Analysis of Sabine river flow data using semiparametric spline modeling

Soutir Bandyopadhyay [*], Arnab Maity [†]

November 4, 2010

Running Head: Analysis of Sabine river flow

ABSTRACT

In this article, a modeling approach for the mean annual flow in different segments of Sabine river, as released in the NHDPlus data in 2007, as a function of five predictor variables is described. Modeling flow is extremely complex and the deterministic flow models are widely used for that purpose. The justification for using these deterministic models comes from the fact that the flow is governed by some explicitly stated physical laws. In contrast, in this article, this complex issue is addressed from a completely statistical point of view. A semiparametric model is proposed to analyze the spatial distribution of the mean annual flow of Sabine river. Semiparametric additive models allow explicit consideration of the linear and nonlinear relations with relevant explanatory variables. We use a conditionally specified Gaussian model for the estimation of the univariate conditional distributions of flow to incorporate auxiliary information and this formulation does not require the target variable to be independent.

Keywords and phrases: Sabine River, semiparametric model, spline.

[*]Department of Mathematics, Lehigh University, Bethlehem, PA 18015

[†]Department of Statistics, North Carolina State University, Raleigh, NC 27695

1 Introduction

One of the primary challenges for the professionals in water sectors is to meet multiple water demands within the constraint of limited freshwater supply. The necessity to integrate the ecosystem needs is also pronounced in water management. Proper ecosystem management is paramount to protect the ecological processes and biodiversity. It has been noted in some literature that demands for surface water are not expressed freely but rather controlled by water rights specifying the location and type of each allowed usage, the amount to be used and the priority date when the right is established (see for example, `http://www.oregonexplorer.info/willamette/`). Therefore, a good understanding of available water resources is needed for water professionals to achieve a sustainable water system that enriches both this generation and future, while considering the expected future climate and other relevant geographical and hydrological parameters.

As Mylevaganam and Srinivasan (2008) note, contemporary efforts in planning, designing and implementing resource management efforts are now at the catchment scale. The reason to exploit at the catchment scale is to allow management actions to be carried out unhindered until the magnitude of effect reaches to a point where regulation becomes necessary. It has also been mentioned in Ziemer (1994) that generalized regulations are usually not efficient and usually a higher level of regulation results in more streams being overprotected. The closer that the regulations can be tailored to the variables associated with the risk, the less likely that proposed management actions are curtailed needlessly, or, conversely, the less likely that the regulations are inadequate to protect a desired resource. Added to this, the effect of water resources allocation in the upstream of a river basin plays a crucial role in determining the state of the downstream water availability. The spatial connectivity of stream networks often plays a big role to avoid upstream-downstream conflict. Reliability of a catchment is also indirectly linked to the mean annual flow it conveys.

Further, the availability of hydrological data is also critical for water resources planning. Most drainage basins in this world do not have these data because of poorly developed hydrological networks (Oyebande (2001) and Rodda (2001)). It is also not feasible to establish a flow measuring station on every drainage basin (Chiang et al. (2002)) and in addition the sheer sizes of some countries make it impossible to develop adequate hydrological networks and therefore most drainage basins are ungauged (Tucci et al. (1995)). Therefore, the need for hydrological data has greatly increased as water resources which are in some cases scarce have to be shared among competing uses.

Therefore, potential to predict water availability, in other words, mean annual flow at a catchment scale considering all the influencing hydrological and geographical parameters is paramount. This also greatly enhances the knowledge on hydrological characteristics of ungauged basins for water resources planning purposes given the prevailing climate and other conditions are of similar nature.

1.1 Dataset

In this section we give a brief description of the NHDPlus data. A more detailed description of the NHDPlus can be found in the website of Center for Research in Water Resources (`http://www.crwr.utexas.edu/gis/gishydro08/ArcHydro/NHDPlus.htm`).

According to NHDPlus Users Guide, NHDPlus (Horizon Systems, 2007) is an integrated suite of application-ready geospatial data products, incorporating many of the best features of the National Hydrography Dataset (NHD), the National Elevation Dataset (NED) and the National Watershed Boundary Dataset (WBD) (Holtschlag, 2009). NHDPlus dataset is distributed for each region as shown in Figure 1. NHDPlus includes a stream network based on the medium resolution NHD (1:100,000 scale), improved networking, feature naming and "value-added attributes" (VAA). NHDPlus also includes elevation-derived catchments which are produced using a drainage enforcement technique. The VAAs include greatly enhanced capabilities for upstream and downstream analysis and modeling. VAA-based routing techniques are used to produce the NHDPlus cumulative drainage areas and land cover, temperature and precipitation distributions. These cumulative attributes are used to estimate mean annual flow and velocity. The objective of the study is to investigate and propose a lattice based mean annual flow predictor for the NHDPlus dataset as released in 2007.

1.2 Study Area and Sabine Basin Hydrology

Detailed description of origin and flow of Sabine river and its hydrology is provided in Comprehensive Sabine Watershed Management Plan Report (1999), available at the official website of Sabine River Authority of Texas. Below we briefly summarize some of the key points. We refer the interested readers to the original report (located at `http://www.sra.dst.tx.us/srwmp/comprehensive_plan/default.asp`) for more detailed description of origin, background and hydrology of Sabine river.

Sabine River, a river in the southwestern United States, rises in northeastern Texas, flows southeast and south, broadening near its mouth to form Sabine Lake and continues from Port Arthur through Sabine Pass, a dredged navigable channel, to the Gulf of Mexico after a course of 578 mi (930 km). It drains 10,400 sq mi (26,950 sq km) entirely in Texas and the Louisiana Coastal Plain. The Sabine is a flat-water river that pumps about 6.8 million acre-feet into the Gulf and is the single largest volume river in Texas in terms of its discharge. The water has the tannin acid brown color that is common in East Texas rivers and streams.

The Sabine River Authority of Texas was created by the Legislature in 1949 as an official agency of the State of Texas. The main purpose of this agency was to act as conservation and reclamation district with responsibilities to control, store, preserve and distribute the waters of the Sabine River and its tributary streams for useful purposes. The boundaries were established by the Act of the Legislature and it comprise all of the area lying within the watershed of the Sabine River and its tributary streams within the State of Texas. The watershed area includes all parts of twenty-one counties. Figure 2(a) shows the total number of catchments available in Texas. We consider the data set of catchments only in the Sabine river basin (Figure 2(b)) containing 5,654 catchments.

The hydrology of Sabine river basin is characterized by diverse climatological, topographical and geological features as well as several climatological factors such as temperature, rainfall and humidity. It is known that topography and geologic factors can affect runoff, evaporation, sedimentation rates, reservoir storage capacity and water quality and define the river system within the basin. As mentioned in Comprehensive Sabine Watershed Management Plan (1999), the hydrology of the northern region of the basin is significantly different from the southern region. These distinct regions are commonly referred to as the "Upper basin" in the north and the "Lower basin" in the south, the division between the two areas being the headwaters of Toledo Bend Reservoir. The Upper basin is characterized by cool winters, hot summers and seasonal rainfall patterns. The Lower basin has a coastal climate with mild winters, high annual rainfall and moderate to high humidity. However, in this paper, for the modeling purpose we have not considered these two regions separately. We assume that even if the two regions are distinct from the hydrological point of view but the flow at any catchment can be modeled in the same way for both the regions. The flow at any catchment only depends on a small number of the neighboring catchments and we assume that the hydrological properties of a particular catchment is not significantly different from the properties of its neighboring catchments. The geological factors affect the neighboring catchments similarly in each region, and hence affecting the covariates (precipitation, temperature etc) similarly, but not necessarily changing the dependence structure among the neighboring catchments.

In this paper we are going to model the mean annual flow of Sabine river based on the NHDPlus data set released on 2007 in its different catchments based on several relevant variables such as length, stream order, temperature, precipitation and slope. Detailed distributions of these variables based on our data set are shown in tables 1 and 2.

The article is organized as follows. In Section 2, we discuss our methodology and implement it to model the data. In Section 3 we discuss the implications of the fitted model.

2 Data analysis

The goal of the analysis performed here and the features of the data at hand give precise indications about the model to be used. First, it is clear that if an event occurs in a region, it is likely to affect the neighboring regions as well, *i.e.*, the events are spatially dependent. The second aim consists in estimating the flow distribution as a function of explanatory variables because flow can be related to a number of factors, for example, precipitation at the specified catchment, temperature, slope of the region etc. For modeling purpose, the logarithmic transformation of the flow values are considered as a function of relevant explanatory variables. By doing this we implement the constraint that the response variable, flow of the river in a catchment, is always a non-negative quantity.

Complex functional relations characterizing the flow of the river and their spatially dependent structure lead to the adoption of a semiparametric lattice model. In this data we have five covariates, namely, precipitation, temperature, slope, length and stream order of the catchment. For this study, instead of taking the original values of the first four covariates, we take their logarithmic values. The stream order is a variable only taking the values 1 - 11.

Considering that the logarithm of river flow is a continuous variable and the model should include the auxiliary variables, it seems natural to resort to Gaussian models. However, the classical Gaussian regression may not be completely adequate since the classical models require the target variables to be independent. Thus some modifications are required to incorporate the spatial dependence of the flow data. More specifically, we can use the well-known *conditionally specified Gaussian models* (see e.g., Cressie (1993)) so that the spatial dependence of the response variable can be taken into account by means of a "conditional specification" model of spatial correlation. In such models one incorporate the fact that an event observed in a certain geographic region depends on what happens in the neighboring regions.

A model is said to be of a conditional specification type if the joint distribution function of the units is built on the basis of the univariate conditional distributions. The conditionally specified Gaussian model approach was first proposed by Besag (1974, 1977). However, the conditionally specified Gaussian model defines a structure of spatial dependence but does not allow to incorporate auxiliary variables. We have used a semiparametric additive model for the mean part in the conditionally specified Gaussian model. Linear effects of the length of the river in the catchment and the stream order of the catchment as well as the nonlinear effects of precipitation, temperature and slope at a given catchment are included in the model.

2.1 The semiparametric lattice model

The spatial models on lattices are analogues of time-series autoregressive models. In time domain the dependence relies upon the unidirectional flow of time where the spatial conditional approach expresses the dependence of a variable on its nearest neighbor regions. Let $\mathbf{Y}$ be the $n \times 1$ vector of the dependent variable. The model can now be formalized by explicitly writing down the conditional distribution of the dependent variable at i-th catchment:

$$f(Y_i|\{Y_j : j \neq i\}) = (\sqrt{2\pi}\sigma)^{-1} \exp\left[-\{Y_i - \mu_i - \gamma \sum_{j \in N_i}(Y_j - \mu_j)\}^2/2\sigma^2\right],$$

where, $E(Y_i) = \mu_i$ for all $i = 1, \cdots, n$, γ is the spatial dependence parameter, σ^2 denotes the conditional variance of Y_i given $\{Y_j : j \neq i\}$. In the above equation, N_i is the set of neighbors for the i-th catchment. For more detailed discussion about the conditionally specified Gaussian models, we refer the readers to Cressie (1993). To find the neighborhood structures we look at the "ToNode" and "FromNode" of each catchment. "ToNode" is a nationally unique ID for the to node (with correct coordinate direction, this is the downstream node) endpoint of the flow line. "FromNode" is the same with the upstream node. A number of catchments are said to be neighbors if the "ToNode" of the catchments are same as the "FromNode" of a particular catchment. The dependence parameter γ is estimated from the neighborhood structure and the mean effect is modeled as

$$\begin{aligned}\mu_i &= \beta_0 + \beta_1 \log(\text{length})_i + \sum_{k=2}^{11} \beta_{2k}\, I(\text{stream}_i = k) \\ &\quad + f_1(\log(\text{precip})_i) + f_2(\log(\text{temp})_i) + f_3(\log(\text{slope})_i), \\ &= \beta_0 + \beta_1 X_{1i} + \sum_{k=2}^{11} \beta_{2k}\, I(X_{2i} = k) + f_1(X_{3i}) + f_2(X_{4i}) + f_3(X_{5i}),\end{aligned}$$

where $f_1(\cdot)$, $f_2(\cdot)$ and $f_3(\cdot)$ are unknown functions describing the effects of precipitation, temperature and slope, respectively. We will use penalized splines (see Wand, 2003) to model these functions. The penalized regression splines representation of the smooth functions is given by:

$$f_\ell(t_i) = \alpha_{1\ell} t_i + \alpha_{2\ell} t_i^2 + \cdots + \alpha_{p\ell} t_i^p + \sum_{j=1}^{K} \delta_{j+p,\ell} |t_i - \kappa_{j,\ell}|_+^p$$

for $\ell = 1, 2$ and 3, where each knot $\kappa_{j,\ell}$ is associated to a coefficient $\delta_{j+p,\ell}$ and $x_+ = \max(0, x), x \in I\!R$, where the coefficients $\delta_{j+p,\ell}$, $j = 1, \ldots, K$ are to be penalized (Wand, 2003). The number of knots and their positions can be obtained in an adaptive way although the sensitivity to this choice is quite low (Ruppert, 2001).

Remark 1. It is worth mentioning that there may be a situation where one encounters a dry season with significant occurrences of no precipitation leading to a number of zero values for the flow. In such a situation, one can use the two-stage model for non-negative variables with a mass point at zero, as described in Velarde et al.(2004). In this approach, a binary model is introduced to describe the presence or not of a zero level and then, conditional on observing a level different of zero, the quantity of the variable will be modeled. The probabilistic description will be a mixture of a discrete and a continuous distribution, generically represented as $(1 - p) + pf(y|y \neq 0)$, where $p = Pr(Y > 0)$ denotes the probability of Y being greater than zero. For more detailed description of the zero-inflated model, see Lambert (1992), Ainsworth (2007).

2.2 Fitting the model

The penalized pseudo log likelihood is given by

$$\mathcal{L} = -\sum_{i=1}^{n} \log(\sigma^2)/2 - \sum_{i=1}^{n} \Big[-\{Y_i - (X_i\beta + Z_i\delta) - \gamma \sum_{j\in N_i}(Y_j - (X_j\beta + Z_j\delta))\}^2/(2\sigma^2)\Big]$$
$$-\delta^T D\delta/2,$$

where X and β denote the unpenalized part of the covariates and corresponding parameters in the model; Z and δ denote the penalized part of the covariates associated with the penalized spline model (Wand, 2003) and corresponding parameters; the matrix $D = diag(\lambda_1 D_1, \lambda_2 D_2, \lambda_3 D_3)$ denotes the penalty matrix associated with δ with D_1, D_2 and D_3 being the penalty matrices corresponding to individual functions $f_1(\cdot), f_2(\cdot)$ and $f_3(\cdot)$, and λ_1, λ_2 and λ_3 are respective smoothing parameters.

We first discuss the model fitting when σ^2 is known. To estimate the parameters, we will adopt a profiling approach as described in Cressie (1993). For a fixed value of γ, the score equations for β and δ are

$$0 = \sum_{i=1}^{n} X_i^{\#T}(\gamma)\{Y_i^{\#}(\gamma) - X_i^{\#}(\gamma)\beta - Z_i^{\#}(\gamma)\delta\},$$
$$0 = \sum_{i=1}^{n} Z_i^{\#T}(\gamma)\{Y_i^{\#}(\gamma) - X_i^{\#}(\gamma)\beta - Z_i^{\#}(\gamma)\delta\}/\sigma^2 - \lambda D\delta,$$

where we define $X_i^{\#}(\gamma) = X_i - \gamma\sum_{j\in N_i} X_j$ and similarly $Y_i^{\#}$ and $Z_i^{\#}$. We can rewrite the score equation in a matrix form

$$0 = \begin{bmatrix} X^{\#}(\gamma) \\ Z^{\#}(\gamma) \end{bmatrix} V^{-1}\{Y^{\#}(\gamma) - X^{\#}(\gamma)\beta - Z^{\#}(\gamma)\delta\} - D^{\#}(\beta^T, \delta^T)^T,$$

where $Y^{\#} = [Y_1^{\#}, \ldots, Y_n^{\#}]^T$, $V = diag(\sigma^2, \ldots, \sigma^2)$, $X^{\#} = [X_1^{\#T}, \ldots, X_n^{\#T}]$ and similarly for $Z^{\#}$ and $D^{\#} = diag(0, D)$. Defining $W^{\#}(\gamma) = \begin{bmatrix} X^{\#}(\gamma) \\ Z^{\#}(\gamma) \end{bmatrix}$, we have

$$\begin{bmatrix} \widehat{\beta}(\gamma) \\ \widehat{\delta}(\gamma) \end{bmatrix} = [W^{\#}(\gamma)V^{-1}W^{\#}(\gamma)^T + D^{\#}]^{-1} W^{\#}(\gamma)V^{-1}Y^{\#}(\gamma).$$

To estimate γ, we first construct the profile likelihood of γ:

$$\mathcal{L}_{\text{prof}}(\gamma) = -\{Y^{\#}(\gamma) - X^{\#}(\gamma)\widehat{\beta}(\gamma) - Z^{\#}(\gamma)\widehat{\delta}(\gamma)\}^T V^{-1}\{Y^{\#}(\gamma) - X^{\#}(\gamma)\widehat{\beta}(\gamma) - Z^{\#}(\gamma)\widehat{\delta}(\gamma)\}.$$

The estimate of γ is then constructed as

$$\widehat{\gamma}_{\text{prof}} = \text{argmax}_{\gamma}\mathcal{L}_{\text{prof}}(\gamma). \tag{2.1}$$

Since γ is an scalar parameter, this maximization problem is easy to solve in any standard software. The final estimates are given by $\widehat{\beta}_{\text{prof}} = \widehat{\beta}(\widehat{\gamma}_{\text{prof}})$ and $\widehat{\delta}_{\text{prof}} = \widehat{\delta}(\widehat{\gamma}_{\text{prof}})$.

To estimate the variance, we first fit the model with $V = I$, that is, using working independence assumption. Let the resulting centered residuals be $\widehat{\epsilon}_i$, $i = 1, \ldots, n$. Then the estimate $\widehat{\sigma}^2$ can be obtained by taking the mean of the squares of the centered residuals, *i.e.*, $\widehat{\sigma}^2 = n^{-1}\sum_{j=1}^{n}\widehat{\epsilon}_j^2$.

Remark 2. Instead of using the three-step approach above to estimate the model components, one can use the maximum likelihood estimators where one maximizes the full likelihood with respect to all parameters. This is reasonable from the theoretical point of view. However, we encounter some computational problems and numerical instability issues while maximizing the full likelihood with respect to all parameters possibly due to the fact that the maximization needs to be jointly done on a large number of parameters. In contrast, for the profiling method described above, estimation of β and δ is only a one-step procedure (with closed forms) for each value of γ and is computationally much more efficient and fast. Thereafter, estimation of γ is only a one-dimensional estimation problem. For further detail on this approach, see Cressie (1993).

2.2.1 Smoothing parameter selection

Most smoothing parameter selection methods do not perform well in the presence of correlated errors, as extensive research in the one dimensional case has shown; see Hart (1996) and Opsomer, Wang and Yang (2001) for overviews. We adopt the approach as in Francisco-Fernandez and Opsomer (2005). In that article the authors propose a smoothing parameter selection method based on the generalized cross-validation (GCV) criterion (Craven and Wahba (1978)), suitably adjusted for the presence of spatial correlation. They consider selecting the smoothing parameter $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ that minimizes the following "bias-corrected" GCV criterion

$$GCV(\lambda) = n^{-1} \sum_{i=1}^{n} \left(\frac{Y_i - \widehat{Y}_i}{1 - n^{-1}\text{tr}(\mathbf{S}\Sigma)} \right)^2, \tag{2.2}$$

where $\mathbf{S}$ is the $n \times n$ smoother matrix such that, $\widehat{\boldsymbol{Y}} = \mathbf{S}\boldsymbol{Y}$, where $\widehat{\boldsymbol{Y}} = [\widehat{Y}_1, \ldots, \widehat{Y}_n]^T$ and $\boldsymbol{Y}$ is defined similarly, and Σ the correlation matrix of the observations which is given by, $\Sigma = (I - C)^{-1}$, where C is a $n \times n$ matrix with (i, j)-th entry $C_{ij} = \gamma$, if the i-th and the j-th catchments are neighbors and $C_{ij} = 0$ otherwise. In general, Σ is unknown and we replace it with its estimate $\widehat{\Sigma}$ in (2.2). To find $\widehat{\Sigma}$, first note that the correlation matrix depends on the unknown parameter γ. We can first estimate γ using the profiling approach described earlier and then we can simply plug in the estimate in the expression for Σ to finally get an estimate of the correlation matrix. Thereafter, finding the minimizer of this function can be performed using numerical algorithms and can be easily implemented in standard statistical softwares.

2.2.2 Knot selection

The number of knots suitable to represent the nonlinear effect can be obtained as (Ngo and Wand (2004)). A reasonable default rule for the knot locations is: $\kappa_j = \{(j+1)/(j+2)\}$th sample quantile of the unique $\mathbf{x}_i$'s, for $j = 1, \cdots, K$. A simple default choice of K that usually works well is:

$$K = \max \left\{ 5, \min \left(\frac{1}{4} \times \text{ number of unique } \mathbf{x}_i' s, 35 \right) \right\}.$$

See Ruppert (2002) for further discussion on default knot specification. The number of knots and their positions can also be obtained in an adaptive way although the sensitivity to this choice is quite low (Ruppert (1997)).

2.3 Results

In order to analyze the spatial distribution of log(flow) of Sabine river, we model the mean part with a quadratic spline ($p = 2$). The entire analysis is based on the standardized variables. The number of knots suitable to represent the nonlinear effects of log(precip), log(temp) and log(slope) are fixed to 35 and is obtained as Ngo and Wand (2004). The knots are considered to be equidistant. The summary statistics for different variables used in this study are presented in Table 1. Table 2 describes the mean and standard deviations of different variables for each stream order. The penalty parameters for functions of precipitation, temperature and slope are calculated as 0.53, 1.02, 0.06, respectively using the GCV criterion as described earlier. For the fitting purposes, we select a grid of 51 equidistant points in the interval $[q_2(x), q_{98}(x)]$ for each $x =$ precipitation, temperature and slope, where $q_2(x)$ and $q_{98}(x)$ denotes the 2nd and 98th quantiles of x. We estimate the functions on this grid and center the estimates so that $\sum_{k=1}^{51} \widehat{f}_\ell(g_{k,\ell}) = 0$, $\ell = 1, 2, 3$, where $g_{k,\ell}$, $k = 1, \ldots, 51$ denotes the grid points corresponding to that function. We plot each of the covariate's effect over the grid along with a 95% point-wise confidence band. The estimated effects are presented in Figures 3 - 6. The slope parameter associated with log(length) is estimated to be 0.89 with a standard error is 0.01.

From the results, it is evident that mean annual flow increases as log length increases. This justifies the spatial pattern of precipitation, draining capacity and their influence on stream flow. From the effect of slope, we see that the significance of steepness of catchment slope on river flow is also pronounced. From Figures 3 and 4, the logarithm values of precipitation and temperature illuminate that these may alone not play a role in determining the stream flow. It is evident that the estimated effects are showing nonlinear patterns within small limits in the vertical axes. From Figure 3, we see a slight upward trend of effect of precipitation on flow. However, the effect becomes flat at the right tail. The connotation is that the type of land use pattern plays a remarkable role in abstracting the precipitation before it eventually contributes to the stream flow. The stream order effects portray that the basin is of mixed nature when it comes to its primary source. There is a possibility of streams being dried in some sections of the basin (higher stream orders).

Regarding model diagnostics, Figures 7(a) and (b) represent the location-spread plot, that is, $|\widehat{\Sigma}^{-1/2}(Y_i - \widehat{Y}_i)/\widehat{\sigma}|^{1/2}$ against $\widehat{Y}_i$, and residual versus predicted values plot for our fitted model. It seems there might be heteroscedasticity present in the model. In view of this, we refit our model using σ_i^2 in place of σ^2, where the model is,

$$log(\sigma_i^2) = h_1(log(precipitation_i)) + h_2(log(temperature_i)) + h_3(log(slope_i)).$$

We fit this additive model using squared centered residuals from an working independence fit of data as response variable and specifying Gaussian likelihood and log-link function. The estimates of σ_i^2 for each individual catchments $i = 1, \dots, n$ are presented in Figure 8 with the horizontal dashed line denoting the estimated variance in the homoscedastic case. We refit our model using this updated variance estimates. The results are very similar to those in Figures 3 - 6 and hence we do not present them. It is interesting to mention that the spread-location and residual-predicted values plots of the updated model (not shown here) still show some signs of heteroscedasticity. We believe this is due to various other physical factors and variables unaccounted in the data. For instance, there are various deterministic relationships/physical models describing the relationship between precipitation, temperature and slope to river flow. We only look at their relationship from a purely statistical point of view and thus do not account for any such physical relationships. This is certainly an area of interest and we hope to pursue this as a future direction of our research.

We also investigate several models apart from the above model. Table 3 describes these models and their corresponding AIC values in the homoscedastic case. First column of Table 3 describes the model, for example, the first entry of first column 'stream.order + length + f(precip)' corresponds to the model where we include stream order and standardized log(length) as linear covariates and standardized log(precipitation) as nonparametrically modeled covariate. The second column of the table provides corresponding AIC values. It is evident that among the models investigated, the model we fit above with all the variables produces least AIC.

3 Discussion

This study defines a lattice based additive model relating catchment properties such as channel slope, precipitation, temperature, length of the stream and stream order to mean annual flow. Though the model is applied to analyze flow of Sabine rive, this type of model have general applicability to other types of such flow network data. Other covariates can also be included in the model if available.

There are several important implication of this model. As noted in Arnold et al. (2000), base flow characteristics are essential for efficient development of groundwater resources, and for minimizing pollution risks to connected surface water. Therefore the integrated approach is necessary to enhance the sustainability of both surface water and ground water. It has been also noted in Adane and Foerch (2006) that river systems are often augmented by their base flows during lean seasons. The fitted values of stream order intercepts could be used to form Base flow Index (BFI) providing a systematic way of assessing the proportion of base flow in the total runoff of a catchment. It indicates the influence of soil and geology on river flows and is important for low flow studies. In addition, extreme low flow events are gradually earning more importance in the emerging field of ecohydrology and are more diligently analyzed nowadays (Adane and Foerch, 2006). However, it is often difficult to get recorded data on base flows of rivers because many of the catchments in developing countries remain ungauged. Our work may provide an indication of the underlying baseflow given the climatic and geographical conditions are similar.

In addition, our model can be used to estimate the rainfall elasticity. Typically, the rainfall elasticity of stream flow is defined as the proportional change in mean annual stream flow divided by the proportional change in mean annual rainfall (Chiew, 2006). However, this definition assumes that the rate of change in flow relative to change in precipitation is the same for any level of precipitation, that is, the relationship between flow and precipitation is linear. One can use our model to estimate the relationship between flow and precipitation and estimate the rainfall elasticity without being constrained by the linearity assumption and also taking into account the change in flow due to other geographical and climatological factors.

The main limitations of this computation are that it does not consider changes in the rainfall frequency and distribution, changes in vegetation characteristics under different climatic conditions and potential feedbacks between the atmosphere and the land surface. We also look at the problem from a purely statistical standpoint and do not take into account the different deterministic models relating flow to other variables. One may take into account these deterministic models into the statistical formulation to borrow strength and information from them. This is one of the future directions of our research.

Acknowledgments. The authors thank S. Mylevaganam of the Spatial Sciences Laboratory in Texas A&M University for providing with the dataset and a number of constructive suggestions regarding the hydrology of Sabine river.

References

Adane, A. and Foerch, G. (2006), "Catchment characteristics as predictors of base flow index (BFI) in Wabi-shebele river basin, East Africa", Proceedings of TROPENTAG 2006, October 11-13, 2006, Bonn, Germany.

Ainsworth, L. (2007), "Models and methods for spatial data: detecting outliers and handling zero-inflated counts", PhD Thesis, Simon Fraser University.

Arnold, J. G., Muttiah, R. S., Srinivasan, R., Allen, P. M. (2000), "Regional estimation of base flow and groundwater recharge in the Upper Mississippi river basin", Journal of Hydrology, 227, 21 - 44.

Besag, J. (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems", Journal of the Royal Statistical Society, Series B, 36(2), 192-236.

Besag, J. (1977), "Efficiency of pseudo likelihood estimation for simple Gaussian fields", Biometrika, 64, 616-618.

Comprehensive Sabine Watershed Management Plan Report (1999), available at http://www.sra.dst.tx.us/srwmp/comprehensive_plan/default.asp.

Chiang, S. M., Tsay, T. K. and Nix, S. J. (2002), "Hydrologic regionalization of watersheds. I: Methodology", Journal of Water Resources Planning and Management 128(1), 3-11.

Chiew, H. S. F. (2006), "Estimation of rainfall elasticity of streamflow in Australia", Hydrological Sciences Journal, 51(4), 613-625

Craven, P. and Wahba, G. (1978) , " Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation ", Numerische Mathematik 31(4), 377-403.

Cressie, N. (2003), "Statistics for spatial data", New York: Wiley.

Francisco-Fernandez, M. and Opsomer, J. (2005), "Smoothing Parameter Selection Methods for Nonparametric Regression with Spatially Correlated Errors", The Canadian Journal of Statistics, 33(2), 279-295.

Hart, J. D. (1996), "Some automated methods of smoothing time-dependent data", Journal of Nonparametric Statistics, 6, 115-142.

Hastie, T. and Tibshirani, R. (1990), "Generalized Additive Models", Chapman and Hall.

Holtschlag, D.J. (2009), "Application guide for AFINCH (analysis of flows in networks of channels) described by NHDPlus," U.S. Geological Survey Scientific Investigations Report 2009-5188, 106 p.

Horizon Systems. (2007). "National Hydrography Dataset Plus." http://www.horizon-systems.com/nhdplus/

Lambert, D. (1992), "Zero-inflated Poisson regression, with an application to defects in manufacturing", Technometrics, 34(1), 1-14.

Mylevaganam, S. and Srinivasan, R. (2008), "Effect of grid sizes as subbasins on SWAT model hydrologic and water quality predictions", Research project (project ID: 2008TX306B). Available at http://water.usgs.gov/wrri/08grants/progress/2008TX306B.pdf

Ngo, L. and Wand, M. P., (2004), "Smoothing with Mixed model software", Journal of Statistical software, 9:1.

Opsomer, J., Wang, Y. and Yang, Y. (2001), "Nonparametric Regression with Correlated Errors", Statistical Science, 16(2), 134-153.

Oyebande, L. (2001), " Water problems in Africa-how can sciences help?", Hydrological Sciences Journal 46(6), 947-961.

Rodda, J. C. (2001), "Water under pressure", Hydrological Sciences Journal 46(6), 841-853.

Ruppert, D. (1997), "Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation", Journal of American Statistical Association 92, 1049-1062.

Ruppert, D. (2002), "Selecting the Number of Knots for Penalized Splines", Journal of Computational and Graphical Statistics, 11(4), 735-757.

Tucci, C., Silveira A. and Sanchez, J. (1995), "Flow regionalization in the upper Paraguay basin, Brazil", Hydrological Sciences Journal, 40(4), 485-497.

Velarde, L. G. C., Migon, H. S., Pareira, B. D. B. (2004), " Space-time modeling of rainfall data", Environmetrics, 15, 561–576.

Wand, M. P. (2003), "Smoothing and mixed models", Computational Statistics 18, 223-249.

Wood, S. N. (2006), "Generalized Additive Models: An Introduction with R", Chapman and Hall/CRC Press.

Ziemer, R. (1994), "Cumulative effects assessment impact thresholds: myths and realities", Kennedy, Alan J., ed. Cumulative Effects Assessments in Canada: From Concept to Practice. Alberta Association of Professional Biologists. Edmonton, Alberta, Canada.

variables	min.	Q_1	median	mean	Q_3	max.	st. dev.
log(flow)	-8.294	-1.791	-0.399	-0.820	0.560	3.304	2.009
log(precipitation)	6.912	6.985	7.111	7.111	7.230	7.307	0.119
log(temperature)	5.136	5.168	5.195	5.197	5.217	5.293	0.033
log(slope)	0	0.0002	0.002	0.004	0.005	0.211	0.007
log(length)	-4.510	-0.488	0.561	0.296	1.240	3.582	1.288

Table 1: *The summary statistics (minimum, first quartile (Q_1), median, mean, third quartile (Q_3) and maximum) for different variables.*

stream order	log(flow)	log(precip.)	log(temp.)	log(slope)	log(length)
1	-0.58(1.89)	7.10(0.11)	5.19(0.03)	0.010(0.006)	0.41(1.27)
2	-0.98(2.07)	7.11(0.12)	5.20(0.03)	0.002(0.005)	0.23(1.27)
3	-1.02(2.06)	7.12(0.12)	5.20(0.03)	0.001(0.009)	0.18(1.22)
4	-1.49(2.14)	7.14(0.13)	5.21(0.04)	0.001(0.002)	0.13(1.36)
5	-1.36(2.22)	7.08(0.13)	5.18(0.03)	0.001(0.005)	0.17(1.40)
6	-1.05(2.27)	7.13(0.10)	5.20(0.03)	0.001(0.006)	0.18(1.45)
7	-0.81(2.15)	7.24(0.05)	5.23(0.03)	0.002(0.010)	0.29(1.27)
8	-1.68(1.78)	7.30(0.01)	5.26(0.003)	0.000(0.001)	0.07(1.35)
9	-2.38(2.35)	7.30(0.01)	5.26(0.004)	0.0001(0.0002)	-0.28(1.42)
10	-1.92(1.91)	7.29(0.005)	5.27(0.005)	0.0002(0.0001)	-0.15(1.22)
11	-1.30(1.73)	7.11(0.12)	5.20(0.03)	0.002(0.005)	-0.04(1.02)

Table 2: *Means and standard deviations (in parentheses) of different variables for each stream order.*

Model	AIC
stream.order + length + f(precip)	6200.72
stream.order + length + f(temp)	6192.73
stream.order + length + f(slope)	6211.16
stream.order + length + f(precip) + f(temp)	6169.06
stream.order + length + f(precip) + f(slope)	6178.41
stream.order + length + f(temp) + f(slope)	6169.43
stream.order + length + f(precip) + f(temp) + f(slope)	6152.28

Table 3: *AIC for different models investigated in the data analysis section.*

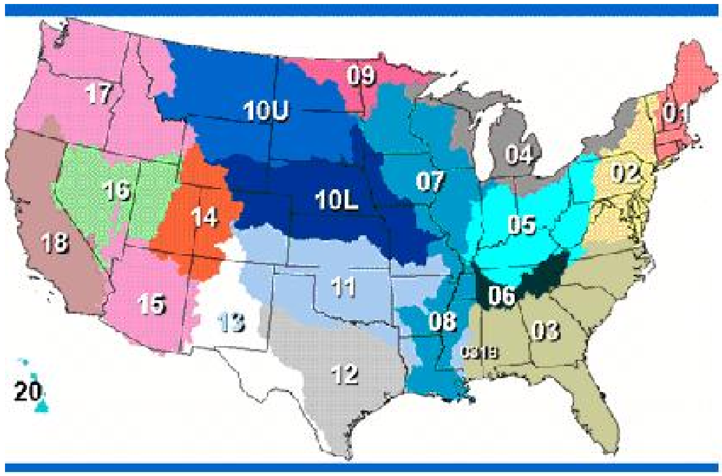
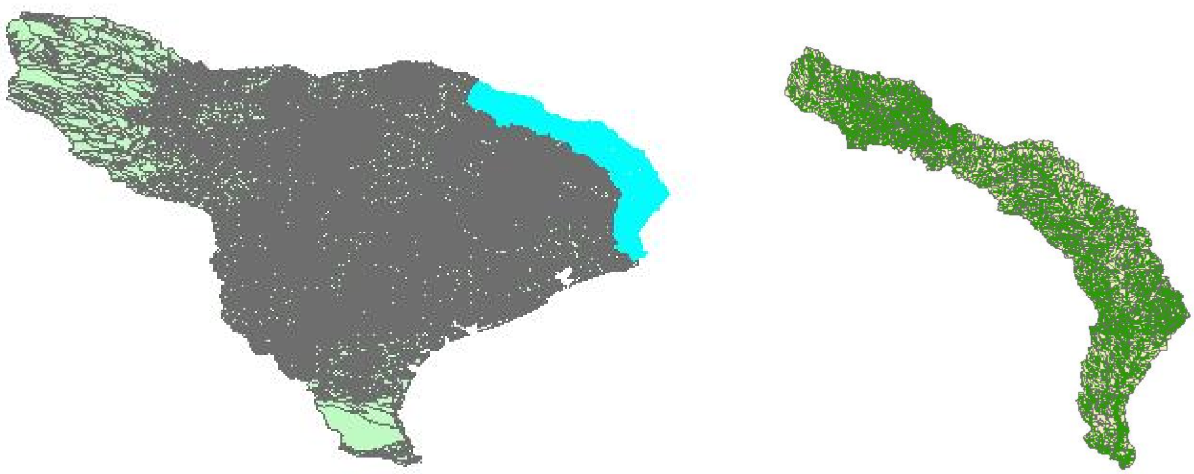


Figure 1: NHDPlus Region



(a)

(b)

Figure 2: (a) River catchments in Texas, (b) Sabine river basin

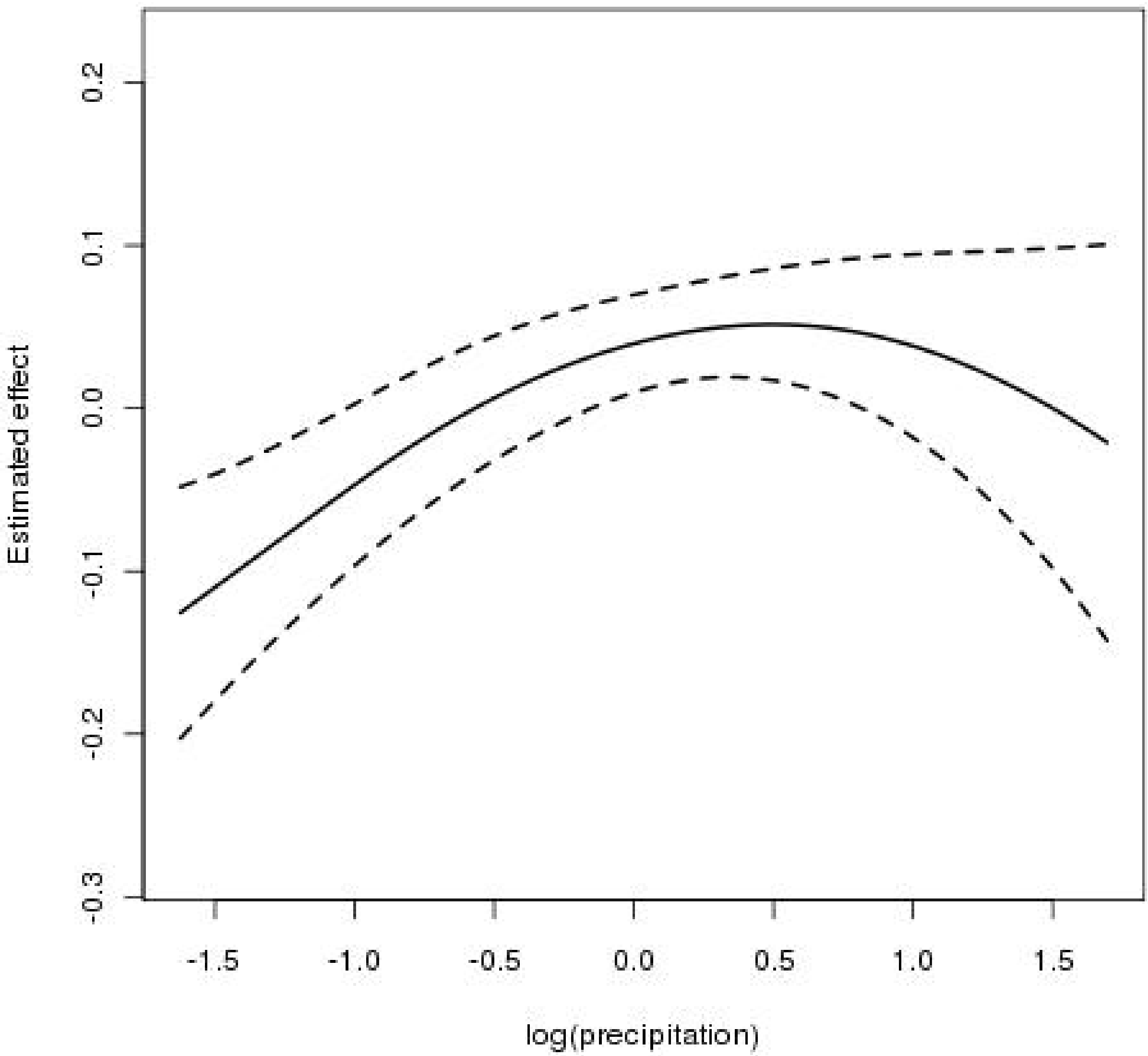


Figure 3: Estimated effect of the logarithm of the precipitation values

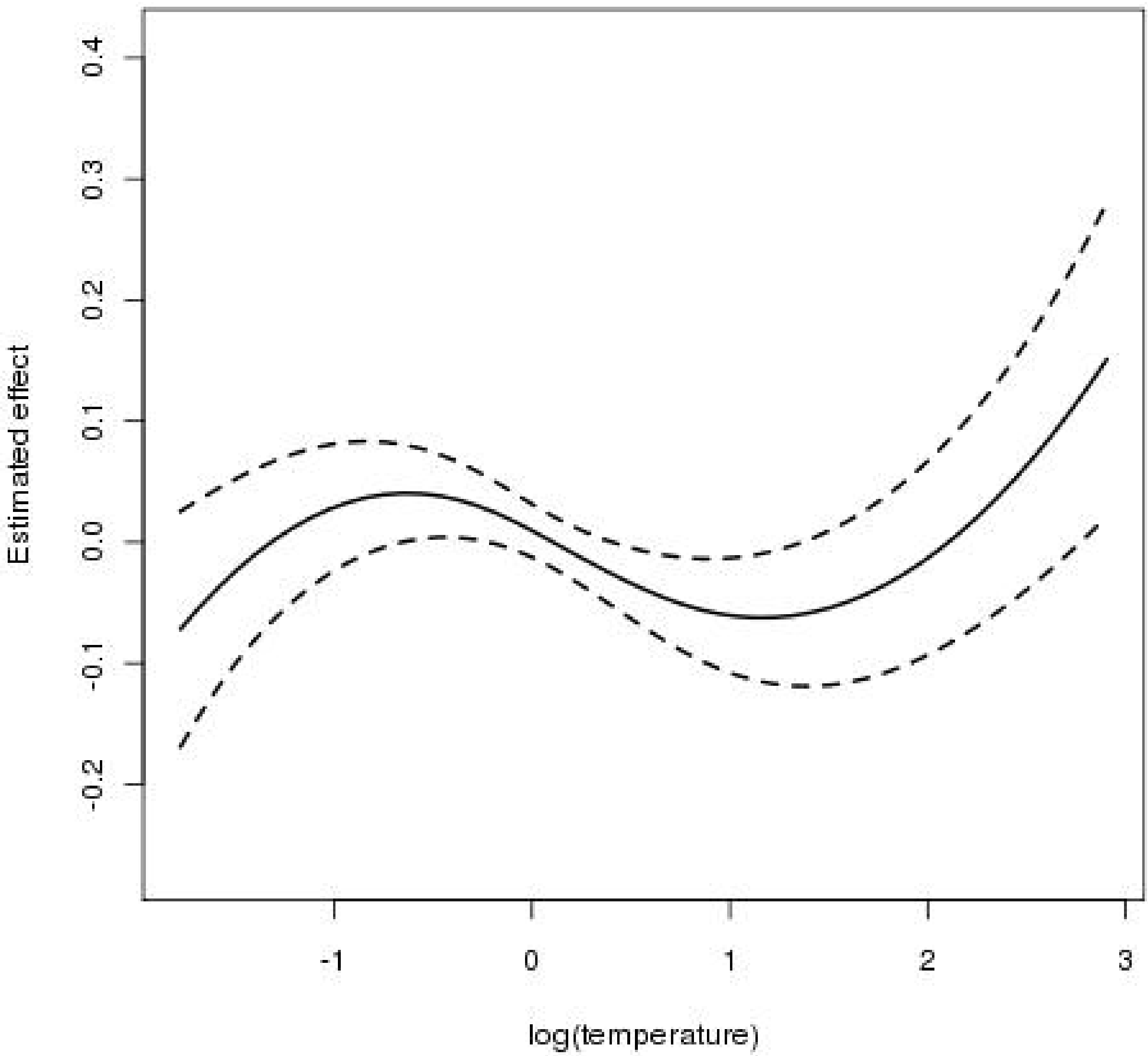


Figure 4: Estimated effect of the logarithm of the temperature values

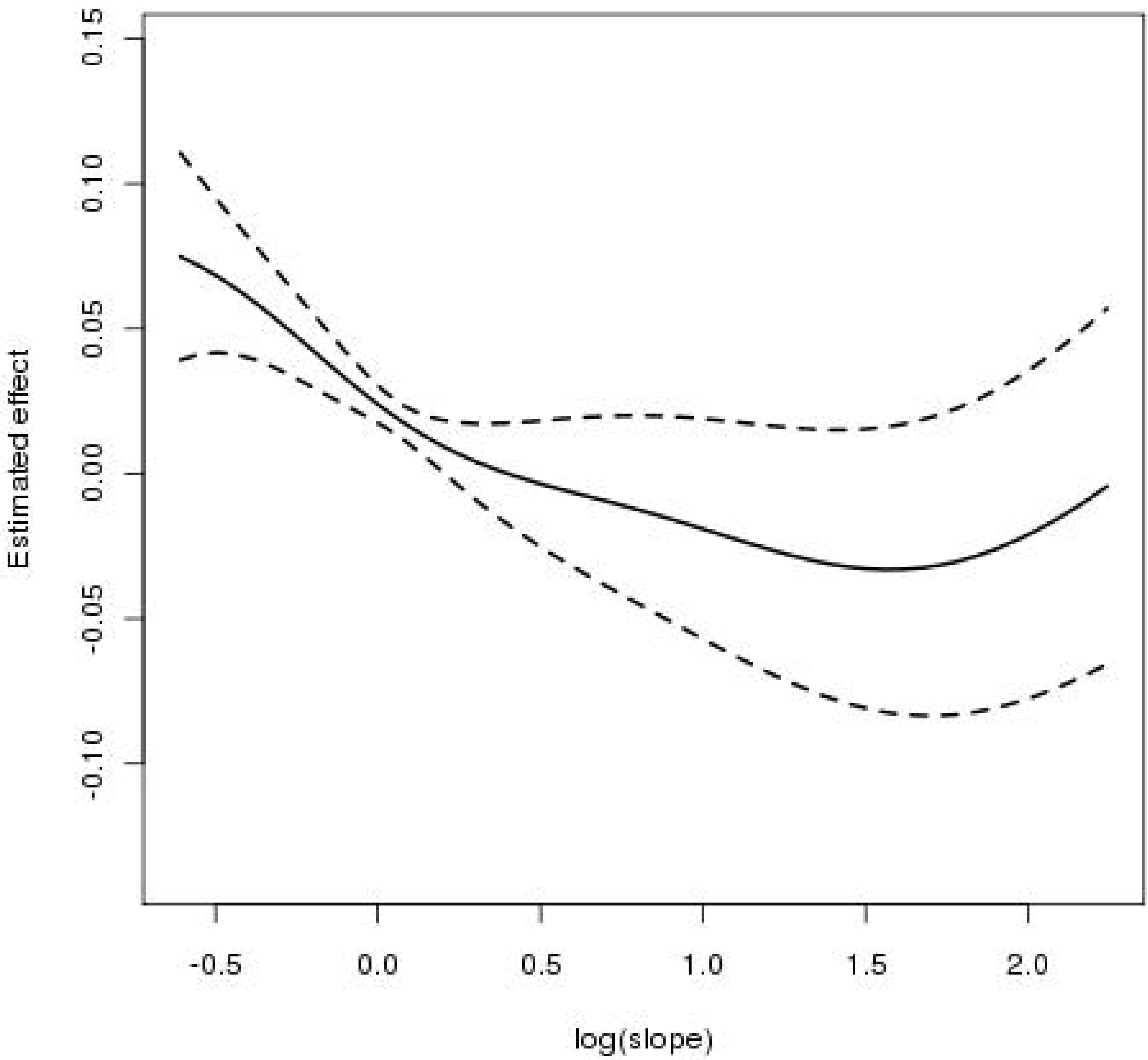


Figure 5: Estimated effect of the logarithm of the slope values

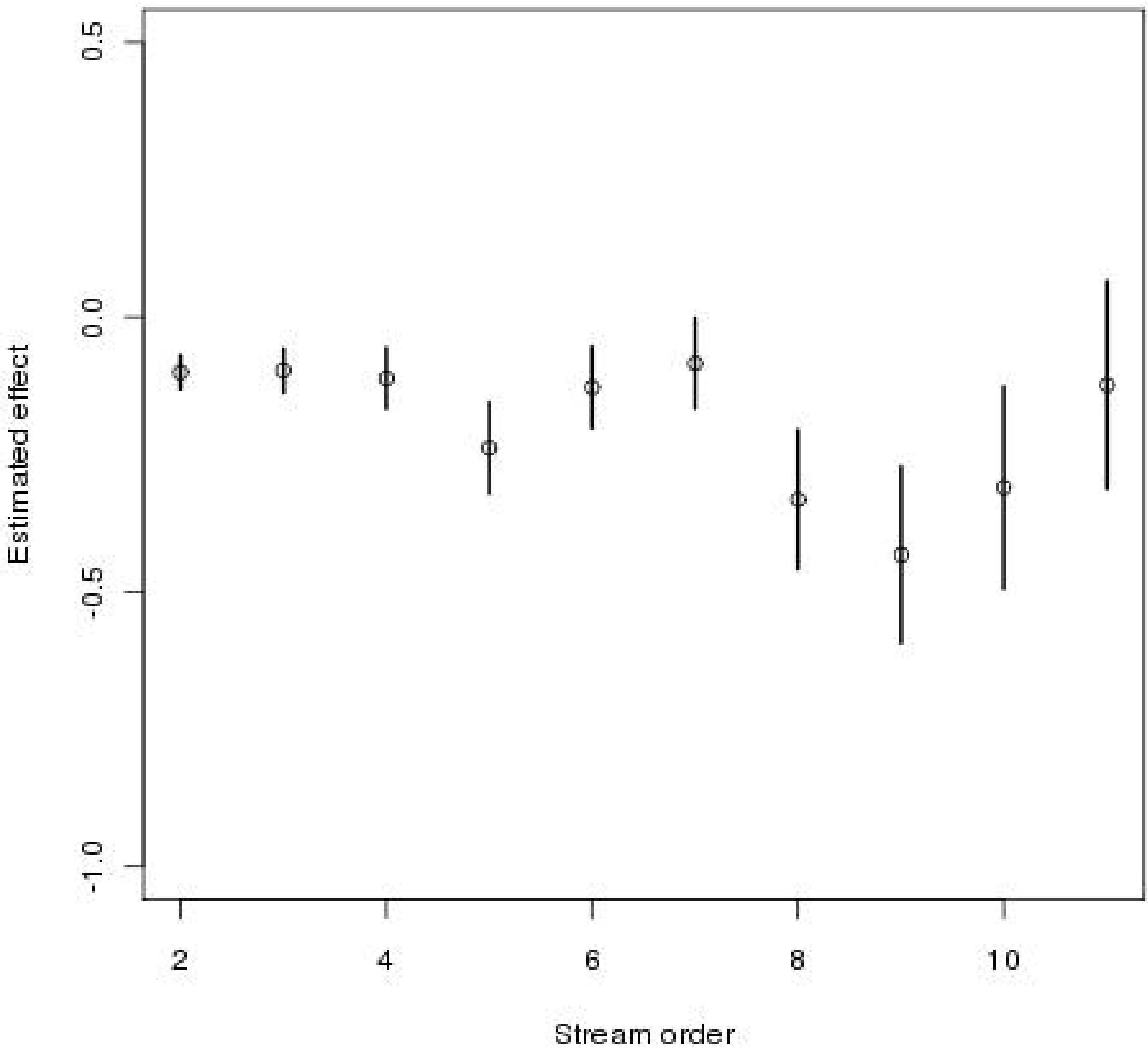
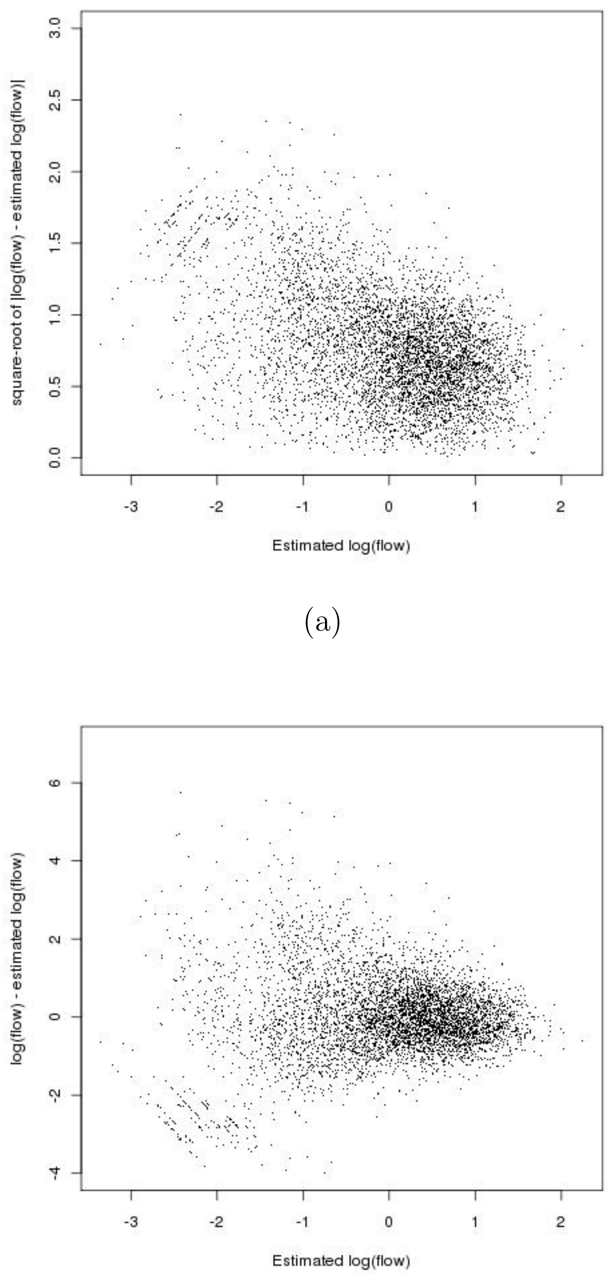


Figure 6: Fitted values of the flow values with the stream order



(a)

(b)

Figure 7: Results from data analysis. Presented are (a) the location-spread plot and (b) plot of the scaled residual versus predicted values.

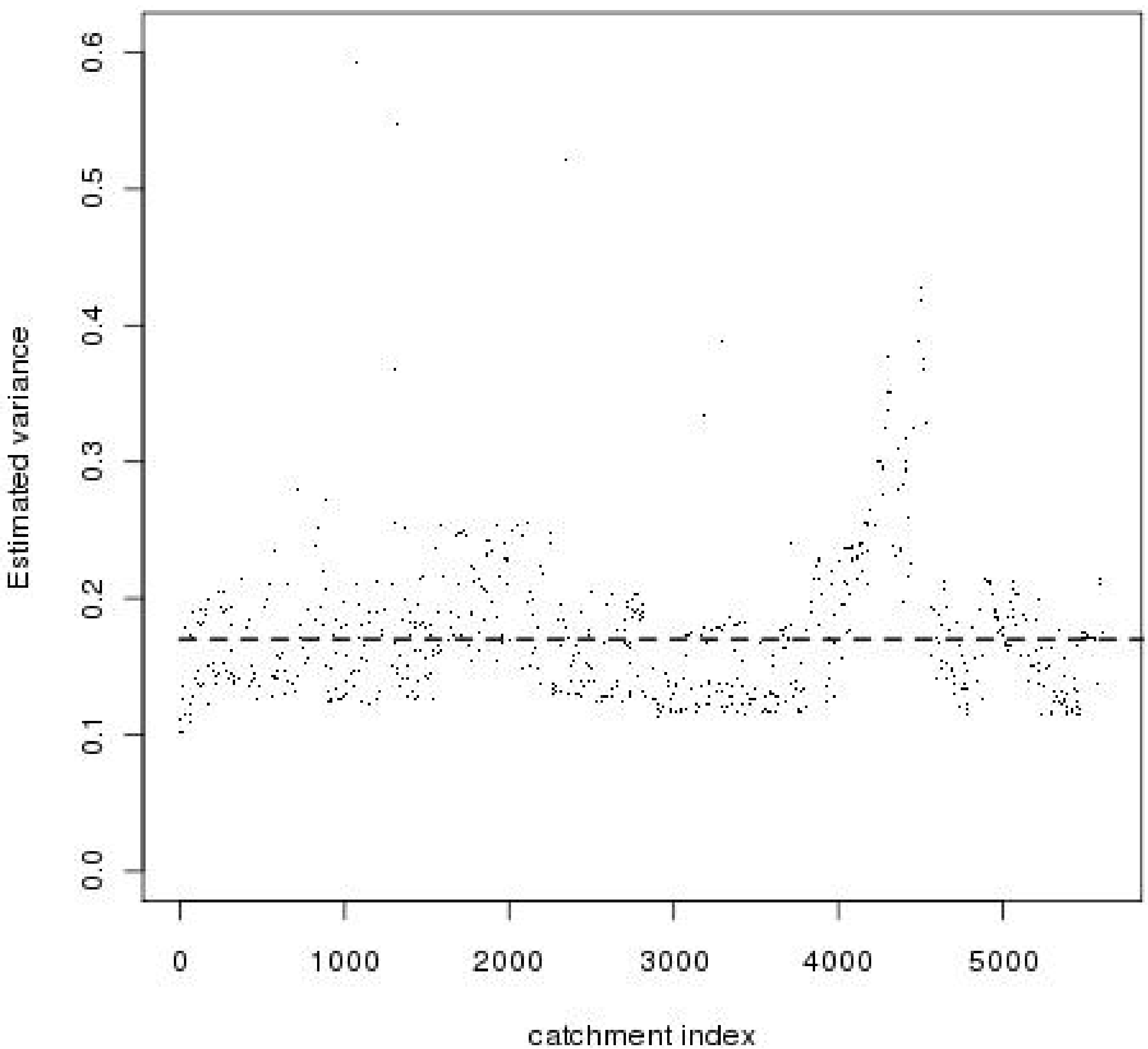


Figure 8: Results from data analysis. Plotted are the estimated variances for each catchment (points) as estimated from the heteroscedastic model and estimated common variance as derived from fitting the homoscedastic model (dashed line).