Forecasting River Runoff through Support Vector Machines

Bryan Bell, Brian Wallace, Du Zhang
Department of Computer Science
California State University, Sacramento
Email: {bryan.w.bell, bw.tech} @gmail.com, zhangd@ecs.csus.edu

***Abstract*—How "wet" or "dry" a year is predicted to be has many impacts. Public utilities need to determine what percentage of their electric energy generation will be hydro power. Good water years enable the utilities to use more hydro power and, consequently, save oil. Conversely, in a dry year, the utilities must depend more on steam generation and therefore use more oil, coal, and atomic fuel. Agricultural interests use the information to determine crop planting patterns, ground water pumping needs, and irrigation schedules. Operators of flood control projects determine how much water can safely be stored in a reservoir while reserving space for predicted inflows. Municipalities use the information to evaluate their water supply and determine whether (in a dry year) water rationing may be needed.**

Currently a combination of linear regression equations and human judgment is used for producing these forecasts. In this paper, we describe a Support Vector Machine based method for river runoff forecasting. Our method uses Smola/Scholkopf's Sequential Minimal Optimization algorithm for training a Support Vector Machine with a RBF kernel. The experimental results on predicting the full natural flow of the American River at the Folsom Dam measurement station in California indicates that our method outperforms the current forecasting practices.

***Keywords:* river runoff forecasting, Support Vector Machines, sequential minimum optimization**

I. Introduction

Our focus is forecasting the unimpaired river runoff for the American River watershed at the Folsom Dam measurement station in California. Water forecasts lead to better planning and management of California's water resources. Our goal, which we achieved, was to outperform the current forecast practices. The main water resources for California are (1) river runoff due to rain, (2) river runoff due to snow-melt, and (3) groundwater.

Due to California's Mediterranean climate the summer months are crucial for water usage and allocation. This makes the unimpaired river runoff forecast for the April-July river flow the most important river flow forecast. The forecast made on April 1st for the April-July unimpaired river flow is used by the Department of Water Resources to allocate water from the river for the different users e.g. agriculture, hydro, etc. The American River Basin is 2,150 sq mi (5,568 km) or about half the size of the state of Connecticut [9]. Our methods use no feature that is specific to the American River Watershed, therefore they easily generalize to other watersheds in California. It is more difficult to determine if our methods would work as well for other states since we are only familiar with the format and methods of rain/snow measurement for California.

Our methods improve upon the current predictions by reducing the average error of the April-July unimpaired river-flow forecast made on April 1st. Engineers at California Department of Water Resources create the unimpaired river flow forecasts at the beginning of February, March, April, and May, the engineers estimate that each forecast takes approximately a total of 100 man hours. Each forecast includes all of the 30 river basins in California. The American River is one of these river basins.

Our work focused on a sub-section of the Bulletin 120 for the American River. We forecasted the April through July full natural flow runoff of the American River measured at Folsom. The April - July "full natural flow" runoff is measured in acre-feet. The measuring stations report their data in snow water equivalent for the snow measurement stations or inches for the rainfall measurement stations.

Issues with the Current Forecast Practices

The current practice uses regression equations to determine the April thru July runoff forecasts for statewide river basins. Regression formula variables include:

- October March Precipitation Index
- April July Precipitation Index
- High and Low Snow Indices
- Previous fall and spring runoff
- 50 year historic database (1956-2005)

The precipitation variables of the regression equation are determined by calculating a precipitation index based on yearly and monthly averages over a mix of stations representing the basin. Similarly, the snow pack variables of the regression equation are determined by calculating a snow index based on yearly and monthly averages over a mix of stations representing the basin. In some cases, the basin is split into a high elevation and low elevation snow pack index to give ultimate consideration of basin snow.

Some of the issues with the current forecast practices are bad snow water content data, over or under sampling of snow water content (usually snow water content data and precipitation are both under sampled), bad or inaccurate reservoir storage data, inaccurate flow data, inaccurate evaporation data.

Other issues are (a) the equations were made using the entire data set. Therefore, some accuracy may be lost when a

Fig. 1. American River Watershed

general equation is used during a very wet or very dry year, (b) there is no mechanical way to check the precipitation, snow, or full natural flow data. This means there is not a mechanism whereby a flag is put on data to alert us of a questionable value, (c) there is no formal way to make an estimate for missing or bad data, (d) for some rivers, the three main equations can produce values that are not close to each other, hence it is hard to get a feel for the true value and other methods are used.

A. Primary forecasting tool: Support Vector Machine using Sequential Minimum Optimization

The dependent variable: April-July cumulative unimpaired runoff.

The independent variables used

- Prior Year April-July Cumulative Unimpaired Runoff
- October-March Cumulative Unimpaired Runoff
- Snow Index (High Elevation)
- Snow Index (Low Elevation)
- October-March Precipitation Index
- April-June Precipitation Index.

B. Data Issues

Snow and Precipitation Index Issues:

- Snow course or precipitation station mix may be limited
- Lack of quality snow course measurements requires alternate source data or re-measurement
- Harsh weather will often delay or prohibit snow course data collection
- May have elevation or east-west bias
- May have out of basin station bias
- Data collection may not be thorough or have quality assurance
- Missing stations cause unintentional biases

American River
April 1, 2008 Forecasts
April-July Unimpaired Runoff

HYDROLOGIC RGN & WATERSHED	Unimpaired Runoff in 1,000 Acre-Feet					
	HISTORICAL			FORECAST		
	50 YR Avg	Max of Record	Min of Record	Apr-Jul Forecasts	Pct of Avg	80% Prob. Range
North Fork	262	716	43	180	69%	
Middle Fork	522	1,406	100	390	75%	
Silver Creek	173	386	37	130	75%	
Below Folsom Dam	1,240	3,074	229	940	76%	660-1,590

Fig. 2. Bulletin 120, April 2008 Forecast

Resolution: Eliminate parameters that have missing or faulty entries.

The rest of this paper is organized as follows: Section II discusses the relevant learning algorithms and approach used, related work and how our proposed approach differs from them; Section III describes the dataset design and issues with missing data values; Section IV presents the experimental results; Section V compares our results with the forecasts produced by the California Department of Water Resources; Section VI gives a brief description of the tools used in our models; Section VII suggests possible areas of improvement; Finally Section VIII concludes the paper with remarks on future work.

II. Approach and Learning Algorithms

We used monthly precipitation and snow data gathered from 10 precipitation monitoring stations and 28 snow monitoring stations located in the American River basin. We also made use of the historical full natural flow data for the American River at the Folsom measurement station.

The data is from the California Department of Water Resources, the precipitation/snow data is available on their website at http://cdec.water.ca.gov. We formatted the raw CSV data from the Department of Water Resources as an Attribute-Relation File Format (ARFF) file.

With the data set of 222 input parameters from the 40 monitoring stations we then further narrowed down the data to create each forecast.

Using the Machine Learning tool WEKA [2], we used an ad-hoc method of trying different learning algorithms and different parameters for each learning algorithm to find the best algorithm and set of algorithm parameters. WEKA (Waikato Environment for Knowledge Analysis) provides a large toolbox of learning algorithms. Additionally WEKA makes it easy to try individual learning algorithms and vary the algorithm parameters to determine which set of parameters yields the best results.

We tried several algorithms for creating river flow forecasts. In particular, we tried simple linear regression, neural networks, an ensemble of neural networks, least median of squares linear regression, and least median of squares linear regression with bagging [1]. The best results were from using SMOreg (Support Vector Machine with Sequential Minimum Optimization) with a RBF kernel function. Version 3.6 of WEKA was used for all the calculations. From our results we

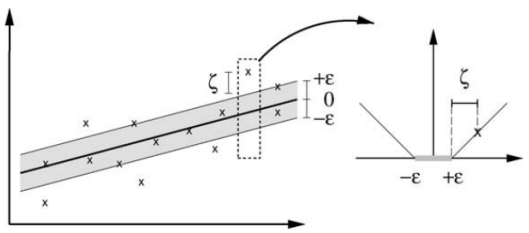

Fig. 3. The soft margin loss setting for a linear SVM (from Scholkopf and Smola, 2002)

cannot conclude that using SVM with Sequential Minimum Optimization is the best algorithm. We can only conclude that it is the best of the algorithms we tried with the specific algorithm parameters we used and it outperforms the existing forecasting methods..

A. Description of using SVM (Support Vector Machine) for Function Estimation

Alex Smola and Bernhard Scholkopf have an excellent tutorial [12] on using Support Vector Machines for function estimation. We will go over only the relevant details for our model.

The idea of a Support Vector Machine was initially developed in Russia in the 60's by Vapnik and Lerner [13], and Chervonenkis [14]. Vapnik further developed the field and wrote the definitive book [15] on the subject.

Here we only cover the basics of SVM and how the algorithm parameters we selected affected the learning process.

A SVM consists of a set of support vectors and a kernel function. The support vectors are a set of vectors from the training data. The support vectors together with the kernel create the function approximation.

The SVM formulation for function estimation is as follows.

Suppose we are given training data

$$\{(x_1, y_1), \ldots, (x_\ell, \ldots, y_\ell)\} \subset X \times \mathbb{R}, \text{ where } X = \mathbb{R}^d. \quad (1)$$

For the linear formulation of a support vector machine we want to find a function, $f(x) = \langle w, x\rangle + b$, that satisfies Equation 2.

$$\text{minimize } \frac{1}{2}\|w\|^2 + C\sum_i^l (\xi_i + \xi^*{}_i)$$

$$\text{subject to } \begin{cases} y_i - \langle w, x_i\rangle - b \leq \epsilon + \xi_i \\ \langle w, x_i\rangle + b - y_i \leq \epsilon + \xi^*{}_i \\ \xi_i, \xi^*{}_i \geq 0 \end{cases} \quad (2)$$

Equation 2 is the formulation stated in Vapnik (1995) [15].

The non-linear formulation uses a kernel function instead of, $\langle\cdot,\cdot\rangle$, the inner product. For sake of brevity please refer to Smola and Scholkopf's tutorial [12] for the full exposition. Briefly stated the relevant parameters are:

1) C, The complexity parameter.
2) kernel, The kernel to use.
3) regOptimizer, The specific learning algorithm used to solve the optimization problem.

Some example kernels are $\langle x, x'\rangle^p, p \in \mathbb{N}$ and $e^{-\frac{\|x-x'\|^2}{2\cdot\sigma}}$.

The specific regOptimizer we used is from Shevade, Keerthi, et al. [16] and is based on the Sequential Minimum Optimization (SMO) algorithm.

After disappointing results with a polynomial kernel we tried a radial basis function (RBF) kernel and had very satisfying results. We varied the complexity parameter, C, and also the epsilon parameter for the regOptimizer (used for the epsilon insensitive loss function). We did not try a full exponential grid search of the parameter space, which could give better results.

The SVM with SMO algorithm is named "SMOreg" in WEKA. The final output from SMOreg is of Equation 3 form.

$$10.0\cdot k_0 + 10.0\cdot k_1 + \cdots - 5.343220517106478 * k_{99} + 0.024 \quad (3)$$

The $k_0, k_1, \ldots, k_{99}$ are the input vectors. To compute the estimated full natural April-July natural flow of a given year Equation 4 is used.

$$f(x) = \sum_{i=0}^{99} a_i \cdot K(x_i, x) + b, \quad (4)$$

where $\{a_0 = 10, a_1 = 10, a_2 = -10, \cdots, a_{99} = -5.343220517106478\}$, $b = 0.024$, the x_i are the training data, and

$$K(x, y) = e^{-\gamma\cdot\langle x-y, x-y\rangle^2}, \gamma = 0.01. \quad (5)$$

Figure 4 shows the full set of support vector coefficients. Equation 4 operates on the normalized input and the output of equation 4 must be unnormalized to get the actual flow in acre-feet. The Weka documenation on the SMOreg algorithm implementation has details on the normalizeation process. In particular the input values are normalized to the range $[0, 1]$ and to have a standard deviation of 1.

For an example of the input vectors x_i (before normalization) see Figure 5.

B. Related Work

Michaela Bray's and Dawei Han's 2004 paper [10] on using SVM's for river flow forecasting gives a detailed description and analysis of the advantages and disadvantages of SVM's. Their paper used data from the Bird Creek basin of Oklahoma, USA. The data was derived from 12 rain gauges in the basin with a training period from October 1955 to September 1963 and a verification period from November 1972 to November 1974. Their paper goes into detail on the techniques they used to optimize their results e.g. choice of kernel function. The best results from SVM were slightly worse than the best results from a transfer function model.

Behzada et al. in a 2009 paper [11] used SVM's to do one-day lead runoff flow prediction. They noted that SVM makes

$a_0 = +10.000$ $a_1 = +10.000$ $a_2 = -10.000$
$a_3 = +10.000$ $a_4 = -10.000$ $a_5 = +10.000$
$a_6 = +9.582$ $a_7 = -5.619$ $a_8 = +3.652$
$a_9 = -6.709$ $a_{10} = +10.000$ $a_{11} = +10.000$
$a_{12} = +10.000$ $a_{13} = +10.000$ $a_{14} = +10.000$
$a_{15} = -7.710$ $a_{16} = +10.000$ $a_{17} = -10.000$
$a_{18} = -0.737$ $a_{19} = -0.416$ $a_{20} = -5.020$
$a_{21} = +10.000$ $a_{22} = +10.000$ $a_{23} = -10.000$
$a_{24} = +10.000$ $a_{25} = +3.190$ $a_{26} = +7.902$
$a_{27} = +2.799$ $a_{28} = -10.000$ $a_{29} = -4.334$
$a_{30} = -10.000$ $a_{31} = -1.009$ $a_{32} = +10.000$
$a_{33} = -7.117$ $a_{34} = +10.000$ $a_{35} = -10.000$
$a_{36} = +3.354$ $a_{37} = +8.859$ $a_{38} = -10.000$
$a_{39} = -10.000$ $a_{40} = -8.760$ $a_{41} = +10.000$
$a_{42} = -10.000$ $a_{43} = -5.602$ $a_{44} = -3.824$
$a_{45} = -7.755$ $a_{46} = -10.000$ $a_{47} = +10.000$
$a_{48} = -0.236$ $a_{49} = +1.652$ $a_{50} = -3.034$
$a_{51} = +1.240$ $a_{52} = +10.000$ $a_{53} = -10.000$
$a_{54} = +10.000$ $a_{55} = -0.361$ $a_{56} = +10.000$
$a_{57} = +10.000$ $a_{58} = -2.028$ $a_{59} = -10.000$
$a_{60} = +5.498$ $a_{61} = -1.857$ $a_{62} = +4.263$
$a_{63} = -4.226$ $a_{64} = -1.374$ $a_{65} = -0.320$
$a_{66} = +8.607$ $a_{67} = -10.000$ $a_{68} = +5.564$
$a_{69} = -10.000$ $a_{70} = -4.404$ $a_{71} = -2.984$
$a_{72} = -10.000$ $a_{73} = -2.313$ $a_{74} = +1.626$
$a_{75} = -3.130$ $a_{76} = +10.000$ $a_{77} = -1.880$
$a_{78} = +6.721$ $a_{79} = -5.394$ $a_{80} = -0.500$
$a_{81} = +9.217$ $a_{82} = +3.958$ $a_{83} = +3.682$
$a_{84} = -4.202$ $a_{85} = -10.000$ $a_{86} = -2.692$
$a_{87} = -4.160$ $a_{88} = -10.000$ $a_{89} = +10.000$
$a_{90} = -5.987$ $a_{91} = -10.000$ $a_{92} = -9.393$
$a_{93} = +0.598$ $a_{94} = -1.016$ $a_{95} = -1.222$
$a_{96} = -3.100$ $a_{97} = +5.877$ $a_{98} = -2.066$
$a_{99} = -5.343$ $b = +0.024$

Fig. 4. Support Vector Coefficients

use of a convex quadratic optimization problem; hence the solution is unique and globally optimal. They demonstrated one-day lead stream flow forecasting of Bakhtiyari River in Iran using the local climate and rainfall data. Compared with artificial neural network (ANN) and ANN integrated with genetic algorithms (ANN-GA) models they saw improvements in root mean squared error (RMSE) and squared correlation coefficient (R^2) by SVM over both ANN models. They concluded that the prediction accuracy of SVM is at least as good as that of the other models and in some cases better.

An interesting paper on using WEKA was Ozlem Terzi's "Monthly River Flow Forecasting by Data Mining Process" [3], which complements our paper very well. He did use SVM's in his paper but like us he tried a large number of algorithms using WEKA and analyzed the results. His dataset consisted of monthly flow data from two stations and monthly rainfall data from three measurement stations. He used data for the years 1972 through 2002, which he separated

```
@relation 'AmericanRiv_Train'
@attribute CPT_RAIN_OCT numeric
@attribute CPT_RAIN_NOV numeric
@attribute CPT_RAIN_DEC numeric
@attribute CPT_RAIN_JAN numeric
@attribute CPT_RAIN_FEB numeric
@attribute CPT_RAIN_MAR numeric
@attribute CPT_RAIN_APR numeric
@attribute TAC_RAIN_OCT numeric
@attribute TAC_RAIN_NOV numeric
@attribute TAC_RAIN_DEC numeric
@attribute TAC_RAIN_JAN numeric
@attribute TAC_RAIN_FEB numeric
@attribute TAC_RAIN_MAR numeric
@attribute TAC_RAIN_APR numeric
⋮
@attribute APR_JUL_SUM_FNF numeric
@data
⋮
1.75,3.76,0.48,12.41,11.18,1.68,...
40630,316077,677762,430604,437786,1108933
```

Fig. 5. Monitoring Station Data

into 20% for testing and 80% for training. He analyzed multilinear regression, multilayer perceptron, RBF network, decision table, REP tree, KStar as possible forecast algorithms and concluded that multilinear regression performed the best.

Our paper differs from the above related work in that we use Smola/Scholkopf's Sequential Minimal Optimization algorithm. Michaela Bray's and Dawei Han's 2004 paper used only twelve measurement sites whereas our dataset contains 40 measurement sites. Ozlem Terzi's paper used WEKA, but it includes data from only five sources. The most significant difference between our work and Behzada, et al.'s 2009 paper is they did one-day lead runoff flow prediction whereas our predictions are for three month runoff flows.

III. Design of Datasets

We focused on comparing our results with the current forecasts and changed either the learning algorithm used or the algorithm parameters. We spent a significant amount of time filtering and narrowing down the dataset to exclude measuring stations that had sparse data.

Our initial data set had 481 input parameters from 40 monitoring stations. Each rain monitoring station had input parameters for October through September precipitation; similarly the snow monitoring stations had input parameters for the October through September snow water equivalent. As a side note each water year runs from October through September, which is why the data measurements are from October through September instead of a calendar year. We removed input parameters that had missing data. For example in our initial

data set one of the monitoring stations, CPF, had missing data for every month except April. Hence we removed all of the input parameters for CPF except the one for April. For example we went from including

```
"CPF_SNOW_OCT,CPF_SNOW_NOV,
CPF_SNOW_DEC,CPF_SNOW_JAN,
CPF_SNOW_FEB,CPF_SNOW_MAR,
CPF_SNOW_APR,CPF_SNOW_MAY,
CPF_SNOW_JUN,CPF_SNOW_JUL,
CPF_SNOW_AUG,CPF_SNOW_SEP"
```

as input parameters to only including "CPF_SNOW_APR" as an input parameter. This filtering of the data set resulted in 222 input parameters.

Please refer to Figure 6 for the complete list of monitoring stations.

The data set contains 110 instances for the water years 1901 through 2010. For testing purposes we kept out the last 10 instances for the years 2001 through 2010 and trained on the years 1901 through 2000.

With the data set of 222 input parameters from the 40 monitoring stations we then further narrowed down the data to create each forecast.

IV. Experimental Results

The single output parameter was the total unimpeded river flow for the April-July period of the American River at the Folsom measurement station, this data point is measured in acre-feet. In Figure 7 we show the American River April-July flow for 2001 through 2010. Figure 8 shows the cross validation and correlation numbers.

V. Performance Evaluation and Comparison

To evaluate our performance we compared the human ensemble forecasts with the results of SMOreg and additionally compared both with the actual river flows for the test years.

In 8 out of the 10 test years the SMOreg and human forecast errors had the same sign e.g. in 2010 both the SMOreg and the human forecast were low. Both the SMOreg forecasts and the human ensemble forecasts predicated less extreme river flows than the actual river flows. This is expected, since they are both minimizing the RMSE. Two exceptions to this were the years 2007 and 2009. In 2007 the human ensemble forecasted less extreme river flows than the actual river flow but the SMOreg forecast was for a more extreme river flow than the actual river flow. In 2009 the SMOreg forecast was for less extreme river flow than the actual river flow but the human forecast was for more extreme river flow than the actual river flow.

VI. Description of development tools/methodologies used

For our learning algorithms we used WEKA. The data was obtained from the Oracle database that the Department of Water Resources maintains, which has the precipitation data for the state of California. We extracted the relevant data using SQL from the database and then used the CSV to ARFF converter to convert it into an ARFF file.

Station	Longitude	Latitude	Elv(ft)
CPT CAPLES LAKE	120.033	38.7	8000
TAC TAHOE CITY	120.133	39.167	6230
TKE TRUCKEE RS	120.183	39.333	6020
BYM BLUE CANYON	120.7	39.283	5280
LSP LAKE SPAULDING	120.633	39.317	5156
SSR SALT SPRINGS	120.219	38.498	3700
PCF PACIFIC HOUSE	120.5	38.765	3400
GRG GEORGETOWN RS	120.8	38.933	3001
FDD FIDDLETOWN	120.7	38.533	2160
PCV PLACERVILLE	120.82	38.7	1850
APH ALPHA	120.215	38.805	7600
CAP CAPLES LAKE	120.042	38.71	8000
CPF CARPENTER FLAT	120.643	39.303	5300
CC5 CASTLE CREEK 5	120.353	39.353	7400
CCO CISCO	120.543	39.303	5900
DRR DARRINGTON	120.053	38.825	7100
DNS DONNER SUMMIT	120.338	39.31	6900
HYS HUYSINK	120.527	39.282	6600
ABN LAKE AUDRAIN	120.037	38.82	7300
LLL LAKE LUCILLE	120.112	38.86	8200
SPD LAKE SPAULDING	120.642	39.317	5200
LCR LOST CORNER MTN	120.215	39.017	7500
LCP LOWER CARSON PS.	119.998	38.693	8400
LYN LYONS CREEK	120.243	38.812	6700
ONN ONION CREEK	120.358	39.275	6100
PHL PHILLIPS	120.072	38.818	6800
RBV ROBBS VALLEY	120.38	38.922	5600
RP1 RUBICON PEAK 1	120.142	38.992	8100
RP2 RUBICON PEAK 2	120.14	39.001	7500
SIL SILVER LAKE	120.118	38.678	7100
SXV SIXMILE VALLEY	120.6	39.315	5750
SQ2 SQUAW VALLEY 2	120.248	39.188	7700
STW STRAWBERRY	120.145	38.793	5700
TBC TALBOT CAMP	120.377	39.193	5750
TMF TAMARACK FLAT	120.103	38.807	6550
UCP UPPER CARSON PS.	119.983	38.695	8500
WBM WABENA MDS.	120.402	39.227	6300
WR2 WARD CREEK 2	120.225	39.142	7000
WRG WRIGHTS LAKE	120.233	38.847	6900

Fig. 6. Monitoring Stations

VII. Critique of Learning Algorithms Used

As noted in Section V, the forecasted river flows by SMOreg were less extreme than the actual river flows. Ideally the forecasted river flows would not be any more or less extreme than the actual river flows. This is constrained by the fact that the algorithm minimize the RMSE. Another area of possible improvement is more accurate forecasts, of dry years. Dry years necessitate water conservation, which means they should have greater weight when optimizing the forecast model.

An area that is completely absent from our analysis is the use of more qualitative data such as the La Nina or El Nino conditions. Our methods used only data local to the American River Basin and used no regional data such as the La Nina or

Year	River Flow
2001	552,626
2002	973,817
2003	1,354,434
2004	632,159
2005	2,003,878
2006	2,622,387
2007	522,651
2008	674,287
2009	1,068,327
2010	1,486,780

Fig. 7. American River Unimpeded River Flow Apr-July (Acre-Feet)

Training Data	
Correlation coefficient	0.923
Mean absolute error	130,554
Root mean squared error	269,111
Relative absolute error	23.85 %
Root relative squared error	39.21 %
Total Number of Instances	100

Cross Validation	
Correlation coefficient	0.7875
Mean absolute error	303,946
Root mean squared error	430,722
Relative absolute error	54.77%
Root relative squared error	61.74%
Total Number of Instances	100

Fig. 8. Correlation & Cross Validation

Year	Actual	Predicted	|Error|
2001	552,626	689,472	136,846
2002	973,817	1,028,681	54,864
2003	1,354,434	459,476	894,957
2004	632,159	713,440	81,281
2005	2,003,878	1,844,360	159,517
2006	2,622,387	2,315,193	307,193
2007	522,651	293,256	229,394
2008	674,287	800,080	125,793
2009	1,068,327	1,253,523	185,196
2010	1,486,780	1,023,649	463,130
Mean	1,189,135	1,042,113	263,817
		Root mean squared error	355,856
		Relative absolute error	48.65%
		Root relative squared error	54.14%

Fig. 9. SMOreg Forecasts 2001-2010

El Nino ocean conditions. The addition of data representing more general conditions than just the American River Basin measurements could yield improved results.

Year	Actual	Predicted	|Error|
2001	552,626	580,000	27,374
2002	973,817	1,100,000	126,183
2003	1,354,434	680,000	674,434
2004	632,159	940,000	307,841
2005	2,003,878	1,510,000	493,878
2006	2,622,387	1,630,000	992,387
2007	522,651	590,000	67,349
2008	674,287	940,000	265,713
2009	1,068,327	1,000,000	68,327
2010	1,486,780	1,050,000	436,780
Mean	1,189,135	1,002,000	346,026
		Root mean squared error	454,492
		Relative absolute error	63.82%
		Root relative squared error	69.15%

Fig. 10. Human Forecasts 2001-2010

VIII. Conclusion

Our current results using SMOreg with a RBF kernel yield a relative absolute error 48.65% versus 63.82% for the human ensemble forecast. This is a significant improvement over the current forecasts and yields a good model for producing future forecasts.

Our most promising line of future work is to apply our methods on other river basins in California and determine if the SMOreg algorithm consistently yields better results than the current forecast methods. A promising possibility is adjusting the SMOreg parameters to optimize forecasts of dry years, since with the current parameters the forecasts weight wet and dry years equally, see Section VII.

Acknowledgment

We thank the California Department of Water Resources for making their data available and the authors of WEKA.

References

[1] Leo Breiman (1996). Bagging predictors. Machine Learning, 24(2):123-140.

[2] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[3] zlem Terzi. "Monthly River Flow Forecasting by Data Mining Process, Knowledge-Oriented Applications in Data Mining", Kimito Funatsu (Ed.), ISBN: 978-953-307-154-1, InTech (2011), Available from: http://www.intechopen.com/articles/show/title/monthly-river-flow - forecasting-by-data-mining-process

[4] Lin, Gwo-Fong, and Lu-Hsien Chen. "A Non-linear Rainfall-runoff Model Using Radial Basis Function Network." Journal of Hydrology, 289.1-4 (2004): 1-8.

[5] Hopson, Thomas, and Peter Webster. "A 1-10-Day Ensemble Forecasting Scheme for the Major River Basins of Bangladesh: Forecasting Severe Floods of 2003-07*." Journal of Hydrometeorology, 11.3 (2010): 618-641.

[6] Wang, Wen-Chuan, Kwok-Wing Chau, Chun-Tian Cheng, and Lin Qiu. "A Comparison of Performance of Several Artificial Intelligence Methods for Forecasting Monthly Discharge Time Series." Journal of Hydrology, 374.3/4 (2009): 294-306.

[7] Firat, M. "Comparison of artificial intelligence techniques for river flow forecasting." Hydrology and Earth System Sciences, 12.1 (2008): 123-139.

[8] 2010 Census: California Profile, http://www2.census.gov/geo/maps/dc10_thematic/2010_Profile/2010_Profile_Map_California.pdf
[9] "Boundary Descriptions and Names of Regions, Subregions, Accounting Units and Cataloging Units". U.S. Geological Survey. Retrieved 2011-08-12. http://water.usgs.gov/GIS/huc_name.html
[10] Michaela Bray and Dawei Han "Identification of support vector machines for runoff modelling"

Journal of Hydroinfomatics | 06.4 | 2004
[11] Mohsen Behzad, Keyvan Asghari, Morteza Eazi, and Maziar Palhang. 2009. Generalization performance of support vector machines and neural networks in runoff modeling. Expert Syst. Appl. 36, 4 (May 2009), 7624-7629. DOI=10.1016/j.eswa.2008.09.053 http://dx.doi.org/10.1016/j.eswa.2008.09.053
[12] A tutorial on support vector regression
Alex J. Smola and Bernhard Scholkopf
[13] Vapnik V. and Lerner A. 1963. "Pattern recognition using generalized portrait method", Automation and Remote Control, 24: 774780.
[14] Vapnik V. and Chervonenkis A. 1974. "Theory of Pattern Recognition" [in Russian]. Nauka, Moscow. (German Translation: Wapnik W. & Tscherwonenkis A., Theorie der Zeichenerkennung, Akademie-Verlag, Berlin, 1979).
[15] Vapnik V. 1995. "The Nature of Statistical Learning Theory". Springer, New York.
[16] S.K. Shevade, S.S. Keerthi, C. Bhattacharyya, K.R.K. Murthy. "Improvements to the SMO Algorithm for SVM Regression", IEEE Transactions on Neural Networks, 1999.
[17] "A Look at California Agriculture", United States Department of Agriculture, http://www.agclassroom.org/kids/stats/california.pdf