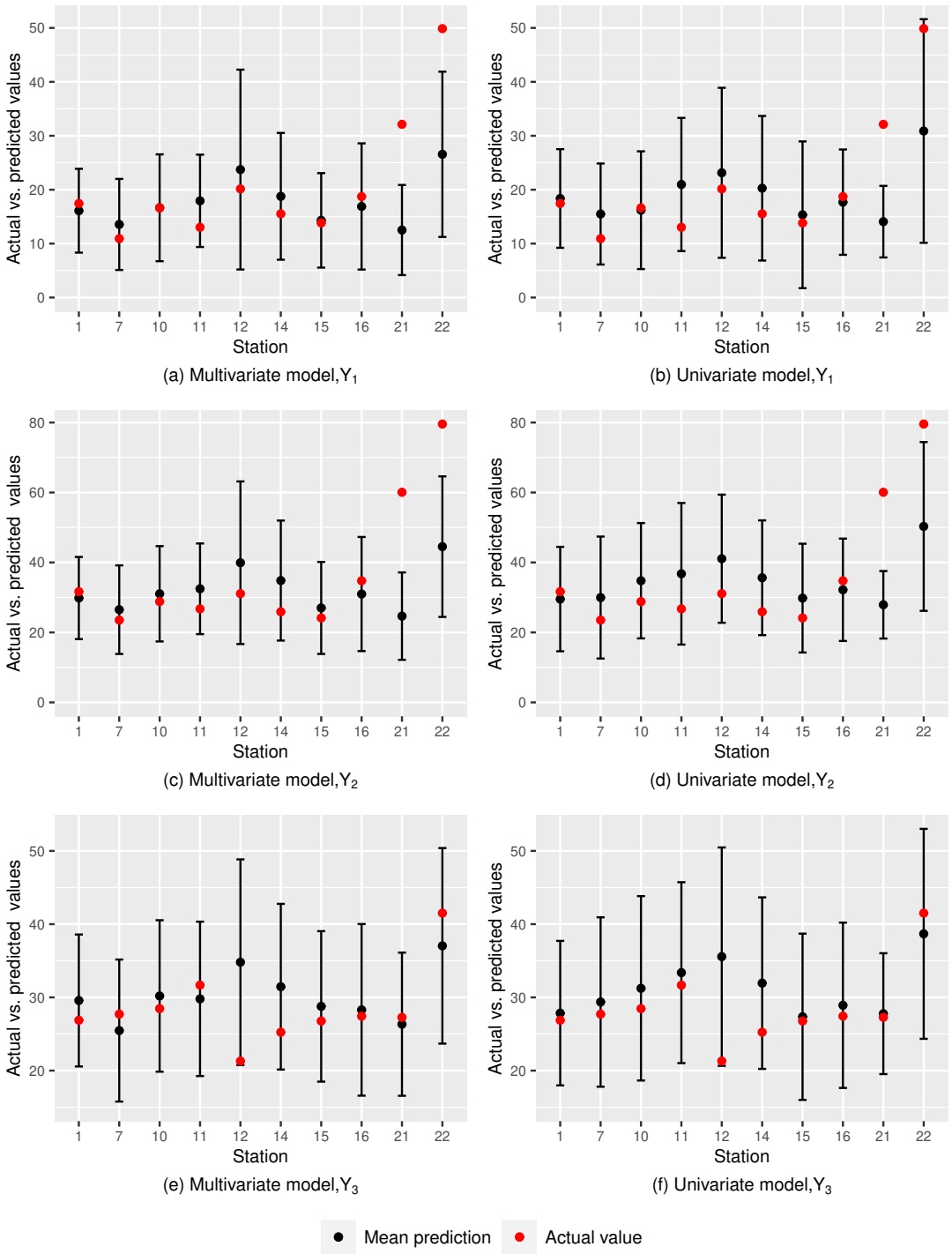

Figure 7: Predictions (mean±standard deviation) from the multivariate and univariate STAMs for the randomly selected ten locations.

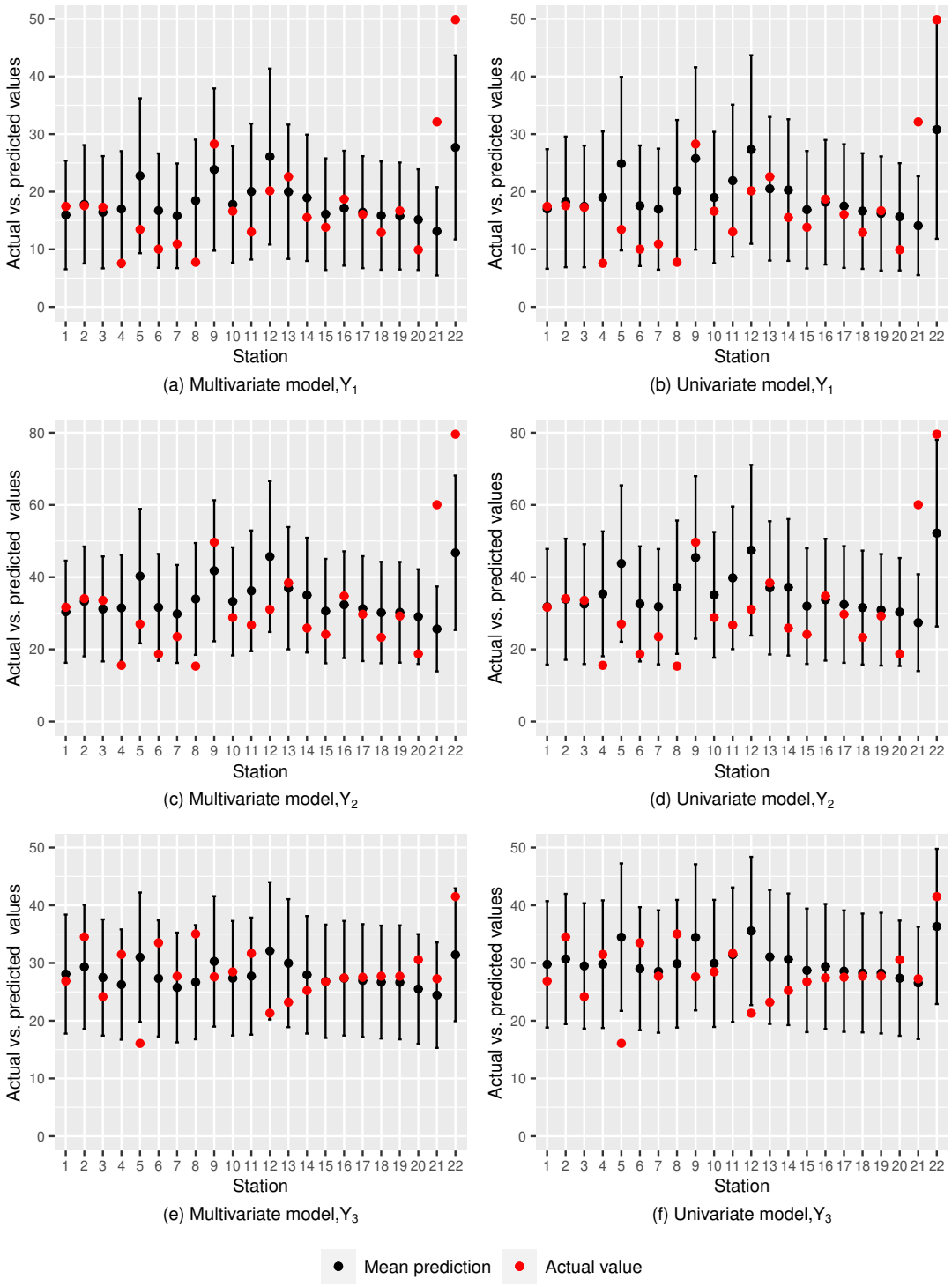

Figure 8: Prediction summary (mean±standard deviation) from 500 simulations in each with multivariate and univariate STAMs were used to predict the randomly selected five locations.

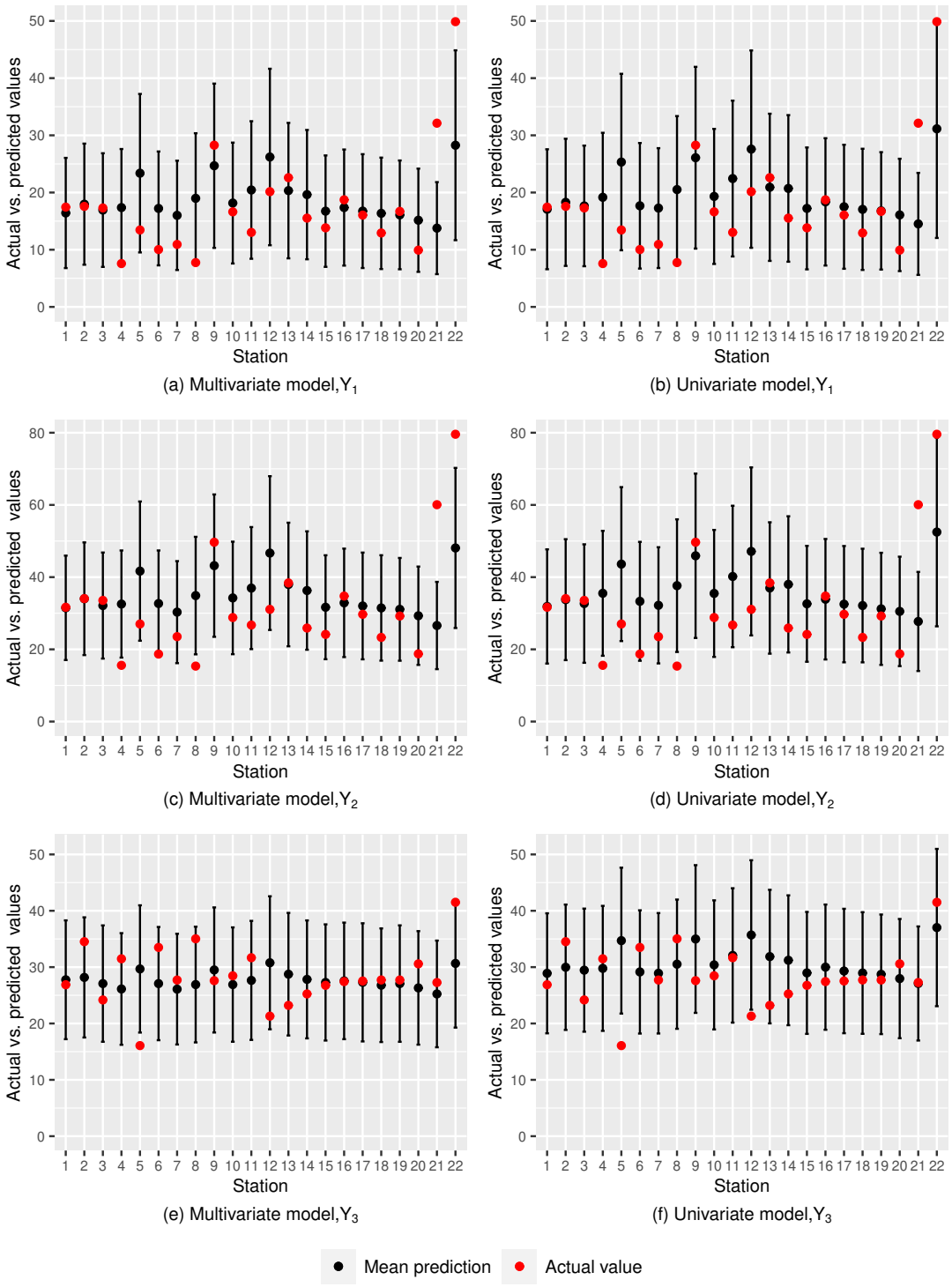

Figure 9: Prediction summary (mean±standard deviation) from 500 simulations in each with multivariate and univariate STAMs were used to predict the randomly selected ten locations

Table 4: Comparison of multivariate and univariate models using two samples Hotelling's T-square test.

Number of prediction locations	Response	Test statistic	p-value
n=5	Y_1	1103.3653	0.0000
	Y_2	843.8431	0.0000
	Y_3	1759.3320	0.0000
n=10	Y_1	589.2464	0.0000
	Y_2	417.9147	0.0000
	Y_3	2041.2734	0.0000

5 Discussion

The present study was designed to develop a multivariate modelling approach to minimise the prediction error of STAMs. The approach was based on the derivation of a covariance structure where spatial auto-correlation was modeled using hydrologic distances, which may effectively capture specific aspects of streams and rivers. It was shown that the variability of spatial process over time could be captured using a smoothing function. Furthermore, we extended univariate STAMs to include multivariate spatio-temporal responses, allowing for a more flexible modeling approach in the geostatistical contexts. In addition, we extended our methodology to handle time dependent data at each location, which was inspired by what is observed in real-world studies. Overall, these improvements demonstrate the potential for enhancing water quality modeling in order to reduce associated water quality surveying work as well as other ecological monitoring.

Two scenarios were explored in relation to the case study in order to evaluate our proposed approach and multivariate model under various conditions based on multiple responses and number of prediction locations. Overall, the proposed multivariate models outperformed their univariate counterparts. However, as the dimension of the multivariate models increases, the predictive performance of the model approaches that of univariate models. This is primarily due to the enormous number of additional parameters that must be evaluated in the multivariate model, necessitating a large number of data points to accurately estimate them. As such, it is essential to consider the dimension of the multivariate model based on the availability of the data and the application.

The responses or water quality parameters used in this investigation were chosen because of their known detrimental health effects. Despite this, all of these parameters must be determined under laboratory settings, making the measurement process time consuming and costly. Using the novel approach proposed in this paper, we were able to predict these parameters with high accuracy over the entire network based on a selection of stations. As a result, our approach has the potential to dramatically enhance overall water quality monitoring.

Kriging is the most extensively used statistical method for modelling spatially dependent data, which assumes inter-point distances to be Euclidean distances. This assumption further guarantees that the spatial covariance matrix corresponding to the variable of interest is positive-definite (Cressie, 2015). However, several attempts have been made to validate the use of non-Euclidean distances for spatial prediction via kriging (Boisvert and Deutsch, 2011; Ver Hoef, 2018; Lu et al., 2014; Curriero, 2006). Most of these efforts concentrated on transforming geo-

logical distances, for example through multi-dimensional scaling (Murphy et al., 2015), which is known to have a significant amount of bias in prediction variance (Davis and Curriero, 2019).

Despite the theoretical underpinnings of the covariance structure of multiple spatial responses proposed in this study, the normal distribution assumption of the responses is a drawback of this approach. In fact, such an approach might not be appropriate in general. Because, if bivariate continuous and binary data are observed, a multivariate normal approximation is unlikely to be appropriate for estimating the covariance structure of the data. As such, further work is required to explore alternative methods such as Copula based models for mixed outcomes (Senarathne et al., 2020). In studies where high-dimensional multivariate data are observed, it may be important to use certain Vine-Copulas (Brechmann et al., 2013). Other areas we hope to pursue into the future include extensions of the proposed covariance structure to quantify the variability spatio-temporal processes where temporal variability is added to the model as a random effect (Liu and Vanhatalo, 2019).

References

Boisvert, J. B. and Deutsch, C. V. (2011), 'Programs for kriging and sequential gaussian simulation with locally varying anisotropy using non-euclidean distances', *Computers & Geosciences* **37**(4), 495–510.

Bowes, M., Armstrong, L., Wickham, H., Harman, S., Gozzard, E., Roberts, C. and Scarlett, P. (2017), 'Weekly water quality data from the river thames and its major tributaries (2009-2013) [CEH Thames Initiative]'.
URL: *https://doi.org/10.5285/e4c300b1-8bc3-4df2-b23a-e72e67eef2fd*

Brechmann, E. C., Schepsmeier, U. et al. (2013), 'Modeling dependence with C-and D-vine Copulas: The R-package CDVine', *Journal of Statistical Software* **52**(3), 1–27.

Crainiceanu, C. M., Ruppert, D. and Wand, M. P. (2005), 'Bayesian analysis for penalized spline regression using WinBUGS', *Journal of Statistical Software* **14**(14), 1–24.

Cressie, N. (2015), *Statistics for spatial data*, John Wiley & Sons.

Cressie, N., Frey, J., Harch, B. and Smith, M. (2006), 'Spatial prediction on a river network', *Journal of agricultural, biological, and environmental statistics* **11**(2), 127–150.

Curriero, F. C. (2006), 'On the use of non-euclidean distance measures in geostatistics', *Mathematical Geology* **38**(8), 907–926.

Davis, B. J. and Curriero, F. C. (2019), 'Development and evaluation of geostatistical methods for non-euclidean-based spatial covariance matrices', *Mathematical geosciences* **51**(6), 767–791.

Friedman, J. H. and Stuetzle, W. (1981), 'Projection pursuit regression', *Journal of the American statistical Association* **76**(376), 817–823.

Ganio, L. M., Torgersen, C. E. and Gresswell, R. E. (2005), 'A geostatistical approach for describing spatial pattern in stream networks', *Frontiers in Ecology and the Environment* **3**(3), 138–144.

Gardner, B., Sullivan, P. J. and Lembo, Jr, A. J. (2003), 'Predicting stream temperatures: geostatistical model comparison using alternative distance metrics', *Canadian Journal of Fisheries and Aquatic Sciences* **60**(3), 344–351.

Hallenbeck, W. H., Brenniman, G. R. and Anderson, R. J. (1981), 'High sodium in drinking water and its effect on blood pressure', *American journal of epidemiology* **114**(6), 817–826.

Isaak, D. J., Peterson, E. E., Ver Hoef, J. M., Wenger, S. J., Falke, J. A., Torgersen, C. E., Sowder, C., Steel, E. A., Fortin, M.-J., Jordan, C. E., Ruesch, A. S., Som, N. and Monestiez, P. (2014), 'Applications of spatial statistical network models to stream data', *WIREs Water* **1**(3), 277–294.

Korashey, R. (2009), 'Using regression analysis to estimate water quality constituents in bahr el baqar drain', *Journal of Applied Sciences Research* **5**(8), 1067–1076.

Kumar, M. and Puri, A. (2012), 'A review of permissible limits of drinking water', *Indian Journal of Occupational and Environmental Medicine* **16**(1), 40–44.

Liu, J. and Vanhatalo, J. (2019), 'Bayesian model based spatiotemporal survey designs and partially observed log Gaussian cox process', *Spatial Statistics* **35**, 1–27.

Lu, B., Charlton, M., Harris, P. and Fotheringham, A. S. (2014), 'Geographically weighted regression with a non-euclidean distance metric: a case study using hedonic house price data', *International Journal of Geographical Information Science* **28**(4), 660–681.

Lyon, S. W., Seibert, J., Lembo, A. J., Steenhuis, T. S. and Walter, M. T. (2008), 'Incorporating landscape characteristics in a distance metric for interpolating between observations of stream water chemistry', *Hydrology and Earth System Sciences* **12**(5), 1229–1239.

Mueller, D. K., Ruddy, B. and Battaglin, W. (1997), 'Logistic model of nitrate in streams of the upper-midwestern united states', *Journal of Environmental Quality* **26**, 1223–1230.

Murphy, R. R., Perlman, E., Ball, W. P. and Curriero, F. C. (2015), 'Water-distance-based kriging in chesapeake bay', *Journal of Hydrologic Engineering* **20**(9), 05014034.

Nandy, S., Lim, C. Y. and Maiti, T. (2017), 'Additive model building for spatial regression', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**(3), 779–800.

Olea, R. A. (1991), *Geostatistical glossary and multilingual dictionary*, number 3, Oxford University Press on Demand.

Peterson, E. E., Theobald, D. M. and VER HOEF, J. M. (2007), 'Geostatistical modelling on stream networks: developing valid covariance matrices based on hydrologic distance and stream flow', *Freshwater biology* **52**(2), 267–279.

Peterson, E. E. and Ver Hoef, J. M. (2010), 'A mixed-model moving-average approach to geostatistical modeling in stream networks', *Ecology* **91**(3), 644–651.

Peterson, E. E., Ver Hoef, J. M., Isaak, D. J., Falke, J. A., Fortin, M.-J., Jordan, C. E., McNyset, K., Monestiez, P., Ruesch, A. S., Sengupta, A. et al. (2013), 'Modelling dendritic ecological networks in space: an integrated network perspective', *Ecology letters* **16**(5), 707–719.

Qiu, P. (1998), 'Discontinuous regression surfaces fitting', *The Annals of Statistics* **26**(6), 2218–2245.

Senarathne, S. G. J., McGree, J. M. and Müller, W. G. (2020), 'Bayesian design for minimising

uncertainty in spatial processes'.

Seya, H., Tsutsumi, M., Yoshida, Y. and Kawaguchi, Y. (2011), 'Empirical comparison of the various spatial prediction models: in spatial econometrics, spatial statistics, and semiparametric statistics', *Procedia - Social and Behavioral Sciences* **21**, 120–129.

Smith, R. A., Schwarz, G. E. and Alexander, R. B. (1997), 'Regional interpretation of water-quality monitoring data', *Water Resources Research* **33**(12), 2781–2798.

Ver Hoef, J. M. (2018), 'Kriging models for linear networks and non-euclidean distances: Cautions and solutions', *Methods in Ecology and Evolution* **9**(6), 1600–1613.

Ver Hoef, J. M., Cressie, N., Fisher, R. N. and Case, T. J. (2001), Uncertainty and spatial linear models for ecological data, *in* 'Spatial uncertainty in ecology', Springer, pp. 214–237.

Ver Hoef, J. M., Peterson, E. and Theobald, D. (2006), 'Spatial statistical models that use flow and stream distance', *Environmental and Ecological statistics* **13**(4), 449–464.

Vercelloni, J., Caley, M. J., Kayal, M., Low-Choy, S. and Mengersen, K. (2014), 'Understanding uncertainties in non-linear population trajectories: A Bayesian semi-parametric hierarchical approach to large-scale surveys of coral cover', *PloS one* **9**(11), e110968.

Vercelloni, J., Mengersen, K., Ruggeri, F. and Caley, M. J. (2017), 'Improved coral population estimation reveals trends at multiple scales on Australia's Great Barrier Reef', *Ecosystems* **20**(7), 1337–1350.

Wood, S. N. (2003), 'Thin plate regression splines', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**(1), 95–114.

Appendix

Table A.1: Post-hoc comparisons results to evaluate station-wise prediction accuracy of bivariate and univariate models based on mean predictions of 500 simulations

Station	Response 1		Response 2	
Number	d=5	d=10	d=5	d=10
1	(-2.4108, 0.3247)	**(-3.9731, -1.1167)**	**(-6.0102, -0.4795)**	**(-9.5114, -3.3545)**
2	(-3.6080, 0.2625)	**(-4.8872, -0.9479)**	(-7.1029, 0.7452)	**(-10.1478, -2.7671)**
3	(-1.9675, 0.3428)	**(-4.1884, -0.7413)**	(-4.3191, 0.5229)	**(-8.0956, -1.6714)**
4	(-2.2120, 0.4523)	**(-3.1719, -0.2645)**	**(-5.4299, -0.2360)**	**(-7.3911, -1.3583)**
5	**(-4.9621, -0.2483)**	**(-4.5647, -0.5055)**	**(-10.6878, -1.0766)**	**(-8.9460, -1.6151)**
6	(-2.0457, 0.5021)	(-1.9207, 1.0814)	**(-5.0630, -0.2559)**	(-4.9541, 0.8807)
7	(-1.7077, 0.3475)	**(-2.5724, -0.0348)**	**(-4.9147, -0.4513)**	**(-5.5109, -0.5965)**
8	**(-3.7667, -0.1555)**	**(-2.9335, -0.1965)**	**(-8.3572, -1.1839)**	**(-6.4677, -1.5094)**
9	**(-2.9596, -0.1015)**	(-2.4903, 0.1797)	**(-8.0569, -1.0547)**	**(-6.1179, -0.6786)**
10	(-2.1363, 0.2811)	**(-6.4700, -1.5305)**	**(-5.2014, -0.1170)**	**(-13.2253, -4.6544)**
11	**(-3.4385, -0.2104)**	(-2.4045, 0.2291)	**(-8.0560, -1.8071)**	**(-5.0787, -0.2038)**
12	(-2.4409, 0.1660)	**(-2.8302, -0.0440)**	**(-6.7585, -0.9077)**	**(-7.0348, -1.1735)**
13	(-1.8216, 0.4542)	**(-3.9429, -0.3326)**	(-4.4621, 0.6697)	**(-8.1657, -1.9741)**
14	(-1.7525, 0.8997)	(-3.4944, 0.0336)	(-4.7793, 0.7634)	**(-7.7706, -1.5777)**
15	**(-3.3503, -0.2602)**	**(-2.4004, -0.1750)**	**(-7.9021, -2.0324)**	**(-5.5528, -0.8965)**
16	**(-4.3399, -0.2901)**	(-2.0789, 0.4575)	**(-8.9079, -1.2550)**	(-4.8483, 0.1101)
17	(-2.0909, 0.4470)	(-2.3828, 0.4457)	(-4.5451, 0.2527)	**(-5.7115, -0.0667)**
18	(-1.5310, 0.5324)	**(-4.8289, -0.4678)**	(-4.7427, 0.1683)	**(-10.4414, -2.8806)**
19	(-1.8382, 0.7058)	(-1.5531, 1.1257)	(-3.6544, 0.4834)	(-4.4620, 0.6366)
20	(-1.8376, 0.4581)	(-2.1854, 0.3244)	**(-6.1298, -0.3389)**	**(-5.0305, -0.2024)**
21	(-1.4881, 0.8070)	(-1.7080, 0.7333)	(-4.9044, 0.1607)	**(-5.1559, -0.0014)**
22	(-1.7669, 0.5539)	(-2.1127, 0.4101)	(-4.4512, 0.0619)	(-5.4764, 0.0976)

Table A.2: Post hoc comparisons results to evaluate station-wise prediction accuracy of multivariate and univariate models based on mean predictions of 500 simulations

Station	Response 1		Response 2		Response 3	
Number	d=5	d=10	d=5	d=10	d=5	d=10
1	(-2.5055, 0.0021)	**(-3.1720, -0.2094)**	(-3.6596, 0.2642)	**(-5.1121, -0.6419)**	**(-4.4626, -0.8420)**	**(-5.1191, -1.7640)**
2	(-3.2760, 0.6833)	**(-3.5522, -0.0224)**	(-4.9045, 1.2430)	(-4.2601, 1.0266)	**(-7.0136, -1.3919)**	**(-7.9013, -2.4116)**
3	(-1.8965, 0.3032)	(-3.0266, 0.0421)	(-3.2409, 0.2711)	**(-5.5746, -0.2353)**	**(-3.2940, -0.5164)**	**(-5.3030, -1.6533)**
4	(-2.0963, 0.2135)	(-2.5021, 0.1678)	(-3.0148, 0.7868)	(-3.4461, 1.1087)	**(-4.2442, -0.2568)**	**(-5.3523, -1.6285)**
5	**(-4.8825, -1.1987)**	(-3.4302, 0.1671)	**(-8.2825, -1.8811)**	(-3.6974, 2.4690)	**(-8.2490, -2.6935)**	**(-8.1461, -1.7650)**
6	(-2.4078, 0.1371)	(-1.9233, 0.8093)	(-3.3865, 0.5901)	(-1.2938, 3.1904)	**(-3.4775, -0.0241)**	**(-5.1049, -1.0495)**
7	(-2.1064, 0.0992)	(-1.8107, 0.6538)	**(-3.5902, -0.1382)**	(-3.3987, 1.1298)	(-4.3438, 0.1694)	**(-3.0875, -0.2083)**
8	**(-3.6028, -0.5901)**	(-2.1221, 0.2201)	**(-6.0275, -0.8661)**	(-2.9679, 1.1279)	**(-6.4506, -1.1484)**	**(-4.3815, -0.8284)**
9	**(-3.1621, -0.3954)**	(-1.8702, 0.6368)	**(-5.4180, -1.1923)**	(-2.5485, 1.4324)	**(-5.2930, -1.3083)**	**(-3.5620, -0.4091)**
10	(-2.3858, 0.0220)	**(-4.9469, -1.0947)**	(-3.9952, 0.0025)	**(-7.5518, -1.2026)**	**(-4.2409, -1.1643)**	**(-9.8667, -3.0167)**
11	**(-3.1746, -0.4011)**	(-1.9037, 0.5758)	**(-5.9257, -1.1746)**	(-2.4873, 1.6730)	**(-5.7038, -2.1724)**	(-2.5649, 0.2065)
12	(-2.5628, 0.2297)	(-2.6099, 0.0901)	(-4.1234, 0.2802)	(-3.9908, 0.4541)	**(-4.6734, -1.1714)**	**(-4.4457, -1.0893)**
13	(-1.6274, 0.7326)	**(-3.4355, -0.4703)**	(-2.3602, 1.6133)	**(-5.6017, -0.7804)**	**(-3.3619, -0.0847)**	**(-6.5131, -2.1038)**
14	(-1.9619, 0.6136)	(-2.5936, 0.4669)	(-2.2615, 1.8022)	(-4.6950, 0.8512)	(-2.9722, 0.6700)	**(-5.2835, -1.2195)**
15	**(-3.4133, -0.9137)**	(-1.7020, 0.3124)	**(-6.0430, -1.6641)**	(-2.8104, 0.7532)	**(-5.3585, -1.5760)**	**(-3.6037, -0.4166)**
16	**(-3.7414, -0.1465)**	(-1.8901, 0.4533)	**(-6.2488, -0.5151)**	(-2.6194, 1.3609)	**(-7.3355, -1.8601)**	**(-4.1356, -0.5694)**
17	**(-2.5771, -0.0723)**	(-1.7477, 0.7694)	(-3.1078, 0.7801)	(-2.7239, 1.5241)	**(-3.1346, -0.2265)**	**(-3.7074, -0.5380)**
18	(-1.7705, 0.5420)	(-3.2873, 0.5009)	(-3.1520, 0.6857)	(-5.6284, 0.5407)	**(-3.5025, -0.3062)**	**(-8.1949, -2.8129)**
19	(-1.5853, 0.8359)	(-1.5205, 0.9097)	(-2.2566, 1.5167)	(-1.7816, 2.1794)	(-3.6218, 0.4617)	**(-3.5760, -0.0739)**
20	(-1.8596, 0.3259)	(-1.8940, 0.3439)	(-2.9193, 0.8380)	(-2.0948, 1.8025)	(-3.1201, 0.1453)	**(-2.8887, -0.2799)**
21	(-1.9853, 0.1767)	(-2.0022, 0.2441)	(-2.5760, 1.1013)	(-3.2275, 0.7981)	**(-3.3092, -0.0794)**	**(-3.0768, -0.2001)**
22	(-2.1299, 0.1171)	(-1.8548, 0.4686)	(-3.1513, 0.4729)	(-2.3862, 1.3972)	**(-3.8314, -0.1782)**	**(-3.5414, -0.4247)**