Simulation of California's Major Reservoirs Outflow Using Data Mining Technique

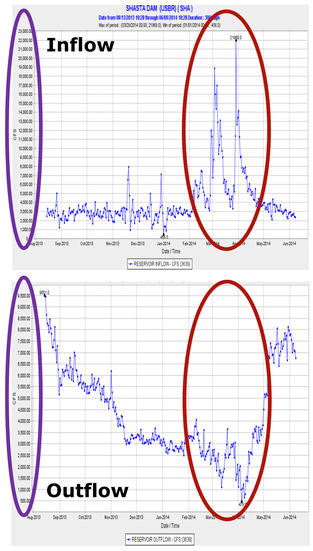
Tiantian Yang, Xiaogang Gao, and Soroosh Sorooshian

Center for Hydrometeorology and Remote Sensing (CHRS) : University of California, Irvine

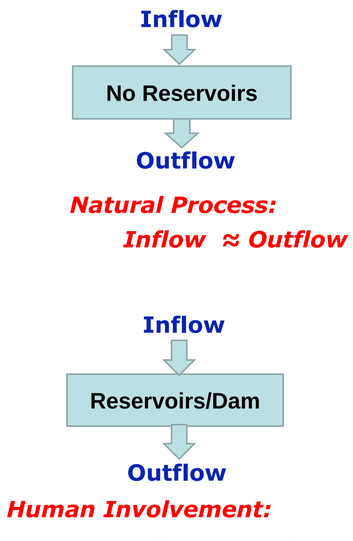
Poster Number:H13A-1054

Introduction

- After a reservoir is built, there is a huge difference between the natural upstream inflows to the reservoirs and the controlled outflows from the reservoirs. The outflows are the most important water input to the downstream users for multiple purposes.
- In order to simulate the complicated outflow decision making process, and extract the reservoir operation patterns for the California's 12 major reservoirs, we build a reservoir outflow simulation tool, which incorporates one of the well-developed statistical and graphical models (decision regression and classification tree) in the field of data mining.



Figure(1) Inflow and outflow for Shasta Lake

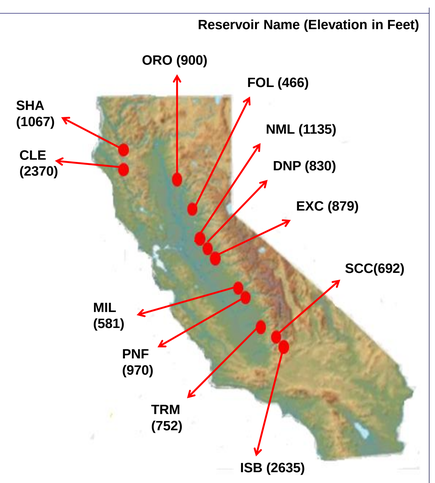


Motivation

- ✓ Use the patterns derived from data mining technique to predict medium to long term (1-3 years), high temporal resolution (daily) reservoir outflows.
- ✓ Utilize the shuffled validation to improve the model reliability and accuracy in predicting the future reservoir operation scenarios.
- ✓ Take the advantage of CART to efficiently extract the importance factors in the decision making process for 12 major reservoirs in California.
- ✓ Provide the downstream users with confident the water availability reference for better water planning and management.

Data

- Source: California Data Exchange Center (CDEC)
- Stations: California's 12 Major reservoirs with varying elevations
- Types of inputs:
  - A. Natural Indicators:
    - Inflow
    - Precipitation
    - Evaporation
  - B. Climate Indicator:
    - 8 River Indexes from DWR River Forecast Branch
  - C. Policy Indicator:
    - State Water Project (SWP) Planned Allocation Amount from the SWP administrative office.
- Temporal Resolutions: Daily
- Length: 10/01/1997 to 12/31/2013



Figure(2) 12 Major Reservoirs in California

Methodology

- Classification and Regression Tree (CART): a decision support tool that uses a tree-like graph model to classify, and predict continuous target variable based on the selected dependent variables or decision variables.
- Advantages: the transparency of modelling framework, the simplicity for understanding and interpreting, and the computational efficiency

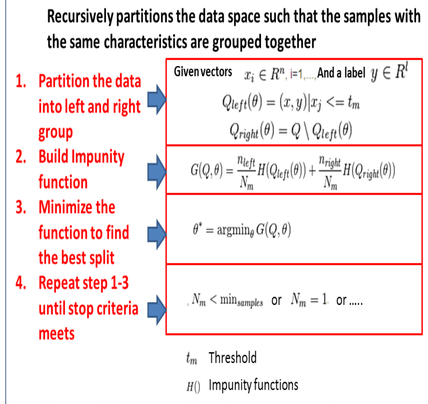


Figure (3): Generalized Pressures for CART

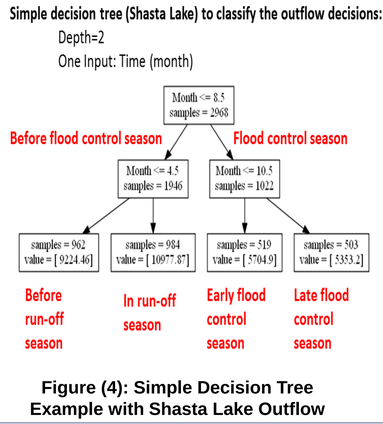
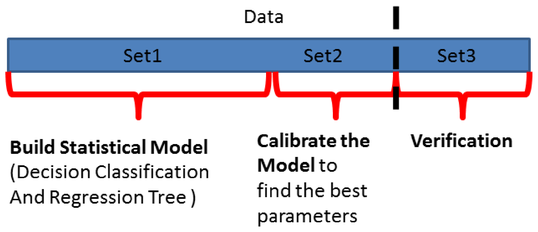


Figure (4): Simple Decision Tree Example with Shasta Lake Outflow

Shuffled Cross Validation

- Cross Validation: is a model validation technique for evaluating predictive performance of a statistical model on an independent or unseen data set.
- Purpose of Outer Loop : To create enough unseen data , which is being tested for selecting proper model parameter against the "over-fitting" phenomena. (1-50,000 in the experiment)
- Purpose of Inner Loop :To find the maximum like-hood of CART parameter (maximum decision tree depth) that gives reliable predictions in the possibility density distribution. (2-15 in the experiment)



To overcome the over-fitting disadvantage:

1. Hold on ¼ data out for verification (Set 3)
2. Shuffle the rest data points (Set 1+ Set 2)
3. Build a decision tree with depth = k (Set 1)
4. Test on the model performance (Set 2)
5. k=k+1 , repeat step 3 and 4
6. Terminate if k > max depth
7. Find the best performance and store the associated depth
8. Repeat step 2- 7 for large number of time (say 5000)
9. Use the possibility density function to find the best depth
10. Test on Set 3

Outer Loop

Inner Loop

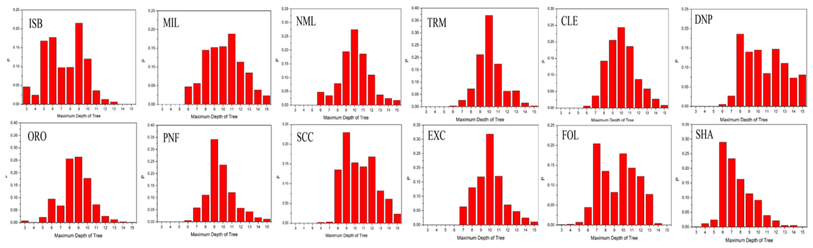


Figure (6) : The possibility density function of best decision tree depth for 12 major reservoirs in California.

Results

- Using the optimal decision tree depth, we derive the outflow/storage change predictions (red) for 2009/12/31-2013/12/31, and compare with the "real" (black)

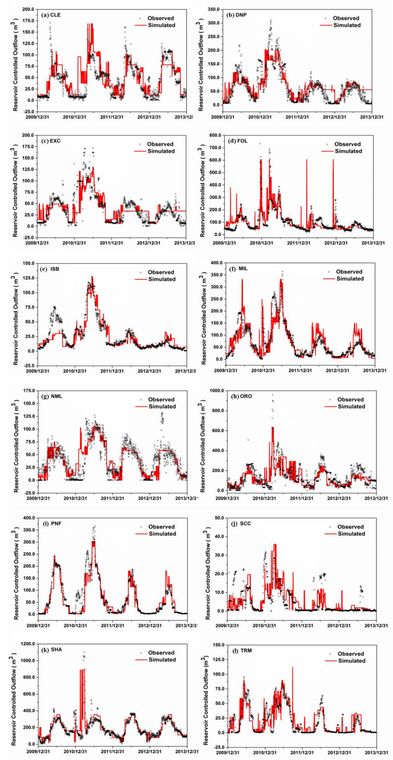


Figure (7) : Predicted vs. observed outflows

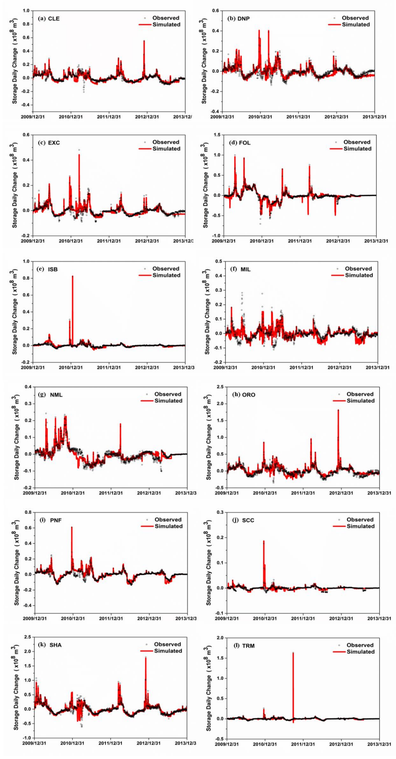


*Figure (8) : Predicted vs. observed storage changes

- The statistical test (Nash-Sutcliffe Coefficient and Standard Deviation)

	ORO	FOL	SHA	NML	DNP	EXC	MIL	PNF	TRM	CLE	SCC	ISB
Nash	0.517	0.511	0.534	0.568	0.628	0.557	0.611	0.851	0.634	0.347	0.351	0.764
Std for 30 runs	0.159	0.119	0.198	0.137	0.059	0.096	0.025	0.052	0.037	0.161	0.034	0.018

- The Predictors and importance factors (I=Inflow, SA=SWP Alloc, ID=River Index)

	ISB	CLE	NML	SHA	PNF	ORO	EXC	DNP	TRM	SCC	MIL	FOL
1st (%)	I (56%)	SA (56%)	M (86%)	I (74%)	I (69%)	I (39%)	I (68%)	I (75%)	I (57%)	I (61%)	I (89%)	I (90%)
2nd (%)	ID8 (31%)	M (28%)	SA (6%)	M (21%)	M (29%)	SA (34%)	M (11%)	M (17%)	M (39%)	M (21%)	M (6%)	M (5%)
3rd (%)	M (10%)	ID4 (4%)	ID1 (5%)	SA (3%)	SA (1%)	M (19%)	SA (10%)	ID2 (4%)	ID8 (2%)	SA (3%)	SA (4%)	SA (3%)
Elev (ft)	3370	2370	1135	1067	970	900	879	830	752	692	581	466

Conclusion

- The proposed data-driven, statistical model is able to provide accurate, efficient and reliable the long term (1-3 years), daily outflow simulation.
- The 12 reservoir's outflow decision making patterns and rules are analyzed based on the predictability of natural/climate/policy indicators.
- The method is universally flexible to other reservoirs in the world.

References and Acknowledgement

[1] Breiman, L., Friedman, J., Stone, C.J. and Olshen, R.A. (1984) Classification and regression trees, CRC press.

[2] Bessler, F.T., Savic, D.A. and Walters, G.A. (2003a) Water reservoir control with data mining. Journal of Water Resources Planning and Management-Asce 129(1), 26-34.

The authors would like to thank the UC-Laboratory Fees Research Program (Award # 09-LR-116849-SORS), the NASA Decision Support System Program (Award # NNX09AO67G), the CDWR Seasonal Forecasting via Database Enhancement Program (DWR Agreement No.4600010378), NSF CyberSEES project (Award # CCF-1331915) and University of California, Irvine for their contribution and support.