Generalized Additive Mixed Modelling of River Discharge in the Black Volta River

Wahab A. Iddrisu[1], Kaku S. Nokoe[1], Albert Luguterah[2], Eric O. Antwi[3]

[1]Department of Mathematics and Statistics, University of Energy and Natural Resources, Sunyani, Ghana
[2]Department of Statistics, University for Development Studies, Navrongo, Ghana
[3]Department of Energy and Environmental Engineering, University of Energy and Natural Resources, Sunyani, Ghana
Email: iddrisu.wahab@uenr.edu.gh

Abstract

River discharge data offer a rich source of information for reservoir management and flood control, if modelling can separate out the effects of rainfall, land use, soil type, relief, and weather conditions. In this paper, we model river discharge data from the Black Volta River, using Generalised Additive Mixed Models (GAMMs) with a space-time interaction represented via a tensor product of continuous time and discrete space. River discharge data from January 2000 to December 2009 for the four gauge stations along the Black Volta River namely, Lawra, Chache, Bui and Bamboi were obtained from the hydrological services department of Ghana and used for model fitting. Four GAMMs were explored, two with space-time interactions and two without space-time interactions. The comparison of the performance of the models with space-time interactions and those without space-time interactions based on Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) suggests that in this application, the former is better overall and in particular for modelling local variations. Further, a model with space and time main effects performed better compared with one without space and time main effects. After model selection, checking and validation, there is evidence for increasing river discharge from the most upstream gauge station to the most downstream gauge station for the study period.

Keywords

River Discharge, GAMM, Tensor Product Smooth, Space-Time Interaction, Black Volta River

1. Introduction

Most important tasks in problem solving in hydrology have been taken over by

mathematical models [1]. According to [2], modelling in environmental science is the representation of a complex natural system in a simplified form through the use of logical mathematical statements. Most hydrologic systems are extremely complex, and we cannot hope to understand them in detail without modelling [3].

Many different reasons account for the development of hydrologic models for a catchment. They therefore have many different forms despite the fact that they are in general developed to meet at least one of two primary objectives [3]. One objective is to gain a better understanding of the hydrologic phenomena operating in a catchment and of how changes in the catchment may affect these phenomena whiles the other objective is to generate synthetic sequences of hydrologic data for facility design or for use in forecasting [3].

River discharge and other components of the hydrologic system are affected by many variables. Key among them is rainfall and its variation in space and time in response to various climatic factors. Other variables that potentially can affect river discharge include rock and soil type, land use, relief and weather conditions such as temperature and humidity. Establishing a relationship among these variables is the central focus of hydrological modelling from its simple form of unit hydrograph to rather complex models based on fully dynamic flow equations [4].

Hydrologic models can be classified into two broad classes, namely physical and abstract models [5]. Physical models can further be categorised into two categories, namely scale and analog models. A scale model refers to a scaled down model of a real system whiles an analog model refers to a physical system having the same characteristic as the original sample. Abstract models on the other hand, are used to show a system in a mathematical form. The model is operated with a set of equations, input and output data. These models are data-driven in nature, as they do not require knowledge of the underlying process beforehand and are solely based on empirical equations calibrated to field data [6].

Quite recently, [7] argued that hydrologic models may be seen as black-box, conceptual or deterministic models. Black-box models explain the relationship between the input and output data mathematically [8] and are often good for modelling with available and analyzed data for a specific catchment. Deterministic models have complex physical theory and need to have a large amount of data and computational time. Conceptual models are formulated with a number of conceptual elements which are simple representations of a reference system [9].

A significant number of physically based and data-driven models have been developed and implemented. Examples include [10]-[21]. Although it is easier understanding the separate hydrological processes that govern the whole system using the physically based models, in many occasions the input data may be unavailable, expensive or time consuming to collect [22]. Also, a number of vari-

ables still need to be determined through model calibration. This makes the operation of physically-based models difficult and time consuming as opposed to data-driven models [23].
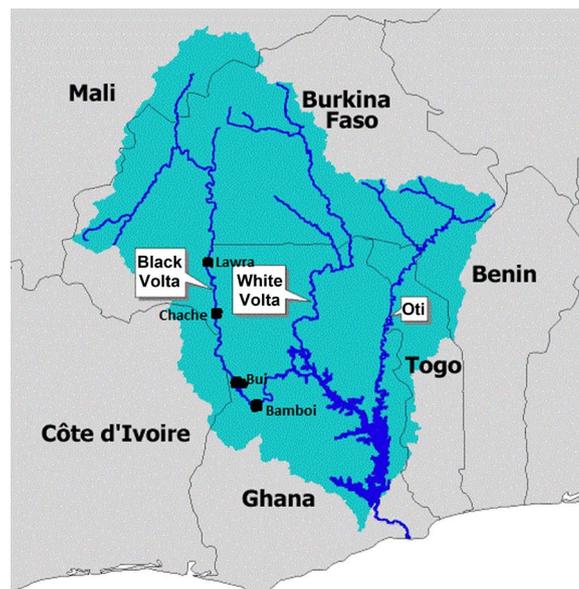
According to [24], the various physical mechanisms governing the river discharge dynamics act on a wide range of spatial and temporal scales. However, an important observation that can be made from the studies conducted thus far on the applications of both the physically-based models and the black-box models for river discharge forecasting is that none of these studies has looked at the influence of both spatial and temporal variability on river discharge forecasting simultaneously. This forms the basis of the present study. Giving the peculiar location of the Black Volta River, quantifying changes of river discharge both in space and time is fundamental in addressing issues of flooding, power generation and survival of ecosystem downstream [25].

In this study, we propose generalized additive mixed models (GAMMs) [26] [27] [28] incorporating a smooth interaction of space and time for modelling space-time variations in river discharge in the Black Volta River, to extract space-time signals for the entire study area. GAMMs are appealing for their flexibility and the straight forward way in which smooth effects of covariables can be incorporated along-side the smooth space time effect and random effects [29].

2. Materials and Methods

2.1. Study Area

The Black Volta river basin (**Figure 1**) stretches from North to South through Mali, Burkina Faso, Ghana and Cote d'Ivoire, and from West to East through
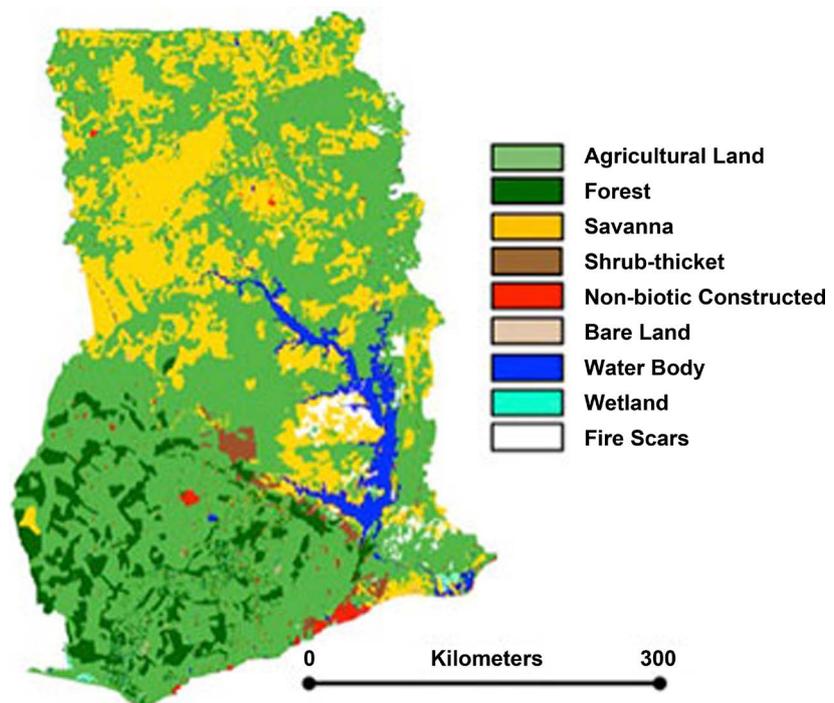

Source:
http://exploringafrica.matrix.msu.edu/curriculum/unit-five/module-twenty-four/module-twenty-four-activity-one/

Figure 1. Map of Study Area.

Burkina Faso, Cote d'Ivoire and Ghana. Geographically, it lies between 7˚00'00"N and 14˚30'00"N and Longitude 5˚30'00"W and 1˚30'00"W. The watershed has an area of about 130,400 km^2 constituting about 32.6% of the Volta basin which occurs when some portion of Bamboi which belongs to the Lower Volta is added to the basin. The portion of the watershed within Ghana has an area of about 18,384 km^2 which is about 14% of the basin. The annual rainfall varies from 1043 mm to 1270 mm. The Wettest month in the basin is September whiles the driest month is March. The estimated mean runoff in the basin is at 7 km^3 per annum. The mean monthly temperature in the basin is around 26˚C. The hottest month is March and the coolest is August [30]. Four gauge stations in the Black Volta basin namely Lawra, Chache, Bui and Bamboi, were all used for modelling and analysis.

2.2. Data and Variables

The data contains information on the four gauge stations along the Black Volta River namely, Lawra, Chache, Bui, and Bamboi. For each gauge station, latitude and longitude, year, month, elevation, land use, soil type, rainfall, humidity, and discharge are reported. Land use data is obtained from the land use map of Ghana in Figure 2. The lands in the Black Volta basin are mostly used for agriculture with bush fallow food crop cultivation. Except in the dry season, where livestock owners/herdsmen migrate with their animals in search of water and feed in nearby communities, animal grazing in the basin is mostly done on free range [31].

Source: https://geog.sdsu.edu/Research/Projects/IPC/research/ids.html.

Figure 2. Land Use Map of Ghana.
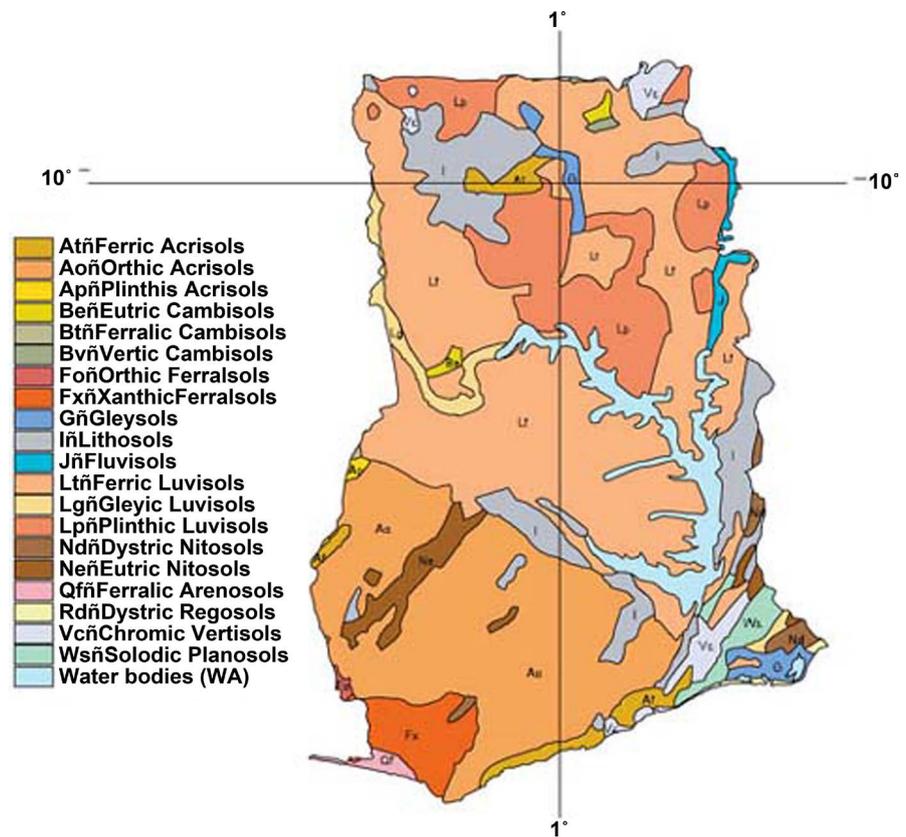
Figure 3. Soil Map of Ghana.

River discharge data from January 2000 to December 2009 for the four gauge stations was obtained from the hydrological services department of Ghana whiles rainfall and humidity data for the same period was obtained from the meteorological services department of Ghana. Data on Soil type was extracted from the soil map of Ghana in Figure 3. Ferric Luvisols is the most dominant soil in the Ghana portion of the basin. It is characterized mainly by Savannah Ochrosols and patches of Savannah Ochrosols-Lithosols [30].

The response variable for these analyses was river discharge (*disch*) measured in cubic metre per second whiles the independent variables were time (*month* & *year*) and space (*loc*) which are the various gauge stations along the Black Volta River considered in this study, namely Lawra, Chache, Bui and Bamboi. The covariates included rainfall (*rain*) measured in millimetres, relative humidity (*humid*), elevation (*elev*) measured in meters, soil type (*soil*) and land use (*luse*) which was considered as a random effect. Interactions between some of these variables were also considered especially the space-time interactions.

2.3. Models and Analyses

After checking the relationship between river discharge (*disch*) and all predictors, independent models were constructed for all covariates to determine their effect on *disch* and, if it resulted significant, its nature (linear or nonlinear) was

also determined. If a covariate was significant as a unique predictor, but not significant jointly with other predictors, it was removed. For instance, it was observed in the case of elevation (*elev*), which had a significant parameter as a unique predictor, but was eliminated in models which had several predictors because its effect resulted no more significant. This preliminary analysis is omitted from the results for the sake of brevity.

The response variable 'river discharge in gauge station *i*' ($disch_i$) is modelled using a generalized additive mixed model (GAMM) [26] [27] [28], as shown in **Table 1**.

2.4. Parameter Estimation

The GAMMs in **Table 1** can be expressed as generalized linear mixed models (GLMMs);

$$\log\left(\mu_i\right) = \boldsymbol{X}_i\boldsymbol{\theta} \tag{1}$$

where $\boldsymbol{X}_i$ is a row of the model matrix containing all components of the model. That is, all explanatory variables of fixed and random effects, and all the basic functions evaluated at observation *i*. The parameter θ contains the coefficients of fixed terms, the random land use effects and the bases. We estimate parameters with maximum likelihood (ML) estimation of the smoothness parameters, by integrating out the part of θ in the log likelihood function that is in the range space as described in [32].

2.5. Model Selection and Validation

Model selection was based on Akaike information criterion (AIC), Bayesian information criterion (BIC), adjusted R-squared, the root mean squared prediction error (RMSPE), and Nash-Sutcliffe efficiency (NSE). However, the key indicators of performance were the RMSPE which is independent of the likelihood and NSE. The RMSPE and NSE are calculated using Equations ((2) and (3)) respectively.

Table 1. Summary of models considered.

Models	Equation	Description
Model 1	$\log\left(\mu_i\right) = \beta_0 + \beta_1 soil_i + luse_{k(i)} + f_1\left(rain_i\right) + f_2\left(humid_i\right)$	GAMM without space and time effects
Model 2	$\log\left(\mu_i\right) = \beta_0 + \beta_1 soil_i + \beta_2 loc_i + luse_{k(i)} + f_1\left(rain_i\right)$ $+ f_2\left(humid_i\right) + f_3\left(month_i\right) + f_4\left(year_i\right)$	Model 1 including space and time main effects only
Model 3	$\log\left(\mu_i\right) = \beta_0 + \beta_1 soil_i + luse_{k(i)} + f_1\left(rain_i\right) + f_2\left(humid_i\right)$ $+ f_5\left(year_i, month_i, loc_i\right)$	Model 1 including space-time interaction effect only
Model 4	$\log\left(\mu_i\right) = \beta_0 + \beta_1 soil_i + \beta_2 loc_i + luse_{k(i)} + f_1\left(rain_i\right) + f_2\left(humid_i\right)$ $+ f_3\left(month_i\right) + f_4\left(year_i\right) + f_5\left(year_i, month_i, loc_i\right)$	Model 1 including both space and time main effects and interaction effect

where $\mu_i = E\left(disch_i\right)$ and $\log\left(disch_i\right)$ is assumed to follow the Gaussian distribution. β_0 is the intercept parameter, β_1 and β_2 are parameter estimates of fixed effects while f_{1-4} are smooth functions of the covariates which are represented using a cyclic cubic regression spline. $k(i)$ indexes the land use at the ith gauge station and $luse_{k(i)}$ is a random land use effect. f_5 is a tensor product of cyclic cubic regression splines.

$$\text{RMSPE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \qquad (2)$$

$$\text{NSE} = 1 - \left[\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - y^{mean})^2}\right] \qquad (3)$$

where y_i and $\hat{y}_i$ are the observed and predicted river discharges for n months, y^{mean} is the mean of the observed data.

For model checking and investigating whether the final selected model has disentangled spatial and temporal correlation in residuals, several diagnostics were used. The QQ-plot and histogram of residuals were used to check normality of residuals. Homoscedasticity of residuals was checked using scatter plot of residuals versus predictors. Also, the relationship between response and fitted values was checked as a visual goodness of fit verification using scatter plot of response versus fitted values.

2.6. Software

Data analysis was done in the R programming environment version 3.2.4 [33] and models were fit using the MGCV package [28].

3. Results and Discussion

3.1. Descriptive Statistics

In generalized regression models, it is important and necessary to study the distribution function of the response variable (*disch*) in order to select both response distribution and link fuction. Boxplots for *disch* and log (*disch*) are reported in the upper panels of Figure 4 while the normal QQ-plots for *disch* and log (*disch*) are reported in the lower panels. We observe from the figure that log (*disch*) gives a good approximation to the normal distribution. Hence the Gaussian distribution was considered as the underlying theoretical distribution of *disch* in the GAMMs with log link function.

The time series plots of river discharge at the various gauge stations are shown in Figure 5, which indicates an obvious seasonality in discharge at all four gauge stations. This suggests that smooth functions may be represented using cyclic cubic regression splines.

3.2. Model Selection

The first GAMM (Model 1) included fixed effect of soil type, random effect of land use, and smooth functions of rainfall and humidity but excluded space and time effects, and was able to explain only about 19.2% (R-sq. (adj) = 0.192) of the variability in river discharge as shown in Table 2. The second GAMM (Model 2) added the main effects of space and time to model 1 which resulted in explaining about 40.1% of the variability in river discharge. The third GAMM (Model 3) added only the space-time interaction effect to model 1 and resulted in explain-
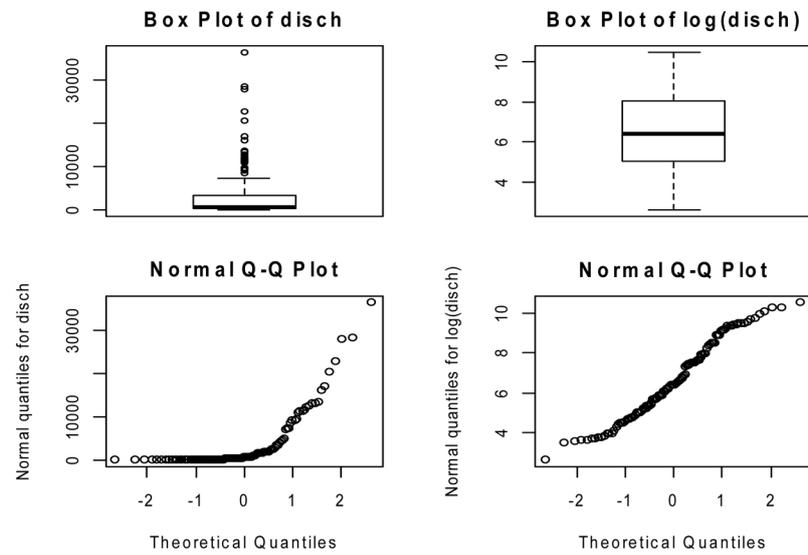
Figure 4. In the upper panels: Boxplots for river discharge (on the left) and the log transformation of river discharge (on the right); in the lower panels: Normal QQ-plots for river discharge (on the left) and the log transformation of river discharge (on the right).
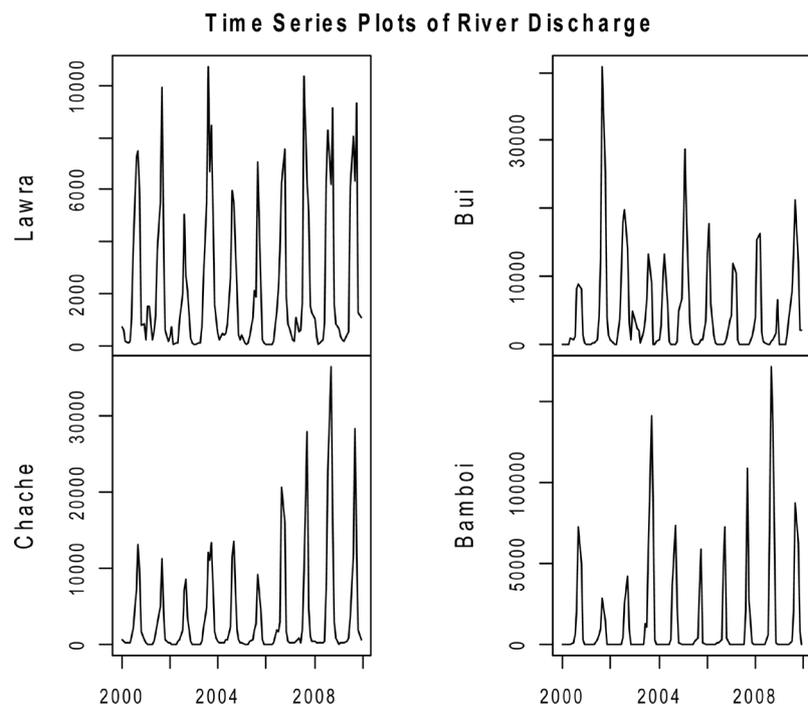

Figure 5. Time series plots for river discharge at the various gauge stations.

Table 2. Model selection criteria.

Model Number	AIC	BIC	R-sq. (adj)	RMSPE	NSE
Model 1	2526.891	2551.934	0.192	3.660	0.501
Model 2	2383.495	2437.755	0.401	2.689	0.527
Model 3	2089.51	2147.943	0.724	2.241	0.542
Model 4	1896.261	1983.911	0.821	2.043	0.601

ing about 72.4% of the variability in river discharge while the final GAMM (Model 4) added both the main and interaction effects of space and time to model 1 and resulted in explaining about 82.1% of the variability in river discharge. This provides an indication of the very significant role space-time effects play in modelling river discharge, but are usually ignored.

Furthermore, the RMSPEs and NSE values in Table 2 indicate satisfactory performance for all GAMMs considered. However, a comparison among them using AIC and BIC values clearly indicate a much better performance by the GAMM which included both the main and interaction effects of space and time (Model 4).

3.3. Parameter Estimates of the Selected Model

Parameter estimates of the selected GAMM are reported in Table 3 and Table 4. We observe from those Tables that, parameter coefficients (both smooth terms and non-smooth terms) were all significant at the 0.05 level.

3.4. Diagnostic Checks

Basic diagnostic plots of the selected GAMM are reported in Figure 6. The QQ-plot of residuals shows an evident arrival of residual quantiles at the theoretical normal quantiles and a near symmetry observed in the histogram of residuals as well. The scatter-plot of residuals versus the linear predictor indicates

Table 3. Parameter coefficients.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.48230	0.21581	16.136	<2e−16
soil	0.04488	0.01837	2.443	0.015
locBui	2.56965	0.21230	12.104	<2e−16
locChache	2.89817	0.21092	13.741	<2e−16
locLawra	3.13176	0.23467	13.345	<2e−16

Table 4. Approximate significance of smooth terms.

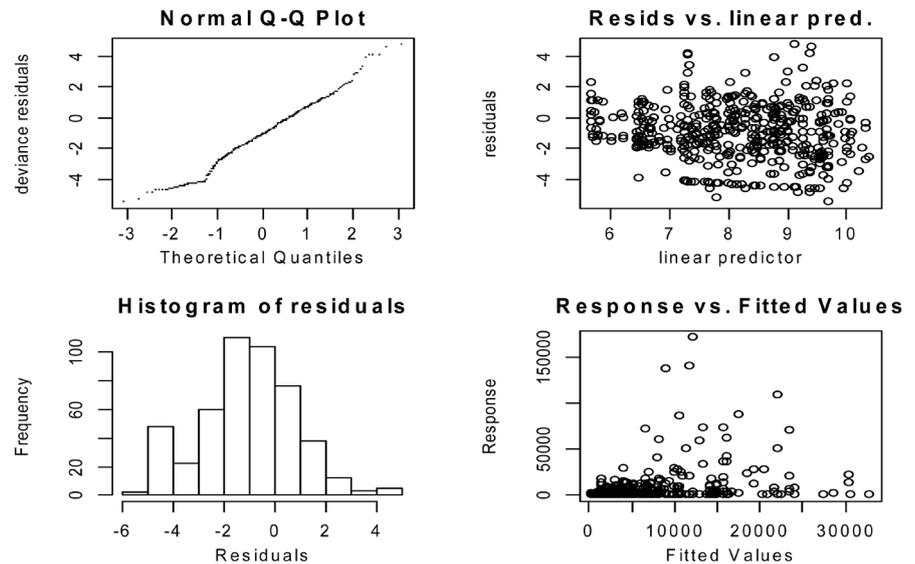|	edf	Ref.df	F	p-value
s (month)	1.000e+00	1	17.627	3.23e−05
s (year)	0.847	1	6.420	0.006
s (luse)	0.962	1	26.334	1.36e−07
s (rain)	9.786e−01	10	3.183	0.013
s (humid)	5.553e−06	10	2.001	0.036
te (year, month):locBamboi	1.155e+01	15	68.670	<2e−16
te (year, month):locBui	1.305e+01	15	22.433	<2e−16
te (year, month):locChache	2.609e+00	15	6.775	<2e−16
te (year, month):locLawra	2.515e+00	15	3.588	1.67e−12

Figure 6. Diagnostic for selected GAMM. In the upper panels: QQ-plot of residuals (on the left) and plot of residuals vs. Linear predictor (on the right); In the lower panels: Histogram of the residuals (on the left) and plot of the response vs. Fitted values (on the right).

an accentuated homoscedasticity of residual variance while that of the response versus fitted values shows independence of the residuals. All in all, diagnostics of the selected GAMM are quite good.

4. Conclusions

We have effectively used GAMMs for modelling space-time river discharge data in this paper. GAMMs provide a flexible framework which allows for smooth effects of covariates and smooth effects of space and time. In other applications such as repeated observations of weather station data, the use of spatio-temporal dynamic models or state-space models have been proposed. Four GAMMs were explored, two with space-time interactions and two without space-time interactions. The comparison of the performance of the models with space-time interactions and those without space-time interactions based on AIC and BIC suggests that in this application, the former is better overall and in particular for modelling variations in river discharge data. Further, a model with space and time main effects performed better compared with one without space and time main effects.

Acknowledgements

References

[1] UNESCO (1985) Teaching Aids in Hydrology. Universitaires de France, Vendome.

[2]   Ampadu, B., Chappell, N.A. and Kasei, R.A. (2013) Rainfall-Riverflow Modelling Approaches: Making a Choice of Data-based Mechanistic Modelling Approach for Data Limited Catchments: A Review. *Canadian Journal of Pure & Applied Sciences*, **7**, 2571-2580.

[3]   Xu, C. (2002) Textbook of Hydrologic Models. Uppsala University, Sweden.

[4]   Getachew, H.E. and Melesse, A.M. (2012) The Impact of Land Use Change on the Hydrology of the Angereb Watershed, Ethiopia. *International Journal of Cosmetic Science*, **1**, 1-7.

[5]   Chow, V.T., Maidment, D.R. and Mays, L.W. (1988) Applied Hydrology.

[6]   Shrestha, R.R. and Nestmann, F. (2009) Physically Based and Data-Driven Models and Propagation of Input Uncertainties in River Flood Prediction. *Journal of Hydrologic Engineering*, **14**, 1309-1319.
https://doi.org/10.1061/(ASCE)HE.1943-5584.0000123

[7]   Gosain, A.K., Mani, A. and Dwivedi, C. (2009) Hydrological Modelling-Literature Review. *Advances in Fluid Mechanics*, **339**, 63-70.

[8]   Nor, N.I., Harun, S. and Kassim, A.H. (2007) Radial Basis Function Modeling of Hourly Streamflow Hydrograph. *Journal of Hydrologic Engineering*, **12**, 113-123.
https://doi.org/10.1061/(ASCE)1084-0699(2007)12:1(113)

[9]   Jajarmizadeh, M., Harun, S.B. and Salarpour, M.M. (2011) A Concept of Classification for Hydrological Models. *Proceedings of 1st Iranian Studies Scientific Conference. Malaysia*, Universiti Putra Malaysia, Kuala Lumpur.

[10] Beven, K.J. and Kirkby, M.J. (1979) A Physically Based, Variable Contributing Area Model of Basin Hydrology. *Hydrological Sciences Journal*, **24**, 43-69.
https://doi.org/10.1080/02626667909491834

[11] Vieux, B.E., Cui, Z. and Gaur, A. (2004) Evaluation of a Physics-Based Distributed Hydrologic Model for Flood Forecasting. *Journal of Hydrology*, **298**, 155-177.

[12] Marsik, M. and Waylen, P. (2006) An Application of the Distributed Hydrologic Model CASC2D to a Tropical Montane Watershed. *Journal of Hydrology*, **330**, 481-495.

[13] Shamseldin, A.Y. (2010) Artificial Neural Network Model for River Flow Forecasting in a Developing Country. *Journal of Hydroinformatics*, **12**, 22-35.
https://doi.org/10.2166/hydro.2010.027

[14] Kisi, Ö. (2004) River Flow Modeling Using Artificial Neural Networks. *Journal of Hydrologic Engineering*, **9**, 60-63.
https://doi.org/10.1061/(ASCE)1084-0699(2004)9:1(60)

[15] Cigizoglu, H.K. (2003) Estimation, Forecasting and Extrapolation of River Flows by Artificial Neural Networks. *Hydrological Sciences Journal*, **48**, 349-361.
https://doi.org/10.1623/hysj.48.3.349.45288

[16] Taormina, R., Chau, K.-W. and Sethi, R. (2012) Artificial Neural Network Simulation of Hourly Groundwater Levels in a Coastal Aquifer System of the Venice Lagoon. *Engineering Applications of Artificial Intelligence*, **25**, 1670-1676.

[17] Nayak, P.C., Sudheer, K.P. and Ramasastri, K.S. (2005) Fuzzy Computing Based Rainfall-Runoff Model for Real Time Flood Forecasting. *Hydrological Processes*, **19**, 955-968. https://doi.org/10.1002/hyp.5553

[18] Liong, S.-Y., Lim, W.-H., Kojiri, T. and Hori, T. (2000) Advance Flood Forecasting for Flood Stricken Bangladesh with a Fuzzy Reasoning Method. *Hydrological Processes*, **14**, 431-448.

https://doi.org/10.1002/(SICI)1099-1085(20000228)14:3<431::AID-HYP947>3.0.CO;2-0

[19] McKerchar, A.I. and Delleur, J.W. (1974) Application of Seasonal Parametric Linear Stochastic Models to Monthly Flow Data. *Water Resources Research*, **10**, 246-255. https://doi.org/10.1029/WR010i002p00246

[20] Noakes, D.J., McLeod, A.I. and Hipel, K.W. (1985) Forecasting Monthly River Flow Time Series. *International Journal of Forecasting*, **1**, 179-190.

[21] Rosenberg, E.A., Wood, A.W. and Steinemann, A.C. (2011) Statistical Applications of Physically Based Hydrologic Models to Seasonal Streamflow Forecasts: Statistical Applications of Physically Based Models. *Water Resources Research*, **47**, n/a.

[22] Chau, K.W., Wu, C.L. and Li, Y.S. (2005) Comparison of Several Flood Forecasting Models in Yangtze River. *Journal of Hydrologic Engineering*, **10**, 485-491. https://doi.org/10.1061/(ASCE)1084-0699(2005)10:6(485)

[23] Veiga, V.B., Hassan, Q.K. and He, J. (2014) Development of Flow Forecasting Models in the Bow River at Calgary, Alberta, Canada. *Water*, **7**, 99-115. https://doi.org/10.3390/w7010099

[24] Sivakumar, B., Jayawardena, A.W. and Fernando, T. (2002) River Flow Forecasting: Use of Phase-Space Reconstruction and Artificial Neural Networks Approaches. *Journal of Hydrology*, **265**, 225-245.

[25] Iddrisu, W.A., Nokoe, K.S., Osei, F.B. and Antwi, E.O. (2016) Spatial Bayesian Methods of Flow Forecasting in the Black Volta River. *European Journal of Scientific Research*, **137**, 89-105.

[26] Lin, X. and Zhang, D. (1999) Inference in Generalized Additive Mixed Modelsby Using Smoothing Splines. *Journal of the Royal Statistical Society*, **61**, 381-400. https://doi.org/10.1111/1467-9868.00183

[27] Fahrmeir, L. and Lang, S. (2001) Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors. *Journal of the Royal Statistical Society. Series C, Applied Statistics*, **50**, 201-220. https://doi.org/10.1111/1467-9876.00229

[28] Wood, S. (2006) Generalized Additive Models: An Introduction with R. CRC Press, Boca Raton.

[29] Augustin, N.H., Trenkel, V.M., Wood, S.N. and Lorance, P. (2013) Space-Time Modelling of Blue Ling for Fisheries Stock Management. *Environmetrics*, **24**, 109-119. https://doi.org/10.1002/env.2196

[30] Allwaters Consult Limited. Diagnostic Study of the Black Volta Basin in Ghana. Global Water Initiative (GWI), CARE International, Catholic Relief Services (CRS), and the Regional Office for Central and West Africa of the International Union for Conservation of Nature (IUCN-PACO).

[31] Barry, B., Obuobie, E., Andreini, M., Andah, W. and Pluquet, M. (2005) Comprehensive Assessment of Water Management in Agriculture. Comparative Study of River Basin Development and Management. International Water Management Institute IWMI.

[32] Wood, S.N. (2011) Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models. *Journal of the Royal Statistical Society*, **73**, 3-36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

[33] R Core Team (2016) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.