HIERARCHICAL BAYESIAN ANALYSIS AND STATISTICAL LEARNING THEORY II: WATER MANAGEMENT APPLICATION

Abedalrazq Khalil†
Mac McKee‡

ABSTRACT

Water scarcity and uncertainties in forecasting future water availabilities present serious problems for basin-scale water management. These problems create a need to design intelligent prediction models that learn and adapt to their environment in order to provide water managers with decision-relevant information related to the operation of river systems. State-of-the-art techniques fused into a model paradigm (described in Part I of this manuscript) will be demonstrated as decision tools to enhance real-time water management. The framework previously discussed in Part I will be able to diagnose abnormality in the system. Abnormality in this context is referred to as outliers, false signals (e.g., the result of sensor failure) and system behavior "drift" (i.e., non-stationarity or "concept drift"). The proposed versatile adaptive paradigm might be utilized in any control process of a dynamical system in which a quantitative characterization of uncertainty is required. The utility and practicality of this proposed approach is demonstrated here with an application in a real case study river basin.

APPLICATION FOR REAL TIME MANAGEMENT

Description of the Study Area

The Sevier River Basin in rural south-central Utah is one of the state's major drainages. A closed river basin, it encompasses 12.5 percent of the state's total area. From the headwaters 250 miles south of Salt Lake City, the river flows north and then west 255 miles before reaching Sevier Lake (Berger et al., 2002). Irrigation is the primary use of water in the basin. The average amount of water diverted annually for cropland irrigation is 903,500 acre-feet. Of this amount, approximately 135,000 acre-feet are pumped from groundwater. About 40 percent of the diversions are return flows from upstream use (Berger et al., 2002). (For a detailed description of the basin and much of the real-time database utilized in this research, refer to http://www.sevierriver.org).

---

† Graduate Research Assistant, Department of Civil and Environmental Engineering, Utah Water Research Laboratory, Utah State University, Logan, Utah 84322-8200.
‡ Professor, Civil and Environmental Engineering; Director, Utah Water Research Laboratory, Utah State University

Effective real-time management of the water of the Sevier River Basin is seen by the managers of the irrigation systems in the basin to be extremely important in achieving an optimal allocation of their scarce water resources. The main functions of an integrated real-time water resources management system are: water resources real-time monitoring and data collection, information and knowledge mining, and prediction for real-time decision support. Real-time water resources management requires a heavily instrumented basin to monitor climatic indices, streamflow, and water demands. The Sevier River is a heavily instrumented basin of gauges that provide real-time data on all these variables. It is therefore a suitable study area to test tools that are not physically based, but that "let the data speak". These tools will ultimately be integrated into a water resources information management system to be used by the operators of the Sevier River water systems, especially in terms of prediction tools to help manage irrigation canal diversions and reservoir releases.

In physically based models, the necessary climatic, hydrologic, and hydraulic processes must be represented to provide near real-time forecasts of river and canal flows and, ultimately, required reservoir releases. Such models need a substantial amount of data that are often costly to obtain, skilled modelers, powerful computing devices, and they require solution of a complex system of non-linear, partial differential equations. As a result, provision of the necessary information base through the development and use of sophisticated physically based models is often prohibitively costly in terms of data collection and acquisition of modeling capabilities.

In the Sevier River Basin, real-time operations of the Piute Reservoir must face downstream uncertainties in the form of variations in losses and gains on the river mainstem. These result in travel times from the reservoir to downstream canal diversion points that are uncertain and vary as a function of the quantity of flow in the river and antecedent flow conditions. In addition, downstream canal operators and farmers change their water management decisions on a day-to-day basis to reflect knowledge of current economic and hydrologic conditions. These human behavioral factors place additional uncertainty on the shoulders of the reservoir operator. To contend with these uncertainties, the Piute Reservoir operator would benefit from a tool that would help decide on a near real-time basis how much water to release to meet water orders to canal operators located downstream of the reservoir. In other words, a common requirement for managing the reservoir that is operated on an "on-demand" basis is the anticipation of the quantity of water that must be released while accounting for losses or gains along the river and changing travel times to each downstream canal diversion point. Operation of the Piute Reservoir constitutes a case study were fine resolution decisions have to be made to meet downstream demands that will change in uncertain ways over the period between the time an action is taken to release water from the reservoir and the time when the downstream diversion takes place. A highly instrumented and controlled river basin, such as the Sevier, could be operated on an hourly (or more

frequent) basis if sufficient detail is available in the information describing the present and desired future system state. For the operator of the Piute Reservoir, therefore, the desired output of the model is simply the hourly quantity of water that should be released from the reservoir. The information that must be made available to the model should include the data that describes current, and perhaps recent historical, flow conditions, various climate indices in the basin, and desired downstream canal diversions.

Identification of Inputs

The degree to which a dataset is judged to be of use to better predict optimal system operations decisions could be measured by many different functions, such as mutual information. In this study the relevancy evaluations were subjective as they depend on a perception of the relatedness of the given dataset.

Streamflow is the result of interactions between many hydrologic events, such as precipitation, snowmelt, evapotranspiration, infiltration, and groundwater recharge, and anthropogenic influences, such as reservoir releases and irrigation diversions. In this paper the short-term predictions of required reservoir releases are supported by hourly streamflow data that are available from 2000 to 2003.

Irrigation demands represent the quantities of water that farmers request be delivered to their head gates. Such requests are made one day in advance of when the deliveries are to take place, and are expressed to the reservoir operator by the various canal operators in the form of hourly measurements of canal diversions. Hourly data on these irrigation diversions for all the canals in the study area of the Sevier River are available for the years 2000 through 2003.

Weather information can directly influence the behavior of farmers and canal operators in the basin. The inclusion of temperature, relative humidity, wind speed, solar radiation, and total precipitation data as predictors can enhance the model performance. Historical daily and hourly weather data are available at many weather stations in the Sevier River Basin.

Management Approach

Establishing a reliable dynamic model for providing the necessary information to support advanced water resources decision making has always been an important concern for researchers and field operators. A real-time management approach within the context of data fusion is employed here. Figure 1 provides an abstract description of the set of functions and processes that may comprise a useful paradigm of machine learning and data modeling implementation for real time management of river basin facilities (modified from Mackay, 1992). Having a suitable database, one aims to estimate or predict entity states that provide insightful information to decision makers. This level of induction, which is the very first bold box in Figure 1, could be obtained by non-Bayesian methods, e.g.,

artificial neural networks (ANNs) and support vector machines (SVMs), or with Bayesian solutions such as relevance vector machines (RVMs).

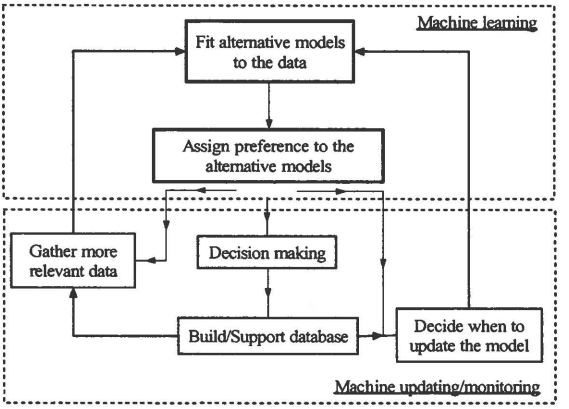

Figure 1: Abstraction of the data modeling process

Owing to its remarkable features and superior performance over non-Bayesian algorithms, RVMs have been adopted to the purpose of the models discussed here. Since there are a considerable number of design choices for RVMs, the second step of machine learning is to assign preferences for the set of plausible models, the second bold box of Figure 1. In this manuscript and for the objective of a fully automated machine, adaptive simulated annealing (ASA) is used to perform this level of inference (Ignber et al., 2004). This is known as models comparison, yet Bayesian methods are also used in this level by employing a quantitative Occam's Razor to penalize the selection of complex models and infer the most plausible model given the data; in other words, Bayesian methods have been implemented to determine the most plausible model in the sense that one maximizes the marginal likelihood of the data $\mathcal{D}$ given the set of hypotheses $\mathcal{H}$ (i.e. different model structure) $p(\mathcal{D}\,|\,\mathcal{H})$.

Having built the most plausible model, the second level of the framework is to use the machine as a decision support system. Hence its performance must be monitored and the model should be updated when needed. There are two events that might occur that indicate the model should be updated. One is the presence of a new data set that has not been previously exploited; the other is the case where there is concept drift, i.e., new trends in the data that the machine has not learned. For this purpose, SVMs have been used to detect an abnormality or novelty and thereby trigger the machine to adapt to these events by retraining.

Figure 2 illustrates the process flow from the raw data to the decision-making level, and where the concepts of model building (RVM), model selection (ASA), and novelty detection (SVM) fit in a fully automated data-driven paradigm.

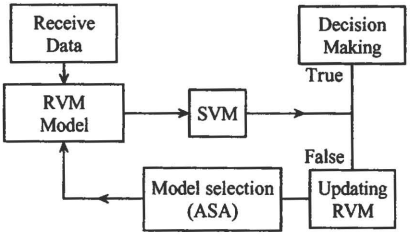

Figure 2: Model structure and process flow to achieve high-level inference.

RESULTS AND DISCUSSION

Reservoir releases must be made to address the conflicting goals of satisfying downstream demands with high certainty, while at the same time conserving water in the reservoir for use later in the season. This problem evolves from the hydraulic and hydrological complexity of the system. The proposed fully automated machine has been implemented for real-world data describing the Sevier River Piute Reservoir water delivery system. The collection of hourly data in the Sevier River basin is ongoing since 2001. Data from the 2001 and 2002 irrigation seasons were used to build the machine, while data from the 2003 irrigation season were used to check the validity of the machine when functioning in real-world conditions. The irrigation season in the basin generally extends from April to the end of October. There are eight inputs for the machine in the form of diversion orders at the different downstream canals, three inputs of streamflow from measurements made along the river downstream of the reservoir, one input representing inflow from Clear Creek (a major gaged tributary that is approximately one day travel time downstream of the reservoir), and one input in the form of a climate index. Clear Creek is an example of an uncontrolled tributary stream that discharges into the river in such a way that its spring and early summer diurnal fluctuations make downstream water management more difficult. The climate index is the first principal component of temperature, relative humidity, solar radiation, wind speed and precipitation.

Model selection involves the selection of values for the parameters so that the model output matches the measured behavior of the system as closely as possible. Kernels present nuisance parameters that are generally determined heuristically.

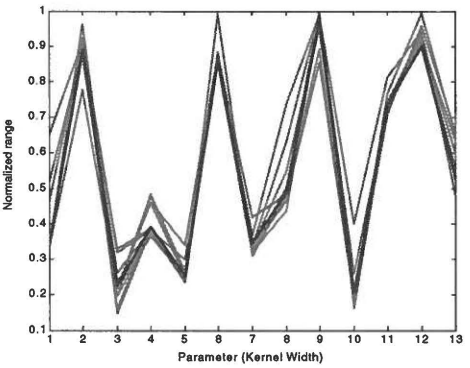

Figure 3: Normalized parameter space, curves corresponded to a set of Pareto solutions (gray), the selected set (solid line).

Arguably, ASA resulted in a sub-optimal kernel parameter data set, as is shown by the model performance illustrated in Figure 3. Douglas et al. (2000) argued that for the purpose of using the model for on-line forecasting, it is desirable to select a single representative parameter set that provides an acceptable trade-off in fitting of the different parts. Thus the bold line in Figure 3 represents the selected kernel parameter set.

Figure 4 shows RVM results for the irrigation season of 2003. Figure 4 also illustrates the 95% confidence interval of the predicted values. The machine removes the redundant features to improve the generalization abilities and it only utilizes 32 relevance vectors (RVs) from the full data set that was used for training (2001 and 2002 irrigation seasons). The RVM ignores the irrelevant inputs to reduce complexity and spurious overfitting. Therefore, it can be used to summarize the information by maintaining the major features of the data set via RVs. Some statistics of interest have been evaluated (Table 1) to test the machine performance (for more details about goodness-of-fit measures, see David and Gregory, 1999). Here, the predicted reservoir releases significantly deviated from the observed releases for the 2003 irrigation season over some time periods, and the confidence interval on the prediction was often very wide. In order to maximize the marginal likelihood of the data, ASA chose a narrow kernel parameter for some input dimensions, thereby suppressing their influence on the forecast release quantity.

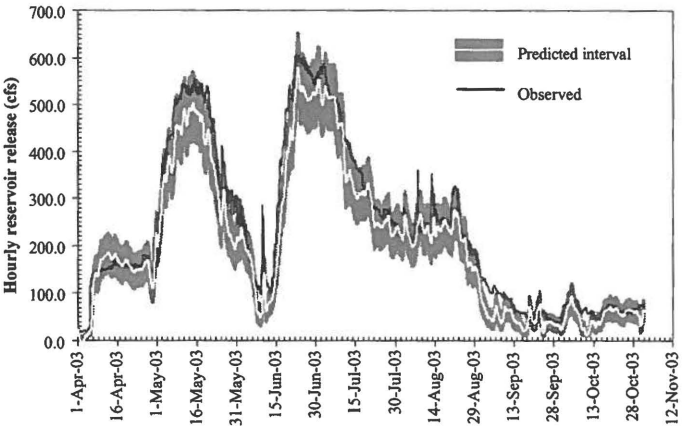

Figure 4: Time series plot of the actual versus predicted releases with confidence intervals.

Again, a SVM has been reformulated to provide a new algorithm, which is in line with Vapnik's principle, for detecting outliers and novelty (Scholkopf, 2000). Figure 5 shows a plot of the output $\rho = \sum_j \alpha_j k(\mathbf{x}, \mathbf{x}_j)$ on the test data set of the Piute Reservoir releases. We used a Gaussian kernel, which has the advantage that the data are always separable from the origin in feature space (Burges, 1998; Cristianini and Shawe-Taylor, 2000; Vapnik, 1995).

Table 1: Machine performance using different statistics. Robust performance measures have been evaluated of the smallest 85% residuals; raw ones are for all the data.

Statistic	RVM	RVM+SVM	RVM	RVM+SVM
	Raw Performance		Robust Performance	
Correlation coefficient (r)	0.983	0.991	0.990	0.997
Coefficient of efficiency (E)	0.932	0.981	0.964	0.993
Bias	-29.762	-4.783	-19.834	-2.627
Root mean square error, cfs	44.080	23.436	29.904	14.951
Mean absolute error, cfs	33.270	17.219	23.961	12.211
Index of agreement (d)	0.982	0.995	0.991	0.998

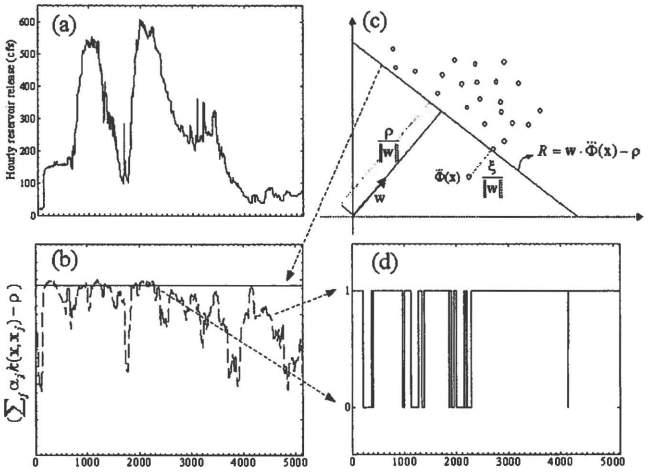

Figure 5: SVM results on the testing data. I (f (x)-ρ) takes the value +1 in a region capturing most of the data points and 0 at the input patterns that show either new trends or outliers.

As shown in Figure 5, the algorithm returns a value of zero to identify outliers and new trends in the data; this triggers the machine to retrain while exploiting all the new data that hasn't been used for training before. Figure 6 shows the results of the machine when linked to the SVM where it has been retrained to account for novelty. Due to this adaptation, the number of RVs increased to 36. The model performs remarkably well and Table 1 provides some statistics of interest for the full paradigm. These statistics further show that the new combination achieves a low error rate.

It is known that abundant data provide robustness (global robustness) for machine learning applications. To ensure good generalization of the inductive learning algorithm given scarce data, the machine has been built on many bootstrap samples from the original dataset to explore the implications of the assumptions made about the nature of the data. Figure 7 shows the results of training using different bootstrap samples.

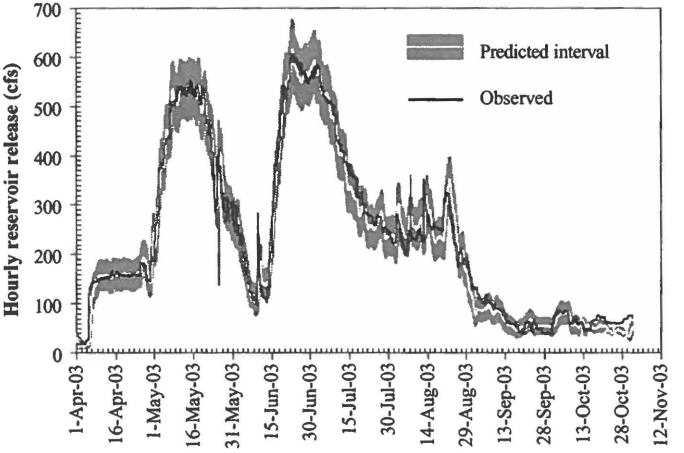

Figure 6: Time series plot of the actual versus predicted releases with confidence intervals for the RVM+SVM machine.

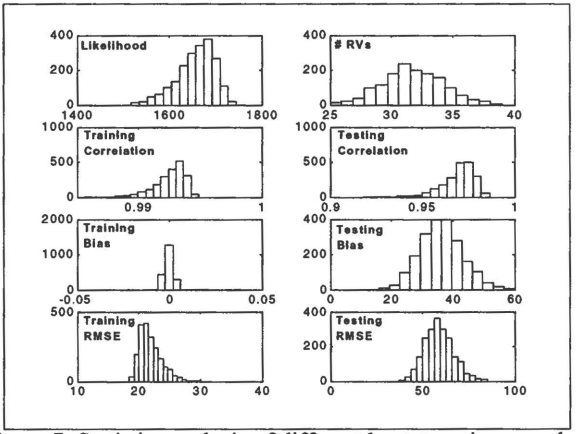

Figure 7: Statistics analysis of different bootstrapping samples.

To utilize the model in near real-time, the predicted reservoir releases can be provided to the reservoir operator, and then it is possible for the operator and and other experts to analyze, judge, and evaluate the results of the machine according to their own knowledge and experience. All in all, the performance results have demonstrated the successful implementations of Bayesian principles (RVM), model selection (ASA), and novelty detection (SVM).

SUMMARY AND CONCLUSIONS

In the Sevier River Basin the principal water problem is twofold: inadequacy and uncertainty of supplies. In this context, a reliable water supply planning policy, specifically during the irrigation season, necessitates acceptably accurate predictions of future water states. In this paper we have presented an operational approach to a decision aid for managing near real-time reservoir releases. While machine-learning techniques have great potential to be used in decision support systems, we believe they have not been fully exploited in water related issues. Growing evidence that there are streamflow variations during the season and from season to season, as well as shifts in climate indices (Kahya and Dracup, 1993; Lins, and Slack, 1999), has lead us to incorporate a novelty detection algorithm in the real-time decision support system. This paper explored the use of unsupervised support vector machines in a sequenced learning technique to recognize behaviors that are outside the norm, and then to trigger the RVM to learn to recognize new patterns.

One could view the present work as an attempt to provide a framework where different algorithms have been fused to better estimate future decisions and detect novelty trends. The approach presented uses a concrete paradigm with well-behaved computational complexity. Beven and Binley (1992) suggested that many models are over-parameterized and therefore result in equifinality. Equifinality is associated with the multiplicity of different possible combinations of values of model parameters. We argue, therefore, that the model structure proposed in this manuscript was formulated so as to avoid equifinality and ensure parameter uniqueness. Parametrically efficient RVMs (sparseness and parsimony), inclusion of many measures each emphasizing a different aspect of model behavior in model selection, and use of structural risk minimization via SVMs, collectively do not conclude equifinality.

Finally, one of the shortcomings of this approach is that, regardless of the parsimony of the model structure that reflects the most dominant characteristics of the system, it cannot be seen how a meaningful physical interpretation can be extracted from the resulting model definition. In spite of this, we believe that the imposed novelty detection tool ensures the persistence of the basic dominant characteristics and reflects any abnormality. One might be able to interpret such events in physically meaningful contexts. Another seemingly unavoidable disadvantage of the algorithm that handles novelty detection is that it detects an abnormality only after new data are available, that is after the learning algorithm performance starts to depreciate (Klinkenberg and Joachims, 2002; Olivier et al., 1999).

REFERENCES

Berger, B., R. Hansen, A. Hilton. 2002. Using the World-Wide-Web as a Support System to Enhance Water Management. The 18th ICID congress and 53rd IEC meeting Montréal, Canada.

Beven, K. J., and A. M. Binley. 1992. The Future of Distributed Models: Model Calibration And Predictive Uncertainty, Hydrological Process, 6, 279–298.

Burges, C. J. C. 1998. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2 (**2**), 121-167.

Cristianini, N., and J. Shawe-Taylor. 2000. An Introduction to Support Vector Machines. Cambridge: Cambridge University Press.

David, R. L., and M. J. Gregory. 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resources Research, 35 (1), 233–241.

Douglas, P. B., H. V. Gupta, and S. Sorooshian. 2000. Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods. Water Resources Research, 36, (12), 3663–3674.

Ignber, L., and contributors, 2004. Adaptive Simulated Annealing, C-Source code, http://www.ingber.come/#ASA-CODE.

Kahya, E., and J. A. Dracup. 1993. U.S. streamflow patterns in relation to the El Niño/southern oscillation, Water Resources Research, 29 (8), 2491-2504.

Klinkenberg, R., and T. Joachims. 2000. Detecting Concept Drift with Support Vector Machines. In Langley, P., Proceedings of ICML-00 (Eds.), 17th International Conference on Machine Learning, 487–494. Stanford, US: Morgan Kaufmann Publishers, San Francisco.

Lins, H. F., and J. R. Slack. 1999. Streamflow Trends In The U.S, Geophysical Research Letters, **26**, 227–230.

MacKay, D. J. 1992. Bayesian Methods for Adaptive Models, Ph.D. thesis, Dept. of Computation and Neural Systems, California Institute of Technology, Pasadena, CA.

Olivier, C., V. Vapnik, and J. Weston. 1999. Transductive Inference for Estimating Values of Functions. Neural Information Processing Systems, **12**, 421-427

Scholkopf, B., R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt. 2000. Support Vector Method For Novelty Detection, In Solla S. A., T. K. Leen, and K.-R. Mller, editors, Advances in Neural Information Processing Systems, 12,582-588, MIT Press, Cambridge.

Vapnik, V. N. 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York.