Expert Systems with Applications 40 (2013) 7407–7414

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Stream water temperature prediction based on Gaussian process regression

Ratko Grbić *, Dino Kurtagić, Dražen Slišković

Faculty of Electrical Engineering University of Osijek, Kneza Trpimira 2b, HR-31000 Osijek, Croatia

ARTICLE INFO

Keywords:
Stream water temperature
Prediction
Gaussian process regression
Variable selection
Mutual information

ABSTRACT

The prediction of stream water temperature presents an interesting topic since the water temperature has a significant ecological and economical role, such as in species distribution, fishery, industry and agriculture water exploitation. The prediction of stream water temperature is usually based on appropriate mathematical model and measurements of different atmospheric factors. In this paper, a probabilistic approach to daily mean water temperature prediction is proposed. The resulting model is a combination of two Gaussian process regression models where the first model describes the long-term component of water temperature and the other model describes the short-term variations in water temperature. The proposed approach is developed even further by modeling the short-term variations with multiple Gaussian process regression models instead with a single one. Apart from that, variable selection procedure based on mutual information is presented which is suitable for input variable selection when nonlinear models for stream water prediction are developed. The proposed approach is compared with traditional modeling approaches on the measurements obtained on the Drava river in Croatia. The presented methodology can be used as a basis of the predictive tools for water resource managers.

1. Introduction

Water temperature represents one of the most important physical properties of streams since it has great influence on the overall health of the aquatic ecosystem. Water quality and biotic conditions in the rivers are heavily determined by the water temperature. Many chemical and biological processes, such as dissolved oxygen and suspended sediment concentration, are influenced by the stream water temperature. The thermal regime of the rivers has great impact on the aquatic organisms, their distribution, growth, production and mortality. Apart from its ecological importance, it also has a significant economic role in water requirements for industry or agriculture (Caissie, 2006; Webb, Hannah, Moore, Brown, & Nobilis, 2008). Therefore, the ability to predict stream water temperature can be of great help in monitoring and protection of the river ecosystems, in the proper planning of water management and utilization or in simulation of the different climate scenarios.

There are many factors which are related to the river temperature and cause its fluctuations. These can be categorized in four groups: atmospheric conditions, topography, stream discharge and streambed (Caissie, 2006). The atmospheric conditions like solar radiation, air temperature, wind speed and humidity are one of the most important factors since they mostly determine the heat exchange that takes place at stream surface. Since the numbers of atmospheric indicators are continuously monitored, these measurements can be used to develop a model which will be used as a basis for stream water temperature prediction.

The models that can be found in the literature can be classified either as a deterministic models or statistical/stochastic models (Ahmadi-Nedushan et al., 2007; Benyahya, Caissie, St-Hilaire, Ouarda, & Bobée, 2007; Caissie, 2006). Deterministic models are based on mathematical approximation of the underlying relationship that exists between water stream and its surroundings in form of energy balances. Although this approach provides an insight into mechanism that drives the stream water temperature, it is frequently impractical and time consuming due to its complexity because a great number of input variables are examined including stream geometry, hydrology and meteorology. A number of statistical/stochastic models was proposed in literature in order to alleviate the need for high number of different explanatory variables which are often not available. These models are used to predict stream water temperature typically using air temperature measurements only. Simple statistical models assume that linear relationship exists between water and air temperature, resulting in linear regression models that are used to produce predictions of stream water temperature based on air temperature on the weekly or monthly basis (Caissie, El-Jabi, & Satish, 2001; Webb & Nobilis,

* Corresponding author. Tel.: +385 31 224750; fax: +385 31 224605.
E-mail addresses: ratko.grbic@etfos.hr (R. Grbić), dino.kurtagic@etfos.hr (D. Kurtagić), drazen.sliskovic@etfos.hr (D. Slišković).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.06.077

1997). To produce predictions on a daily basis, stochastic models are often preferred due to smaller prediction error (Ahmad, Khan, & Parida, 2001; Ahmadi-Nedushan et al., 2007; Caissie et al., 2001; Kothandaraman & Evans, 1972). The stochastic models take into account the property that water temperature exhibit long-term (annual) component and short-term component. The long term component is usually modeled with a simple sinusoidal function while the short-term component, i.e. the departure from the long term component, is modeled with some kind of dynamic model (Ahmad et al., 2001). With the enormous advances in artificial intelligence and machine learning last three decades (Bishop, 2006; Haykin, 1999), the so-called nonparametric approaches such as artificial neural networks (Benyahya et al., 2007; Maier & Dandy, 2000) became increasingly popular for prediction of different water resource variables. The artificial neural networks, which can be viewed as a general nonlinear regression models, provide similar or better prediction accuracy of stream water temperature compared to the classical statistical/stochastic approaches (Chenard & Caissie, 2008; Sahoo, Schladow, & Reuter, 2009). Recently nonparametric approach based on k nearest neighbors model appeared in literature, providing the predictions of water temperature on daily basis (Benyahya et al., 2007; St-Hilaire, Ouarda, Bargaoui, Daigle, & Bilodeau, 2012).

Gaussian processes present a formal way for solving regression and classifications problems (Bishop, 2006; Rasmussen & Williams, 2006). Although Gaussian process are a quite mature modeling technique, a significant increase of their usage in predictive model development only recently can be observed in various fields of scientific research (see for example (Ažman & Kocijan, 2007; Gregorčič & Lightbody, 2009; Petelin, Grancharova, & Kocijan, 2013; Vanhatalo, Veneranta, & Hudd, 2012; Wu, Law, & Xu, 2012)).

In this paper we present the probabilistic approach for stream water temperature prediction based on Gaussian Process Regression (GPR). The idea of the proposed method is the similar to the stochastic models such that prediction is obtained as a combination of two GPR models. First GPR model is used for prediction of the periodic component of water temperature while the other GPR model is modeling the influence of input variables which are believed to cause the short-term (non periodic) changes of the stream water temperature. The latter model can be further divided into multiple GPR models which can effectively model the nonlinear relationship that exists between input variables and water temperature. The appropriate input variables and their lags are determined by the estimation of mutual information between (lagged) input variables and the stream water temperature. The proposed method is applied to the measurements obtained at the Drava river in Croatia and the results are compared with those obtained by the linear regression model, logistic model and stochastic modeling approach.

This paper is organized as follows. Section 2 gives the brief overview of standard approaches to stream water temperature prediction. In Section 3 method for stream water temperature prediction based on Gaussian process regression is proposed. The case study is described in Section 4. The results of the proposed method applied to the given case study are presented in Section 5, followed by the accompanying discussion. Conclusions are drawn in Section 6.

2. Stream water temperature prediction by statistical/stochastic models

In order to make predictions of the stream water temperature, the model has to be developed which approximates natural functional dependence that exists between input (explanatory, independent) variables and stream water temperature. This general dependence can be represented as:

$$\widehat{T}_w(kT_s) = f_m(\mathbf{x}(kT_s), \boldsymbol{\Theta}), \tag{1}$$

where $\widehat{T}_w(kT_s)$ stands for predicted value of the stream water temperature, $f_m(\bullet)$ is a model, $\mathbf{x}(kT_s)$ is a vector of available measurements of input variables and $\boldsymbol{\Theta}$ is vector of model parameters. The sample time T_s represents the time between two consecutive measurements of the input or output measurements and it is usually omitted for the sake of clarity. This sample time can be quite different in models for water temperature prediction, resulting in models at hourly, daily, weekly, monthly or even yearly basis. According to (1), model development consists of: (i) the selection of the appropriate input variables, (ii) the selection of the appropriate model structure (i.e. type of functional dependence f_m) and (iii) the estimation of the model parameters $\boldsymbol{\Theta}$ based on the available (recorded) measurements of stream water temperature and input variables. The measurements upon which model development is based is usually called the training dataset. While the selection of input variables is limited with the number of variables actually monitored and recorded, there are a plethora of possible model structures, ranging from simple linear functions to quite complex like those obtained by neural networks.

The simplest models are linear regression models which are used for prediction of stream water temperature based on air temperature:

$$\widehat{T}_w(k) = b_0 + b_1 T_a(k), \tag{2}$$

where $T_a(k)$ is the air temperature and b_0, b_1 are regression coefficients, i.e. model parameters. If there are more explanatory variables, multiple linear regression is used for making stream water temperature predictions:

$$\widehat{T}_w(k) = \mathbf{x}(k)\mathbf{b}. \tag{3}$$

In order to better describe the nonlinear relationship between air and water temperature, a model based on logistic function was proposed in (Mohseni & Stefan, 1999):

$$\widehat{T}_w(k) = \frac{\alpha}{1 + e^{\gamma(\beta - T_a(k))}}. \tag{4}$$

The model parameters are usually determined by a least squares approach, i.e. by minimizing the sum of squared differences between predicted and true value of the stream water temperature.

Linear regression or logistic function based models provide satisfactory prediction on weekly or monthly basis, but on smaller time scales can exhibit poor performance. On the smaller time scales the water temperature measurements are time dependent, i.e. thermal inertia of water streams should be taken into account. Standard statistical time models such as autoregressive or autoregressive moving average (Ahmad et al., 2001; Benyahya et al., 2007) can be used to model water temperature time series at daily basis. However, since the stream water temperature generally consists of long-term periodic component and short-term (component) variations, the stochastic models that separate these components can provide satisfactory prediction performance. The prediction of a stochastic model can be written as:

$$\widehat{T}_w(k) = \widehat{S}_w(k) + \widehat{R}_w(k), \tag{5}$$

where $\widehat{S}_w(k)$ is a prediction of long-term model and $\widehat{R}_w(k)$ is a prediction of residual model which models the variations about long-term stream water temperature component. In (Kothandaraman & Evans, 1972) it was shown that long-term periodic component can be successfully modeled with simple sinusoidal function with a period of one year:

$$\widehat{S}_w(k) = a + b\sin\left(\frac{2\pi}{365}(k + t_0)\right), \tag{6}$$

where a, b and t_0 are model parameters which are obtained by minimization of the appropriate criterion (e.g. least squares) applied to

the stream water temperature measurements of the training dataset. The short-term component, i.e. non-seasonal variations of the stream water temperature are mainly influenced by meteorological conditions so dynamical models with the air temperature as an input variable are generally used. In (Ahmadi-Nedushan et al., 2007; Kothandaraman & Evans, 1972), multiple linear regression model with lagged air temperature residuals $R_a(k)$ is used for making predictions of water temperature residuals:

$$\widehat{R}_w(k) = \beta_1 R_a(k) + \beta_2 R_a(k-1) + \beta_3 R_a(k-2), \quad (7)$$

where air temperature residual is obtained by subtracting the long-term air temperature component from the measured value of air temperature:

$$R_a(k) = T_a(k) - \widehat{S}_a(k), \quad (8)$$

where the long-term air temperature component can be obtained in the same way as the long-term component of stream water temperature. The regression coefficients β_1, β_2 and β_3 are determined from the water and air temperature residuals of the training dataset by a standard least squares approach. Other dynamic model structure can be used for modeling residuals, such as second order Markov model or Box Jenkins model (Ahmadi-Nedushan et al., 2007; Caissie et al., 2001).

3. Proposed methodology for stream water temperature prediction

The proposed approach for model development for a stream water temperature prediction consists of two steps. Initially, the input variables and their appropriate lags are selected according to the mutual information between input variables and stream water temperature. After the most influential variables/lags are selected, the model based on Gaussian process regression is built. In that way a nonlinear model is developed which gives the prediction of the stream water temperature in the form of a predictive distribution. This means that apart from the most probable value of stream water temperature, the reliability of prediction is also provided which can be valuable information in water utilization planning or in simulation of different climate scenarios.

3.1. Input variable selection by mutual information

To our best knowledge the selection of the appropriate model inputs and the appropriate lags gained very little attention in the literature regarding prediction of stream water temperature. The selection of the appropriate model inputs and the appropriate lags is usually based on a priori knowledge about water behavior and its dependence on different meteorological factors. When dealing with linear regression models or stochastic models, this selection can be done even by trial and error, by training and testing models with different number of variables or its lagged instances. However, when developing nonlinear regression models for stream water temperature prediction, the selection of input variables and the appropriate lags is very important since selection can greatly affect the overall prediction performance of the built model. For example, the number of parameters of artificial neural network drastically increases as the number of input variables increases which can lead to unreliable estimation of the model parameters due to limited number of samples in the training dataset and thus poor prediction performance of the model (Maier & Dandy, 2000). Simple trial and error procedure for variable selection can be quite time-consuming since the training of the neural network should be performed number of times with different initial parameters to ensure that global optimum of error function is reached (Haykin, 1999). Keeping in my mind that optimal number of neurons should be also determined during neural network learning, it is getting clear that the appropriate variable selection procedure should be implemented when using flexible structures like neural networks as a model basis. Therefore, the input variable selection procedure is proposed which can select the most important variables or their lagged instances upon which the stream water temperature model is going to be developed. The proposed procedure for variable selection is based on the information theory and it is independent from the modeling method that is applied.

Mutual Information (MI) is a measure of the amount of information that one continuous random variable X contains about another continuous random variable Y (Cover & Thomas, 2006):

$$I(X;Y) = -\int\int p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\mathrm{d}x\mathrm{d}y. \quad (9)$$

MI is a nonparametric, nonlinear measure of relevance between variables and it can identify relations of any type between variables. The MI will be larger for variables which have a stronger relation and vice versa. Input variable selection based on value of MI (9) of each input variable with the output variable can be used to identify the most important input variables for output variable prediction (Battiti, 1994). However, this cannot detect which variables are highly correlated, i.e. which carry the same information about output variable and thus necessary complicate model development. For example the measurements of different meteorological variables and its lagged instances can be highly correlated. Therefore, following variable selection procedure is used. If all the available measurements are presented as M input variables $V = \{x_1, \ldots, x_M\}$ and stream water temperature as T_w, then the first variable that is selected is the variable which has the highest MI with the stream water temperature:

$$x_{S1} = \arg\max_{x_i} I(x_i; T_w), \quad i = 1, \ldots, M. \quad (10)$$

After the first variable x_{S1} is selected, the next variable (from the remaining) which has the highest MI with the stream water temperature is selected, but taking into account that one variable is already selected. This can be done by searching the variable x_i which maximizes the following MI:

$$x_{S2} = \arg\max_{x_i} I(x_{S1}, x_i; T_w), \quad i = 1, \cdots\cdots, M, \ i \neq S_1. \quad (11)$$

The procedure is repeated in similar fashion until the MI between selected variables and stream water temperature stops its increase or starts to decrease, indicating that the last added variable does not provide additional information about stream water temperature and should not be used in modeling. In that way only the most important input variables are selected and the redundancy of the input space is kept as low as possible. In the end only m variables from available M are used in model development.

Unlike the autocorrelation or cross-correlation functions which are usually used for selecting the appropriate variable lags, the MI takes into account also nonlinear correlations that exists between variables and therefore is suitable for variable selection in nonlinear (dynamic) model development.

3.2. Stream water temperature model based on GPR

Gaussian Process Regression (GPR) presents a probabilistic, nonparametric approach for solving nonlinear regression problems. GPR assumes that the measurements of the output variable y are generated in the following way:

$$y = f(\mathbf{x}(k)) + \varepsilon, \quad (12)$$

where $\mathbf{x}$ stands for a measurement of input variables, f is the unknown functional dependence and ε is a Gaussian noise with

variance σ_n^2. Instead of parameterizing the unknown function f like in (1), the prior probability over the space of functions is defined in terms of Gaussian process (Rasmussen & Williams, 2006). The Gaussian process is fully defined by mean $m(\mathbf{x})$ and covariance function $\mathrm{cov}(\mathbf{x}, \mathbf{x}')$ which encode our assumptions about the underlying process that generated the data. Once the mean and covariance function are selected, the prediction of the output variable for a sample of input variables $\mathbf{x}_*$ is given in the form of predictive Gaussian distribution $p(y_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)$ with mean and variance:

$$\begin{aligned} \hat{y}_* &= m(\mathbf{x}_*) + \mathbf{k}_*^{\mathrm{T}}(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}(\mathbf{y} - m(\mathbf{x}_*)), \\ \sigma_{y_*}^2 &= k_* + \sigma_n^2 - \mathbf{k}_*^{\mathrm{T}}(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{k}_*, \end{aligned} \tag{13}$$

where $\mathbf{K}$ is a covariance matrix with elements $[\mathbf{K}]_{i,j} = \mathrm{cov}(\mathbf{x}_i, \mathbf{x}_j)$, vector $\mathbf{k}_*$ is defined as $[\mathbf{k}_*]_i = \mathrm{cov}(\mathbf{x}_i, \mathbf{x}_*)$ and $k_* = \mathrm{cov}(\mathbf{x}_*, \mathbf{x}_*)$. From (13), it can be seen that the prediction is based on the training dataset $\mathbf{X}$, $\mathbf{y}$ which is in contrast with traditional regression approaches where only the parameters are used to obtain the prediction.

In order to make accurate predictions, the parameters of the mean and covariance function have to be estimated from the available data. These parameters are called hyperparameters since they define properties of the predictive probability distribution (13). The values of the hyperparameters are usually obtained by maximization of $\log p(\mathbf{y}|\mathbf{X})$ which is log-likelihood function of the training data (Rasmussen & Williams, 2006):

$$\log p(\mathbf{y}|\mathbf{X}) = -\frac{1}{2}\mathbf{y}^{\mathrm{T}}(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y} - \frac{1}{2}\log(|\mathbf{K} + \sigma_n^2\mathbf{I}|) - \frac{n}{2}\log(2\pi), \tag{14}$$

where n is the number of training samples.

Our proposed approach for stream water temperature prediction is based on the observation that water temperature has a long-term seasonal component and short-term variations which are departures from the long-term component mainly associated with current meteorological conditions. Both components are modeled with GPR models. Hence, the probabilistic prediction of the stream water temperature is given in the form:

$$p(T_w(k)|k, \mathbf{x}(k), S_a(k)) = p(S_w(k)|k) + p(R_w(k)|\mathbf{x}(k), S_w(k), S_a(k)). \tag{15}$$

The first GPR model gives the predictive distribution $p(S_w(k)|k)$ and presents the time-series model of the stream water temperature. This seasonal component can be successfully modeled with the constant mean function and the periodic type of covariance function:

$$\begin{aligned} m(k) &= c, \\ \mathrm{cov}(k, k') &= s_f^2 \exp\left(-2\sin^2\left(\frac{k - k'}{p}\right)\Big/ l\right), \end{aligned} \tag{16}$$

where c, s_f, p and l are hyperparameters of the mean and covariance function. Similarly, the seasonal component of the air temperature can be modeled as the time series with the same type of mean and covariance function, resulting in predictive probability distribution $p(S_a(k)|k)$. Once the seasonal components of the water and air temperatures are obtained, the residuals of air $R_a(k)$ and water temperature $R_w(k)$ are obtained for the available training dataset by subtracting the mean predictions made by the seasonal component models from the measured values of air and water temperature. Then, the GPR model which describes the relationship between water and air temperature residuals is derived and it provides the second predictive distribution on the right side of Eq. (15). This second model can have other input variables as well (like wind speed, pressure, stream discharge and so on) which are together with air temperature residuals denoted as vector $\mathbf{x}$. Hereby, constant mean function and squared exponential covariance function is a good choice:

$$\mathrm{cov}(\mathbf{x}, \mathbf{x}') = s_f^2 \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^{\mathrm{T}}\mathbf{P}(\mathbf{x} - \mathbf{x}')\right), \tag{17}$$

where $\mathbf{P}$ is a diagonal matrix $\mathbf{P} = diag\{l_1, .., l_m\}$ and m is the number of input variables.

The stream water has different properties at very low air temperatures (freezing can appear), mid air temperatures and at very high air temperatures (evaporative cooling appears). To take into account this nonlinear behavior, the residual part of model (15) can be expanded even further by modeling this part with J (local) GPR models where each model covers some specific part of a year like winter or summer. The partitioning (clustering) of the training data can be done according to the water and air temperatures using Gaussian mixture model (Bishop, 2006; Hastie, Tibshirani, & Friedman, 2008) which is a finite mixture of J Gaussian functions:

$$p(\mathbf{x}|\mathbf{\Theta}_{\mathrm{GM}}) = \sum_{j=1}^{J} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j). \tag{18}$$

The vector $\mathbf{\Theta}_{\mathrm{GM}} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_J, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_J, \pi_1, \ldots, \pi_J\}$ is the vector of parameters of the Gaussian mixture model, where $\boldsymbol{\mu}_j$ is the mean vector, $\boldsymbol{\Sigma}_j$ is the covariance matrix of the jth Gaussian distribution and π_j are mixing coefficients. The parameters of the Gaussian mixture model are usually obtained by the maximum likelihood approach (Bishop, 2006; Hastie et al., 2008). Once the parameters of the Gaussian mixture model are determined, the training data can be clustered and for each cluster appropriate local GPR model is built. Then, the prediction of the residual part of the stream water temperature is obtained as weighted sum of predictive distributions of the local GPR models:

$$p(R_w(k)|\mathbf{x}(k), S_w(k), S_a(k)) = \sum_{j=1}^{J} p(j|\mathbf{x}(k)) p(R_{wj}(k)|\mathbf{x}(k), S_w(k), S_a(k)), \tag{19}$$

where $p(R_{wj}(k)|\mathbf{x}(k), S_w(k), S_a(k))$ is predictive distribution of the jth local model and $p(j|\mathbf{x}(k))$ is the probability that sample $\mathbf{x}(k)$ belongs to the jth cluster which is obtained by the Gaussian mixture model.

4. Case study

The measurements of hydrological and meteorological parameters for this analysis were collected on the Drava river on the territory of the Republic of Croatia. The total length of the Drava river is 749 km of which 305 km are inside the Republic of Croatia and making, on the north, a natural border with Republic of Hungary. This part of Croatia has a continental climate with a clear exchange of the four seasons: summer, autumn, winter and spring.

The available data were collected at the hydrological and meteorological stations Donji Miholjac, Croatia, in the period 1993–1998 and consists of water temperature measurements, daily mean air temperature measurements and daily mean river flow measurements. The period from 1993–1996 was selected as a training dataset while the prediction capabilities of the models were tested on the rest of the available data, i.e. the period 1997–1998. The training (blue) and test data (red) samples are shown in Fig. 1. It can be noticed that temperature measurements have long-term periodic component with a period of approximately one year. The mean values of the water and air temperature are similar (air temperature mean value for the train dataset is 11.41 °C, while the mean value of water temperature is 11.96 °C). The maximal values of air temperature and water temperature are 28.6 °C and 26.2 °C respectively, while the minimal values are −14.8 °C and 0 °C which clearly indicates that water temperature

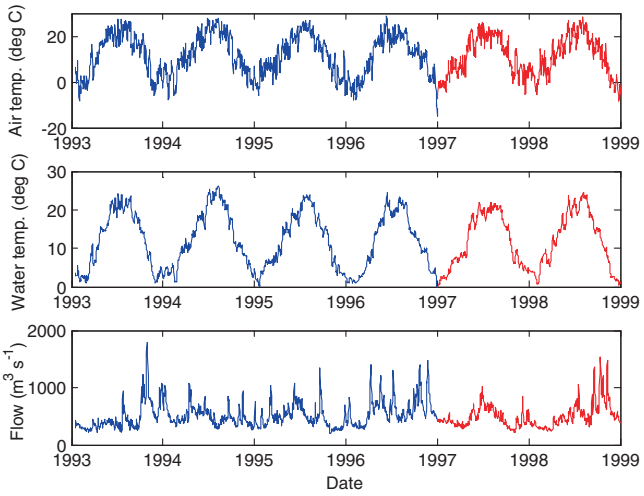

Fig. 1. Water temperature, air temperature and river flow for the period 1993–1998.

differently responds to the air temperature changes during the winter period.

5. Results and discussion

The prediction capabilities of the built models were evaluated on the test dataset using three performance indicators. The Nash–Sutcliffe coefficient of efficiency has been widely used to evaluate the performance of hydrological models. It provides a good insight into reliability of the models because it measures how much of the total variance in the output variable data can be explained by the model:

$$\mathrm{NSC} = 1 - \sum_{i=1}^{n_t}\left(\widehat{T}_w(i) - T_w(i)\right)^2 / \sum_{i=1}^{n_t}\left(\overline{T}_w - T_w(i)\right)^2, \tag{20}$$

where n_t is the number of samples in the test dataset. The NSC coefficient takes the values from $-\infty$ to +1. The higher the value of NSC coefficient the better is the model. The root mean square error on the test data:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n_t}\sum_{i=1}^{n_t}\left(\widehat{T}_w(i) - T_w(i)\right)^2}, \tag{21}$$

gives an overall insight into how well the model fits the test data since its incorporates both the variance and the bias of the prediction error. However, because the RMSE heavily weights the outlying observations, the mean absolute error is used as an additional indicator:

$$\mathrm{MAE} = \frac{1}{n_t}\sum_{i=1}^{n_t}|\widehat{T}_w(i) - T_w(i)|. \tag{22}$$

5.1. Linear regression, logistic and stochastic models

Classical models for water temperature prediction were developed according to Section 2. For all models parameters were obtained by least squares criterion. The prediction capabilities of different models for the test dataset are presented in Table 1. The simple linear regression model (2) and logistic model (4) cannot satisfactory describe the relationship that exists between air and water temperature. This can be seen on Fig. 2 where blue dots represent training data, solid black line linear regression model and green dashed line logistic model. These models have high values

Table 1
Prediction capabilities of linear regression, logistic and stochastic models for the test dataset.

Model	Input variables	NSC	RMSE [°C]	MAE [°C]
Linear regression	$T_a(k)$	0.8661	2.5485	2.0367
Multiple linear regression	$T_a(k)$, $T_a(k-1)$, $T_a(k-2)$	0.9242	1.9177	1.5471
Logistic model	$T_a(k)$	0.8769	2.4436	1.9126
Sin. time series model	k	0.9341	1.7985	1.4537
Stochastic model	k, $T_a(k)$, $T_a(k-1)$, $T_a(k-2)$	0.9748	1.1047	0.8609

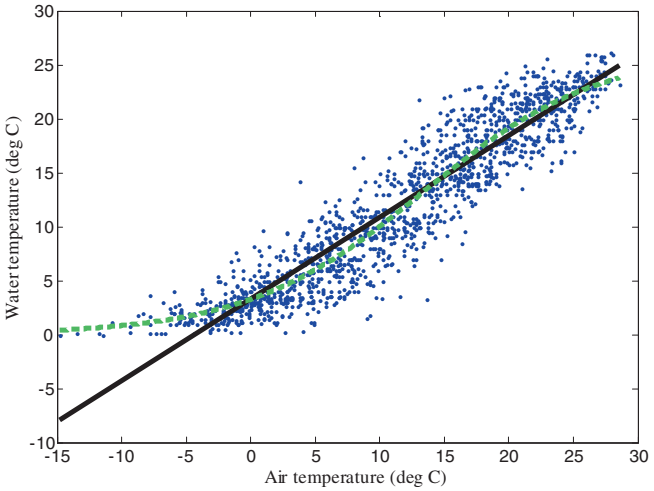

Fig. 2. Linear regression model, logistic model and the training data.

of RMSE and MAE with NSC coefficient lower than 0.9. This comes from the fact that seasonality appears when temperatures are measured at the daily basis and this phenomenon cannot be described with such models. Multiple linear regression model (3), which has two additional inputs corresponding to the lagged air temperature measurement, performs much better than linear regression model or logistic model since it takes into account dynamics that exists between air and water temperature relationship. Interestingly, simple sinusoidal time series model (6) performs better than regression models or logistic models, clearly indicating that periodic component dominates in the stream water temperature. This time series sinusoidal model of stream water temperature was used as a basis of the stochastic model (5), where residuals were modeled with multiple linear regression (7). According to the Table 1, this model has the best performance which was expected since it models both components of the stream water temperature series. The test samples of the stream water temperature and the predictions obtained by the stochastic model are shown in Fig. 3 where it can be seen that the biggest prediction errors are during summer and winter time.

5.2. Stream water temperature prediction by GPR models

The seasonal components of the air and water temperature were modeled with a GPR model with the constant mean function and periodic covariance function (16). The hyperparameters are obtained by maximization of the log-likelihood function (14) of the training dataset. The resulting values of the hyperparameters are given in Table 2. The values of the constant of the mean function for both models are around 5 °C which is quite different from the mean of air and water temperature of training dataset. The period of the covariance function is around 365 days as expected

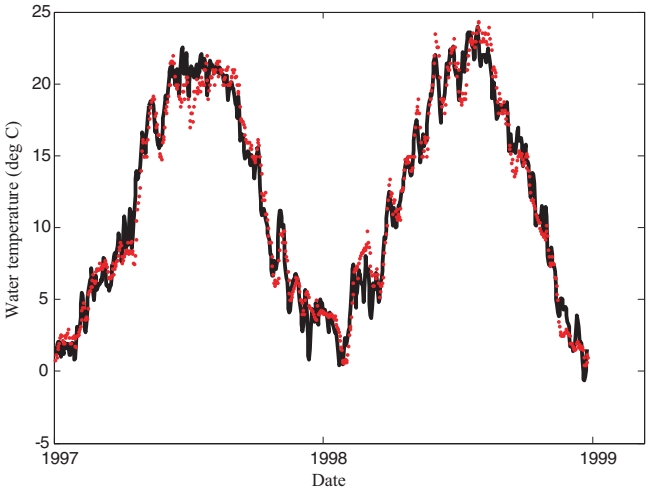

Fig. 3. Water temperature prediction for test dataset by the stochastic model.

Table 2
Hyperparameters of GPR models of water and air temperature seasonal components.

Model	Hyperparameter			
	c	s_f^2	p	l
GPR air temperature	5.0019	20.8674	366.01	3.7503
GPR water temperature	5.0069	6.4021	365.60	1.0124

while the parameter s_f^2, which can be seen as the magnitude, is much higher for air temperature model than for water temperature model.

Fig. 4 shows the obtained predictive distribution $p(S_w(k)|k)$ for the seasonal components of the stream water temperature for the period 1993–1996. The black solid line represent the mean of the predictive distribution while the grey shaded region corresponds to the ±2 standard deviations of the prediction and can be interpreted as a level of confidence of the model. It can be noticed that confidence region expands during winter and summer indicating that water temperature exhibits (short-term) behavior which cannot be modeled with the current GPR. The prediction capabilities of this time series GPR model for the test dataset are shown in Table 3. This model performs much better than simple

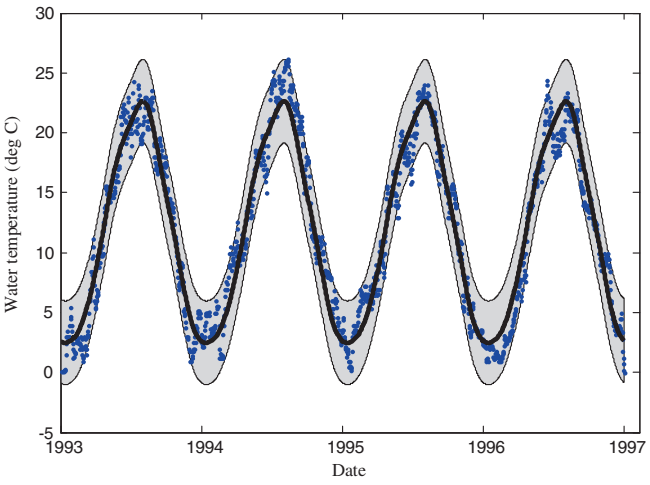

Fig. 4. Predictive distribution of the water temperature seasonal component by time series GPR model for training dataset.

Table 3
Prediction capabilities of different GPR based models for the test dataset.

Model	Input variables	NSC	RMSE [°C]	MAE [°C]
Time series GPR	k	0.9370	1.7582	1.4173
Single GPR	k, $T_a(k-1)$, $T_a(k-2)$, $T_a(k-3)$	0.9750	1.1017	0.8702
Single GPR	k, $T_a(k-1)$, $T_a(k-2)$, $T_a(k-6)$, $Q(k)$	0.9817	0.9426	0.7559
Multiple GPR ($J=2$)	k, $T_a(k-1)$, $T_a(k-2)$, $T_a(k-6)$, $Q(k)$	0.9842	0.8758	0.6979
Multiple GPR ($J=3$)	k, $T_a(k-1)$, $T_a(k-2)$, $T_a(k-6)$, $Q(k)$	0.9845	0.8658	0.6813

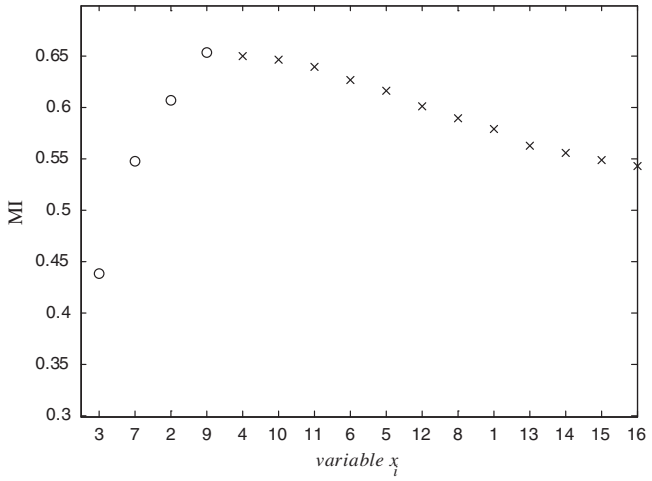

Fig. 5. Values of MI during variable selection procedure.

linear regression model and logistic model (it has much higher value of NSC coefficient and smaller values of the RMSE and MAE) and has similar performance to the sinusoidal time series model (see Table 1). However, this model provides the confidence of the prediction which traditional models (Section 2) do not provide.

Before residual model development, variable selection procedure based on mutual information (MI) is performed in order to select the most informative variables for stream water temperature prediction. The previously described time series GPR models of seasonal components of air temperature and water temperature (see Table 2) are used to obtain the residuals of water and air temperatures for the training dataset. To the available (training) measurements of the air temperature residuals $R_a(k)$ and measurement of river flow $Q(k)$ their lagged instances were added. The maximal lag of these two explanatory variables was equal to 7 days to ensure that only short-term variations are examined by the variable selection procedure. This resulted in the following set of possible input variables $V = \{R_a(k), R_a(k-1), \ldots, R_a(k-7), Q(k), Q(k-1), \ldots, Q(k-7)\}$. From all available variables in set V only the most influential were selected according to Section 3.1. The result of the variable selection procedure is presented in Fig. 5. The abscissa axis shows which variable from set V is selected in each step of selection procedure while the values at ordinate show the MI between selected input variables and water temperature residuals. It can be noticed that the MI between the selected variables and residuals of stream water temperature is starting to decrease after the 4th variable from set V is added. This indicates that there is no new information regarding output variable in the remaining variables and the selection procedure should be stopped. Therefore, the following input vector is used in residual model training $\mathbf{x}(k) = [R_a(k-2) \quad R_a(k-6) \quad R_a(k-1) \quad Q(k)]$.

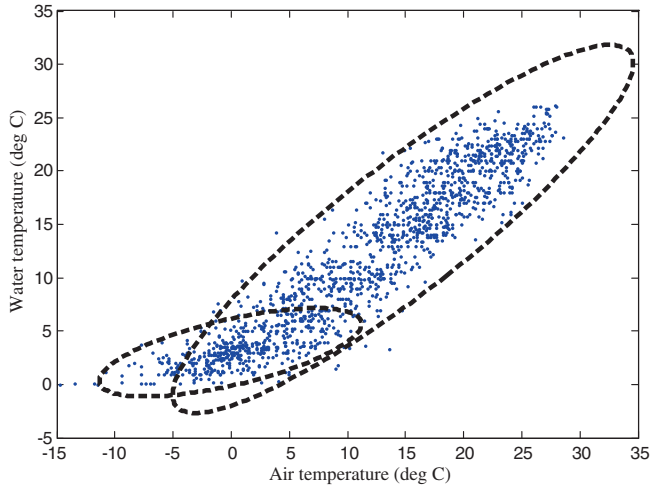

Fig. 6. Training dataset and confidence region of each component in Gaussian mixture model ($J = 2$).

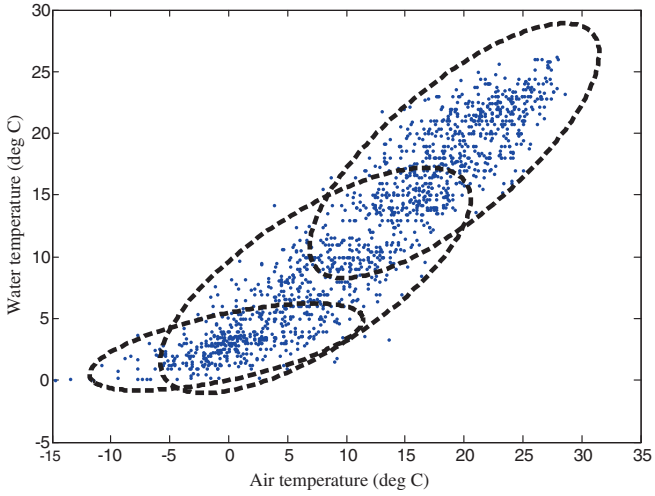

Fig. 7. Training dataset and confidence region of each component in Gaussian mixture model ($J = 3$).

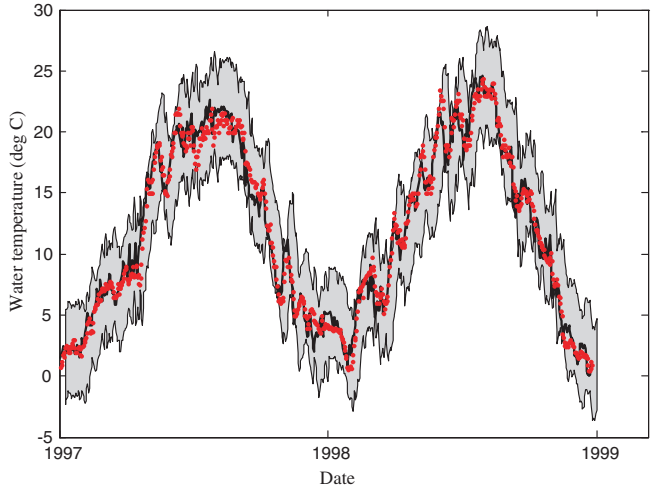

Fig. 8. Predictive distribution of the water temperature by the multiple GPR model for test dataset.

The results of stream water temperature prediction of GPR based models given by Eq. (15) are shown in Table 3. The models consist of time series GPR model which captures seasonal component of water temperature and single or multiple GPR models which capture relationship between input variables and residuals of water temperature. Latter GPR models are based on constant mean function and squared exponential covariance function (17). To compare the single GPR model with the stochastic model, in which the seasonal component is modeled with a sinusoidal function and residuals by multiple linear regression model (see Subsection 5.1), the single GPR model was trained on the same input variables as the stochastic model. According to the obtained results, this single GPR models has similar prediction capabilities as the stochastic model but can provide the confidence in prediction. However, the single GPR model developed on the input variables obtained by the variable selection procedure performs better, suggesting that variable selection procedure plays important role in prediction model building.

In order to model the residuals with a multiple GPR models, training data clustering has to be performed. Figs. 6 and 7 show the results of Gaussian mixture model fitting to the training data for the number of components equal to 2 and 3. The ellipses represent the 98% confidence region of the each component in the Gaussian mixture model (18). In the first case, one component is covering the low temperature area while the other one the rest of the temperature span. In the second case, the first component is associated with lower temperatures, the second one with mid temperature interval and the third with higher temperature regime. Training data were clustered according to the value of $p(j|\mathbf{x}(k))$ and then a local GPR model was associated with each cluster. The prediction of the test dataset by these multiple GPR models was made according to (19). As seen in Table 3, the approach based on multiple GPR models has the best prediction capabilities with the RMSE below 0.9 °C, MAE below 0.7 °C and NSC over 0.984. Both multiple GPR models have similar prediction performance from which it can be concluded that the two local models are sufficient to capture the nonlinearity which exists between the input variables and water temperature. The stream water temperature prediction of the multiple GPR model ($J = 3$) for the test dataset is shown in Fig. 8. If this figure is compared with Fig. 3, where prediction of the stochastic model is shown, it can be concluded that these models perform similarly during mid temperature season (autumn/spring) but during winter and summer period the multiple GPR model predicts the stream water temperature better.

6. Conclusion

The paper presents a Gaussian process regression approach for model development in stream water temperature prediction. The predictive model consists of two GPR models where the first model captures the periodic (seasonal) component of water temperature time series, while the second GPR model models the short-term (non-periodic) component, i.e. the relationship between water temperature residuals and input variables. This GPR approach to modeling is examined even further by modeling the relationship between water temperature residuals and input variables with multiple GPR models. In this case, the available measurements of air temperature and water temperature are clustered by a Gaussian mixture model where a local GPR model is developed for each cluster. The methodology is applied to measurements obtained at the hydrological and meteorological station situated on the Drava river in Donji Miholjac, Croatia. Multiple GPR model approach combined with input variable selection procedure obtained the best result in stream water temperature prediction with RMSE around 0.87 °C and MAE below 0.7 °C. This clearly outperforms traditional model-

ing approach and predictive model with single GPR model for short-term water temperature component where RMSE is around 1.10 °C and MAE is around 0.86 °C. In addition, GPR based models provide the prediction of stream water temperature in the form of a predictive distribution which means that a level of confidence of the model prediction is also provided.

The input variables were selected according to the mutual information which measures the amount of information that two variables share. According to the results, the most important variables for stream water temperature prediction are lagged air temperature measurements and the current flow of the river. It would be interesting to see the influence of other meteorological variables (wind speed, pressure, moisture, solar radiation) on the results of variable selection procedure and overall prediction performance of the proposed models.

The models are interesting from the practical view because they provide the confidence in prediction which traditional models do not provide. Apart from that, the presented methodology can be used for prediction of other important factors in industry where periodicity can be recognized, like in electrical energy consumption, drinking water consumption, natural gas consumption, CO_2 concentration, etc.

Acknowledgements

The authors would like to thank Croatian National Meteorological and Hydrological Service for providing the hydrological and meteorological measurements that are used in this study.

References

Ahmad, S., Khan, I. H., & Parida, B. (2001). Performance of stochastic approaches for forecasting river water quality. *Water Research, 35*(18), 4261–4266. http://dx.doi.org/10.1016/S0043-1354(01)00167-1.

Ahmadi-Nedushan, B., St-Hilaire, A., Ouarda, T. B. M. J., Bilodeau, L., Robichaud, É., Thiémonge, N., et al. (2007). Predicting river water temperatures using stochastic models: case study of the Moisie River (Québec, Canada). *Hydrological Processes, 21*(1), 21–34. http://dx.doi.org/10.1002/hyp.6353.

Ažman, K., & Kocijan, J. (2007). Application of Gaussian processes for black-box modelling of biosystems. *ISA Transactions, 46*(4), 443–457. http://dx.doi.org/10.1016/j.isatra.2007.04.001.

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks, 5*(4), 537–550. http://dx.doi.org/10.1109/72.298224.

Benyahya, L., Caissie, D., St-Hilaire, A., Ouarda, T. B. M., & Bobée, B. (2007). A review of statistical water temperature models. *Canadian Water Resources Journal, 32*(3), 179–192. http://dx.doi.org/10.4296/cwrj3203179.

Bishop, C. M. (2006). *Pattern recognition and machine learning* (second ed.). Springer.

Caissie, D. (2006). The thermal regime of rivers: A review. *Freshwater Biology, 51*(8), 1389–1406. http://dx.doi.org/10.1111/j.1365-2427.2006.01597.x.

Caissie, D., El-Jabi, N., & Satish, M. G. (2001). Modelling of maximum daily water temperatures in a small stream using air temperatures. *Journal of Hydrology, 251*(1–2), 14–28. http://dx.doi.org/10.1016/S0022-1694(01)00427-9.

Chenard, J.-F., & Caissie, D. (2008). Stream temperature modelling using artificial neural networks: application on Catamaran Brook, New Brunswick, Canada. *Hydrological Processes, 22*(17), 3361–3372. http://dx.doi.org/10.1002/hyp.6928.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory*. Hoboken, NJ, USA: John Wiley & Sons, Inc.. http://dx.doi.org/10.1002/047174882X.

Gregorčič, G., & Lightbody, G. (2009). Gaussian process approach for modelling of nonlinear systems. *Engineering Applications of Artificial Intelligence, 22*(4–5), 522–533. http://dx.doi.org/10.1016/j.engappai.2009.01.005.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The elements of statistical learning: data mining, inference and prediction* (second ed.). Springer.

Haykin, S. (1999). *Neural networks: a comprehensive foundation* (second ed.). Prentice hall.

Kothandaraman, V., & Evans, R. (1972). *Use of air–water relationships for predicting water temperature*. Urbana: Illinois State Water Survey.

Maier, H. R., & Dandy, G. C. (2000). Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling & Software, 15*(1), 101–124. http://dx.doi.org/10.1016/S1364-8152(99)00007-9.

Mohseni, O., & Stefan, H. G. (1999). Stream temperature/air temperature relationship: a physical interpretation. *Journal of Hydrology, 218*(3–4), 128–141. http://dx.doi.org/10.1016/S0022-1694(99)00034-7.

Petelin, D., Grancharova, A., & Kocijan, J. (2013). Evolving Gaussian process models for prediction of ozone concentration in the air. *Simulation Modelling Practice and Theory, 33*, 68–80. http://dx.doi.org/10.1016/j.simpat.2012.04.005.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes in machine learning*. MIT Press.

Sahoo, G. B., Schladow, S. G., & Reuter, J. E. (2009). Forecasting stream water temperature using regression analysis, artificial neural network, and chaotic non-linear dynamic models. *Journal of Hydrology, 378*(3–4), 325–342. http://dx.doi.org/10.1016/j.jhydrol.2009.09.037.

St-Hilaire, A., Ouarda, T. B. M. J., Bargaoui, Z., Daigle, A., & Bilodeau, L. (2012). Daily river water temperature forecast model with a k-nearest neighbour approach. *Hydrological Processes, 26*(9), 1302–1310. http://dx.doi.org/10.1002/hyp.8216.

Vanhatalo, J., Veneranta, L., & Hudd, R. (2012). Species distribution modeling with Gaussian processes: a case study with the youngest stages of sea spawning whitefish (Coregonus lavaretus L. s.l.) larvae. *Ecological Modelling, 228*, 49–58. http://dx.doi.org/10.1016/j.ecolmodel.2011.12.025.

Webb, B. W., Hannah, D. M., Moore, R. D., Brown, L. E., & Nobilis, F. (2008). Recent advances in stream and river temperature research. *Hydrological Processes, 22*(7), 902–918. http://dx.doi.org/10.1002/hyp.6994.

Webb, B. W., & Nobilis, F. (1997). Long-term perspective on the nature of the air–water temperature relationship: A case study. *Hydrological Processes, 11*(2), 137–147. http://dx.doi.org/10.1002/(SICI)1099-1085(199702)11:2<137::AID-HYP405>3.0.CO;2-2.

Wu, Q., Law, R., & Xu, X. (2012). A sparse Gaussian process regression model for tourism demand forecasting in Hong Kong. *Expert Systems with Applications, 39*(5), 4769–4774. http://dx.doi.org/10.1016/j.eswa.2011.09.159.