water

MDPI

Article

Applicability of a Nu-Support Vector Regression Model for the Completion of Missing Data in Hydrological Time Series

Jakub Langhammer * and Julius Česák

Department of Physical Geography and Geoecology, Faculty of Science, Charles University in Prague, Albertov 6, Praha 2 128 43, Prague, Czech Republic; julius.cesak@natur.cuni.cz

* Correspondence: jakub.langhammer@natur.cuni.cz; Tel.: +420-221-951-415

Academic Editor: Marco Franchini

Received: 7 October 2016; Accepted: 24 November 2016; Published: 30 November 2016

Abstract: This paper analyzes the potential of a nu-support vector regression (nu-SVR) model for the reconstruction of missing data of hydrological time series from a sensor network. Sensor networks are currently experiencing rapid growth of applications in experimental research and monitoring and provide an opportunity to study the dynamics of hydrological processes in previously ungauged or remote areas. Due to physical vulnerability or limited maintenance, networks are prone to data outages, which can devaluate the unique data sources. This paper analyzes the potential of a nu-SVR model to simulate water levels in a network of sensors in four nested experimental catchments in a mid-latitude montane environment. The model was applied to a range of typical runoff situations, including a single event storm, multi-peak flood event, snowmelt, rain on snow and a low flow period. The simulations based on daily values proved the high efficiency of the nu-SVR modeling approach to simulate the hydrological processes in a network of monitoring stations. The model proved its ability to reliably reconstruct and simulate typical runoff situations, including complex events, such as rain on snow or flooding from recurrent regional rain. The worst model performance was observed at low flow periods and for single peak flows, especially in the high-altitude catchments.

Keywords: data-driven model; SVR; runoff; precipitation; snowmelt; sensor network

1. Introduction

Among the technologies for monitoring the dynamics of runoff processes, automated sensor networks have played an increasingly important role in experimental research and water management practices [1]. The automated monitoring of surface and groundwater runoff processes allows the acquisition of data with high levels of frequency and accuracy [2,3]. This type of data enables research into the highly dynamic processes in catchments with unprecedented level of detail and provides deeper insight into the mechanisms of runoff generation. The coupling of monitoring devices with communication modules using Global System for Mobile communications (GSM) or satellite telemetry enables online access to the observed data in the near real-time regime [4].

The large amounts of data from automated monitoring network sensors represent new opportunities for research and new challenges for data analysis and results in higher vulnerability of monitoring systems to the occurrence of issues related to data quality [5,6].

Because modern automated sensors and stations are complicated sets of electronic devices that are exposed to extreme environmental conditions, the monitoring systems are vulnerable to the occurrence of data outages or quality problems. The data outages typically occur in consequence of damage to the sensors or control stations related to extreme weather, e.g., periods of extreme cold, electrical shock after lightning, physical damage after flooding, mud flows or freefall, and energy shortages.

In stations without online access to data or with limited physical accessibility, e.g., due to climatic conditions over winter season or restrictions of access for nature conservation reasons, outages in data recording or breaks in energy supply are detected with delays, which can result in data gaps, devaluing the time series of monitoring. For such cases, the ability to reconstruct the incomplete time series is of vital importance.

Manifold approaches for the reconstruction of missing data in hydrological time series have been developed, ranging from conventional statistical methods to physical-based modeling. The stochastic nature of hydroclimatic processes, the complexity of relations among the meteorological drivers, the state of the environment and the runoff response make the use of the conventional hydrological models complicated for practical applications [7–9].

Because monitoring sensors are often organized into networks, where the stations observe interdependent processes, there is a potential for the use of data-driven models, such as artificial neural networks (ANN) and support vector machines (SVM), to simulate the observed processes [10]. These models can be used to complete missing data in the monitoring network. Progress in the development of machine learning algorithms and data-driven models with the availability of high-performance computing enabled the growing application of machine learning techniques, such as ANN, SVM and fuzzy logic [11,12]. The application of machine learning approaches to fill the gaps in time series hydrological data is beneficial, especially in areas with limited quality or availability of supporting spatial and qualitative data, which are necessary for conventional hydrological models [8].

This study aims to test the ability of the SVM model to fill in the missing data in time series hydrological data resulting from an automated sensor network, operating in a set of experimental montane watersheds. The SVM approach was selected as a proven robust and reliable technique that is suitable for modeling continuous data series. However, SVM is not frequently used in hydrological research compared to other models, e.g., ANN.

The particular goals of the study are: (i) to design a self-learning network model based on SVM and suitable configuration of the input variables to enable reconstruction of missing hydrological data from an automated sensor network; (ii) to test the SVM model performance in conditions of variable physiographic conditions; and (iii) to test the SVM model applicability and performance on select runoff scenarios, including storm flows, flood events, snowmelt events and periods of drought.

The study is based on data from five years of continuous hydrological monitoring in a network of experimental catchments in the Sumava Mountains, featuring automated water level and weather stations operated by Charles University in Prague [3]. The Konstanz Information Miner (KNIME) computing framework with the Library for Support Vector Machines (LIBSVM) module was used for the modeling and statistical treatment of the data.

2. Materials and Methods

2.1. Study Area

The study area of the Roklanský Brook and the experimental subcatchments is located at the headwaters of the Sumava Mountains, Central Europe (Figure 1), and features an elevated montane plain with moderate hillslopes [13]. The area is homogeneous in terms of physiography. The upper part of the montane range is developed on the crystalline core of the Bohemian Massif. The bedrock is mostly composed of metamorphic gneiss and schists with intrusions of granite; Entic Podzols, Histosols, and Gleysols are the common soil types [14]. The region features a typical mid-latitude montane climate, with an average annual precipitation of 1370 mm and a mean air temperature of 3.6 °C [3].

The Roklanský Brook basin (RKM) has an area of 47.8 sq km, with the outlet at Modrava Village at an altitude of 978 m a.s.l. The average discharge at the RKM monitoring station is 1.66 $m^3 \cdot s^{-1}$ [13]. For the purpose of this study, this station was supplemented by the data from the water level monitoring stations at the basin headwaters: Roklanský Brook at Hajenka (HAJ), Rokytka (ROK)

and Javoří Brook (JAV), plus the information from the weather station at Rokytka (ROK) and the weather and snow monitoring station at Ptaci Brook (PTA).

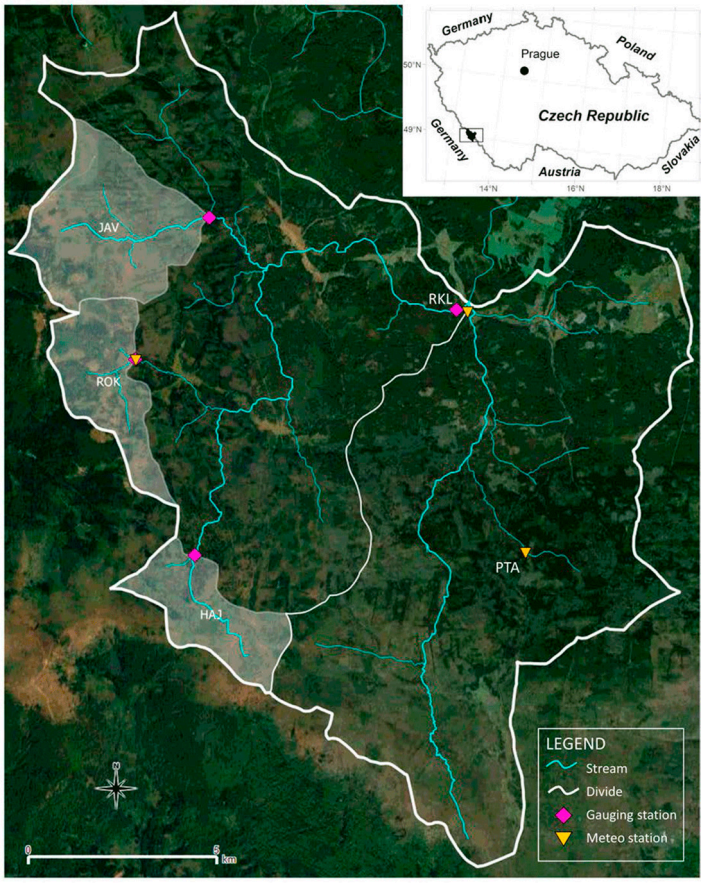

Figure 1. Study area. Upper Vydra basin with experimental catchments and automated sensor network stations.

Although the three experimental catchments (HAJ, ROK, and JAV) representing the headwater zone of the Roklanský Brook basin are in the same vicinity, their particular physiographic properties and environmental status vary. The most elevated catchment (HAJ) is the smallest, has the least dense river network and the highest share of the forest, which has been damaged by bark beetle infestation since the 1990s (Table 1). In contrast, the lowest altitude catchment (JAV) is the largest, with the longest river network and the lowest share of damaged forest. Such differences in basic physiography and in the land cover and vegetation cover quality create environments with different runoff responses.

Table 1. Physiographic properties of the experimental headwater catchments. Data: Czech Hydrometeorological Institute (CHMI), Czech Geological Service (CGS), Forest Management Institute (FMI) and Charles University (CUNI).

Catchment	JAV	ROK	HAJ
Catchment area (sq km)	6.33	3.60	3.321
Mean altitude (m a.s.l)	1117 ± 69	1125 ± 48	1233 ± 70
River network length (km)	16.55	12.36	7.53
Drainage density	2.61	3.43	2.27
Bedrock—share of granite %	64.9%	31.9%	55.8%
Bedrock—share of sedimentary material %	26.8%	45.6%	8.9%
Soils—share of peat land %	8.3%	22.5%	35.3%
Land cover—share of decayed forest %	37.0%	48.1%	68.2%

2.2. Sensor Network

The sensor network employed for this study consists of high-frequency monitoring of water levels and basic hydrochemistry parameters (water temperature, electric conductivity, pH) at the catchment outlets performed by the weather stations (Figure 1). The monitoring network in the Upper Vydra basin was established in 2006 by Charles University.

The water levels at the monitoring stations are observed using two major technologies—determination of the water level based on changes in hydrostatic pressure (HAJ station) and measurement of the water level by ultrasonic beam (ROK, JAV, and RKM stations).

The measurement based on changes in hydrostatic pressure uses a sensor fixed at the river bottom to detect changes in the water column height (Fiedler AMS TSH22, Figure 2a). This approach is beneficial for small and unregulated streams with no structures, where a direct measurement sensor can be attached. The pressure probes are beneficial due to the ability to capture a wide range of water levels. They do not demand a large amount of energy and are reliable; however, they are vulnerable during floods and elevated material transport. The ultrasonic gauge is based on direct measurement of changes in water level from a given vertical distance using an ultrasonic beam (Fiedler AMS US3200, Figure 2b). The sensor is fixed to a structure above the water level, typically a bridge. Ultrasonic beam measurement offers the ultra-high precision, reaching even sub-millimeter values.

(**a**) (**b**)

Figure 2. Water level sensors: (**a**) Hydrostatic pressure gauge fixed to the bottom of a montane creek; and (**b**) ultrasonic water level sensor fixed to the top of a dam culvert. Photo by Jakub Langhammer.

The monitoring frequency is set to 10-min intervals, with daily automated transmission of the monitoring data to cloud-based storage (Fiedler M4016). The water level data are then converted to streamflow using rating curves established by hydrometric measurements using an acoustic Doppler velocimeter (Flow Tracer, SonTek Inc., San Diego, CA, USA) over the study period. For this study, data were collected for the period covering 5 hydrological years from 1 November 2010 to 30 October 2015.

2.3. Data-Driven Models for the Simulation of Hydrological Processes

Machine learning is a rapidly evolving field of data-driven modeling techniques with the extensive potential for application in hydrology. Among the various machine learning approaches, ANNs, Bayesian networks (BN) and support vector machines (SVMs) are the most frequently applied techniques [15,16]. These approaches are algorithms with significant importance for processes where the application of conventional models is complicated by the complexity and unknown conditions affecting the process or by changing environmental conditions. In hydrological research, their ability to learn from the patterns of the input data is beneficial, especially for predictions in ungauged basins, for the development of hydrological models in complex physiographic conditions and as a tool for reconstructing incomplete meteorological and hydrological data.

Despite their similarities, the principles of the three approaches differ significantly in nature and underlying mathematical principles. The design of ANN models was inspired by the structures of biological neural networks. The computation networks are structured in terms of an interconnected group of artificial neurons that process information using a connectionist approach to computation [17–19]. In hydrological research, ANN models are typically used to simulate the complex relationships between inputs and outputs, to find patterns in data or to find parameters for conventional models in complicated environmental conditions [20].

BN models are based on different principles, stemming from a probabilistic graphical model that represents a set of random variables and their conditional independencies via a directed acyclic graph [21,22]. BN models have proven to be an efficient tool for different applications in hydrology, e.g., for probabilistic hydrological forecasting as an approach to quantify model uncertainty [22] and to support decision-making process in environmental applications [23].

SVMs, developed by Vapnik in the 1960s [24,25], are a set of related supervised learning methods that are used for classification and regression. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that predicts which category a new example falls into [26]. SVM classifiers became widely popular in recent decades in scientific, industrial and real-world applications, including i.e., Optical Character Recognition, financial modeling, medical imaging [27–29] due to its robustness and availability of computing environments that could efficiently handle the demanding learning phase of classification [30].

A complex overview of recent hydrological applications of SVMs was given by Raghavendra and Deka [31]. SVMs were applied in rainfall–runoff modeling studies across different environments, temporal and spatial scales. Rainfall–runoff modeling at longer time scales using SVMs was demonstrated, e.g., by Lin et al. [32], while Yu et al. [33] have applied SVM model in real-time flood forecasting. Granata et al. [34] have tested the performance of support vector regression (SVR) model for rainfall–runoff simulations in an urban environment. Shahraini et al. [15] have compared the SVM runoff model in a large scale complex basin. There was found a lack of studies focused on SVM applications in small montane catchments, comparable to the sites, examined by our study. The robustness of SVM modeling approach, proven by the mentioned studies, however, indicates the suitability of application of this modeling technique to such conditions [35].

2.4. Support Vector Machines

SVM models represent still a relatively new concept, although the mathematical foundations of the method were set in the 1960s, similar to the concept of neural networks [28]. SVM is a supervised learning method based on a set of training examples that builds a model that can assign classified samples into separate categories. SVM is a non-probabilistic classifier based on the separation of examples into distinct classes by defining a hyperplane that separates the categories.

The key principle of SVM classification is the conversion of the original input space to another, with higher dimensionality [36]. Augmenting dimensionality of the data space enables to find a linear solution for separating the data which does not exist in the original data space [37]. This mechanism can be demonstrated in an example of two-dimensional space with binary data, containing e.g., positive and negative training samples, whose distribution does not allow linear separation of the classes (Figure 3a). The initial two-dimensional space, where each vector is defined by two attributes, can be remapped into three-dimensional space, by adding the third attribute, based on the transformation of the initial ones. Such transformation allows shifting the data points along the new axis so we can find a plane, separating the data classes (Figure 3b). The transformation, here represented by simple exponentiation, is in the SVM model performed by the kernel, based on different principles—linear, polynomial, radial basis or sigmoid [38]. This example illustrates a general principle of the SVMs—it is assumed that when the data are mapped into space with a sufficient number of dimensions, we should be able to find a linear plane, separating the samples [36].

Location of the separating hyperplane is based on the position of points, defining the bounds of the plane—the support vectors (Figure 3c). Identification of these support vectors significantly simplifies the classification. This is because for identification of the separating plane it is not necessary to take into account all of the data points but just the support vectors (Figure 3c). This principle significantly improves the performance of the classification, especially on large datasets (citation). Finding of the optimal solution is then aimed at maximizing the margin around the separating hyperplane [25].

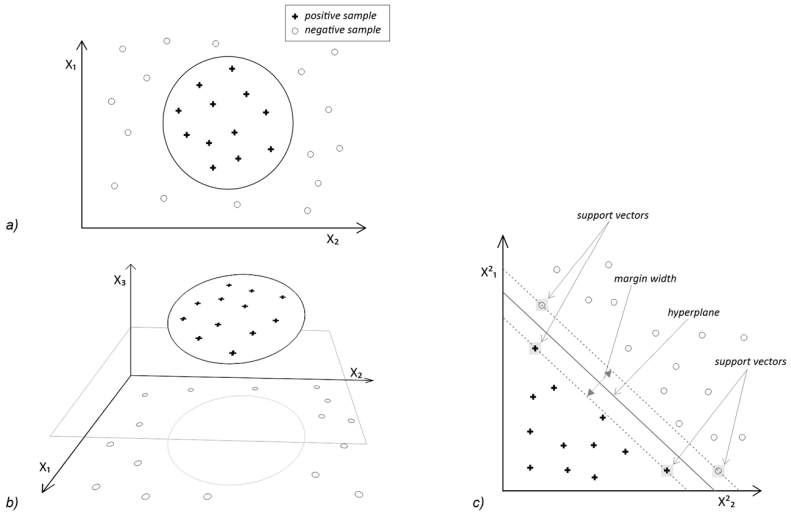

Figure 3. Principle of finding the linear solution of classification by augmenting the dimensionality: (**a**) sample of training data with positive and negative values, where linear separation is not possible; (**b**) data space with augmented dimensionality, enabling to find a plane, separating the data; and (**c**) separating hyperplane with margins, defined by the support vectors.

The generic Vapnik's concept of support vector machines is applicable for classification purposes as support vector classification (SVC) as well as for solving regression tasks as support vector regression (SVR) [39,40]. SVR keeps the basic SVM idea to map the data into a high dimensional feature space via a nonlinear mapping and to do linear regression in this space. The kernel functions transform the data into a higher dimensional feature space that makes it possible to perform the linear separation. Thus, the linear regression in a high dimensional space corresponds to nonlinear regression in the low-dimensional input space [39].

The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with just some differences. The main idea remains the same: to minimize error, individualizing the hyperplane which maximizes the margin, keeping in mind that part of the error is tolerated. Because the output is a real number it becomes very difficult to predict the information. Hence, a margin of tolerance epsilon is set in approximation to the SVM.

There are two basic versions of SVM regression, epsilon-SVR and nu-SVR, which differ in the way the algorithm handles with margin control and penalty parameter. In epsilon-SVR, there is no control on how many data vectors from the dataset become support vectors unlike in nu-SVR. The algorithm has a control on how much error is allowed in the model by the regularization parameter C. The main motivation for the nu versions of SVM is that it has a more meaningful interpretation. This is because nu represents an upper bound on the fraction of training samples which are errors (badly predicted) and a lower bound on the fraction of samples which are support vectors.

Schölkopf et al. [41] proposed an SVR algorithm, called nu-SVR, adjusting automatically the parameter epsilon. The nu-SVR introduces a parameter nu, enabling to control the number of support vectors, by setting the proportion of the number of support vectors kept in the solution with respect to the total number of samples in the dataset. The epsilon parameter, defining the margin of tolerance is here introduced into the optimization problem formulation and it is estimated automatically [42].

The use of kernel functions make the SVR applicable to both linear and non-linear approximations, while it features good generalization performance as a result of the use of only the support vectors for prediction the absence of local minima because of the convexity property of the objective function and its constraints, and the fact that the methodology is based on structural risk minimization that seeks to minimize the generalization rather than the training error [42]. It makes SVR suitable for use in time series analysis and forecasting with manifold applications, e.g., in geosciences or financial forecasting [43].

2.5. Application of the SVM Model to the Study Area

2.5.1. Applied Computing Environment

LIBSVM [38] was used as the environment for model learning and forecasting. LIBSVM comprises several types of SVM models, enabling classification of discrete as well as continuous data: (i) algorithms for support vector classification (C-SVC and nu-SVC), which are applicable to the classification of discrete datasets; (ii) algorithms for support vector regression (epsilon-SVR and nu-SVR), enabling the construction of SVM models for continuous data; and (iii) distribution estimation (one-class SVM) [38].

The nu-SVR model, which is applicable for modeling continuous time series and simulations, was selected for this study. Based on experience from applications in different disciplines, the nu-SVR model was found to be robust and reliable, even for simulations based on a small training sample or data burdened by noise [44]. This ability is especially important for hydrological data originating from sensor networks, where noise occurs for various reasons [4], and for simulations of highly dynamic events of short duration, where the number of training samples is limited by the nature of the process.

The KNIME Analytics Platform 3.12 was applied as the framework for the experiment. KNIME is a computing platform that integrates a set of data mining, statistical, machine learning and plotting modules to design the model workflow [45,46]. The KNIME platform is based on the visual design of the model structure, combining key parts of the model—the source data input and pretreatment, the definition of datasets for training, validation and simulation scenarios, model learner, model predictor and post-processing (Figure 4).

The model in the KNIME platform is composed of a set of nodes, performing different types of operations—input/output, data transformation, calculation, statistics, plotting or data output. Each node has its properties, defining the action performed, used data and parameters. The nodes are interlinked into a functional workflow, which can be run in step-by-step basis from individual nodes or as a batch. The workflow, defined for this study comprised the following key structures: (i) data input node (Figure 4a), uploading the database with records from the sensor network, stored as plain text; and (ii) block of nodes, performing selection and pretreatment of the data. In particular, is performed a selection of data columns with parameters, used further in the model, exclusion of the data rows with missing observations or incorrect values, setting of the bounds in time series for analysis (Figure 4b). From these data, the cross-correlation analysis was performed (Figure 4c). From the selected dataset, there were selected time periods, used for model training (Figure 4d) and validation (Figure 4e). The node with training data is connected with the LIBSVM learner modules, set for each of the four catchments (Figure 4f). The learner uses the whole input data matrix with the target column, set to the water level values of the given station. The learner node parameters for all of the four catchments were set identically.

The trained model network for each catchment is then applied for calculating predictions, using the input data from the validation period (Figure 4e) or simulation scenarios SC1–SC6 (Figure 4g, showing network configuration of the model for JAV catchment and SC3 scenario). The values, simulated by the LIBSVM predictor (Figure 4h) are then post-processed by the nodes, performing plotting of values, R^2 statistics and data output (Figure 4i).

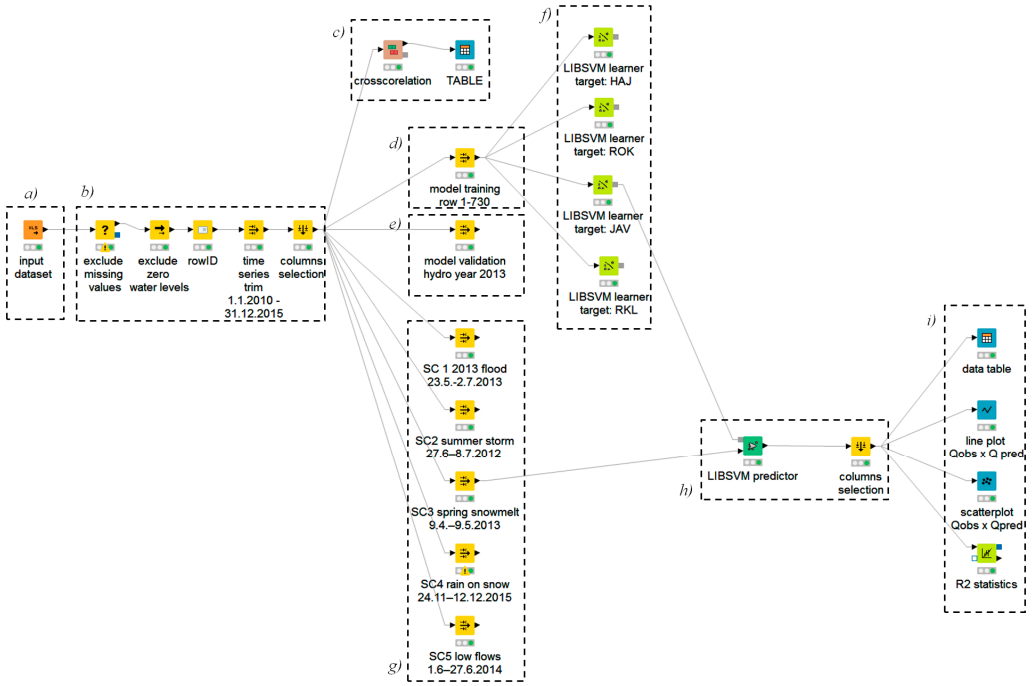

Figure 4. Structure and components of the model as defined in the KNIME workbench. The annotations highlight the principal functional blocks of the model, based on nodes, linked into in the workflow: (**a**) input data; (**b**) data selection and pre-treatment; (**c**) cross-correlation analysis; (**d**) selection of data for model training period; (**e**) selection of data for model validation; (**f**) model learning modules for individual catchments; (**g**) selection of data for simulation scenarios; (**h**) model predictor; and (**i**) post-processing of simulation outputs.

2.5.2. Input Data

The data-driven model is based on the matrix of daily values of the meteorological and hydrological parameters available from the measurements of the sensor network at four experimental subcatchments (HAJ, ROK, JAV and RKM) in the Roklanský Brook basin (Figure 1). As the subcatchments at the basin headwaters have different physiography and environmental status, the model was developed and run for each catchment and all scenarios to identify the effects of different conditions.

For the meteorological drivers, the model applied the daily precipitation total (P), precipitation total for the preceding day (PD-1), the antecedent precipitation index calculated for time spans of 2.5 and 7 days (API2, API5, and API7), snow depth in a given day, (SNW), mean daily air temperature (T), minimum and maximum daily temperature (Tmin and Tmax, respectively) and mean air temperature on the two preceding days (TD-1 and TD-2).

For the runoff parameters, the model applied information on the daily mean water levels at the observed stations (H_HAJ, H_ROK, H_JAV, and H_RKM). This combination of potential input parameters that are relevant to explaining the hydrological variability was tested for cross-correlations to identify the dependencies among the input data and to reduce the input data matrix by redundant parameters, which significantly decreased the time for model learning. The linear correlation matrix was calculated using Pearson correlation coefficient, while the missing observations were excluded and only the complete records were taken into account. The KNIME cross-correlation node was used for calculation.

The cross-correlation matrix (Figure 5) indicates strong positive correlations among the parameters of air temperature and positive correlations among the API values calculated for different time spans. Therefore, it was possible to reduce these data without the risk of a significant loss of information.

From the correlation values for the water levels at the monitoring stations, there is a high level of similarity for the two neighboring catchments TMA and JAV; however, due to different physiography, the runoff process in these catchments cannot be considered identical. Both stations are therefore included in the model input. Based on these assumptions, a final set of parameters was selected and applied in the study: T, T-2, P, SNW, API2, API7, H_HAJ, H_ROK, H_JAV, and H_RKM.

	T PTA oC	T-1	T-2	T RKL	T JAV	T ROK	SNW_PTA	P PTA mm	P-1 PTA	API2 PTA	API5 PTA	API7 PTA	H_RKL mm	H_HAJ mm	H_ROK mm	H_JAV mm
T PTA oC	1.000	0.951	0.898	0.924	0.984	0.984	-0.561	0.098	0.071	0.092	0.134	0.152	0.053	0.169	0.228	0.219
T-1	0.951	1.000	0.951	0.897	0.948	0.940	-0.572	0.114	0.098	0.107	0.141	0.159	0.047	0.165	0.229	0.220
T-2	0.898	0.951	1.000	0.853	0.900	0.889	-0.568	0.099	0.114	0.134	0.155	0.170	0.042	0.164	0.228	0.222
T RKL	0.924	0.897	0.853	1.000	0.937	0.930	-0.588	0.116	0.091	0.113	0.158	0.178	0.034	0.177	0.201	0.151
T JAV	0.984	0.948	0.900	0.937	1.000	0.988	-0.583	0.108	0.081	0.102	0.143	0.161	0.059	0.171	0.230	0.194
T ROK	0.984	0.940	0.889	0.930	0.988	1.000	-0.577	0.082	0.065	0.085	0.125	0.144	0.050	0.174	0.229	0.192
SNW_PTA	-0.561	-0.572	-0.568	-0.588	-0.583	-0.577	1.000	-0.068	-0.073	-0.093	-0.120	-0.127	0.143	-0.096	-0.139	-0.104
P PTA mm	0.098	0.114	0.099	0.116	0.108	0.082	-0.068	1.000	0.245	0.222	0.207	0.195	-0.030	-0.025	0.003	0.141
P-1 PTA	0.071	0.098	0.114	0.091	0.081	0.065	-0.073	0.245	1.000	0.806	0.576	0.523	-0.049	-0.035	-0.007	0.134
API2 PTA	0.092	0.107	0.134	0.113	0.102	0.085	-0.093	0.222	0.806	1.000	0.757	0.687	-0.060	-0.044	-0.009	0.169
API5 PTA	0.134	0.141	0.155	0.158	0.143	0.125	-0.120	0.207	0.576	0.757	1.000	0.932	-0.068	-0.061	0.003	0.239
API7 PTA	0.152	0.159	0.170	0.178	0.161	0.144	-0.127	0.195	0.523	0.687	0.932	1.000	-0.079	-0.069	0.001	0.265
H_RKL mm	0.053	0.047	0.042	0.034	0.059	0.050	0.143	-0.030	-0.049	-0.060	-0.068	-0.079	1.000	0.414	0.516	0.164
H_HAJ mm	0.169	0.165	0.164	0.177	0.171	0.174	-0.096	-0.025	-0.035	-0.044	-0.061	-0.069	0.414	1.000	0.522	0.134
H_ROK mm	0.228	0.229	0.228	0.201	0.230	0.229	-0.139	0.003	-0.007	-0.009	0.003	0.001	0.516	0.522	1.000	0.292
H_JAV mm	0.219	0.220	0.222	0.151	0.194	0.192	-0.104	0.141	0.134	0.169	0.239	0.265	0.164	0.134	0.292	1.000

Figure 5. Cross-correlation of parameters.

2.5.3. Model Setup and Learning

The model was developed up in the KNIME computing environment using the LIBSVM library.

The key parts of the model network are represented by the following computing blocks: (i) source data selection; (ii) definition of modeling scenarios; (iii) SVM network learner; (iv) SVM predictor for selected scenarios; and (v) post-processing of simulated values (Figure 4).

The source data were pretreated in terms of database consistency and by removing the database rows with missing observations. The model scenarios were defined according to the types of runoff situations described above: long rain, storm on a saturated catchment, storm on a dry catchment, snowmelt and dry periods.

The learner block of modules (Figure 3) consisted of two nodes: the selection of training data and the core node with the SVM model learner. The learning period of the model consisted of the first two hydrological years of the monitoring period (1 November 2010–31 October 2012), covering approximately 40% of the dataset. The simulation events were selected from the rest of the time series to not overlap with the model training data.

The model learner was consecutively set up for all four stations: HAJ, ROK, JAV and RKM. The nu-SVM model with linear kernel was selected, enabling continuous simulation.

The choice of the kernel type in the LIBSVM (linear, polynomial, radial basis function and sigmoid) was affected by the fact that the dataset, used for the study provided a large number of training samples. In such there is no necessity to map the data to a higher dimensional space as the non-linear mapping here does not improve the model performance, while significantly affects the requirements for computing power and time [47]. This assumption was confirmed by Bray and Han who tested the performance of different kernels in the LIBSVM toolkit for runoff modeling [48] and identified nu-SVR with the linear kernel as an optimal configuration in terms of learning capabilities. Hence the linear

kernel was used as a basic configuration in our study for model training. While the results of validation indicated satisfactory fitness of the predicted data to the observations (see Section 3.1), the nu-SVR with linear kernel was applied in model learner for the whole study. The default parameters of the algorithm, suggested by the LIBSVM framework, were applied for the validation of the model: degree parameter, 3; Cost, 1.0; Nu, 0.5; Loss-Epsilon, 0.1; and Epsilon, 0.001. These default values of parameters were also kept for the simulation of the selected scenarios.

The predictor module (Figure 4) was consecutively run for all of six scenarios at all five catchments. Hence, the modeling framework comprised 24 models.

The post-processing of results in the KNIME modeling framework consisted of line plots, scatterplots and robust statistical calculations of the observed and simulated values.

2.5.4. Model Validation, Simulation Scenarios

The model validation was performed on a complex hydrological year comprising all basic types of runoff events that typically occur in the study area. The hydrological year in the study area is considered from November to October to cover the period of the closed hydrological balance. The hydrological year of 2013, comprising the period of 1 November 2012–31 October 2013, was selected for model validation. The validation model was developed for all the analyzed catchments.

The model learner, calculated for the validation period, was further directly applied for simulations of the selected model scenarios and no further calibrations or parameter tweaking of the model learner was applied.

For the simulation scenarios, five basic types of the runoff situations occurring in the area were selected. The aim was to test the performance of the SVM model under varying runoff situations with different complexity and in variable physiographic conditions. The simulation scenarios were based on the events in the selected periods when all the monitoring stations had available reliable data covering the whole extent of the event in the assessed periods. The tested situations were following:

(i) Flooding from long-term and recurrent rain (SC1): This scenario was based on flooding in June 2013 resulting from recurrent, intense regional-scale precipitation. The flood magnitude reached the level of a 20-year flood in the study area and a 50–100 year flood in the consequent lowland streams. The simulation period for this event was 23 May–2 July 2013 to cover the whole span of the pre- and post-flood conditions.

(ii) Single peak storm (SC2): To analyze the model performance to simulate highly dynamic events, a typical summer storm with short duration, which frequently occur in the area, was considered. A single peak summer storm in June 2012, covering the period of 27 June–8 July 2012 was selected as the simulation event.

(iii) Snowmelt (SC3): The runoff response to the snowmelt processes was tested on the typical late spring snowmelt situation, driven primarily by the rise of air temperatures with no or unimportant precipitation. The runoff response to such snowmelt episode was tested for the period of 9 April–9 May 2013.

(iv) Rain on snow (SC4): The effect of snowmelt, driven and accelerated by intense precipitation, represents an event that is often difficult to simulate by conventional hydrological models due to its complexity. The performance of the SVM model for a rain on snow event was tested in the case of a 20-year flood in the late fall of 2015, resulting from the heavy precipitation on the newly formed snow cover in the period of 24 November–12 December 2015.

(v) Low flows (SC5): The runoff response to an extended period of drought was tested on a model situation from the early summer of 2014. The low flows resulted from a period of no precipitation for more than 20 consecutive days with above-average temperatures. The simulation period was 1 June–27 June 2014.

3. Results

For all four catchments, an SVR network was developed and trained based on the set of parameters described above. The simulation scenarios comprised the simulation of a regional flood from recurrent rain (SC1), single summer storm (SC2), spring snowmelt (SC3), rain on snow (SC4) and low flow period (SC5).

3.1. Model Validation

The hydrological year of 2013, consisting of the time period between 1 November 2012 and 31 October 2013, was used as the validation dataset to verify the model performance in the long-term time span representing complex conditions occurring during a typical year.

The model validation results indicate fairly good overall performance for a complex hydrological year for all catchments, with R^2 values ranging from 0.89 to 0.91 (Figure 6). Despite the overall good model performance, the performance is different in the individual catchments with underestimation or overestimation of different types of events. In case of the topmost catchment of HAJ, the model underestimated the spring peak flow and (Figure 6d).

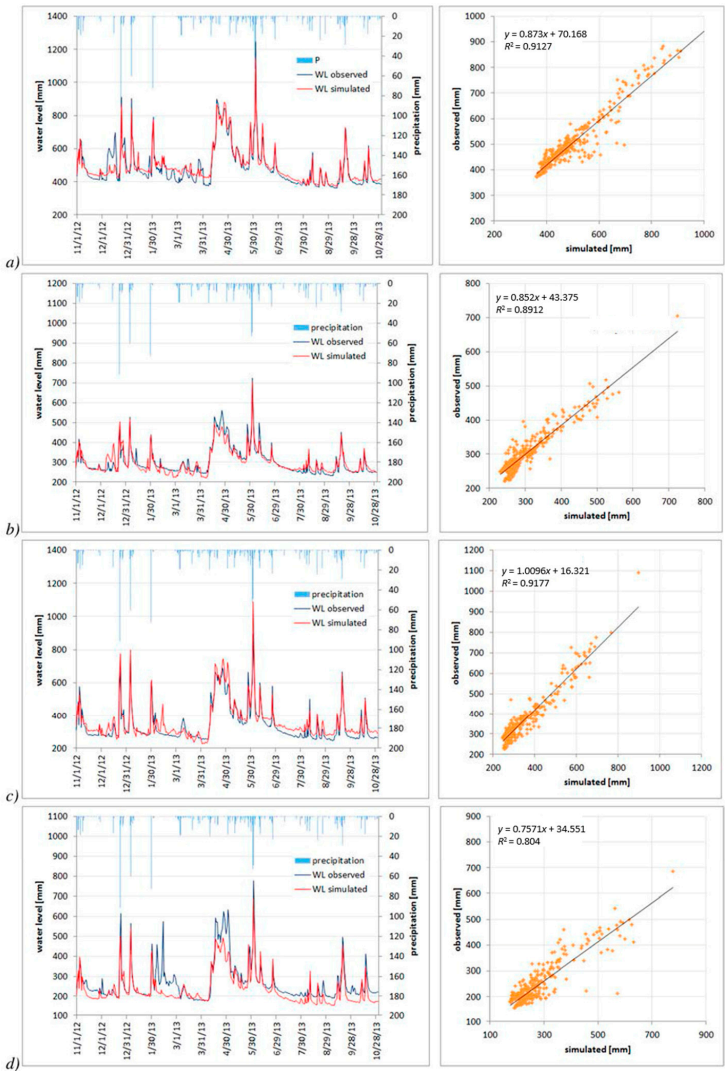

Figure 6. Model validation in a complex hydrological year, 1 November 2012–31 October 2013 at stations: (**a**) RKM; (**b**) JAV; (**c**) ROK; and (**d**) HAJ.

The validation for the basin outlet at Modrava (RKM) indicates a good fit to the observed values (R^2 = 0.913). During the snowmelt period, there is an apparent slight overestimation of the minor peak flows, however, the general trend is fitted well. The consequent waves of the spring and early summer flood events are simulated with very high reliability, including the low flow period during the fall season. The best fitness is achieved for the ROK catchment (R^2 = 0.918, Figure 6c) located at the middle range of altitudes of the basin. There is slight overestimation in the low flow period of the year, with a fairly good fit of the trend and all peak flows captured. The worst results (R^2 = 0.804) were achieved in the HAJ catchment located in the top part of the basin (Figure 6d). In all key periods, there is apparent underestimation of the values compared to the observed time series.

3.2. Simulation of the Typical Runoff Event Scenarios

3.2.1. Regional Flooding from Recurrent Precipitation

The regional flood in June 2013, resulting from heavy precipitation events with a 50-year flood magnitude, was selected as the case study to simulate a complex flood event.

The simulated period was 23 May–2 July 2013, when the precipitation-driven flood occurred in the period after late spring snowmelt. The flood was initiated by recurring precipitation, which resulted in three discharge peaks, hitting the area during a period of two weeks at the turn of June 2013.

The best fit of the simulation to the observed values was achieved at the outlet of the basin, at the Roklanský Brook at Modrava (RKM, R^2 = 0.96). Despite the slight overestimation of the values, the trend and timing of the events by the model fits the observed values (Figure 7a).

A very solid fit is observed at the Javoří Brook (JAV) station, located in the lower part of the basin (R^2 = 0.92, Figure 7b). Except for the third peak flow, where the simulation underestimated the real values and did not fit the shape of the wave, there is a very good fit of the trend and the timing. Similar performance was achieved at the ROK catchment (R^2 = 0.91, Figure 7c). The worst results are apparent at the HAJ catchment, located in the top part of the basin, where a time shift of the peak flow during the first event occurred. The major peak flow is simulated well; however, the third event was partially overestimated (Figure 7d).

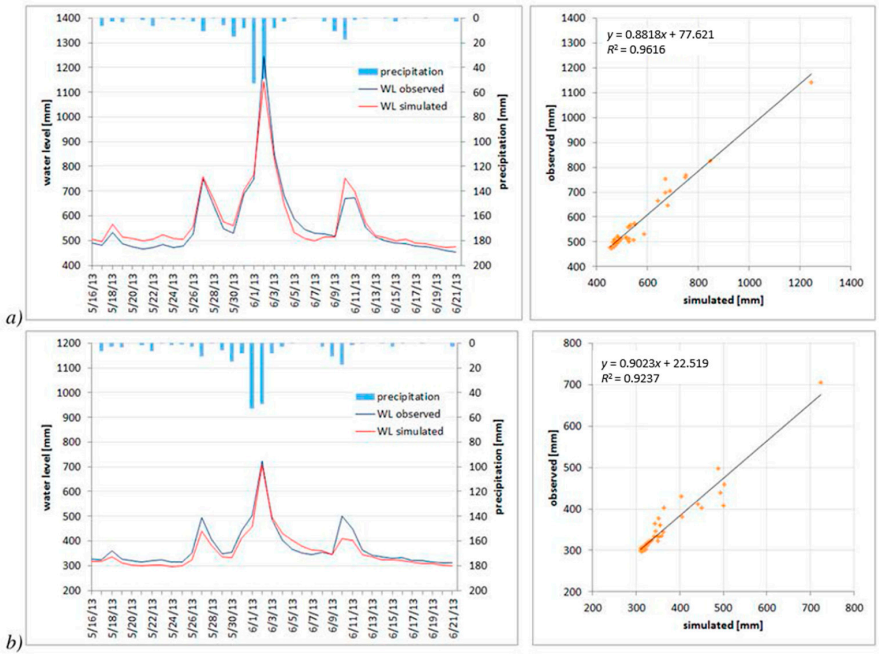

Figure 7. *Cont.*

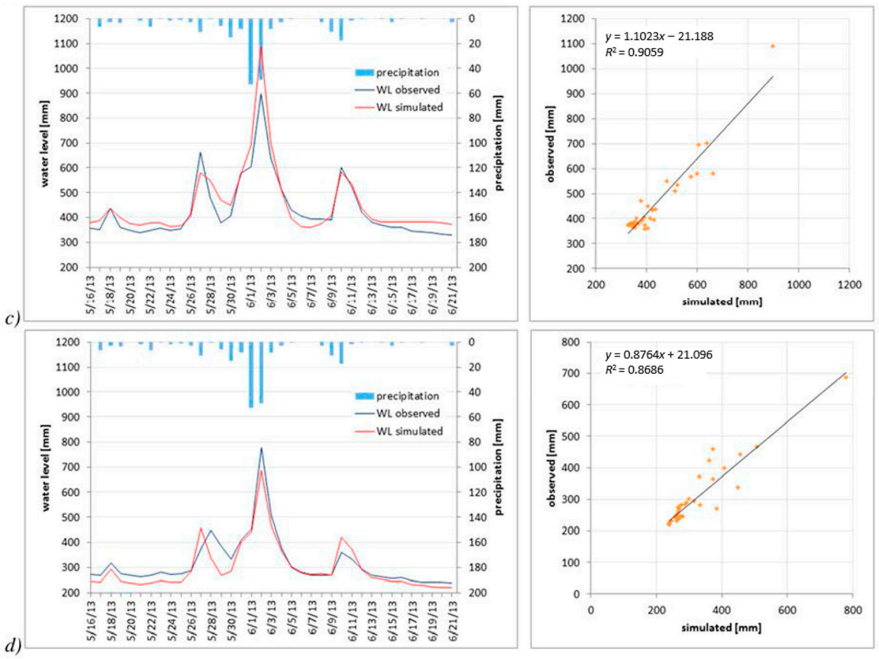

Figure 7. Simulation scenario SC1—Regional flood from recurrent heavy precipitation, 16 May–21 July 2013 at stations: (**a**) RKM; (**b**) JAV; (**c**) ROK; and (**d**) HAJ.

3.2.2. Summer Storm

This scenario represented a typical summer storm resulting from a single event occurring on an unsaturated basin. A storm at the beginning of July 2012 was selected as the model situation.

Despite the uncomplicated initial conditions triggering the runoff event, the simulation was reliable only in the lower part of the basin, while it was poor at the headwater catchment.

The outlet station of the Roklanský Brook basin at Modrava (RKM, R^2 = 0.96) displays a very good fit of the values, trend, peak flow and timing (Figure 8a). A good fit of the simulated to the observed values is reached in Javoří Brook (JAV) at the lower part of the basin (R^2 = 0.92), with only a slight overestimation of the values at the recession limb of the wave (Figure 8c).

However, the simulation in the mid-altitude catchment of Rokytka (ROK, R^2 = 0.85) and especially at the high-altitude catchment at Hajenka (HAJ), results in poor value distributions and misestimation of the simulated values (R^2 = 0.66). At ROK station, which drains a catchment with a large share of peat land, there were overestimated initial values and the initial stage of the rising limb. However, the peak flow is underestimated, with a reliable simulation of the recession limb (Figure 8c). At HAJ station, the model simulation resulted in greater differences in the values compared to the observations. The initial period and the recession limb are underestimated, while the peak flow is substantially overestimated, however, with proper timing (Figure 8d).

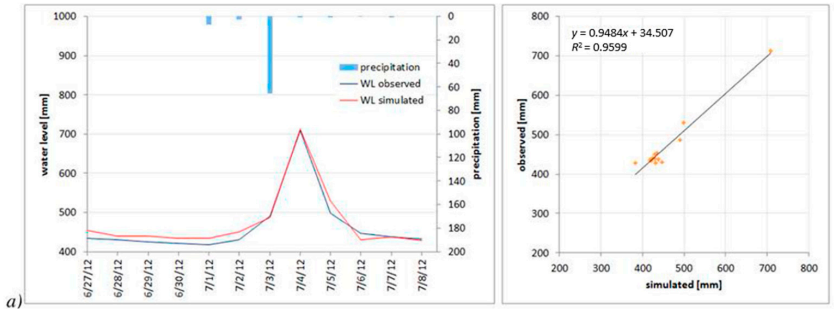

Figure 8. *Cont.*

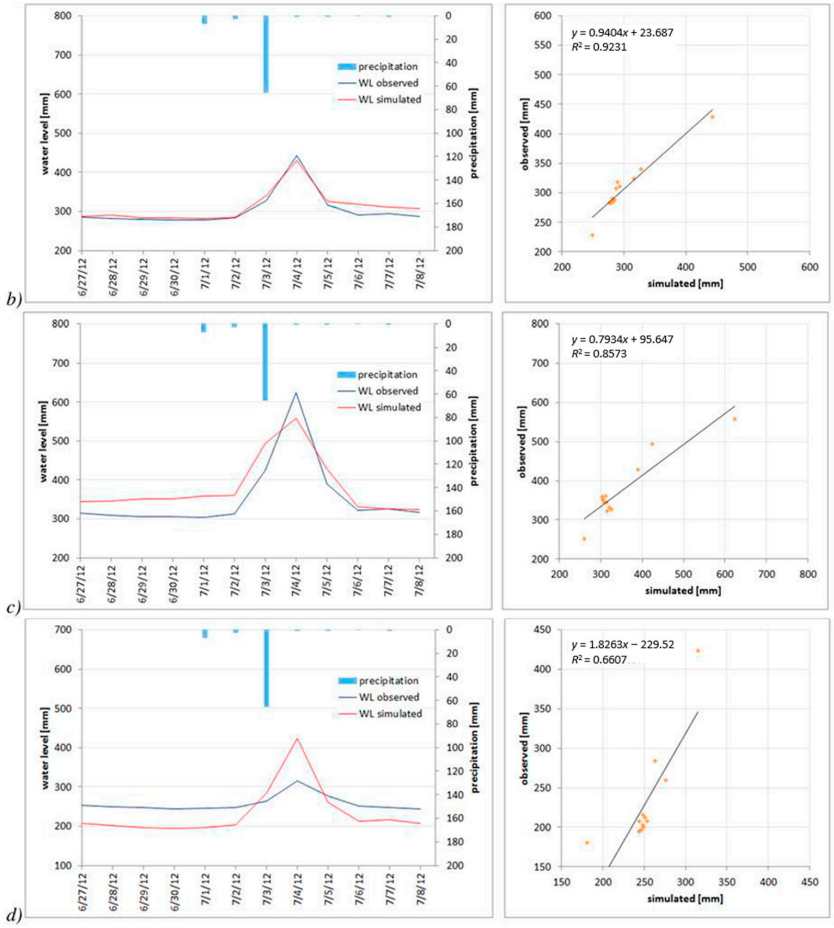

Figure 8. Simulation scenario SC2—single summer storm during 27 June–8 July 2012 at stations: (**a**) RKM; (**b**) JAV; (**c**) ROK; and (**d**) HAJ.

3.2.3. Spring Snowmelt

This scenario describes a high flow period after snowmelt in late spring during 9 April 2013–9 May 2013. In this period, there was only insignificant precipitation at the beginning and at the end of the observed period, which did not accelerate the snowmelt. The snowmelt event resulted in relatively flat peak flow curve, while the water levels remained elevated for more than two weeks.

The model performance for the snowmelt situation was very good, especially in the lower part of the basin. The fit of simulated to the observed values was good in both the basin outlet (RKM) and JAV station (R^2 = 0.96) (Figure 9). The trends, timing, and fit of the values were satisfactory in both cases, with a slight underestimation of values at JAV station in the second half of the period.

In the mid-altitude ROK catchment, the model performance was still very good (R^2 = 0.89) (Figure 9c). Compared to JAV station, the model slightly overestimated the peak values in the second half of the snowmelt period. However, the general fits of the trend, values, and timing of the events were reliable.

The least successful simulation was obtained in the topmost part of the basin at HAJ station (Figure 9d); however, the results were better than the rain storm events (R^2 = 0.84). Despite the accurate fit of the trend and timing of the events, the peak values in the decisive period of the event were significantly underestimated.

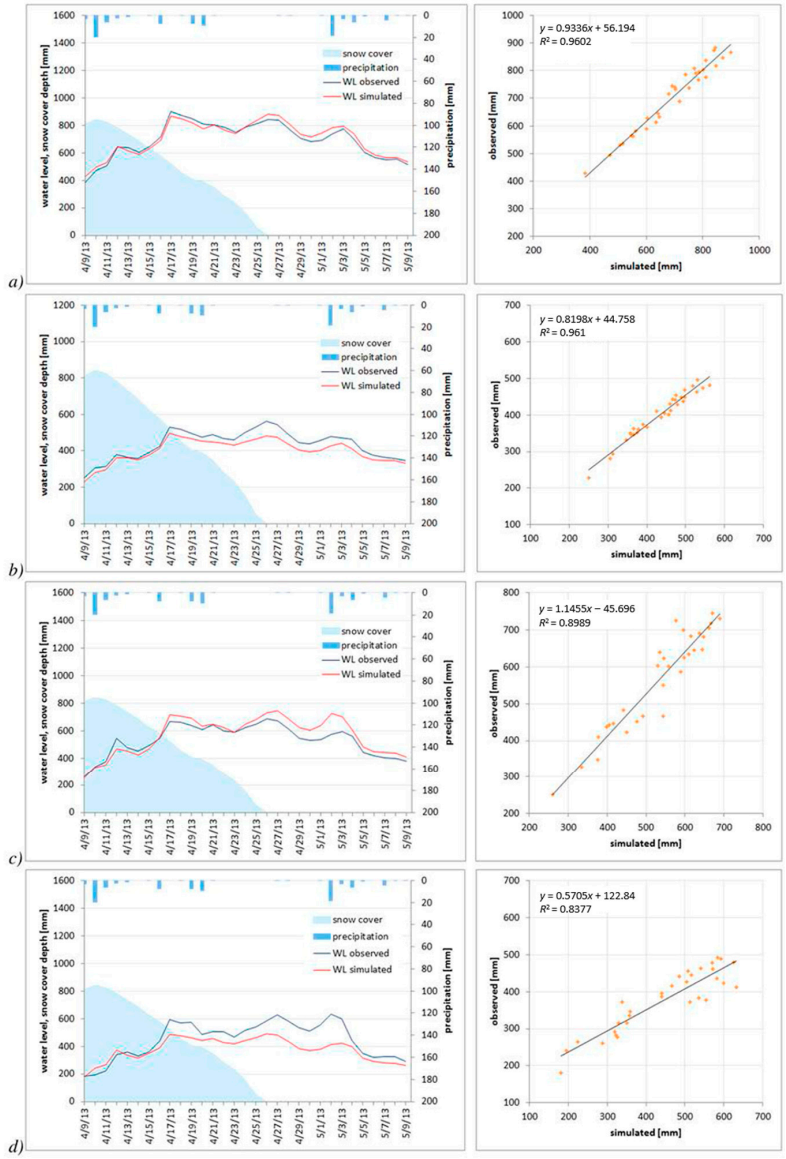

Figure 9. Simulation scenario SC3—snowmelt in late spring during 9 April–9 May 2013 at stations: (**a**) RKM; (**b**) JAV; (**c**) ROK; and (**d**) HAJ.

3.2.4. Rain on Snow

To simulate a rain on snow event, a 20-year flood period was selected that affected the study area in the beginning of December 2015. The flood event was driven by snowmelt, which was accelerated by intense precipitation that completely washed out the snowpack that had accumulated prior the event. The first compact snow cover of the upcoming winter season, reaching 20–30 cm of the snow depth, has melted in one day from 30 November–1 December 2015 and resulted in an intense contribution to the magnitude of the flood event.

The simulation of the rain on snow event by an SVM model showed surprisingly good reliability (Figure 10). In three of the four catchments, the R^2 values reached or exceeded 0.94. The only exception was the highest HAJ station, where the simulation performance was weak (R^2 = 0.62).

At the outlet of the basin (RKM, Figure 10a), the simulated values fit the observations in terms of the trend, timing and values (R^2 = 0.975) with only a slight overestimation of values in the recession limb of the flood hydrograph. The success rate of the simulation was even higher for the JAV catchment (R^2 = 0.986), where there were no substantial differences in the simulated values and the observations.

The fit of the simulation of the rain on snow event to the observations in the ROK catchment was also very good (R^2 = 0.948). The SVM model slightly overestimated the response of the catchment to recurrent precipitation in the recession phase of the runoff but with no substantial difference (Figure 9c).

The simulation in the HAJ catchment (Figure 10d) was, as in the previous scenarios, the least successful in terms of the goodness of fit of the simulated values to the observations (R^2 = 0.621). The key peak flow resulting from snowmelt was captured successfully in terms of values, trend, and timing. However, the response of the basin to the following recurrent precipitation, which was almost neglected by the modeling network, was heavily underestimated.

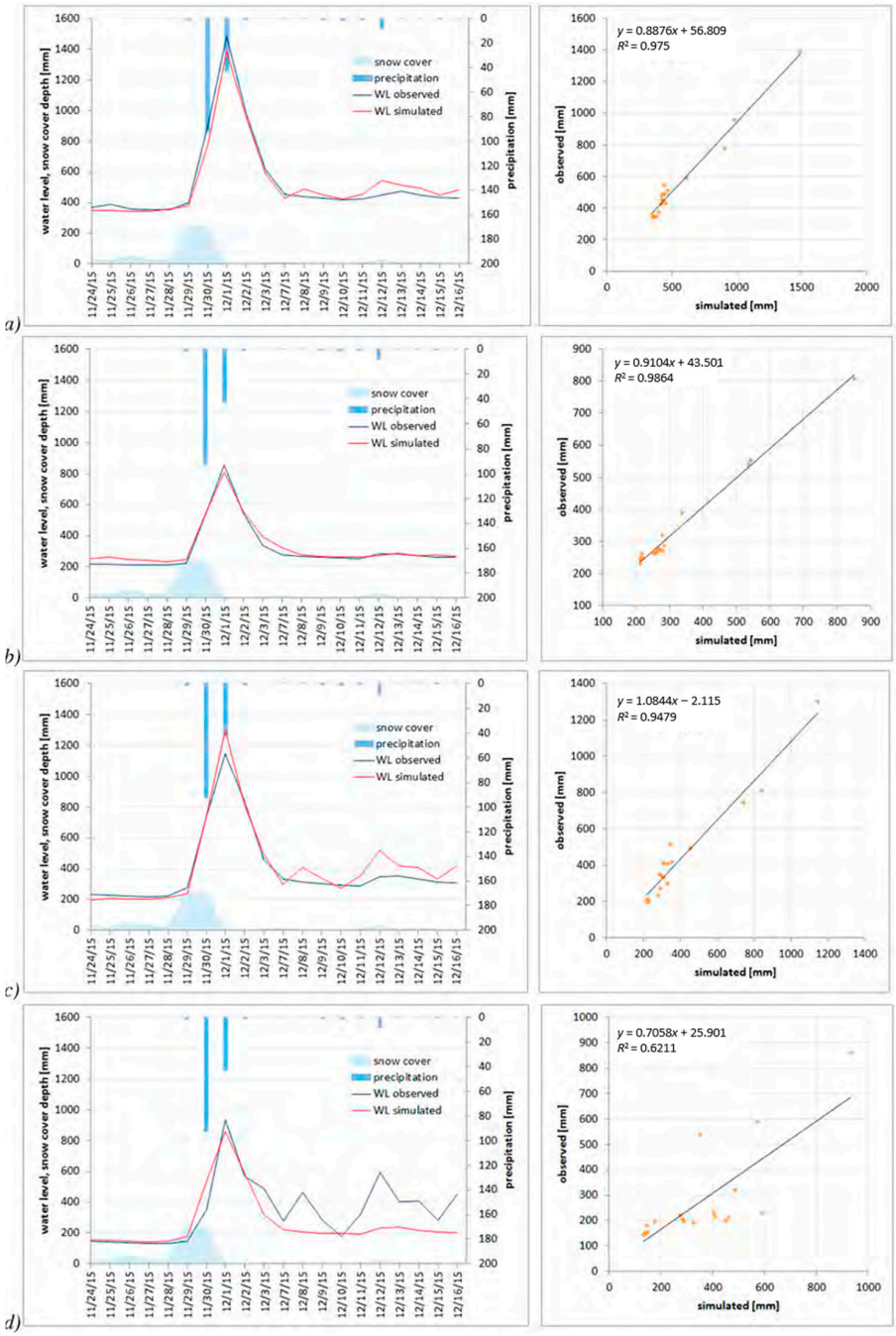

Figure 10. Simulation scenario SC4—rain on snow, 24 November–16 December 2015 at stations: (**a**) RKM; (**b**) JAV; (**c**) ROK; and (**d**) HAJ.

3.2.5. Low Flows

Unlike flood events, where the model can respond to the initial pulses from precipitation or snowmelt, the simulation of the low flow period is more complex because the hydrological regime in dry periods depends on a range of physiographic and environmental factors. The period selected for testing the SVM model performance (1 June–27 June 2014) represents a typical summer situation with an extensive period of almost no precipitation and above-average temperatures.

As in the preceding scenarios, the reliability of the model was very good for the stations located in the lower part of the basin: the outlet station RKM and JAV station. Surprisingly, the worst performance was observed for the model at ROK station (Figure 11c), which had the lowest R^2 values of the entire series of models.

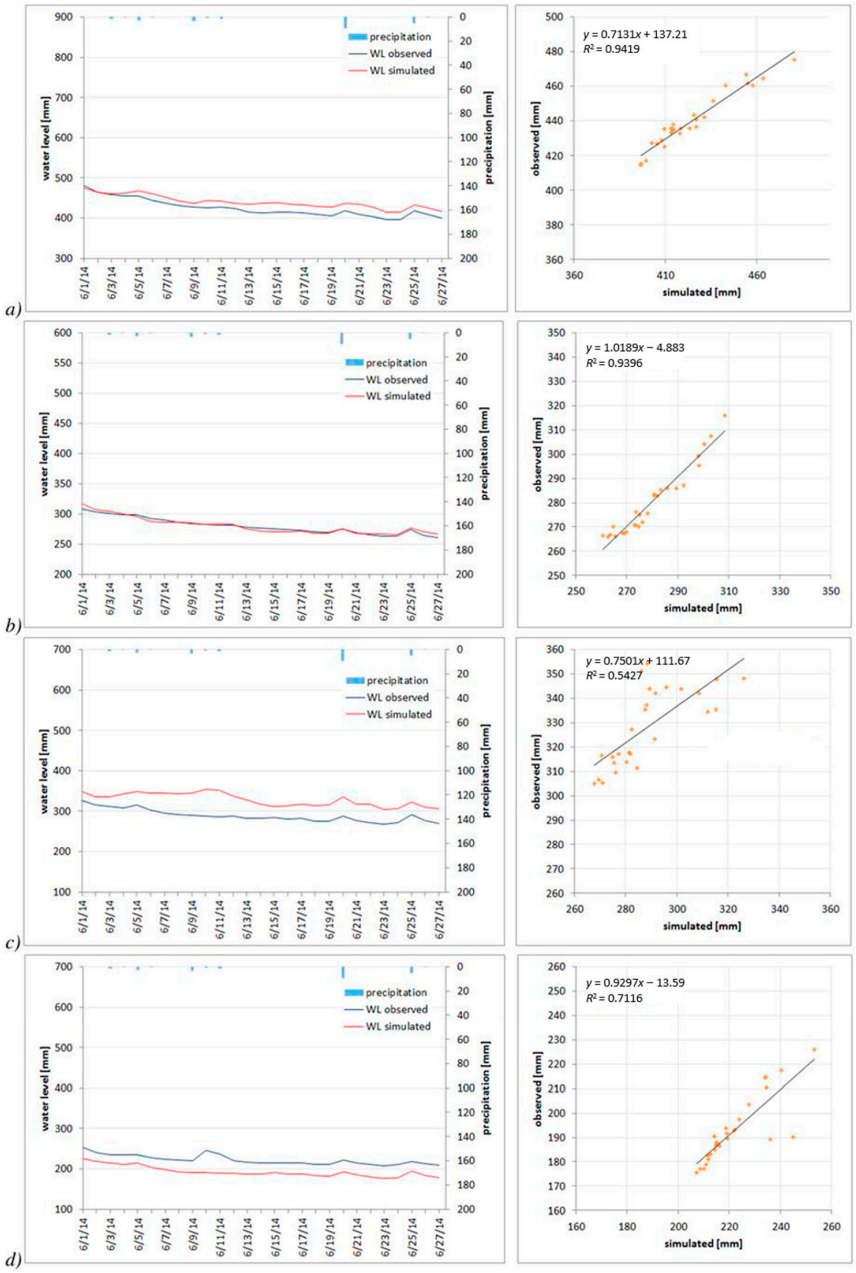

Figure 11. Simulation scenario SC5—low flow period of 1 June–27 June 2014 at stations: (**a**) RKM; (**b**) JAV; (**c**) ROK; and (**d**) HAJ.

At the basin outlet at RKM, the goodness of fit was fairly good (R^2 = 0.94) with a well-fitted general trend and slightly overestimated water level values in the second half of the simulation period (Figure 11a). At JAV station (R^2 = 0.94), the fits of the values and the trend were close to the observations (Figure 11b).

In the ROK catchment (R^2 = 0.54), the model significantly overestimated the values compared to the observations over the whole period (Figure 11c). In contrast, in the HAJ catchment (R^2 = 0.71) the simulated values were underestimated, and the model missed a minor peak flow.

3.3. Model Performance

The application of the support regression model to simulate the runoff for five different scenarios, reflecting typical types of events, showed relatively high goodness of fit of the simulated to the observed values, with significant differences among the catchments and simulation scenarios (Table 2).

Table 2. Model performance for simulation scenarios in the study catchments.

R^2	RKM	JAV	ROK	HAJ	Mean
Validation (hydrological year)	0.9127	0.8912	0.9177	0.8040	0.8814
Regional flood	0.9616	0.9237	0.9059	0.8686	0.9150
Summer storm	0.9599	0.9231	0.8573	0.6607	0.8503
Spring snowmelt	0.9602	0.961	0.8989	0.8377	0.9145
Rain on snow	0.9750	0.9864	0.9479	0.6211	0.8826
Drought	0.9419	0.9396	0.5427	0.7116	0.7840
Mean	0.9519	0.9375	0.8451	0.7506	0.8713

The best performance was achieved at stations located in lower part of the basin, and the performance decreased with increasing mean altitude of the catchments. At the outlet of the Roklanský Brook basin at Modrava (RKM station), the SVM model performed very well in all simulation scenarios; the R^2 values for every scenario exceeded 0.91 with a R^2_{mean} = 0.95. Similar success was achieved for Javoří Brook (JAV station, R^2_{mean} = 0.94), located in lower part of the basin, where only one simulation (hydrological year) had an R^2_{mean} value less than 0.9.

In the Rokytka catchment (ROK station), located in the middle of the basin altitude, the simulations were very reliable for all scenarios but the low flow (R^2_{mean} = 0.84).

The most elevated catchment of Roklanský Brook at Hajenka (HAJ, R^2_{mean} = 0.75), located at the ridge of the Sumava Mts., showed the worst performance across the assessed catchments. For almost all of the simulation scenarios, the simulations in this catchment were least reliable. However, in three scenarios, the R^2_{mean} values exceeded 0.8 and were never less than 0.62.

For the simulation scenarios, the best average performance was achieved for the events resulting from regional flooding (SC2, R^2_{mean} = 0.915) and for events related to the spring snowmelt (SC3, R^2_{mean} = 0.915). The third best scenario was rain on snow (SC4, R^2_{mean} = 0.883). The weakest performance of the SVM model network was achieved for the simulation of the low flow period (SC5, R^2mean = 0.784).

4. Discussion

Recent studies comparing SVMs with other types of models—physically based, conceptual or machine learning, indicate the solid performance of SVMs, especially on nonlinear data series [26]. The examples can be taken from different environments and spatial scales. Dibike et al. have compared the SVMs to ANNs and conceptual runoff models across three distinctly different scales and climate conditions and demonstrated that the SVMs perform well under varying conditions and provide a good alternative to the conventional models [37]. Runoff simulations in long-term time scale, provided by Lin et al. [32] indicated that SVM model gave more accurate results than ARMA and ANN models in a complex basin at the regional scale. The study comparing SVMs to the EPA´s Storm Water

Management Model (SWMM) at a micro-scale in two different urban catchments confirmed that SVMs reach similar performance than a conventional model, even in uneasy conditions of the urban environment with significant share sewage drainage [31].

The good performance of SVMs in simulations on hydrological time series was confirmed also in our study, focused on small homogeneous catchments in the mid-mountain environment. The cross-comparison of the model performance in the model catchments for different scenarios indicated that the data-driven SVM model generally performed relatively very well.

The SVM network was able to reliably simulate most of the typical runoff situations in the montane catchments with varying physiographic conditions without substantial errors in the fits of the trend, timing of the events and peak values.

The efficiency of the SVM algorithm suggests its wide usability. The typical SVM applications are in binary classification, non-linear classification, and SVM clustering. A specific branch of SVM, support vector regression (SVR), can apply a continuous function to data. This feature is particularly useful when the predicted variable is continuous, as in the case of hydroclimatic time series [49].

The model performance in our study was generally better in the complex catchments with stations located in the lower part of the catchment. In contrast, the simulations in the topmost catchment were the least reliable for most scenarios; however, the simulations still had satisfactory fits without substantial errors.

The SVM model proved the ability to reliably simulate complex situations, such as rain on snow events, which are complicated for conventional hydrological models. However, the simulations of some of the simple events, such as a single peak summer storm and the dry period, were less accurate.

The varying model performance in different catchments might be affected by the differences in physiography and environmental status. Although located in close vicinity, the catchments have variable distributions of bedrock, soil, and vegetation cover, as well as different topography and floodplain characteristics. This variation might result in a complex hydrological response that is difficult simulate using data-driven models. Among the factors that can affect the predictability of the runoff response in the simulated catchments, the following should be considered.

The whole area experiences the effect of climate change, which is modifying the rainfall–runoff properties in a larger scale, with an apparent effect on the frequency and magnitude of the peak flows and dry periods [13].

The upper part of the basin has undergone massive forest decay since the 1990s due to a bark beetle outbreak [50,51], which hit the catchments to different extents and with different timing. In the HAJ catchment, deforestation occurred prior to the monitoring period, and the forest has recently undergone restoration of the bottom layer vegetation. The ROK catchment has been heavily affected by forest decay since 2010, and the share of the healthy, damaged and decayed forest is rapidly changing. The JAV catchment remains relatively undisturbed.

Another source of difference among the catchments is the soil properties. Two catchments at high altitude, HAJ and ROK, have significant shares of peatland (see Table 1), which affects the speed of the runoff response under rainfall situations, the retention and transformation potential of the catchments, and the response to dry periods [52].

Moreover, there are remnants of the historical stream regulations and abandoned ponds that were built in the 18th century for timber floating (Figure 12c). The hydrological regime at the outlets of catchments of ROK and HAJ is affected by the abandoned reservoirs that act as polders during the flood events, with a substantial effect on the transformation of the flood wave [53].

Such factors affect the hydrological response and predictability of the processes and values in the catchments, especially as some of them undergo rapid changes.

These factors might be apparent, especially in the catchments that are undergoing the most intense transition of environmental conditions and are affected by the presence of the former large reservoirs, the HAJ and the ROK catchment. Here, the cumulative effect of the transient environmental conditions,

extensive peatland and the presence of reservoirs acting as polders might be the cause of the weaker performance of the SMV network learning and the fit of simulations to the observed values.

However, despite the weaker performance of the model in these two catchments, the overall performance of the SVM model for reconstruction of the hydrological time series can be considered reliable.

Figure 12. The course of the rain on snow flood on 1 December 2015, in different parts of the study area: (**a**) regulated channel at the basin outlet at Modrava (RKM); (**b**) flood spill in the large floodplain of JAV catchment; (**c**) flood filling the polder at ROK; and (**d**) ice jam at a road culvert at the headwaters (BRE) (Photo by Lukáš Vlček).

The solid performance, together with the ease of model setup, is highly beneficial when building models in areas that lack appropriate data for the setup of conventional hydrological models, especially experimental catchments, where the level of detail necessary for the model is often beyond the resolution of the available data. In this study area, this applies to the high level of generalization of the geological and soil maps or outdated land cover maps, which are typical data inputs for hydrological modeling [54,55].

Hence, although the data-driven models can be burdened by uncertainty, their application may be beneficial for different applications, ranging from the reconstruction of past events and reconstruction of missing data in time series to hydrological forecasting. The superior performance of SVMs indicated in several studies [32,37] is seen in avoiding the typical weaknesses of ANNs. First, ANNs often converge on local minima rather than global minima, meaning that they are essentially "missing the big picture" (or missing the forest for the trees). Second, the ANNs often overfit if training goes on too long, so for a given pattern, an ANN might start to consider the noise as part of the pattern [56,57]. SVMs do not suffer from either of these problems [16].

The rising number of applications of machine learning models in different aspects of hydrological indicated their suitability for reliable reconstruction or prediction of hydrological processes [31]. In particular, machine learning techniques seem to be coping well with one of the key problems with noise hydrological data, which represent one of the principal sources of limited performance of conventional modeling techniques [11].

Despite the proven performance of machine learning models including SVMs, their application in hydrological forecasting is still limited. Unlike in conventional conceptual or physically based models, where the uncertainty in forecasts can be explained in relation to the processes and their schematization, the unexpected results of data-driven models can be hard to interpret which prevents their wider use in forecasting [34].

A promising field of application for machine learning models in hydrology is the reconstruction of missing data in observations. Filling the gaps in time series of observations is a long-term problem, e.g., in developing countries, where the good applicability of SVMs has already been proven [11]. With the rapid development of sensor networks, seeking efficient approaches for completion of missing data is of growing importance with applications in research as well as in water management or flood warning systems [1,5]. The sensor networks, although reliable, are burdened by a number of potential technical issues. Ten years of operation of our monitoring network in the montane environment [3] showed that issues related to data loss might be the result of physical damage by natural processes, technical defects or failures in operation. Physical damage to sensors or control stations can result from extreme weather that can damage the sensor itself or the electronics of the control station. Typically, it is related to periods of extreme cold, electrical shocks after lightning, physical damage after flooding, mud flows or freefall, and in rare cases, vandalism. A frequent source of data loss is power outages. Depending on the type of the monitoring devices installed at the station, the sampling frequency and data transfer interval, the energy demand of the station varies significantly. If not properly balanced, battery drain could result in monitoring interruptions and data losses. As the monitoring stations are often placed in remote areas with limited accessibility, e.g., in the winter season, power outages can affect complex time periods.

The presented approach for time series modeling using an SVM data-driven model demonstrated potential benefits but also limitations for applications. The positive aspects of the application of the SVM model to hydrological simulations include the following:

(i) Universal applicability: Data-driven models are independent of the given physical environment or conditions. While the only driver of the algorithm is data, the simulation network can be developed and trained for any type of environment, scale or temporal resolution. In addition to the observed variables used to train the model for a given process, the algorithm does not require additional or specialized data sources required by conventional hydrological models, such as detailed geologic or soil maps, soil parameters, and vegetation maps, which might be unavailable for experimental catchments.

(ii) Transparent model setup and control: The environment of data mining frameworks, such as KNIME and RapidMiner, enables the visual development of the model and transparent control of the workflow. The workflow can be easily modified for different scenarios or reused with different datasets.

(iii) Rapid model setup and learning: The building of the data-driven model in the visual environment of the data mining workflow is rapid. The SVM model learner performs relatively fast; a model with 12 input variables based on daily values over five years can be trained on the scale of hours or less, depending on computer performance.

(iv) Reliability and robustness: Testing of the SVM model proved that the model can learn and predict complex runoff situations that are difficult to handle using conventional hydrological models, such as snowmelt or rain on snow events. The goodness of fit in varying environmental conditions is satisfactory for most of the simulated scenarios. Although the model overestimates or underestimates the values in specific conditions and scenarios, the general fit of the trends, peak values and timing of events is reliable.

The data-driven modeling approach is also burdened by several limitations:

(i) Specificity of the model: The learning of the network is always specific to the given configuration of the selected set of variables and parameters for a given basin. A network trained for given conditions is specific and cannot be applied to different catchments. The model must be re-trained for each configuration of applied variables.

(ii) The risk of selection of inappropriate variables: The quality of the machine learning model relies on the quality and structure of the applied data inputs. The selection of inappropriate variables for model training can deteriorate the model performance and reproducibility of the results.

(iii) The risk of insufficient model training: Reliable model training requires a complex set of events, including samples of the typical events or situations that might occur in simulation scenarios. The use of a limited sample of training data, even for appropriate variables, might result in false signals in the predictor phase and poor quality of the forecasts.

5. Conclusions

This paper analyzed the potential of SVR models for the completion of missing data in hydrological time series observed in a sensor network.

Automated sensor networks are currently experiencing rapid growth of applications in experimental research and monitoring and provide an opportunity to study the dynamics of hydrological processes in previously ungauged or remote areas. Due to physical vulnerability or limited maintenance, networks are prone to data outages, which can devaluate the unique data sources.

Monitoring sensors are often organized into networks in basins, where the processes at individual stations are at least partially interrelated. Hence, there is potential for application of data-driven models, such as SVM, to simulate the observed processes. These models can be used to complete missing data in the monitoring network.

The SVR model was applied to test the applicability in a network of nested experimental catchments in the mid-latitude montane environment of the Sumava Mountains, Central Europe, which are characterized by different physiography and environmental status. The model was applied to a range of typical runoff situations, including a single event storm, multi-peak flood event, snowmelt, rain on snow, low flow period and a complex hydrological year.

The simulations based on daily values proved the high efficiency of the SVM modeling approach to simulate hydrological processes in a network of monitoring stations. The SVM model was developed and run for all stations to analyze the performance under different environmental conditions.

Application of the model to a mid-mountain environment proved the robustness and good performance of the data-driven SVM model to simulate hydrological time series. The SVM network reliably simulated most of the typical runoff situations in montane catchments characterized by variable physiographic conditions without substantial errors in the fit of the trend, timing of the events and peak values. The model reliably reconstructed and simulated even the complex events, such as rain on snow episodes and flooding from recurrent precipitation. The model had weaker performance in simulations of rapid summer storms and low flow periods.

The model performance was generally better in the complex catchments with stations located in the lower part of the catchment. The simulations in the topmost catchment were the least reliable for most scenarios, but the fits were good and without substantial errors.

The study indicated that the data-driven SVM model can be used for reliable reconstruction of missing data from hydrological sensor networks, and this technique has the potential for wider applications in hydrological research and water management.

Acknowledgments: This research was supported by the EU COST Action 1306 project LD15130 "Impact of landscape disturbance on the stream and basin connectivity" and the Czech Science Foundation project GACR 13-32133S. The authors thank Lukáš Vlček for providing photos of the flood in December 2015.

Author Contributions: Jakub Langhammer designed the study, set up and run the nu-SVR model, analyzed and interpreted the results and wrote the manuscript. Julius Česák has designed and developed the monitoring network, provided technical maintenance of the stations and sensors and the data quality checking.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hart, J.K.; Martinez, K. Environmental Sensor Networks: A revolution in the earth system science? *Earth-Sci. Rev.* **2006**, *78*, 177–191. [CrossRef]
2. Xu, N. A survey of sensor network applications. *IEEE Commun. Mag.* **2002**, *40*, 102–114.

3. Langhammer, J.; Hartvich, F.; Kliment, Z.; Jeníček, M.; Bernsteinová, J.; Vlček, L.; Su, Y.; Štych, P.; Mišijovský, J. The impact of disturbance on the dynamics of fluvial processes in mountain landscapes. *Silva Gabreta* **2015**, *21*, 105–116.
4. Kotamäki, N.; Thessler, S.; Koskiaho, J.; Hannukkala, A.O.; Huitu, H.; Huttula, T.; Havento, J.; Järvenpää, M. Wireless in-situ sensor network for agriculture and water monitoring on a river basin scale in southern Finland: Evaluation from a data user's perspective. *Sensors* **2009**, *9*, 2862–2883. [CrossRef] [PubMed]
5. He, B.; Li, Y. Big data reduction and optimization in sensor monitoring network. *J. Appl. Math.* **2014**, *2014*, 294591. [CrossRef]
6. Rettig, A.J.; Khanna, S.; Heintzelman, D.; Beck, R.A. An open source software approach to geospatial sensor network standardization for urban runoff. *Comput. Environ. Urban Syst.* **2014**, *48*, 28–34. [CrossRef]
7. Nayak, P.; Sudheer, K.; Rangan, D.; Ramasastri, K. A neuro-fuzzy computing technique for modeling hydrological time series. *J. Hydrol.* **2004**, *291*, 52–66. [CrossRef]
8. Elshorbagy, A.; Simonovic, S.P.; Panu, U.S. Estimation of missing streamflow data using principles of chaos theory. *J. Hydrol.* **2002**, *255*, 123–133. [CrossRef]
9. Teegavarapu, R.S.V.; Chandramouli, V. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. *J. Hydrol.* **2005**, *312*, 191–206. [CrossRef]
10. Singh, S.; Jain, S.; Bárdossy, A. Training of artificial neural networks using information-rich data. *Hydrology* **2014**, *1*, 40–62. [CrossRef]
11. Dastorani, M.T.; Moghadamnia, A.; Piri, J.; Rico-Ramirez, M. Application of ANN and ANFIS models for reconstructing missing flow data. *Environ. Monit. Assess.* **2010**, *166*, 421–434. [CrossRef] [PubMed]
12. Rodriguez-Iturbe, I.; Febres De Power, B.; Sharifi, M.B.; Georgakakos, K.P. Chaos in rainfall. *Water Resour. Res.* **1989**, *25*, 1667–1675. [CrossRef]
13. Langhammer, J.; Su, Y.; Bernsteinová, J. Runoff Response to Climate Warming and Forest Disturbance in a Mid-Mountain Basin. *Water* **2015**, *7*, 3320–3342. [CrossRef]
14. Kozák, J.; Nemecek, J.; Jetmar, M. The database of soil information system-PUGIS. *Rostl. Vyroba UZPI* **1996**, *42*, 529–534.
15. Shahraiyni, H.; Ghafouri, M.; Shouraki, S.; Saghafian, B.; Nasseri, M. Comparison between active learning method and support vector machine for runoff modeling. *J. Hydrol. Hydromech.* **2012**, *60*, 16–32. [CrossRef]
16. Sahay, T.; Aggarwal, A.; Bansal, A.; Chandra, M. SVM and ANN: A comparative evaluation. In Proceedings of the 1st International Conference on Next Generation Computing Technologies (NGCT), Dehradun, India, 4–5 September 2015; IEEE: New York, NY, USA, 2015; pp. 960–964.
17. Dawson, C.W.; Wilby, R. An artificial neural network approach to rainfall-runoff modelling. *Hydrol. Sci. J.* **1998**, *43*, 47–66. [CrossRef]
18. Tingsanchali, T.; Gautam, M.R. Application of tank, NAM, ARMA and neural network models to flood forecasting. *Hydrol. Process.* **2000**, *14*, 2473–2487. [CrossRef]
19. Abrahart, R.J.; Dawson, C.W. Neural network modelling trade-offs: Small might be beautiful but perhaps bigger is better? *Assembly* **2009**, *11*, 4832.
20. Mas, J. Modelling deforestation using GIS and artificial neural networks. *Environ. Model. Softw.* **2004**, *19*, 461–471. [CrossRef]
21. Pang, A.P.; Sun, T. Bayesian networks for environmental flow decision-making and an application in the Yellow River estuary, China. *Hydrol. Earth Syst. Sci.* **2014**, *18*, 1641–1651. [CrossRef]
22. Zhang, X.; Liang, F.; Yu, B.; Zong, Z. Explicitly integrating parameter, input, and structure uncertainties into Bayesian Neural Networks for probabilistic hydrologic forecasting. *J. Hydrol.* **2011**, *409*, 696–709. [CrossRef]
23. Xue, J.; Gui, D.; Zhao, Y.; Lei, J.; Zeng, F.; Feng, X.; Mao, D.; Shareef, M. A decision-making framework to model environmental flow requirements in oasis areas using Bayesian networks. *J. Hydrol.* **2016**, *540*, 1209–1222. [CrossRef]
24. Vapnik, V.N.; Lerner, A. Pattern Recognition using Generalized Portrait Method. *Autom. Remote Control* **1963**, *24*, 774–780.
25. Vapnik, V.N. *Statistical Learning Theory*; Wiley: New York, NY, USA, 1998; Volume 2, p. 736.
26. Hwang, S.H.; Ham, D.H.; Kim, J.H. Forecasting performance of LS-SVM for nonlinear hydrological time series. *KSCE J. Civ. Eng.* **2012**, *16*, 870–882. [CrossRef]
27. Kim, K. Financial time series forecasting using support vector machines. *Neurocomputing* **2003**, *55*, 307–319. [CrossRef]

28. Ren, J. ANN vs. SVM: Which one performs better in classification of MCCs in mammogram imaging. *Knowl.-Based Syst.* **2012**, *26*, 144–153. [CrossRef]
29. Kahraman, F.; Capar, A.; Ayvaci, A.; Demirel, H.; Gokmen, M. Comparison of SVM and ANN performance for handwritten character classification. In Proceedings of the IEEE 12th Signal Processing and Communications Applications Conference, Kusadasi, Turkey, 28–30 April 2004; pp. 1–10.
30. Scholkopf, B.; Smola, A.J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*; MIT Press: Cambridge, MA, USA, 2002.
31. Raghavendra, S.; Deka, P.C. Support vector machine applications in the field of hydrology: A review. *Appl. Soft Comput. J.* **2014**, *19*, 372–386. [CrossRef]
32. Lin, J.-Y.; Cheng, C.-T.; Chau, K.-W. Using support vector machines for long-term discharge prediction. *Hydrol. Sci. J.* **2006**, *51*, 599–612. [CrossRef]
33. Yu, P.S.; Chen, S.T.; Chang, I. Support vector regression for real-time flood stage forecasting. *J. Hydrol.* **2006**, *328*, 704–716. [CrossRef]
34. Granata, F.; Gargano, R.; de Marinis, G. Support vector regression for rainfall-runoff modeling in urban drainage: A comparison with the EPA's storm water management model. *Water* **2016**, *8*, 69. [CrossRef]
35. Noori, R.; Karbassi, A.R.; Moghaddamnia, A.; Han, D.; Zokaei-Ashtiani, M.H.; Farokhnia, A.; Gousheh, M.G. Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction. *J. Hydrol.* **2011**, *401*, 177–189. [CrossRef]
36. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*; Cambridge University Press: Cambridge, UK, 2000.
37. Dibike, Y.B.; Velickov, S.; Solomatine, D.; Abbott, M.B. Model induction with support vector machines: Introduction and applications. *J. Comput. Civ. Eng.* **2001**, *15*, 208–216. [CrossRef]
38. Chang, C.; Lin, C. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2013**, *2*, 1–39. [CrossRef]
39. Muller, K.; Smola, A.; Ratsch, G.; Scholkopf, B.; Kohlmorgen, J.; Vapnik, V. Predicting time series with support vector machines. In Proceedings of the International Conference on Artificial Neural Networks, Lausanne, Switzerland, 8–10 October 1997; Volume 1327, pp. 999–1004.
40. Smola, A.; Vapnik, V. Support vector regression machines. *Adv. Neural Inf. Process. Syst.* **1997**, *9*, 155–161.
41. Schölkopf, B.; Bartlett, P.; Smola, A.; Williamson, R. Support vector regression with automatic accuracy control. In *ICANN 98*; Springer: Heidelberg, Germany, 1998; pp. 111–116.
42. Ojemakinde, B.T. Support Vector Regression for Non-Stationary Time Series. Master's Thesis, University of Tennessee, Knoxville, TN, USA, 2006.
43. Hao, W.; Yu, S. Support Vector Regression for Financial Time Series Forecasting. In *Knowledge Enterprise: Intelligent Strategies in Product Design, Manufacturing, and Management*; Springer: Heidelberg, Germany, 2006; Volume 207, pp. 825–830.
44. Zhu, D.; Ji, B.; Meng, C.; Shi, B.; Tu, Z.; Qing, Z. The performance of nu-support vector regression on determination of soluble solids content of apple by acousto-optic tunable filter near-infrared spectroscopy. *Anal. Chim. Acta* **2007**, *598*, 227–234. [CrossRef] [PubMed]
45. Berthold, M.R.; Cebron, N.; Dill, F.; Gabriel, T.R.; Kötter, T.; Meinl, T.; Ohl, P.; Thiel, K.; Wiswedel, B. KNIME-the Konstanz information miner: Version 2.0 and beyond. *AcM SIGKDD Explor. Newsl.* **2009**, *11*, 26–31. [CrossRef]
46. Stahl, F.; Gabrys, B.; Gaber, M.M.; Berendsen, M. An overview of interactive visual data mining techniques for knowledge discovery. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2013**, *3*, 239–256. [CrossRef]
47. Chang, C.-C.; Lin, C.-J. Training nu-support vector regression: Theory and algorithms. *Neural Comput.* **2002**, *14*, 1959–1977. [CrossRef] [PubMed]
48. Bray, M.; Han, D. Identification of support vector machines for runoff modelling. *J. Hydroinf.* **2004**, *6*, 265–280.
49. Wang, L. *Support Vector Machines: Theory and Applications*; Springer: Heidelberg, Germany, 2005; Volume 177.
50. Hais, M.; Jonášová, M.; Langhammer, J.; Kučera, T. Comparison of two types of forest disturbance using multitemporal Landsat TM/ETM+ imagery and field vegetation data. *Remote Sens. Environ.* **2009**, *113*, 835–845. [CrossRef]
51. Nováková, M.H.; Edwards-Jonášová, M. Restoration of central-european mountain norway spruce forest 15 years after natural and anthropogenic disturbance. *For. Ecol. Manag.* **2015**, *344*, 120–130. [CrossRef]

52. Janský, B.; Kocum, J. Peat bogs influence on runoff process: Case study of the vydra and křemelná river basins in the šumava mountains, Southwestern Czechia. *Geografie* **2008**, *113*, 383–399.
53. Čurda, J.; Janský, B.; Kocum, J. The effects of physical-geographic factors on flood episodes extremity in the Vydra River basin. *Geografie* **2011**, *116*, 335–353.
54. Váňová, V.; Langhammer, J. Modelling the Impact of Land Cover Changes on Flood Mitigation in the Upper Lužnice Basin. *J. Hydrol. Hydromech.* **2011**, *59*, 262–274. [CrossRef]
55. Jeníček, M. Rainfall-runoff modelling in small and middle-large catchments—An overview. *Geografie* **2006**, *111*, 305–313.
56. Shawe-Taylor, J.; Cristianini, N. *Kernel Methods for Pattern Analysis*; Cambridge University Press: Cambridge, UK, 2004; Volume 47.
57. Suykens, J.A.K.; De Brabanter, J.; Lukas, L.; Vandewalle, J. Weighted least squares support vector machines: Robustness and sparse approximation. *Neurocomputing* **2002**, *48*, 85–105. [CrossRef]