Title
A Framework to Provide Optimal Management Strategies for California’s Reservoirs in Achieving Sustainable Water Supply and High Hydropower Productivity

Permalink
https://escholarship.org/uc/item/9gj589tm

Author
YANG, TIANTIAN

Publication Date
2015

Peer reviewed|Thesis/dissertation
A Framework to Provide Optimal Management Strategies for California’s Reservoirs in Achieving Sustainable Water Supply and High Hydropower Productivity

Dissertation

Submitted in Partial Fulfillment of the Requirements for the degree of

DOCTOR OF PHILOSOPHY

in Civil Engineering

by

Tiantian Yang

Dissertation Committee:
Professor Soroosh Sorooshian, Chair
Professor Xiaogang Gao, co-Chair
Professor Amir AghaKouchak

2015
# TABLE OF CONTENT

LIST OF TABLES ........................................................................................................ v

LIST OF FIGURES ..................................................................................................... vi

LIST OF ACRONYMS AND ABBREVIATIONS ........................................................ viii

ACKNOWLEDGMENTS ............................................................................................. xi

CURRICULUM VITAE ................................................................................................ iii

ABSTRACT OF THE DISSERTATION ...................................................................... vi

1. Introduction .............................................................................................................. 1

   1.1 Water-Energy Nexus in California ................................................................. 1

       1.1.1 Background on California’s Water Supply ........................................... 2

       1.1.2 Background on California’s Hydropower Generation ......................... 4

       1.1.3 New Challenges for Reservoir Operation in California ....................... 6

   1.2 Research Motivation and Research Tasks ....................................................... 9

   1.3 Organization of the Dissertation ..................................................................... 12

2. Simulating Reservoir Releases Using Data-mining Techniques ........................... 13

   2.1 Introduction ..................................................................................................... 13

   2.2 Data and selected reservoirs .......................................................................... 19

   2.3 Methodology .................................................................................................. 26

       2.3.1 Classification and Regression-Tree (CART) algorithm ....................... 26

       2.3.2 Random Forest Algorithm .................................................................... 27

       2.3.3 Enhancement of Model Calibration ....................................................... 28

       2.3.4 Gini Diversity Index ............................................................................. 31

   2.4 Results ............................................................................................................ 33

       2.4.1 Sensitivity Analysis of Decision Tree Depths ....................................... 33

       2.4.2 Comparison of Simulated Controlled Outflows ................................. 33
4.3 Results ........................................................................................................................................93
  4.3.1 Settings ..................................................................................................................................93
  4.3.2 Uncertainty of Model ..............................................................................................................94
  4.3.3 Simulation Results for Dry, Wet, and Normal Years ..............................................................95
4.4 Discussion ....................................................................................................................................98
4.5 Conclusion .................................................................................................................................101
5. Summary and Recommendation ...................................................................................................105
  5.1 Dissertation Summary ..............................................................................................................105
  5.2 Major Findings and Contributions ...........................................................................................105
  5.3 Future Research Recommendations .........................................................................................108
REFERENCES: ................................................................................................................................111
LIST OF TABLES

Table 2-1 Detailed Information on Selected Reservoirs and Observation Stations................20
Table 2-2 Detailed information on the decision and target variables .................................26
Table 2-3 Statistics comparison between the observed reservoir controlled outflow and
simulated results with different methods, including combined CART and Shuffled
Cross Validation (CART+SCV), CART with 2-fold cross-validation (CART Ctrl),
and Random Forest (RF);..................................................................................................38
Table 3-1 Detail Information on Test Functions................................................................69
Table 3-2 Settings for each test function ............................................................................71
Table 3-3 Diversity metric for MODE, MOGA, MOSA, MOSPD, MOCOM and
MOSPD on the test functions .........................................................................................76
Table 3-4 Convergence metric GD for MODE, MOGA, MOSA, MOSPD, MOCOM,
MOCOM and MOSPD on the test functions ................................................................77
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 2-1 Locations of the selected reservoirs with elevations in parentheses (m)</td>
<td>23</td>
</tr>
<tr>
<td>Figure 2-2 Flowchart of the shuffled cross-validation scheme</td>
<td>31</td>
</tr>
<tr>
<td>Figure 2-3 The frequency histograms of CART tree depths using both shuffled cross-validation and standard 2-fold cross validation for the selected 9 major reservoirs in California</td>
<td>35</td>
</tr>
<tr>
<td>Figure 2-4 Reservoir controlled outflow comparison between observed daily releases (black) with the simulated releases with CART combined with shuffled cross-validation scheme (red), CART without shuffling scheme as control run (blue), and random forest (green)</td>
<td>37</td>
</tr>
<tr>
<td>Figure 2-5 Reservoir daily storage changes comparison between observed storage changes (black) with the calculated results with CART combined with shuffled cross-validation scheme (red), CART without shuffling scheme as control run (blue), and random forest (green)</td>
<td>41</td>
</tr>
<tr>
<td>Figure 2-6 Reservoir storage trajectory comparison between the actual storage volume (black) with the calculated results with CART combined with shuffled cross-validation scheme (red), CART without shuffling scheme as control run (blue), and random forest (green)</td>
<td>42</td>
</tr>
<tr>
<td>Figure 2-7 The normalized Gini diversity index or the importance for each decision variable for each reservoir</td>
<td>44</td>
</tr>
<tr>
<td>Figure 3-1 Flowchart of the MOCOM algorithm with enhancement modules (grey-dashed boxes)</td>
<td>60</td>
</tr>
<tr>
<td>Figure 3-2 Possibility-adjustment module</td>
<td>65</td>
</tr>
<tr>
<td>Figure 3-3 Dimension monitoring and restoring module</td>
<td>66</td>
</tr>
<tr>
<td>Figure 3-4 Test results of MODE, MOGA, MOSA, MOPSO, MOSPD, and MOCOM</td>
<td>73</td>
</tr>
</tbody>
</table>
Figure 3-5 Evolution process of MOCOM and MOSPD on the SCH function (a) and (b),
the POL function (c) and (d), the FON function (e) and (f), and the KUR function (g)
and (h). ...........................................................................................................74

Figure 3-6 Evolution process of MOCOM and MOSPD on ZDT1 function (a) and (b),
ZDT2 function (c) and (d), ZDT3 function (e) and (f), and ZDT4 function (g) and (h)
..................................................................................................................................75

Figure 4-1 SWP and OTC (Courtesy of the California Legislative Analyst's Office and
the California DWR) .................................................................................................83

Figure 4-2 The configurations of Oroville Thermalito Complex (OTC).........................88

Figure 4-3 (a) Storage-elevation curve of Lake Oroville; (b) Storage-elevation curve of
Thermalito Forebay including power canal and pool; (c) Storage-elevation curve of
Thermalito Afterbay. Sum of squared residuals for (d) Lake Oroville, (e) Thermalito
Forebay including power canal and pool, and (f) Thermalito Afterbay........................92

Figure 4-4 The comparison of the objective function values between the real operation
scenarios and 25 independent runs of the model generated scenarios using the
randomly sampled initial parameters for April, May June in 1998, 2000, and 2001.......95

Figure 4-5 Simulation results for MOCOM and MOSPD for April, May, and June in
1998, 2000, and 2001 ...............................................................................................96

Figure 4-6 Comparison of different S-E curve-fitting methods for April, May, and June
in 1998, 2000, and 2001 ..........................................................................................97

Figure 4-7 Two Non-dominated solutions with extreme objective function values for the
releases from (a) the Thermalito Forebay area, (b) from the Thermalito Afterbay to
the Feather River. The storages volumes for (c) Thermalito Forebay including
Power Canal and Pool, and (d) Thermalito Afterbay .............................................98
# LIST OF ACRONYMS AND ABBREVIATIONS

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>CART</td>
<td>Classification and Regression Tree</td>
</tr>
<tr>
<td>CDEC</td>
<td>California Data Exchange Center</td>
</tr>
<tr>
<td>CDWR O/M</td>
<td>CA Department of Water Resources/O &amp; M</td>
</tr>
<tr>
<td>CDWR S/S</td>
<td>CA Dept of Water Resources/Snow Surveys</td>
</tr>
<tr>
<td>CVP</td>
<td>Central Valley Project</td>
</tr>
<tr>
<td>DG</td>
<td>Downstream Gauge</td>
</tr>
<tr>
<td>EP</td>
<td>Ecosystem Protection</td>
</tr>
<tr>
<td>EA</td>
<td>Evolutionary Algorithm</td>
</tr>
<tr>
<td>FC</td>
<td>Flood Control</td>
</tr>
<tr>
<td>HP</td>
<td>Hydropower</td>
</tr>
<tr>
<td>MID</td>
<td>Merced Irrigation District</td>
</tr>
<tr>
<td>MOCOM-UA</td>
<td>Multi-Objective Shuffled Complex Evolution Global Optimization Algorithm</td>
</tr>
<tr>
<td>MODE</td>
<td>Multi-Objective Differential Evolution</td>
</tr>
<tr>
<td>MOGA</td>
<td>Multi-Objective Genetic Algorithm</td>
</tr>
<tr>
<td>MOSA</td>
<td>Multi-Objective Simulated Annealing</td>
</tr>
<tr>
<td>MOSPD-UCI</td>
<td>Multi-Objective Shuffled Complex Evolution Global Optimization Algorithm</td>
</tr>
<tr>
<td></td>
<td>with Principal Component Analysis and Crowding Distance</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>USBR</td>
<td>U.S. Bureau of Reclamation</td>
</tr>
<tr>
<td>USGS</td>
<td>US Geological Survey</td>
</tr>
<tr>
<td>WRD</td>
<td>Weaverville Ranger District</td>
</tr>
<tr>
<td>WS</td>
<td>Water Supply</td>
</tr>
<tr>
<td>WYI</td>
<td>Water Year Index</td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENTS

First, I would like to thank my parents, Xinglai Yang and Jin Fan for giving me the opportunity to study in the United States, and my wife, Meiqi Li for her truly love and care during my Ph.D. program. All of this could not been possible without their supports and encouragements.

I also wish to thank Dr. Xiaogang Gao who provided many useful insights and encouragement and co-directed this research. Finally, I wish to thank Dr. Soroosh Sorooshian for his utmost support both intellectually and financially for making this work possible. To all these individuals, I am deeply indebted.
CURRICULUM VITAE

Tiantian Yang

EDUCATION

PhD  University of California, Irvine, Civil Engineering  Dec 2015

Dissertation: “A Framework to Provide Optimal Management Strategies for California’s Reservoirs in Achieving Sustainable Water Supply and High Hydropower Productivity”

Committee:
Soroosh Sorooshian (chair), Xiaogang Gao (Member), and Amir AghaKouchak (Member)

MS University of California, Irvine, Mechanical and Aerospace Engineering  June 2010

BS Tsinghua University, School of Aerospace  June 2009

HONORS AND AWARDS

UCI Association of Graduate Students Travel Grant (2015)
UCI Association of Graduate Students Travel Grant (2013)
UCI Dept. of Mechanical and Aerospace Engineering Graduate Student Fellowship (2010)
Tsinghua University Outstanding Student Fellowship of 2007-2008 (2008)
Tsinghua University Outstanding Student Fellowship of 2006-2007 (2007)
Tsinghua University Outstanding Student Fellowship of 2005-2006 (2006)
China National Merit Student Award for 2004-2005 (2005)

RESEARCH EXPERIENCE

University of California Irvine, Irvine, California  2010 to 2015
Research Assistant, CENTER OF HYDROMETEOROLOGY AND REMOTE SENSING
    ● Reservoir operation and modelling, optimization algorithm development, system decision making support

Tsinghua University, Beijing, China,  2005 to 2009
Student Researcher, CENTRAL LABORATORY OF STRENGTH AND VIBRATION
    ● Studying the composite metal material under erosion of hydrogen environment
    ● Quantifying the dynamical mechanic and cracking processes of bamboo materials
TEACHING EXPERIENCE

University of California Irvine, Irvine, California  September 2014 to December 2014  
**Lecturer**, Civil Engineering Department  
CEE 299, Prof. Xiaogang Gao, Matlab and Python Programming in Engineering Applications

University of California Irvine, Irvine, California  September 2013 to December 2013  
**Grader**, Civil Engineering Department  
CEE 30, Prof. Lizhi Sun, Statics

University of California Irvine, Irvine, California  Jan 2013 to April 2013  
**Teaching Assistant**, Civil Engineering Department  
CEE 283, Prof. Xiaogang Gao, Mathematical Methods of Engineering Analysis

PUBLICATIONS


Yang, T.*, X. Gao, S. Sorooshian, “Simulating California Reservoir Operation Using the Classification and Regression-Tree Algorithm Combined with a Shuffled Cross-Validation Scheme”, *Water Resources Res.*, (In moderate revision)


Yang, Z.*, H. Liu, T. Yang, X. Xu, “A path-based structural decomposition analysis of


**PRESENTATIONS AND INVITED LECTURES**


**COMPUTER SKILLS**

**Programming**: Python, MatLab, R, C++

**Applications**: HEC-GeoHMS, HEC-RAS, ArcGIS, ENVI

**Platforms**: Linux, Windows, Mac OS
ABSTRACT OF THE DISSERTATION

A Framework to Provide Optimal Management Strategies for California’s Reservoirs in Achieving Sustainable Water Supply and High Hydropower Productivity

By

Tiantian Yang

Doctor of Philosophy in Civil Engineering

University of California, Irvine, 2015

Professor Soroosh Sorooshian, Chair

With the increasing demands on freshwater water and clean energy due to population growth and impacts of climate change, the stresses on natural resources are increasing worldwide. Therefore, efficient operation of reservoir systems with the intention of optimizing sustainable water supply and hydropower production is crucially needed by policy and decision makers, and water users. In this dissertation, a framework, including analysis of reservoir controlled outflows, optimization algorithm development, and realistic reservoir modelling, is presented and demonstrated in Chapter 2, 3 and 4, respectively.

In Chapter 2, a Classification And Regression Tree (CART) algorithm with an enhanced cross-validation scheme is applied to simulate the human controlled outflows in 9 major reservoirs in California. The proposed approach is capable of incorporating multiple types of information into decision making and mathematically quantifying how releases are related to many decision variables. A verification study has been carried out in 9 major reservoirs in California. Without any prior
information, the model is able to identify that the historical operation in Oroville Lake, Shasta Lake and Trinity Lake are highly dependent on policy and regulation, while the reservoirs with low elevations are sensitive to reservoir inflows. The approaches developed in this chapter serve as the analytical tool to help understand reservoir operation.

In Chapter 3, an enhanced multi-objective global optimization technique is developed in order to better address multiple conflicting interests from decision makers when water and energy related objectives are jointly considered in reservoir operation. A comparison study has been conducted comparing the enhanced algorithm with multiple cutting-edge multi-objective heuristic search algorithms on various test functions. Results show the enhanced algorithm has superior performance regarding diversity and convergence measures over the other algorithms.

Last, a newly developed cascade reservoir optimization model for the Oroville-Thermalito Complex (OTC) in northern California is presented in Chapter 4. Multiple alternative operation strategies that maximize sustainable water supply and hydropower production are derived and recommended for the OTC’s operation under various dry/wet conditions. The suggested optimal operation alternative will be intuitive for reservoir operators to further adjust and improve current reservoir operation strategy and planning.
1. Introduction

1.1 Water-Energy Nexus in California

In California, as in many other places around the world, water and energy are two indispensable resources for societal prosperity. However, water and energy are essentially and closely related to each other (so called the “water-energy nexus”). In water sectors, a lot of energy is used to pump, transport and treat water. In energy sectors, plenty of water is consumed during energy production, such as cooling unclear and coal-fired electric generators and producing hydropower. Typically, the water and energy sectors are embedded in water transferring projects. For example, California’s State Water Project (SWP), the largest state-built water transferring project providing water for two-thirds of California’s population, consists of 34 reservoirs, 20 pumping plants, 4 pumping-generating plants, and 5 hydroelectric power plants. The SWP is not only the largest single electricity user in California, but also the SWP itself produces millions KWhs of electric power via its impoundment, run-of-river and pumped storage hydropower facilities.

Since the last decade, the water-energy nexus in California and the western U.S. has been acknowledged by many government agencies, research institutions, and industrial organizations [Ackerman and Fisher, 2013; Burt et al., 2003; CEC, 2005a; Dale et al., 2009; Harto et al., 2011; Kenney and Wilkinson, 2011; Larson et al., 2007; Lofman et al., 2002; Nair et al., 2014; Tanaka et al., 2006; Tarroja et al., 2014a; b; Robert Wilkinson, 2000]. Because of the demands for both water and energy are expected to sharply rise due to population growth, and the changing climate imposes great challenges on current water and energy management, researchers suggest conjointly managing these
two resources by coordinating the state’s water and energy policies and facilities in a manner of increasing system sustainability and resiliency.

Therefore, in this dissertation, three major elements in reservoir operation are considered to achieve the sustainable water supply and high hydropower productivity. The first element is the approach that improves the capabilities of simulating and analyzing historical reservoir release patterns to better identify the important factors in current reservoir operations. The second element is the optimization tool in supporting decision making and the last element is the provision of realistic reservoir model, in which multiple reservoir operation alternatives are able to be derived to achieve system sustainability and resiliency.

In this chapter, a detailed background on California’s water supply and hydropower is provided. Then, the new challenges and requirements from government agencies for operating California’s reservoir system are summarized. Last, the research motivation and research tasks are introduced.

1.1.1 Background on California’s Water Supply

California’s water systems are unimaginably complex and linked to every facet of natural resources, the State’s economic activity, and public safety. The challenge for California’s water system is due to the fact that the major water-abundant areas are geographically dislocated from the water demand areas. Association of California Water Agencies (ACWA) [1995] summaries that around 75% of California’s water supply comes from north of Sacramento, while 80% of the water demand occurs in
the central and southern parts of the state. In addition, the state’s ecosystem, agricultural, and urban water users have variable water demands for the quantity, timing, and place of use. Because of these unique characteristics of California’s water supply and demands, many water transferring projects, including the State Water Project, the Central Valley Project, the Colorado River Aqueduct, as well as a number of local projects (Los Angeles Aqueduct, the Tuolumne River Aqueduct, the Hetch Hetchy Aqueduct, and the Mokelumne Aqueduct), have been built to timely store, convey, collect, treat and deliver fresh water to ecosystem, agricultural, and urban water users.

In California, precipitation is the primary source of the state’s water supplies, and it varies from place to place, season to season, and year to year. Most of the snowfall and rainfall occurs in the mountains in the northern and eastern areas, and most water is used in the central and southern valleys and along the coast. Other sources include groundwater extraction, and water imports from Colorado, Oregon and Mexico via rivers and channels. According to California Department of Water Resources (CDWR) statistics [DWR, 2013d], approximately, 50-60 percent of the total supply is used by vegetation, evaporates to the atmosphere, provides managed wetlands, and flows to other states, the Pacific Ocean and salt sinks. The remaining 40-50 percent is distributed among urban and agricultural uses, as well as to ecosystem protecting and restoring the environment, or as storage in surface and groundwater reservoirs for later use.

Within the water projects in California, reservoirs/dams are the most vital infrastructures to provide water sustainability for the system. According to CDWR Division of Safety of Dams statistics, more
than 1200 dams/reservoirs have been built in California to store fresh surface water resources with the purpose of timely irrigating thousands acres of farmland and supplying water to millions of residents and industry entities. Reservoirs play an incredibly important role in providing the flexibility and resilience for California’s water projects in diverting water from one area to another and mitigating the impacts from extreme climate events, such as flood and drought. As the changing climate and population growth in California continue to challenge California’s water delivery projects, effective and adaptable models and tools for reservoir management, which are capable of representing the complex nature of the system and its requirements, are critically needed.

1.1.2 Background on California’s Hydropower Generation

According to [EIA, 2013], hydropower supplies about 7 percent of total electricity generation of the entire U.S. In California, Hydropower contributes on average about 15 percent of California’s total in-state generation, and it is largely dependent on the water supply conditions [Brown et al., 2014]. One of the major benefits of hydropower is the provision of carbon-free, pollutant-free and load-following electric generation. Compared to other renewable energy sources, such as solar and wind, hydropower is not highly dependent on weather condition and day-night shift. This characteristic of hydropower allows the continuous provision of clean energy to the electric grid. Furthermore, hydropower has also provided ancillary services to support and maintain the reliability of the electric grid during normal operation and contingency events, especially in the hot months of the year when electricity demand is at its peak [Brown et al., 2014]. The flexibility of hydropower generation is due to the fact that turbines can be turn on and off quickly, while nuclear, natural gas
and coal-fire thermal generators cannot. In California, large hydropower provided up to 80% of the annual spinning reserve capacity and 30-45% of the annual regulation up and regulation down capacity in non-drought years, according to Federal Energy Regulatory Commission data.

According to my personal communication with CDWR reservoir operators and engineers in operation branches, hydropower normally ranks 3rd or 4th priorities in most of the California’s reservoir operation, followed by water supply and flood control purposes. The hydropower facilities in California are under joint operation by multiple agencies, i.e. CDWR, U.S. Bureau of Reclamation (USBR), California Department of Fish and Wildlife (CDFW), and local utility agencies. The reservoir operation largely relies on empirical regression models, and manually monitoring and adjustment by on-site hydrologist and engineers. An applied linear reservoir model by CDWR is called the Water Resource Integrated Modeling System (WRIMS model engine or WRIMS) (formally named CALSIM), which is a generalized water resources modeling system for evaluating operational alternatives of large, complex river basins. The main purpose of CALSIM is to simulate California State Water Project (SWP)/Central Valley Project (CVP) operations for planning purposes. More detailed introduction about pros and cons of CALSIM and other reservoir operation models can be found in Chapter 4.

In California, most of the large hydro plants are located on dammed lakes. Oroville Lake is the second largest lake in California. According to the statistics of USBR et al. [2013], the hydro facilities in
Oroville Lake and its adjacent Thermalito Forebay and Afterbay together have a total installed capacity of 762 MW. The annual hydropower generation ranks the first among all hydropower plants, which is about 2.2 million kWh. Shasta Lake is the largest lake in California. The Shasta hydropower plant has an installed capacity of 676 MW and provides 1.6 to 2.1 million kWh of electric power annually. The Trinity hydropower plant and Folsom lake hydropower plants rank as the third and fourth largest hydropower plants in California and are all located on the dammed lakes. Even though the involved case studies in this dissertation are only limited to the major reservoirs in California, the analytical approaches developed in Chapter 2 are universally adaptive to other reservoirs; the optimization tools presented in Chapter 3 are flexible to be used reservoir operation and other related field; and in Chapter 4 the non-linear reservoir optimization model applied to the Oroville-Thermalito Complex is able to be used in other cascade reservoir system.

1.1.3 New Challenges for Reservoir Operation in California

Even though the California’s water system is well built, maintained, and operated for decades, there are still many challenges need to be overcome, and improvement can be made. Generally, due to the impacts of climate change and blossom of population and its associated demand increase, the California’s water system is undergoing increasing pressures for maintaining sufficient water supply to ecosystem, agriculture, residential and industry users. For example, because of the persistent drought condition from 2012, on January-31-2014 California Governor Brown made an official drought declaration to public. Followed by the drought declaration, the CDWR announced an amendment to the SWP allocation, in which the SWP allocation to farmers and agricultural water
agencies is dropped to 0%. According to CDWR news release on Jan-31-2014, it means that farmers and 29 SWP agencies will no longer receive water allocation until further notice. Both local water agencies and users are more and more aware of the fact that policy and regulation are playing an extremely important role in the California’s water system operation.

In addition, people begin to realize the fact that water and energy are intrinsically bounded together. Besides the large numbers of government reports and scientific journals mentioned in Section 1.1, many government activities and changing of laws also indicate the importance and necessity of jointly managing water and energy:

1. DWR [2013d] concludes that “California’s water-related assets and services are provided by many interdependent systems that historically have been managed on a project-by-project basis and resources have been managed independently. This lack of systemic planning and management has contributed to an assortment of ongoing and emerging crises, as well as increased probability of large-scale social catastrophes.”

2. U.S. Department of Energy has released a pilot research opportunity and challenge report. According to the report [DOE, 2014], the first three of six Strategic Pillars to Address the Water-Energy Nexus issues are: “1. Optimize the freshwater efficiency of energy production, electricity generation, and end use systems; 2. Optimize the energy efficiency of water management, treatment, distribution, and end use systems; and 3. Enhance the reliability and
(3) One California official bill passed in 2014, which is called the California’s First Update to the Scoping Plan AB32, requires multiple state agencies in California to achieve the near-term 2020 greenhouse gas limit and be well positioned to maintain and continue reductions beyond 2020. Among multiple clean energy sources, hydropower generation is specifically emphasized to mitigate demands from other conventional electric power generation sources.

(4) A special action team, termed the Water-Energy Team of the Climate Action Team (WET-CAT), is formed by multiple state agencies including California Air Resources Board, California Environmental Protection Agency, California Department of Food and Agriculture, California Department of Public Health, California Department of Water Resources, California Public Utilities Commission, California Energy Commission, California Natural Resources Agency, Governor's Office of Planning and Research, State Water Resources Control Board, and Strategic Growth Council. The WET-CAT team is tasked with coordinating efforts to reduce greenhouse gas (GHG) emissions associated with the energy intensity of water use, and with coordinating how such efforts to reduce the energy intensity of water use can help with efforts to address potential climate change impacts to water.

In a nutshell, the new challenges for California’s water system and reservoir operation can be summarized as:
(1) Strengthen and optimize the current system operation toward better resilience and sustainability for water supply and clean energy production.

(2) Develop advanced analytical tools, models and approaches to support reservoir operation and management decision making.

(3) Build the bridges connecting technologies, knowledge, know-how, and theoretical approaches to the real world application.

This dissertation is dedicated to address the above challenges and provide a tentative framework to achieve better sustainability and resiliency for California’s reservoir system. Each chapter is designed to break through one or two specific bottle-necks, and address various sub-challenges associated with a specific problem.

1.2 Research Motivation and Research Tasks

As mentioned above, the motivation for this research is originated from the needs of improving current reservoir operation and building sustainability for both water and energy. The overall objective of this dissertation is to provide a framework for the optimization of the coupled water and energy system where “optimal” water sustainability and hydropower production are set to be the objectives. The proposed procedure is tested on the California’s reservoir system with the goal of closing the “gap” between theoretical approaches and real world application. The effectiveness of the proposed analytical approaches, decision making support tools, and realistic reservoir system
optimization models is demonstrated. Specifically, the following research objectives and tasks are designed for each individual chapter:

**Research Objective 1:**

The objective for Chapter 2 is to develop reservoir operation analytical approaches useful for reservoir operator and policy makers to understand and quantify the influences of hydrological information (i.e. inflow, precipitation, evaporation, snow conditions, etc.) and non-hydrological information (i.e. policy and regulation, dry/wet conditions and flood control stages, etc.) on controlled reservoir outflows. Historical reservoir release patterns will be extracted and investigated in order to provide a comprehensive understanding of individual reservoir operation in California.

**Research Task 1:**

As for the first research task, the focuses will be on applying a data-mining technique, termed Classification and Regression Tree (CART) in reservoir operation, and proposing a novel shuffled-cross validation scheme to attack one of CART’s disadvantages (so called “overfitting”). Task 1 also includes the comparison between proposed approaches with other benchmark decision tree algorithms and implementation of proposed approaches into 9 major reservoirs in California.

**Research Objective 2:**
The objective for Chapter 3 is dedicated to improve a multi-objective global optimization algorithm with the intention of providing software tools to support calibration of models and development of optimal operation strategies.

**Research Task 2:**

As for the second research task, the focus will be improving an efficient multi-objective optimization algorithm to further enhance its capability of solving high-dimensional problems and address several weakness of the current algorithm with regard to its’ premature phenomenon and tendency to stuck in local optimum. Task 2 also includes a comparison study with other cutting-edge heuristic searching algorithms on a number of benchmark test functions.

**Research Objective 3:**

The objective for Chapter 4 is to incorporate some real word non-linearities into cascade reservoir modelling and provide a platform for reservoir operators to adjust the current operation strategies in maximizing water supply and hydropower production.

**Research Task 3:**

As for the final research task, a realistic reservoir operation model will be built. The focus will be formulating the non-linear reservoir storage and elevation relationship, as well as the hydropower generation processes. The linear simplification and assumption will be investigated by comparing several interpolation approaches, including successive interpolation, polynomial interpolation and piece-wise linearization, in order to quantify the associated error and uncertainty. In addition, the artificial reservoir releases produced in Research Task 1 will be
used as initial forcing and the optimal operation alternatives in response to various dry/wet conditions will be derived using the enhanced multi-objective optimization tools from Task 2.

1.3 Organization of the Dissertation

Besides the introduction in Chapter 1, the organization of rest of dissertation is as follows: Chapter 2 describes the development of a data-driven, decision tree based reservoir simulation model, which is capable of incorporating additional information into reservoir release decision making. A verification study is carried out comparing other decision tree algorithm with the proposed approaches. Chapter 3 is devoted to the development and validation of an enhanced multi-objective global optimization algorithm. The improved algorithm is compared with multiple existing multi-objective heuristic search algorithms on several human designed test functions. Chapter 4 presents a newly developed cascade reservoir model for the Oroville-Thermalito Complex, which is the head water source of the State Water Project located in the northern California. Two types of non-linearities are addressed in the modelling framework and optimal operation strategies for various water supply conditions (dry, wet and normal) are produced by implementing the initial solutions derived from Chapter 2 and optimization techniques developed in Chapter 3. Chapter 5 summarizes the conclusions, major findings, and limitations. Discussions about future directions and recommendations are also included.
### 2. Simulating Reservoir Releases Using Data-mining Techniques

#### 2.1 Introduction

Reservoirs and dams are the major infrastructures in California for surface water resources management, flood control, and ecosystem protection. Decision makers in California are under increasing pressure because of the emerging unsustainable water-supply problems caused by population growth, environmental degradation, and climate change. California’s growing challenge of meeting rising water demands with limited resources has been widely recognized by decision makers after the state experienced the recent drought \cite{DWR2013a,b,c,d}. Such a changing situation brought the awareness of policy makers and water management agencies, such as the California Department of Water Resources (CDWR) and the U.S. Bureau of Reclamation (USBR), to timely establish and enforce water regulations and policies to promote water management efficiency in the vast reservoir systems in California. The framework, which is capable of (1) modeling multiple types of hydrological variables (i.e. inflow, precipitation, evapotranspiration, snow depth, river stage) and non-traditional hydrological variables (i.e. policy and regulation, dry/wet condition) and (2) quantifying their impacts on reservoir operation, will allow water agencies to evaluate current operation rules and improve future decisions.

Considering the downstream water users’ water management requirements, the amount of upstream inflows to reservoirs might not be sufficient information to establish proper water management plans for agriculture irrigation, ground water pumping and ecosystem protection, etc. To efficiently manage fresh water resources, downstream users require an estimate of actual amount of controlled releases...
from a reservoir, which are different from the natural inflows to reservoirs. Therefore, the reservoir modeling framework that is capable of discovering the complex human-controlled outflows patterns is critically needed. Such modeling framework will allow downstream users to understand their source reservoir’s operation rules, and enable them to estimate future controlled reservoir release availability based on the most sensitive factor dominating their source reservoir’s operation.

After the classic work of Young [1967], the system approach to simulate reservoir operation has become popular among many researchers. One of the goals in reservoir operation is to establish optimal operation scenarios. Various approaches are designed and applied to reservoir operation, such as Liner Programming, Dynamic Programming, Stochastic Simulation, Heuristic Search, Data-mining, etc. However, as pointed out by ReVelle [1999], Wurbs [1991] and [1993], Yeh and Becker [1982], and Yeh [1985], there is a “gap” between theoretical developments and real-world implementations in the field of reservoir management. Labadie [2004] pointed out that many optimization models are unable to respond to many real-world information, such as new environmental and ecological constraints. Hejazi et al. [2008] evaluated the role of hydrologic information in reservoir operation and further confirmed that reservoir operators rely on some hydrologic information which is not fully considered in optimization models. By comparing different hydrological variables in stochastic dynamic models, Tejada-Guibert et al. [1995] found out that in California one type of valuable real-world information is policy and regulation, which significantly influences reservoir system operation. Giuliani et al. [2014] confirmed that the baseline operating policy, which only considers
historical inflow, significantly overestimates reservoir system’s reliability in meeting demands in using a multi-objective evolutionary algorithm.

With respect to the limitations of an optimization model and the need for adding supplement real-world information to the modeling framework, recently the application of data-mining techniques in reservoir operation research have gained much popularity. Hejazi et al. [2008] evaluated the sensitivities of hydrologic information’s time-scale and seasonality in reservoir historical releases in California and the Great Plain in U.S. Corani et al. [2009] used a Lazy Learning algorithm to reproduce human decisions in reservoir management in Lake Lugano. Bessler et al. [2003] extracted the operating rules for a single-reservoir in U.K. using the decision tree algorithm, linear regression and evolutionary algorithm, and found out that the results with decision tree algorithm was superior over the others. Compared among these three types of approaches, Bessler et al. [2003] also concluded that the decision tree algorithm had an advantage, which allows the derived rules could be audited and further improved by domain experts.

Building on these previous works that focus on applying data-mining techniques to reservoir management, in this study a data-driven reservoir simulation model is built to simulate the controlled reservoir releases, and investigate the impacts of multiple types of operational decision variables on reservoir decision making process. Specifically, the number of decision variables are extended from only hydrological information’s time-scale and seasonality [Hejazi et al., 2008] to multiple sources of information, including precipitation, reservoir inflows, policy and regulation, dry/wet conditions,
runoff conditions, downstream river status, and snow course information. Using data-mining techniques, Corani et al. [2009] achieved high accuracy in reproducing human’s decisions in a single lake in Italy. In this study, the reproduction of controlled outflow decisions in 9 major reservoirs in California is attempted. In order to further extend the application of data-mining technique in reservoir management, the technique belonging to the same algorithm family that Bessler et al. [2003] employed is used with some enhancements incorporated.

The method employed in this research is a white-box and tree-like data-mining technique, termed the Classification and Regression Tree (CART) algorithm, combined with a novel shuffled cross-validation scheme. CART was originally introduced by Breiman et al. [1984], and further developed by Breiman [1996] and [2001] into bagging-tree and random forest, respectively. The applications of CART algorithm and its enhanced versions are numerous. De’ath and Fabricius [2000] employed CART in analyzing ecological data. Lewis [2000] applied CART in developing clinical decision rules. Prasad et al. [2006] used bagging-tree and random forest in ecological prediction. Steinberg and Colla [2009] and Wu et al. [2008] summarized that CART is one of the top 10 algorithms in the field of data mining. Other popular decision tree algorithms includes the ID3 [Quinlan, 1987] and ID4.5 [Quinlan, 1990] algorithms, which use different splitting rules in the tree-growing procedure than CART employs.

Given a set of decision variables (inputs or predictor) and target variables (outputs), the mechanism in CART is to repeatedly find a classification of the target variable associated with its decision variables
based on selected splitting rules so that any new prediction will be the most similar to its observation in terms of the splitting rule defined measurement. The method has been successfully applied to a diversity of fields, such as finance engineering [Fayyad et al., 1996], system-failure detection [Chebrolu et al., 2005], ecosystem modeling [Araújo and New, 2007; Elith and Leathwick, 2009], remote-sensing data analysis [Xu et al., 2005], and reservoir operation [Bessler et al., 2003; Kumar et al., 2013a; Kumar et al., 2013b; Li et al., 2014; Sattari et al., 2012; Wei and Hsu, 2008].

CART has many advantages which could be suitable for reservoir operation and favored by decision makers [Bessler et al., 2003; Wei and Hsu, 2008]. The nature of data-driven mechanism of decision tree model provides the transparency in its model framework, which allows decision maker to audit and improve the simulation quality [Bessler et al., 2003]. CART is a non-parametric algorithm, indicating its simplicity in providing physical interpretation of historical data. Moreover, the CART algorithm is computationally efficient [Breiman et al., 1984; Lewis, 2000]. The low computational cost computation allows the bridging of the modeling framework with the increasing amount of data observed and recorded from the complex processes in nature.

However, a major issue may arise when using the CART model on the data that contains significant levels of random noise. This issue is termed as “over-fitting” [Breiman et al., 1984], in which any statistical model, such as the CART algorithm, tends to give a very good or near “perfect” fitting to the training data instead of an accurate estimate of the relationship between the target and decision variables, resulting in a poor predictive capability on a the test dataset or called model “unseen”
dataset. To address this issue, model ensemble approaches are commonly employed in the development of decision tree algorithm, in which the weak learners associated with poor predictive performances are constantly eliminated, such as the strategy adopted in Random Forest [Breiman, 2001]. Different from the model ensemble approach, in this study the “over-fitting” problem of CART is addressed by shuffling the training data, and maximizing the posterior performances to select the best model structure, i.e., the decision-tree depth. The attempt is to efficiently use limited data to ensure a sufficient number of training samples that contain distinct information to be recognizable to the CART algorithm so that accurate predictions on any independent data can be stabilized.

In order to achieve such a goal, a shuffled cross-validation scheme in the tree growing process is implemented. Using the proposed scheme, the CART algorithm is iteratively used to develop many decision-tree models with different model structures (tree depths). In addition, data are repeatedly shuffled to create many independent training, validating and testing data sets. The performances of these decision trees using different shuffled training data sets are evaluated by the Nash-Sutcliffe model-efficiency coefficient [Nash and Sutcliffe, 1970] and stored in an archive. The reasons for using the Nash-Sutcliffe model-efficiency coefficient are (1) the Nash-Sutcliffe model-efficiency coefficient is a popular measure of the accuracy in evaluating decision-tree models [Arlot and Celisse, 2010; Burnham and Anderson, 2002; Picard and Cook, 1984]; and (2) it is a normalized index that addresses the differences between simulation and observation. The commonly used RMSE criterion may not be the appropriate measure given the magnitudes of outflows that can vary significantly depending on the size of the reservoirs in the system. Then, a maximum-likelihood method is used to
select the best decision-tree model according to model’s predictive performances on a temporary hold-out dataset within the whole training dataset. The selected best model on validating dataset is further used on another independent data for testing. The procedure is identical to the “split sample analysis” in hydrologic modeling, in which part of the data is used for model development and parameter estimation and the other part of the data (not used) serves as independent data for validating and testing. The results show that accurate predictions are achievable as long as the required decision variables are available.

This chapter is organized into six sub-sections: Section 2.2 describes the 9 major reservoirs included in this study and their associated data. The methodologies, including CART decision-tree algorithm, Random Forest algorithm, shuffled cross-validation scheme, and Gini diversity index are introduced in Section 2.3. Section 2.4 provides the simulation results of reservoir controlled outflows, storage daily changes, storage trajectories and the sensitivity analysis on decision variables. Discussion and limitations are presented in Section 2.5. Section 2.6 summarizes the conclusions.

2.2 Data and selected reservoirs

In this study, 9 major reservoirs in California are selected, namely, the Trinity Lake, Don Pedro Reservoir, New Exchequer Reservoir, Folsom Lake, Friant Reservoir, New Melones Reservoir, Oroville Lake, Success Lake, and Shasta Lake. Most of the reservoir operation data and hydrological data are collected from the California Data Exchange Center (CDEC), which is an official data-sharing portal used by water agencies, decision makers and water users in California. CDEC
installs, maintains, and operates a collection of extensive, centralized hydrologic operational and historical data (http://www.water.ca.gov/floodmgmt/hafoo/hb/cdecs/) gathered from various agencies and utilities throughout the United States. Data supporting agencies include the National Weather Service (NWS), the U.S. Army Corps of Engineering (USACE), the U.S. Bureau of Reclamation (USBR), the U.S. Geological Survey (USGS), the California Department of Water Resources (DWR), the Sacramento Municipal Utility District (SMUD), Pacific Gas & Electric (PG&E), the East Bay Municipal Utility District (EBMUD), and multiple local water agencies. In addition, the snow depth data and downstream flow information are retrieved from each reservoir’s nearby snow course station and river gauge station in the downstream service area, respectively. A summary of selected reservoir, snow course station and downstream gauge is provided in Table 2-1. Figure 2-1 shows the corresponding locations of the selected major reservoirs in California.

<table>
<thead>
<tr>
<th>Name</th>
<th>River Basin</th>
<th>Station</th>
<th>ID</th>
<th>Lat.</th>
<th>Lon.</th>
<th>Elev. (m)</th>
<th>Agency</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trinity Lake</td>
<td>Trinity</td>
<td>Res.</td>
<td>CLE</td>
<td>40.801</td>
<td>-122.762</td>
<td>722.4</td>
<td>USBR</td>
<td>WS, FC, EP, Others</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S.C.</td>
<td>BBS</td>
<td>40.967</td>
<td>-122.867</td>
<td>1981.2</td>
<td>WRD</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>D.G.</td>
<td>DGC</td>
<td>40.645</td>
<td>-122.957</td>
<td>487.7</td>
<td>USGS</td>
<td></td>
</tr>
<tr>
<td>Don Pedro Reservoir</td>
<td>Tuolumne</td>
<td>Res.</td>
<td>DNP</td>
<td>37.702</td>
<td>-120.421</td>
<td>253.0</td>
<td>TID</td>
<td>FC, WS, Others</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S.C.</td>
<td>HRS</td>
<td>38.158</td>
<td>-119.662</td>
<td>2560.3</td>
<td>CDWR S/S</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>D.G.</td>
<td>MOD</td>
<td>37.627</td>
<td>-120.988</td>
<td>27.4</td>
<td>USGS</td>
<td></td>
</tr>
<tr>
<td>New Excheque Reservoir</td>
<td>Merced</td>
<td>Res.</td>
<td>EXC</td>
<td>37.585</td>
<td>-120.270</td>
<td>267.9</td>
<td>MID</td>
<td>WS, FC, Others</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S.C.</td>
<td>STR</td>
<td>37.637</td>
<td>-119.550</td>
<td>2499.4</td>
<td>CDWR S/S</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>D.G.</td>
<td>CRS</td>
<td>37.425</td>
<td>-120.663</td>
<td>50.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reservoir</td>
<td>Res.</td>
<td>S.C.</td>
<td>D.G.</td>
<td>Storage</td>
<td>Department</td>
<td>Use</td>
<td>Others</td>
<td></td>
</tr>
<tr>
<td>----------------------------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>---------</td>
<td>------------</td>
<td>-----</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>Folsom Lake American Res.</td>
<td>Res.</td>
<td>S.C.</td>
<td>D.G.</td>
<td>Storage</td>
<td>Department</td>
<td>Use</td>
<td>Others</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FOL</td>
<td>HYS</td>
<td>AMF</td>
<td>38.683</td>
<td>121.183</td>
<td>USBR</td>
<td>WS, HP, Others</td>
<td></td>
</tr>
<tr>
<td></td>
<td>39.282</td>
<td>-120.527</td>
<td>2011.7</td>
<td>USBR</td>
<td>CDWR S/S</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Friant Dam San Quaquin Res.</td>
<td>Res.</td>
<td>S.C.</td>
<td>D.G.</td>
<td>Storage</td>
<td>Department</td>
<td>Use</td>
<td>Others</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MIL</td>
<td>NLL</td>
<td>MEN</td>
<td>37.001</td>
<td>-119.705</td>
<td>USBR</td>
<td>FC, WS, Others</td>
<td></td>
</tr>
<tr>
<td></td>
<td>37.257</td>
<td>-119.225</td>
<td>2438.4</td>
<td>SCE</td>
<td>USGS</td>
<td>51.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>New Melones Stanislaus Reservoir</td>
<td>Res.</td>
<td>S.C.</td>
<td>D.G.</td>
<td>Storage</td>
<td>Department</td>
<td>Use</td>
<td>Others</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NML</td>
<td>BLD</td>
<td>OBB</td>
<td>37.948</td>
<td>-120.525</td>
<td>USBR</td>
<td>WS, HP, Others</td>
<td></td>
</tr>
<tr>
<td></td>
<td>38.450</td>
<td>-120.033</td>
<td>2194.6</td>
<td>USBR</td>
<td>CDWR</td>
<td>35.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oroville Feather Dam</td>
<td>Res.</td>
<td>S.C.</td>
<td>D.G.</td>
<td>Storage</td>
<td>Department</td>
<td>Use</td>
<td>Others</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ORO</td>
<td>KTL</td>
<td>GRL</td>
<td>39.540</td>
<td>-121.493</td>
<td>CDWR O/M</td>
<td>WS, FC, EP, Others</td>
<td></td>
</tr>
<tr>
<td></td>
<td>40.140</td>
<td>-120.715</td>
<td>2225.0</td>
<td>CDWR O/M</td>
<td>CDWR O/M</td>
<td>274.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Success Tule Dam</td>
<td>Res.</td>
<td>S.C.</td>
<td>D.G.</td>
<td>Storage</td>
<td>Department</td>
<td>Use</td>
<td>Others</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SCC</td>
<td>OEM</td>
<td>TRL</td>
<td>36.061</td>
<td>-118.922</td>
<td>USACE</td>
<td>FC, WS, Others</td>
<td></td>
</tr>
<tr>
<td></td>
<td>36.243</td>
<td>-118.678</td>
<td>2011.7</td>
<td>CAL FIRE</td>
<td>USACE</td>
<td>210.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shasta Sacramento Dam</td>
<td>Res.</td>
<td>S.C.</td>
<td>D.G.</td>
<td>Storage</td>
<td>Department</td>
<td>Use</td>
<td>Others</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SHA</td>
<td>SLT</td>
<td>IGO</td>
<td>40.718</td>
<td>-122.420</td>
<td>USBR</td>
<td>WS, HP, EP, Others</td>
<td></td>
</tr>
<tr>
<td></td>
<td>41.045</td>
<td>-122.478</td>
<td>1737.4</td>
<td>USBR</td>
<td>USGS</td>
<td>325.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>40.513</td>
<td>-122.524</td>
<td>205.1</td>
<td>USGS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Res.: Reservoir**

**S.C.: Snow Course**

**D.G.: Downstream Gauge**

**WS: Water Supply**

**FC: Flood Control**

**HP: Hydropower**
EP: Ecosystem Protection

Others: Navigation, Recreation, Groundwater Recharge, etc.

WRD: Weaverville Ranger District

CDWR S/S: CA Dept of Water Resources/Snow Surveys

USBR: U.S. Bureau of Reclamation

USACE: U.S. Army Corps of Engineers

TID: Turlock Irrigation District

MID: Merced Irrigation District

CDWR O/M: CA Department of Water Resources/O & M

SKCNP: Sequoia and Kings Canyon National Parks

SCE: Southern California Edison Company, Big Creek

USGS: US Geological Survey
The reservoir-operation data are categorized into model input (decision variables) and output (target variables). The model inputs and outputs are summarized as follows:

(1) The first type of model input primarily includes mostly the traditionally hydrological data for the daily reservoir operation, such as the reservoir daily inflow, the daily accumulated precipitation (point measurement), snow depth in reservoir’s upstream mountain area, and the flow conditions in the downstream water supplying area.

(2) The second type of model input is the wet/dry conditions. CDWR uses the Water Year Index (WYI) for the Sacramento Valley and the San Joaquin Valley to classify the water-supply conditions in a water year. In California, WYI is an important guideline for water planning and management (DWR 2013c, 2009, 2005). Officially, the WYIs are determined by the State Water Resources Control Board (SWRCB), which categorizes five
types of water-year: (1) wet, (2) above normal, (3) below normal, (4) dry, and (5) critical year. The calculation and classification examples for WYIs for the Sacramento Valley and the San Joaquin Valley are included in Supporting Information (SI) for interested readers.

(3) Besides the WYIs, there are six different river indices commonly used in California, which are believed to be the operational climate predictors forecasted and calculated by CDWR Snow Survey Office in the beginning of each water year. According to the communication with CDWR, these indices are also used by other water agencies as indicators for evaluating the climate conditions and operating their facilities for the entire state. More detailed information of these indices can be found in CDWR Snow Survey office website.

Generally, the calculation of these indices is based on a linear regression method using multiple weather information, ground-based measurements and decades of experiences of on-site hydrologists. Namely, the six river indices are the Sacramento Valley’s October-March runoff, April-July runoff, and water year total runoff sum, and the San Joaquin Valley’s October-March runoff, April-July runoff, and water year total runoff sum.

(4) The fourth type of model input is water regulation and policy. In this study, the official California governor’s announcement for the State Water Project (SWP) allocation amounts is used, which represents the political influence by the California Water Codes (laws), agreements, and water rights among stakeholders. Each year, the SWP is designated to provide specific amounts of agricultural and municipal water supplies to its 29 agencies over the entire state. However, based on California’s current water-supply condition, the SWP water supply is subject to change over time. Once a change of the projected water supply is
authorized by the governor and the SWRCB, the 29 agencies can only receive and use the officially announced amount of water under jurisdictional rights, unless further announcement is released. For example, on January-31-2014, the CDWR announced an amendment to the SWP allocation, in which the SWP allocation to farmers and agricultural water agencies is dropped to 0% due to continuing drought conditions in California. This change of regulation is enforced by California Governor Brown’s drought declaration made on January-17-2014. The infrastructures in California were operating under this action until another announcement was made on April-18-2014, increasing the SWP allocation back to 5%. These changes in the allocation percentages as a result of the governor’s announcements are retrieved from the California Water Control Board and the SWP official document archive. Detailed information about the decision and target variables can be found in Table 2.

(5) The last model input concerns the influence of seasonality on reservoir operation, which is the month of a year.

(6) The model output (target variable) is the controlled reservoir daily outflow.

Most of the data for the selected reservoirs covers sixteen years from October 1, 1997-December 31, 2013, except the Trinity Lake (CLE), New Exchequer Reservoirs (EXC), and Shasta Lake (SHA), of which the data start on March 24, 2003, March 30, 1999, and January 1, 2001, respectively. A summary of the decision variables and target variable is provided in Table 2-2.
Table 2-2 Detailed information on the decision and target variables

<table>
<thead>
<tr>
<th>Decision/Target Variable Names</th>
<th>Unit</th>
<th>Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reservoir Inflow</td>
<td>m$^3$/day</td>
<td>Daily</td>
</tr>
<tr>
<td>Accumulated Precipitation</td>
<td>mm/day</td>
<td>Daily</td>
</tr>
<tr>
<td>Downstream Daily Mean Flow or River Stage</td>
<td>m$^3$/day or m</td>
<td>Daily or 6 hours</td>
</tr>
<tr>
<td>Snow Depth</td>
<td>m</td>
<td>Daily or Monthly</td>
</tr>
<tr>
<td>Month of a year</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sacramento Valley October-March Runoff</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sacramento Valley April-July Runoff</td>
<td>m$^3$</td>
<td>Annually</td>
</tr>
<tr>
<td>Sacramento Valley Water Year Total Runoff Sum</td>
<td>m$^3$</td>
<td>Annually</td>
</tr>
<tr>
<td>Sacramento Valley Water Year Index (WYI)</td>
<td></td>
<td>Annually</td>
</tr>
<tr>
<td>San Joaquin Valley October-March Runoff</td>
<td>m$^3$</td>
<td>Annually</td>
</tr>
<tr>
<td>San Joaquin Valley April-July Runoff</td>
<td>m$^3$</td>
<td>Annually</td>
</tr>
<tr>
<td>San Joaquin Valley Water Year Total Runoff Sum</td>
<td>m$^3$</td>
<td>Annually</td>
</tr>
<tr>
<td>San Joaquin Valley’s Water Year Index (WYI)</td>
<td></td>
<td>Annually</td>
</tr>
<tr>
<td>SWP Allocation Announcement</td>
<td></td>
<td>Occasionally</td>
</tr>
<tr>
<td>Reservoir Outflow (Target Variable)</td>
<td>m$^3$/day</td>
<td>Daily</td>
</tr>
</tbody>
</table>

2.3 Methodology

2.3.1 Classification and Regression-Tree (CART) algorithm

As mentioned above, CART is a non-parametric data-mining algorithm capable of predicting continuous dependent variable (target variable, $\bar{y} \in \mathbb{R}^n$) with categorical and continuous predictor variable (decision variables, $\bar{x} \in \mathbb{R}^n$). CART is originally introduced by Breiman et al. [1984] and it uses a binary tree to recursively partition the decision variable space into subsets in which the distribution of target variable is successively more homogenous [Chipman et al., 1998]. Before each split in CART, the prior data set is called “parent” node and the two split sub-datasets are referred to “children” nodes. The partitioning procedure searches through all values of the decision variable $\bar{x}$ to find the variable $\bar{x}_{\text{jen}}$ that provides the best partition of the target variable $\bar{y}$ [Razi and Athappilly, 2005] by maximizing the homogeneity of target variable $\bar{y} | \bar{x}_i \leq \bar{x}_j$ and $\bar{y} | \bar{x}_i > \bar{x}_j$ in
the “child” nodes. The maximum of homogeneity is governed by the selected splitting rule, such as to minimize the summation of relative errors in “child” nodes (Eqs. 2-1) [Hancock et al., 2005].

\[
\arg \min(RE(d)) = \arg \min\left( \sum_{0}^{L} (y_i - \bar{y}_L)^2 + \sum_{0}^{R} (y_r - \bar{y}_R)^2 \right)
\]

where \(y_i\) and \(y_r\) are the left and right “child” nodes with \(L\) and \(R\) numbers of target variables; \(\bar{y}_L\) and \(\bar{y}_R\) are the mean of resulting target variables in the left and right “child” nodes; and \(d\) is the decision or splitting rule governing the partition of the data the decision variable \(\bar{x}\). The resulting “child” nodes are recursively partitioned into smaller sub-nodes until pre-set stopping criteria are met in the tree-growing procedure. The stopping criteria could be number of iteration, minimum number of samples in final “child” nodes (classes or leaf), or/and maximum of decision tree depth (size). In this study, the minimum number of samples in a leaf is set to 10, the maximum size of decision tree is set to 20, and number of iteration is set to be relaxed.

2.3.2 Random Forest Algorithm

To obtain a good predictive performance, the output classes (tree leaves) have to be high. However, the risk of overfitting the observed data will accordingly increase. One mean to resolve this weakness is to use an ensemble method, such as bagging [Breiman, 1996], boosting [Freund and Schapire, 1996] and random forest [Breiman, 2001] etc. Large-scale empirical comparison has been conducted by Caruana and Niculescu-Mizil [2006], in which the random forest algorithm achieved excellent performances compared to numerous supervised learning algorithms. According to Liaw and Wiener [2002], differs from the standard trees, each node is split using the best among a subset of predictors randomly chosen at that node, instead of all decision variables. This counterintuitive strategy turns out
to perform very well compared to many other classifiers, including discriminant analysis, support vector machines and neural networks, and is robust against overfitting [Breiman, 2001]. Therefore, in this study, random forest algorithm is also employed to predict the reservoir releasing. Experiments are carried out on the 9 major reservoirs, comparing the observed controlled outflows with the results from random forest, the CART algorithm combined with a shuffled cross validation scheme, and a standard CART algorithm with 2-fold cross validation as control run. Some settings for the use of random forest are listed as follows: the number of trees in a forest is set to be 200; the number of variables in the random subset at each node is set to be 10; and the minimum of samples in a leave is 1 with the purpose of obtaining a fully developed tree.

### 2.3.3 Enhancement of Model Calibration

The cross-validation scheme, also called the rotation estimation [Bauer and Kohavi, 1999; Geisser, 1975; 1993], is a model-validation technique for evaluating predictive performance of a statistical model on an independent or unseen data set [Arlot and Celisse, 2010; Breiman et al., 1984; Burnham and Anderson, 2002; Picard and Cook, 1984]. Several commonly used methods are the hold-out method, the K-fold method [Breiman and Spector, 1992; Kohavi, 1995], and the leave-one/p-out cross-validation method [Allen, 1974; Geisser, 1975; Stone, 1977]. In this section, the shuffled cross-validation scheme will be introduced. The goal of the proposed scheme is to randomly shuffle the training data and ensure that the underlying relationship between the decision variables and the target variable can be fully discovered within CART’s mechanism. Differs from the concept in the random forest that in each tree decision variables are randomly selected, the proposed scheme first
creates many CART models using full decision variables but detects the weak learner (model with poor predictive performances) based on the posterior maximum likelihood of model performances on a pre-partitioned datasets from the training datasets.

There are several steps in the shuffled cross-validation scheme: (1) the data is split into a training set and a test set, which is identical to the hold-out method. The training set includes about 80% of the data, and the test set includes the remaining data; (2) The training data set is then shuffled and further split into two subsets, with the first subset containing about another 80% of the training data and the second one containing the remaining data; (3) one of the CART parameters (decision-tree depth) is iterated from 2 to user-defined maximum and build a corresponding decision-tree model using the first subset; (4) The Nash-Sutcliffe model-efficiency coefficient has been calculated for each model using the second subset; (5) The model with the highest Nash-Sutcliffe model-efficiency coefficient is selected, and the corresponding decision tree depth has been stored; (6) The processes of (2)-(5) are repeated for many times (e.g., 50,000 iterations), and the possibility functions of the numbers of candidate models in each tree depth are obtained; (7) The best decision tree depth is chosen based on the highest likelihood of the numbers of candidate models falling into each tree depth; (8) The verification experiment is carried on using the hold-out data set from Step (1). The flowchart of this shuffled cross-validation scheme is shown in Figure 2-2.

The proposed scheme differs from the cross-validation methodology. The scheme combines the strength of (a) the hold-out methodology, (b) the K-fold method that, in my proposed scheme, K
approximately equals to 2, and the data used for training is about 60% of the whole datasets; (c) the leave-p-out method in which the training data set is shuffled and left about 20% of the training data set out, and (d) the maximum-likelihood estimation. The main purpose of combining these techniques is to ensure that the training data contains the proper predictabilities so that the selected model is able to utilize the historical information in predicting any model “unseen” (independent) data.
2.3.4 Gini Diversity Index

In the decision tree growing stage, as the same to all decision tree family, CART relies on the splitting rule that measures how well a split will result in the most homogenous “child” nodes. Two types of
splitting rules, namely, the Gini index of diversity criterion and Twoing criterion, are originally introduced by Breiman et al. [1984]. The Gini diversity index is a standard and broadly used rule in CART. According to Breiman et al. [1984], the GINI diversity index measures the impurity of a node, while the Twoing criterion chooses a split that balances the datasets in the “child” nodes, which is not related to a node impurity measure. The use of Gini diversity index in CART is also favored by many researchers in the feature selection problem [Chandra et al., 2010; Guyon and Elisseeff, 2003; Qi et al., 2006; Sandri and Zuccolotto, 2008; 2010], as well as in the field of reservoir operation [Tsai et al., 2012; Wei, 2012]. Following the mentioned former works, in this study, the Gini diversity index is used as the splitting rule in CART and use it as the measure to quantify decision variable’s contribution. According to Timofeev [2004], Gini diversity index is calculated by the impurity function \( i(t) \) shown in the following Eqs. (2-2):

\[
i(t) = \sum_{k=1}^{K} p(k \mid t)p(l \mid t)
\]

where \( k, l \in 1, 2, \ldots, K \) are the index of the class (leaves); \( p(k \mid t) \) is the conditional probability of class \( k \) provided that is in node \( t \). The maximization of homogeneity of all child nodes will be equivalent to maximization of change of impurity function \( \Delta i(t) \), as shown in Eqs. (2-3):

\[
\text{arg max}(\Delta i(t)) = \text{arg max}\left[-\sum_{k=1}^{K} p^2(k \mid t_p) + P_l \sum_{k=1}^{K} p^2(k \mid t_l) + P_r \sum_{k=1}^{K} p^2(k \mid t_r)\right]
\]

where \( t_p, t_l, \text{ and } t_r \) are the parent, left “child” and “right” child nodes, respectively; and \( P_l \) and \( P_r \) are the probability of left “child” and right “child” nodes, respectively.

The details of other splitting rules and measurements are available in literatures, such as the Twoing rules [Breiman et al., 1984], the Quinlan’s information gain measure (IM) [Quinlan, 1979; Quinlan,
1986], Marshall Correction [Mingers, 1989] and a random selection of attribute for splitting. The comparisons of different measures are also numerous, such as the works by Mingers [1989] and Buntine and Niblett [1992].

2.4 Results

2.4.1 Sensitivity Analysis of Decision Tree Depths

In this section, the experiment settings and simulation results are demonstrated. The simulated reservoir controlled outflows are compared with observation under three different scenarios, including CART combined with shuffled cross validation scheme, original CART with 2-fold cross validation, and random forest. Using the simulated controlled outflows, reservoir daily storage changes and storage trajectories are calculated. The contributions of decision variables are compared using the Gini diversity index.

2.4.2 Comparison of Simulated Controlled Outflows

As introduced in section 2.3.3, though the data lengths for the 9 major reservoirs in California (Table 2) are not same, most of them are from October 1, 1997-December 31, 2013. Based on a commonly accepted “80/20” split rule, the data from Jan 1, 2010 to December 31, 2013 are held out as independent test period and the rest are used for training and cross-validation. In the shuffled cross validation scheme, the tree depth for each iteration are set from 2 to 20 and the training data are shuffled 50,000 times, as shown in Section 2.3.3: steps (2)-(5). This means that there are 700,000 ((15 – 2 +1)×50,000) candidate decision tree models being constructed and validated using the
shuffled training data. The Nash-Sutcliffe model-efficiency coefficient is calculated to evaluate the model performance and compare the candidate models. It is notable that 19 CART models with tree-depth parameters ranging from 2 to 20 are constructed using 50,000 shuffled training data sets, but not all of the models are qualified to be the candidate model. In each shuffle, one candidate model is obtained by selecting the best one among the 19 CART models. Therefore, after 50,000 shuffles, a total of 50,000 candidate models are created, which fall into 19 types of CART models. Each type of tree contains certain numbers of the candidate model. Figure 2-3 plots the frequency histograms of candidate models in the 19 types of CART models applied to California major reservoirs (CLE, DNP, EXC, FOL, MIL, MNL, ORO, SCC and SHA) using the shuffled cross validation scheme (red). The best depth for decision-tree models applied to the reservoir has the maximum frequency, which indicates that this type of decision-tree model is expected to have the most stable and accurate predictive performance on the “unseen” data. The selected depths for reservoirs CLE, DNP, EXC, FOL, MIL, MNL, ORO, SCC, and SHA are 11, 13, 14, 9, 13, 8, 8, 10, and 15, respectively. In addition, the frequency histogram of only using CART without shuffling the data (blue), which is equivalent to a standard 2-fold cross validation are obtained. The corresponding maximum of the best tree depths for the 9 reservoirs are 7, 7, 6, 7, 8, 6, 6, 7, and 9. A slightly higher tree depth and larger size of tree are chosen when using the shuffled cross-validation scheme.
Figure 2-3 The frequency histograms of CART tree depths using both shuffled cross-validation and standard 2-fold cross validation for the selected 9 major reservoirs in California.

2.4.3 Comparison of Reservoir Controlled Outflows

Using the selected decision-tree depths from the previous section, the predictive capability of the decision tree model on the hold-out data set (December 31, 2010-December 31, 2013) are tested. The purpose is to examine the actual model’s predictive performance. Because the hold-out data are never used in any training process, they are considered as an independent future scenario. The closer the
predicted outflow is to the observation, the better the model’s predictive performance. The predicted daily outflows from the 9 major reservoirs in California are shown in Figure 2-4.
Figure 2-4 Reservoir controlled outflow comparison between observed daily releases (black) with the simulated releases with CART combined with shuffled cross-validation scheme (red), CART without shuffling scheme as control run (blue), and random forest (green).
To mathematically quantify and compare models’ performance, 4 statistical measurements suggested by Gupta et al. [1998], namely, Root-Mean-Square-Error (RMSE), correlation coefficient (R), Nash-Sutcliffe Model Efficiency (NSE), and Peak Flow Difference (PDIFF) are chosen. In Table 2-3, the computed statistics for 9 reservoirs’ controlled outflows simulation are summarized. The formula for calculating the selected statistical measurements are listed as follows:

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Q_{\text{obs},i} - Q_{\text{sim},i})^2}{N}}
\]  

(2-4)

\[
R = \frac{\sum_{i=1}^{N} (Q_{\text{obs},i} - \overline{Q_{\text{obs},i}})(Q_{\text{sim},i} - \overline{Q_{\text{sim},i}})}{\sqrt{\sum_{i=1}^{N} (Q_{\text{obs},i} - \overline{Q_{\text{obs},i}})^2 \sum_{i=1}^{N} (Q_{\text{sim},i} - \overline{Q_{\text{sim},i}})^2}}
\]  

(2-5)

\[
NSE = 1 - \frac{\sum_{i=1}^{N} (Q_{\text{obs},i} - Q_{\text{sim},i})^2}{\sum_{i=1}^{N} (Q_{\text{obs},i} - \overline{Q_{\text{obs},i}})^2}
\]  

(2-6)

\[
PDIFF = Q_{\text{obs},m} - Q_{\text{sim},m}, \ m = \arg \max (Q_{\text{obs},i}), i \in 1, 2, ..., N
\]  

(2-7)

where \(Q_{\text{sim}}\) and \(Q_{\text{obs}}\) are the simulated and observed outflow, respectively; \(\overline{Q_{\text{obs}}}\) and \(\overline{Q_{\text{sim}}}\) are the mean of the observed and simulated outflow, respectively; \(j\) is the time when maximum peak flow happens during the verification period; and \(N\) is the total number of days during the verification period.

Table 2-3 Statistics comparison between the observed reservoir controlled outflow and simulated results with different methods, including combined CART and Shuffled Cross Validation (CART+SCV), CART with 2-fold cross-validation (CART Ctrl), and Random Forest (RF):

<table>
<thead>
<tr>
<th>Reservoir</th>
<th>Methods</th>
<th>RMSE (m³/s)</th>
<th>R</th>
<th>NSE</th>
<th>PDIFF (m³/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

38
|      | CART+SCV |  CART Ctrl | RF    | CART+SCV |  CART Ctrl | RF    | CART+SCV |  CART Ctrl | RF    | CART+SCV |  CART Ctrl | RF    | CART+SCV |  CART Ctrl | RF    | CART+SCV |  CART Ctrl | RF    | CART+SCV |  CART Ctrl | RF    | CART+SCV |  CART Ctrl | RF    | CART+SCV |  CART Ctrl | RF    | CART+SCV |  CART Ctrl | RF    | CART+SCV |  CART Ctrl | RF    | CART+SCV |  CART Ctrl | RF    | CART+SCV |  CART Ctrl | RF    | CART+SCV |  CART Ctrl | RF    | CART+SCV |  CART Ctrl | RF    | CART+SCV |  CART Ctrl | RF    | CART+SCV |  CART Ctrl | RF    |
|------|----------|------------|-------|----------|------------|-------|----------|------------|-------|----------|------------|-------|----------|------------|-------|----------|------------|-------|----------|------------|-------|----------|------------|-------|----------|------------|-------|----------|------------|-------|----------|------------|-------|----------|------------|-------|----------|------------|-------|----------|------------|-------|----------|------------|-------|----------|------------|-------|----------|------------|-------|----------|------------|-------|----------|------------|-------|----------|------------|-------|----------|------------|-------|----------|------------|-------|----------|------------|-------|----------|------------|-------|
| CLE  | 28.383   | 0.748      | 0.510 | 5.396    | -161.452   |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |
| DNP  | 23.716   | 0.925      | 0.853 | -41.866  | -92.142    |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |
| EXC  | 14.620   | 0.913      | 0.833 | -65.826  | -65.826    |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |
| FOL  | 50.136   | 0.845      | 0.671 | 140.299  | 158.584    |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |
| MIL  | 29.370   | 0.933      | 0.813 | -9.226   | -53.702    |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |
| NML  | 15.311   | 0.901      | 0.794 | -2.835   | -36.580    |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |
| ORO  | 40.652   | 0.944      | 0.889 | -122.655 | -645.501   |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |
| SCC  | 4.638    | 0.771      | 0.588 | -15.543  | -19.445    |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |
| SHA  | 93.193   | 0.778      | 0.596 | -67.461  | -740.935   |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |          |            |       |

2.4.4 Comparison between Storage Trajectories

The daily-storage changes during the same verification period are calculated through the water-balance equation using the simulated outflow, observed inflow, calculated lake-surface
evaporation, and precipitation for each reservoir. The comparison between the calculated-storage changes and the observed-storage changes is shown in Figure 2-5.

Using the daily storage changes, the storage trajectories from December 31, 2010-December 31, 2013 are calculated as shown in Figure 2-6. In Figure 2-6, the starting storage volumes on December 31, 2010 for are fixed for all scenarios. The storage trajectories are obtained by accumulating all the simulated daily storage changes throughout the whole verification period. The observations (black line) are collected from the actual daily storage volume archived in CDEC.
Figure 2-5 Reservoir daily storage changes comparison between observed storage changes (black) with the calculated results with CART combined with shuffled cross-validation scheme (red), CART without shuffling scheme as control run (blue), and random forest (green)
Figure 2-6 Reservoir storage trajectory comparison between the actual storage volume (black) with the calculated results with CART combined with shuffled cross-validation scheme (red), CART without shuffling scheme as control run (blue), and random forest (green)
2.4.5 Contributions of Decision Variables

It is of particular interest to discover which decision variable has the most influence on the reservoir operators’ decisions. As mentioned above, the Gini diversity index is used to mathematically quantify the contribution of decision variables. Generally, according to Eq. (2-3) and (2-4), the smaller the Gini diversity index, the purer a set of “child” node is. The Gini diversity index for a “parent” node is always higher than that of two descendant nodes (“child” nodes), which indicates that the employed split using one of the decision variables is able to rationally partition the data and result in a better homogeneity in the “child”. Here, the Gini diversity indexes of all resulting “child” nodes for all splits are summed up using each decision variable. Then, the summation of the Gini diversity index for each decision variable is normalized to a value between 0 to 1, following the rule that decision variable with smaller Gini index summation has higher normalized value. Such normalization ensures that the decision variable with higher contribution or sensitivity in generating the model output has a higher value than others. In other words, the closer the normalized Gini diversity to 1, the more efficient and dominating the decision variable is in splitting the target variable. In Figure 2-7, the normalized decision variable importance is presented using a pie-graph for each selected reservoir. For better illustration and discussion, the WYIs for Sacramento Valley and San Joaquin Valley in Table 2-2 are grouped as Dry/Wet conditions, and the 6 decision variables associated with river indices are categorized as runoff condition. Therefore, there are total 9 categories of decision variables are compared in Figure 2-7, namely dry/wet condition, runoff condition, regulation/policy, reservoir inflow, seasonality/month, precipitation, snow depth, and downstream river stage.
2.5 Discussion

2.5.1 Comparison of Simulated Outflows

As shown in Figure 2-3, the highest posterior maximum likelihood suggests that the most proper tree size in CART that allows the model have good predictive performances, and more importantly, guarantees model’s stability on randomly constructed validating data (not used). The distribution
shapes of the best tree depth of both CART combined with shuffled cross validation and CART control run indicate that either a larger or smaller tree might increase the prediction uncertainty. Similar experiments were presented in Breiman et al. [1984], in which Breiman et al. [1984] concluded that too small a tree will not use some of the classification information, and therefore result in a large misclassification rate. On the other hand, the misclassification rate originally decreases as tree size grows, and then climbs after it hits a minimum. For the reservoir cases, the proper tree sizes found by the CART combined with shuffled cross-validation scheme are consistently higher than that using CART only (Figure 2-3), because the shuffling scheme introduces more predictive information which the fixed training data set might not contain.

According to the comparison of simulated controlled outflows (Figure 2-4), in terms of the magnitude and variation, the simulated results from all scenarios (CART+SCV, CART control, and Random Forest) are very similar to the observation, except the CART control run in MIL, in which numbers of stepwise predictions fail to capture the actual releases. Both of the predicted and observed releases (Figure 2-4 b, c, d, e, and h) tend to decrease as the drought conditions in California becomes more severe after 2011. The controlled outflow peaks at the beginning of 2011 (Figure 2-4 a, b, d, e, g, and i) are well predicted with the proposed method and random forest, while the CART control have poor performances on almost all the cases. Notably, all methods fail to predict the first high peak in SCC (Figure 4h). However, another following peak (around spring in 2011) is captured by all methods with certain overestimates when using random forest and CART control run. The unsuccessful prediction on the first peak in SCC might due to some unknown situations that current decision variables do not
consider, such as an emergent water delivery request from the downstream area of SCC, the maintenance of reservoir releasing gates, or special reservoir operation during drought condition [Kelly, 1986; Yang et al., 2015], etc. The model performance could be further improved once proper historical data that have similar reasoning of human’s releases are incorporated into the simulation process. It is believed that another importance data source that has critical influences on reservoir operation is the water demands, which currently is not included in the designed decision variable due to its availability. However, as mentioned in the previous section, the transparency of the simple mechanism of CART allows decision makers and reservoir operators to improve its quality by feeding more influencing and concerned information.

The statistical performances of the simulated outflows are satisfactory. Generally, the best NSEs of each reservoir range from 0.557 to 0.896 (Table 2-3). According to Moriasi et al. [2007], model simulation can be judged as satisfactory if NSE is greater than 0.50 for streamflow. As shown in Table 2-3, CART combined with shuffled cross-validation scheme outperforms the other two methods for 5 out of 9 reservoirs with regard to RMSE and NSE values, while the results for the other 4 reservoirs are very competitive to the best values generated by either random forest or CART control run. For each reservoir, the correlation coefficients (R values) for all methods are similar to each other, and the differences are trivial. No individual method is obviously superior over the others if solely comparing the R values. However, the Peak Flow Difference (PDIFF) calculated with CART combined with shuffled cross-validation scheme is consistently smaller than that with the other two methods. Comparing CART control run with random forest, the latter seems to have better
performances on predicting the peak flows, especially for SHA (Shasta Lake) and ORO (Oroville Lake). In daily reservoir operation, the response to peak flow is always of great importance. When there is a high flow, effective and efficient operation considers the resilience of reservoir so that proper amount of water is stored for future sustainable supply. More importantly, reservoir controlled releases need to prevent the downstream area from flooding even an increase of release or a deliberate spill is needed. It is also the reason that downstream river stage information from each reservoir’s service area is included as one of the model inputs. With respect to the overall performances on RMSE, R, NSE, and PDIFF, CART combined with shuffled cross validation scheme will be able to more effectively predict the responses of human under peak flow condition than the other two methods.

2.5.2 Comparison of Storage Daily Changes and Trajectories

Using the simulated outflows, the storage daily changes and storage trajectories are calculated and compared in Figure 2-5, and Figure 2-6, respectively. In Figure 2-5, the calculated storage daily changes from all methods are very close to the observed values, especially for CLE, DNP, EXC, MNL, ORO, and SHA (Figure 2-5 a, b, c, f, g, and i). The major discrepancies between the simulated values and observation happen in predicting the peak changes in FOL (Figure 2-5d) and some mid-level changes in MIL (Figure 2-5e). The differences are probably because of the errors in quantifying the losses and gains in each reservoir, such as the overestimates on reservoir evaporation, and un-measured tributary inflows to reservoirs, which tend to cause positive changes in daily storage.
Further using the storage daily changes, the storage trajectories, which are much commonly interested by community, are calculated with a forced starting value (on December 31, 2010) for all methods. Figure 2-6 shows the calculated storage trajectories and observed storage volume. As shown in Figure 2-6, CART control run (blue) gives the worst simulation of storage trajectory, and the differences to observation are relatively larger than the other two methods, especially for MIL, ORO, SCC and SHA (Figure 2-6 e, g, h and i). Random forest is able to give better results than the CART control run, but is relatively weak in matching the observation in CLE, DNP, EXC and SCC (Figure 2-6 a,b,c and h). Generally, the results produced by CART with shuffled cross validation scheme are closer to observation than the other two methods. Due to the fact that the storage trajectory is calculated by accumulating the daily storage changes to the starting storage volume, errors in certain days might be cancelled with each other. For example, an overestimate on the first day and an underestimate on the second day might induce a near perfect storage volume on the third day. However, a robust storage trajectory calculation employed in this study is still able to give decision makers certain confidence in evaluating the model performances, especially for seasonal time scale or shorter prediction lead time. As the confidence in the quality of input data deteriorates, shorter prediction or simulation time-scales will become more realistic and operational, such as seasonal-scale or weekly-scale. Shorter time-scale simulation will allow users to recursively test model performances and gradually adjust model settings and forcing data. Longer period of simulation will inevitably bring more uncertainty in the results and hesitations among decision makers, which could prevent the broadly practices of both data-driven and physical-based models. Nevertheless, the validation and test should be conducted in relatively long time-scale, such as years, to expose both pros and cons of any proposed model.
2.5.3 Reservoir Operation Patterns

It is observed that certain reservoirs are extremely sensitive to policy and regulation. As mentioned in Section 2.3.4, the decision variables contributions are measured by the normalized Gini diversity index and presented in Figure 2-7. One interesting finding is that the proposed method automatically detects that the operation of certain reservoirs are highly correlated to policy and regulations without any user defined prior information. For example, model finds out that 22% of the human controlled outflows from Trinity Lake (CLE) can be explained by policy and regulation changes (Figure 2-7a).

The reasoning lies that the Trinity Lake (CLE) is the second-largest reservoir in California’s Central Valley Project (CVP) which is operated by the U.S. Bureau of Reclamation. Water flows from the Trinity Lake to the Sacramento River, which is the most important SWP water-supply source, followed by the Feather River supplies. Policy and regulation’s influences on the outflows from Trinity Lake are extensive, due to the multi-agencies water sharing agreements from SWP and CVP [DWR, 2005; 2009; 2013d; USBR, 2004]. Similar to the Trinity Lake (CLE) case, the model automatically discovers that the historical controlled outflow records in the Shasta Lake (SHA) are also associated with the changes of policy and regulation, accounting for 14.5% of the total variation.

It is because the Shasta Lake is the largest source reservoir for CVP, and same water sharing agreements with SWP apply. However, the percentage of policy and regulation is smaller than that of the Trinity Lake (CLE). The reason is that the compulsory flood control obligation in Shasta Lake is dominating. The larger shares of seasonality/month in the Shasta Lake compares to Trinity Lake indicate that the flood control influence is higher in the Shasta Lake’s operation. Compared among all
9 reservoirs, the reservoir that is most sensitive to policy and regulation found by the model is the Oroville Lake (ORO) (Figure 2-7g). As the most vital fresh head-water supply source for the SWP delivery, the Oroville Lake provides about $4.317 \times 10^9 \text{m}^3$ of water at maximum capacity annually. The political impact of changing SWP water allocations will directly dominate the human controlled outflows amount in order to properly supply water to all stake-holders and water agencies along the SWP.

Another interesting finding is that the reservoir inflows play a more important role in the reservoirs with lower elevations. According to Figure 2-7 (d-e), the contributions of reservoir inflows account for over 50% variation of the human controlled outflows in FOL and MIL. Surprisingly, FOL and MIL are the two reservoirs with the lowest elevations comparing to others, of which the elevations are 142 and 177 m, respectively (Table 2-1). The reason is that, for the low elevation reservoirs, the operation rules turn to be more simplistic when it is compared to those in high/medium elevation reservoirs. With less risk of being suffering from flooding, the priority of water supply for low reservoirs becomes higher. Moreover, low elevation reservoirs have fewer obligations to transfer water to other areas, because they are already closer to water demand areas, such as farmland, residents, and industry, compared with high/medium elevation reservoirs. The operation rules for these reservoirs are mostly to mitigate the deficiency between the available water and demands. In other words, the management of the outflows depends mostly on the available water that these reservoirs receive from upper-stream.
2.5.4 Limitations and Further Improvement

One of the limitations of this study lies in the assumption that decision variables are not significantly dependent to each other, which might not be always be true. For example, the precipitation and snow depth might be correlated to seasonality/months (Figure 2-7). From the model developer point of view, the ideal case for any modelling framework is that model inputs are strictly independent to each other. However, due to the non-linearities brought by coupling of natural process and complexity in human decision making process, to reach the ideal case is extremely hard. The current 14 types of model inputs are designed based on surveys and consultation from USACE reservoir operators and CDWR hydrologists and decision makers. The reality is that the concerned information to make a release decision varies from each reservoir to another, as well as from one specific region to another. Generally, the current model inputs include most of the important information in reservoir operation in California. Users are recommended to employ customized model inputs in order to further close the gap between theoretical approaches and realistic problems.

A further improvement of prediction accuracy involves the consideration of reservoir’s connectivity, as well as some statewide operations. In some cases, one reservoir’s outflow will become the lower reservoir’s inflow when several reservoirs are in series, such as a cascade reservoir system. The connectivity and interactions in reservoir systems are not included in this study. However, it is possible to make the outflows and operation rules from/of an upper-reservoir as the model inputs for a lower-reservoir using the same approach demonstrated in this paper. By properly connecting the inputs and outputs of two independent models, the intrusion of reservoir-connectivity could be
analyzed. In California, another non-physical aspect that influences the daily releases is the systematical and large-scale operation on certain reservoirs controlled by the same agency. For example, in some circumstances, USBR operates its reservoirs in a systematical manner, in which certain patterns of increasing or reducing releases can happen to all USBR controlled reservoirs at the same time. To tackle such kind of state-wide operation needs proper designs of a conceptual indicator to mathematically quantify this agency’s behavior, similar to the handling of policy and regulation indicator or dry/wet condition indicator employed in this study. In order to eliminate both physical and anthropogenic reservoir connectivity issues, the reservoirs in this study are carefully chosen. The selected 9 major reservoirs belong to different river basins (Table 2-1), and are not physically connected as cascade reservoir system. Moreover, the major reservoirs from different operating agencies were sampled so that the systematical influence from a single agency can be reduced. Therefore, the influence of reservoirs independence and joint operations are negligible.

Future works are suggested to employ the predicted outflows to physical-based optimization models to find the optimal strategies, such as minimizing water shortage and flood risks, maximizing hydropower generation and efficiencies for irrigation and water supply, etc. Even though the mathematical definition of objective functions might be subjective, which obstructs the closing of the gap between theoretical and realistic, the predicted outflows from this study can be used as an initial solution and baseline representing humans’ empirical decisions. The model and outflow results developed in this Chapter will be used as forcing to a physical-based optimization model for further optimizing specific operation objectives. The details of optimization tools and enhancements will be
introduced in the following Chapter 3 and corresponding reservoir operation models will be
developed in Chapter 4.

2.6 Conclusion

In this chapter, a data-driven, decision-tree model is proposed to simulate the controlled outflows
from 9 major reservoirs in California. The inputs of the predictive model include precipitation,
reservoir inflows, policy and regulation, wet/dry conditions, runoff conditions, snow depth, and
downstream river stage information. These decision variables are becoming significantly crucial to
understand and predict human’s behavior on reservoir operations. The results obtained from the
proposed approaches are compared with original CART with 2-fold cross validation, random forest,
as well as observations during the verification period. Experiments show that CART combined with
shuffled cross validation scheme is able to generate reasonable simulations of controlled reservoir
outflows. More importantly, the results with the proposed method are superior over that using other
two methods, especially in modelling the peak flow difference. The model is capable of predicting the
controlled outflows based on its learning mechanism of the dynamic characteristics and relations of
the decision variables and historical releases. Furthermore, the proposed model provides an intuitive
framework allowing decision makers to better understand the reservoir operation and to evaluate
whether a policy or regulation is successfully implemented to the daily operation of reservoirs. Based
on the experiment results, the following conclusions can be drawn for the 9 major reservoirs in
California:
(1) Model finds out that the Oroville Lake (ORO), Trinity Lake (CLE), and Shasta Lake (SHA) are the three reservoirs intensively dominated by changing policy and regulations; The contribution of policy and regulation accounts for approximately 30%, 22%, and 14% for ORO, CLE and SHA, respectively.

(2) Model discovers that low elevation reservoirs, such as the Folsom Lake, and New Melones Reservoir (NML), are operated with considerably large influences from the inflow amounts, because of their closeness to demand areas.

The operations on the rest of the 9 reservoirs consider multiple aspects, such as flood control, water supply capabilities, and downstream river status, and the contributions of decision variables vary from one reservoir to another.

More importantly, the generated controlled outflow could be used as initial solutions for regional optimization models if specific operation objectives need to be optimized, such as increase hydropower production, sustain temperature constraints for ecosystem, and maximize the water supply sustainability. The novelty of the data-driven approach introduced in this chapter lies in the transparency and flexibility of incorporating decision makers and reservoir operators to develop favored model inputs and variables. In the following chapters, the artificially generated human-controlled releases will be optimized by an enhanced multi-objective algorithm (Chapter 3) with the intention to reach higher benefits for hydropower production and water supply sustainability (Chapter 4).
3. Building An Enhanced Multi-Objective Optimization Algorithm

3.1 Introduction

As one of the general decision making support tools, optimization algorithm has gained its popularity in academia, industry, and government agencies. The most classical optimization algorithms are designed to solve the linear programing problems. The development of optimization algorithm has been flourishing for decades as more complex and complicated models and problems being developed and formulated by scientists and engineers. With regard to optimization problems in the field of water resources management, studies in applying Evolutionary Algorithms (EAs) have become very popular in the last few decades. The focuses are mainly on incorporating new strategies in development of new algorithms or on verifying the suitability of a particular algorithm in solving different kinds of conceptual water-resources management problems [Maier et al., 2014]. Tremendous efforts have been made to develop EAs with new methodologies, which are inspired by various natural phenomena, such as foraging behavior of ants [Dorigo et al., 1996] and bees [Pham et al., 2011] and social behavior of fish colonies [Kennedy et al., 2001]. These innovations lead to a large number of studies that apply EAs to many fields of water-resources management problems [Nicklow et al., 2010], including ground-water calibration, water treatment and reservoir operation. For reservoir operation problems in particular, EAs are recognized as good decision-making support tools because of their multiple advantages. In general, according to Abraham and Jain [2005], there are many advantages of using EAs because they require little knowledge about the problem being solved, are less susceptible to the shape or continuity of the Pareto front, are easy to implement, robust, and could be implemented in a parallel environment. The performance of EAs on human-designed test functions
with various mathematical properties, such as discontinuous, non-differentiable, non-convex, and multimodal [Reddy and Kumar, 2006; Reed et al., 2013], have been reported in many studies. The convenience of using EAs to find the Pareto optimaums, and their effectiveness for highly complex test functions provide some confidence that EAs will also be successful with complex real-world problems. In addition, EAs have proved to be effective and suitable, especially for solving multi-objective problems. In multi-objective optimization problems, EAs consider all of the objectives simultaneously without any user defined weights to each objective, and the population-based searching mechanism enables EAs to generate a set of equally important solutions, termed non-dominated solutions [Deb, 2001], in a single run instead of performing a series of separate runs [Abraham and Jain, 2005]. The non-dominated solution set forms a Pareto front which is able to provide decision makers with trade-off information between conflicting objectives. These benefits mentioned above allow EAs to provide decision makers with a reliable set of solutions to real-world problems, as well as the confidence about the use of these solutions that are able to consider the often multi-objective nature of decision making in reservoir operations. Therefore, as Maier et al. [2014] concluded, EAs have bolstered the ability to solve problems that are more relevant to real-world systems. Many studies have applied EAs to realistic reservoir management cases, especially to reservoir release scheduling and hydropower scheduling problems [Afshar et al., 2007; Chang and Chang, 2009; Chen et al., 2007; Cheng et al., 2008; Farmani et al., 2005; Reddy and Kumar, 2006; Reddy and Nagesh Kumar, 2007; Wardlaw and Sharif, 1999].
Although EAs have gained much popularity in the field of algorithm development and theoretical application to reservoir operations, there are few reports about the actual uses of EAs on realistic reservoir-operation problems. As Jones et al. [2002] and Oliveira and Loucks (1997) concluded, there are few documented applications in which decision makers actually use EAs. Shepherd and Ortolano [1996] reported on personal communications with system operators, stating that they “don’t like being told what to do” and “want to do it in his (her) own way”. Yeh and Becker [1982], Wurbs [1991], Yeh [1985] and [1993] summarized that there is a gap between theoretical development and actual real-world implementation. In the state-of-the-art review of optimal operation of multi-reservoir systems, Labadie [2004] concluded that one reason for the gap is that many reservoir operators lack the confidence in simplistic problem formulation, which purports decision makers to replace their judgment and prescribe solution strategies. To fill the gap between theoretical development and realistic problem implementations, Labadie [2004] and Goulter [1992] suggested more interactions and involvement of decision makers in the problem formulation as well as better suitable packaging of the optimization approaches along with designed problems. U.S. Army corps of engineering [USACE, 1990] stated that, among all of the applications of optimization techniques, the combined use of simulation models and optimization has been found to be an effective analysis strategy for reservoir operation problems. Maier et al. [2014] also outlined the current status and future research challenges and directions in the development of a fundamental understanding of both real-world problem and algorithms. These suggestions indicate that both problem formulation and algorithm development are equally important to promote the uptakes of EA by decision makers.
In order to enable EAs to be more broadly applied to realistic water resources problems, this Chapter combined with next Chapter will focus on improving an existing EA in order to optimizing a conceptual reservoir operation model that is built on realistic objectives and constraints for the head-water supply region in California.

This chapter is organized into five sub-sections: Section 3.2 describes an existing EA termed the Shuffled Complex Multi-Objective Optimization Algorithm and two enhancements modules with the intentions of improving the diversity and accuracy of the algorithm. Section 3.3 provides the comparison results among the new algorithm, original algorithm and a couple popular optimizers in the field of optimization algorithm development. Discussion and limitations are presented in Section 3.4. Section 3.5 summarizes the conclusions.

### 3.2 Methodology

#### 3.2.1 Shuffled Complex Multi-Objective Optimization

The goal of this Chapter is to improve the performances of a Multi-Objective Complex Evolution Global Optimization Algorithm (MOCOM-UA) using crowding distance and principal component analysis. Here, the original MOCOM-UA algorithm is introduced. Then, the powerfulness and weakness of MOCOM-UA are summarized.

The original MOCOM-UA algorithm [Yapo et al., 1998] is an extension of the successful single-objective Shuffled Complex Evolution (SCE-UA) global optimization algorithm [Duan et al., 1992]. The general steps of MOCOM can be summarized as follows according to the flowchart in
Figure 3-1: (1) A total of $m \times p$ points are randomly sampled in the parameter space to form the initial population, where $m$ is the number of complexes and $p$ is the total number of individuals in a complex; (2) The functions are evaluated for each individual; (3) The entire population is shuffled and split into $m$ groups (complexes). In each of the groups (complexes), the $p$ individuals form the sub-population; (4) The Pareto ranks [Goldberg and Holland, 1988] are calculated for the entire population; (5) A triangular possibility function is used to assign a selection possibility to each individual according to its Pareto ranks; (6) A simplex is constructed by selecting $n + 1$ individuals according to the possibility distribution of the sub-population derived from the previous step; (7) The Nelder-Mead evolution strategy [Nelder and Mead, 1965] is implemented to obtain a new individual, and the population is updated; (8) The steps from (3) to (7) are repeated until the maximum of the Pareto ranks in step (4) becomes 1, which means the individuals in the population are all non-dominated in relation to each other.
Figure 3-1 Flowchart of the MOCOM algorithm with enhancement modules (grey-dashed boxes)
The MOCOM algorithm combines the strength of competitive evolution [Holland, 1975], the Nelder-Mead (simplex) method [Nelder and Mead, 1965], and Pareto ranking [Goldberg and Holland, 1988]. The MOCOM uses a triangular probability distribution to assign the possibilities to the members in the population so that parents that have better objective function values are more likely to be chosen to produce offspring. This strategy is able to generate a fairly uniformed non-dominated solution set. Various studies have tested its usefulness in hydrologic models [Boyle et al., 2000; Gupta et al., 1998; Gupta et al., 1999; Leplastrier et al., 2002; Vrugt et al., 2003; Xia et al., 2002].

However, Gupta et al. [2003] pointed out several weaknesses of MOCOM-UA, in which two major issues are its clustering tendency of non-dominated solutions and premature phenomena in certain cases. Vrugt et al. [2003] concluded that the failing is due to its fitness assignment of the Pareto ranking, in which members having an identical rank are not distinguished when assigning a selection possibility. The second issue is due primarily to the high dimensionality of the optimization problem (the so-called "curse of dimensionality"), which prevents the evolution of the population from exploiting the entire decision variable space.

To quantify the performances of multi-objective optimization algorithms, two distinct metric are commonly used, namely diversity and convergence metric. Diversity and convergence [Deb, 2001] are the two major concerns to evaluate whether a set of non-dominated solutions is beyond another set. The diversity refers to the non-dominated solutions set's coverage along the Pareto front. And the
convergence measures the non-dominated solutions’ closeness towards the Pareto front. The ideal case is when the non-dominated solutions are identical (i.e., exact overlap) to the global Pareto front.

A solution set with more diverse members is preferred by decision makers than the one in which its members are relatively similar. The diversity of the operation alternatives is able to give decision makers various options in response to the changing climate conditions. In dry conditions, various operation alternatives could help decision makers to revise their operations to conserve certain amounts of reservoir storage for future water supply. In wet scenario, diversified solutions are able to provide decision makers with efficiently water releasing strategies for maximizing other non-water-supply objectives, such as hydropower generation.

Equally important, convergence of a non-dominated solution set helps decision makers understand the limitations of the system as well as the range of potential benefits. A non-dominated solution set which fails to approach the higher values of objective function values is not likely to be used by decision makers. In the decision makers’ perspectives, greater benefits can be gained by using a more converged non-dominated solution set. For example, in a dry scenario, the failure to reach a converged solution means the water is not used efficiently in a water supply maximization problem. In addition, the solution that is closer to the global Pareto front could provide greater benefit for both water and power supply in wet scenarios, when the supply is not as critical as in drought conditions.
In response to the diversity and convergence concerns, the goal of this study is to improve the original MOCOM by incorporating two distinct enhancement modules. The first module revises the selection possibilities of the members with identical Pareto ranking so that the generated non-dominated solutions can form a more uniformed distribution along the Pareto front. The second module monitors the diversity of the population during evolution based on principal component analysis, which has been shown to prevent the population from degenerating.

### 3.2.2 Enhancement Modules

As mentioned in previous Section 3.2.1, the original MOCOM algorithm [Yapo et al., 1998] was enhanced with two effective techniques, and entitled as the “Multi-Objective Shuffled Complex Evolution with Principal Component Analysis and Crowding Distance Operator (MOSPD)”. The two enhancement modules to the original MOCOM algorithm are shown as the grey box in Figure 3-1. The first module is called the “possibility-adjustment” module (Figure 3-2) while the second is called the “dimension monitoring and restoring” module (Figure 3-3).

The details of the “possibility-adjustment” module are exhibited in Figure 3-2. As Figure 3-1 and Figure 3-2 show, this module is embedded in the main routine of the MOCOM. The steps for this module can be summarized as follow: (1) the individuals with a Pareto rank of 1 are selected and stored in a temporary set. For better illustration, the total number of individuals in this temporary set is t; (2) the crowding distance vector \( D(d_1, d_2, ..., d_t) \) is calculated according to the Euler distance.
between an individual and its neighbors in the objective function space [Deb, 2001]; (3) the selection possibility of each individual in this temporary set is calculated as

\[ P_l = \frac{D_l}{\sum_{i=1}^{l} D_i} \cdot \sum_{i=1}^{t} P_i, l = 1, 2, \ldots t; \]  

(3-1)

where the \( P_l \) is the selection possibility from the main routine that is calculated using the triangular possibility density function; (4) the selection possibility is updated from \( P_l \) to \( P_t \) for the individuals in the temporary set; (5) the process in steps (2)-(4) is repeated for the individuals with Pareto ranks that equal to 2, 3… and so on, until the maximum Pareto rank is reached.

This module is intended to ensure that the members with the same Pareto ranking are put into the same group as well as their selection possibilities. In each group, the crowding distance [Deb, 2001] is calculated for all members according to the distance between two of the closest neighboring members.

The crowding distance technique has been successfully tested and applied as an enhancement to many evolutionary algorithms [Azadnia and Zahraie, 2010; Reddy and Nagesh Kumar, 2007]. Then, a new selection possibility for each member in this group is computed, which equals the total selection possibility, which is assigned by the main routine of MOCOM multiplied by a distance coefficient. The distance coefficient is calculated as each member’s crowding distance divided by the total crowding distance of all members in this group. The new selection possibility replaces the original one for each member in this group. This adjustment is looped from the group with the lowest Pareto ranking to the one with the highest Pareto ranking until all members in the population are assigned with a new selection possibility. Different from the MOCOM, in this new strategy, members with identical Pareto rankings are assigned different possibilities, with the criterion that a member which
locates remotely from its neighboring members is more likely to be selected than those closely
clustered in a limited objective space.

Figure 3-2 Possibility-adjustment module
The second module, “dimension monitoring and restoring”, is shown in Figure 3-3. The aim of this module is to capture and restore the lost dimension during evolution. The module uses the Principal Component Analysis (PCA), which was invented by Pearson [1901] and further developed by
[Hotelling, 1933; 1936]. The PCA is a statistical procedure that transforms a given dataset into an orthogonal coordinate system, in which the first coordinate – termed first principal component of PC has the largest variance of the projection from the data set and other coordinates have smaller values of variances in descending order. Sometime the lower-ranked PCs have negligible variance, which means that the data set has a dimensionality reduction. In the MOSPD algorithm, once a lost dimension is discovered by PCA, a new point is sampled along the corresponding PC, and the member that results in the dimension lost is replaced with this new point. The lost dimension here is defined as a dimension has an eigenvalue less than 1% of the summation of all eigenvalues of the covariance matrix. This module can be generalized in terms of two steps. The first step is to check the dimensionality of the space spanned by all individuals in the population using the following procedures:

(a) Let $C = [c_{ij}] = [\bar{x}_1 \ldots \bar{x}_{mp}]$ be the matrix with the coordinates of each point as its columns.

Then, $C$ has dimensions of $n \times (m \times p)$, where, $\bar{x}_i, i = 1,2 \ldots mp$ are the points in the population, $n$ is the dimensionality of the problem, $m$ is the number of complexes, and $p$ is the number of individuals in a complex, as mentioned above;

(b) Transform the original coordinate system to a normalized coordinate system by centering and normalizing each row of $C$ and obtain $C' = [c'_{ij}]$, where $c'_{ij} = (c_{ij} - \bar{c}_i) / \sqrt{v_i}$, where $\bar{c}_i$ and $v_i$ are the mean and variance of the $i$ th row of $C$, respectively.

(c) Calculate the covariance matrix of $C'$ and denote it as $R$. Obtain the eigenvector and eigenvalues of $R$. Each eigenvector is a principal component (PC) of the population, and its corresponding eigenvalue measures the variance of the individuals along the direction of that PC.
If the variance along one PC is too small, it means the individuals are not spanning well over that direction, and it is defined as a lost dimension. Mathematically, a threshold of 1% of the total variance as a lost dimension is used to measure whether a variance is too small.

Once a lost dimension is detected, the second step is to restore the lost dimension by randomly sampling a new point along the PC:

(a) Sample a point from the side of centroid of $C'$ along the PC:

$$\bar{x}' = \bar{c}' + a\bar{l}$$  \hspace{1cm} (3-2)

where $\bar{c}'$ is the centroid of $C'$, $a$ is random number from a normal distribution with mean=2 and variance = 1, $r$ is the radius of the entire population in $C'$, and $\bar{l}$ is the unit vector along the PC.

(b) Transform $\bar{x}'$ back to the original coordinates and evaluate the objective function. Then, update the population and terminate the module.

This method has already proved efficient and effective in solving population degeneration problem in high-dimension, single-objective optimization problems [Chu et al., 2010; 2011] and reducing the number of objectives in multi-objective optimization problems [Giuliani et al., 2014]. However, the implementation of this method in the MOCOM algorithm is rarely reported.

### 3.3 Comparison Results

#### 3.3.1 Test Functions and Settings

To demonstrate the superior performance of the newly developed MOSPD, a comparison study was carried out, including the multi-objective differential evolution method (MODE) [Babu and Jehan,
2003; Babu et al., 2005], the multi-objective genetic algorithm (MOGA) [Murata and Ishibuchi, 1995], the multi-objective simulated annealing approach (MOSA) [Ulunlu et al., 1999], the multi-objective particle swarm optimization (MOPSO) [Coello Coello and Lechuga, 2002; Coello Coello et al., 2004], the newly developed MOSPD, and the original MOCOM over eight multi-objective test functions, which are recognized as benchmark functions in past studies. From Veldhuizen and Lamont [2000], the SCH function [Schaffer, 1984] is chosen, the POL function [Poloni et al., 2000], the FON function [Fonseca and Fleming, 1993], and the KUR function [Kursawe, 1991]. From Zitzler and Thiele [1999] and Zitzler et al. [2000], the ZDT1, ZDT2, ZDT3 and ZDT4 function are chosen.

The population sizes for the selected algorithms are identical for each test function. For SCH, POL, and FON, the population size is set to 16, while for KUR, ZDT1, ZDT2, ZDT3, and ZDT4, a population size of 124 is used. Detailed information on these benchmark functions is included in Table 3-1. Note that it is possible that algorithm performance could vary as the population size changes; nevertheless, an identical population size for each algorithm is used in order to conduct a fair comparison. Detailed population sizes for each test problem can be found in the Table 3-2 last column.

Table 3-1 Detail Information on Test Functions

<table>
<thead>
<tr>
<th>Problem</th>
<th>$n$</th>
<th>Variable range</th>
<th>Objective functions</th>
<th>Optimal solutions</th>
<th>Characteristics</th>
</tr>
</thead>
</table>

69
<table>
<thead>
<tr>
<th>SCH</th>
<th>1</th>
<th>$[-10^3, 10^3]$</th>
<th>( f_1(x) = x^2 )</th>
<th>( f_2(x) = (x - 2)^2 )</th>
<th>( x \in [0, 2] )</th>
<th>convex</th>
</tr>
</thead>
<tbody>
<tr>
<td>POL</td>
<td>2</td>
<td>([-\pi, \pi])</td>
<td>( a_1 = 0.5 \times \sin(l) - 2 \times \cos(l) + \sin(2) - 1.5 \cos(2) )</td>
<td>( a_2 = 1.5 \times \sin(l) - \times \cos(l) + 2 \times \sin(2) - 0.5 \cos(2) )</td>
<td>( b_1 = 0.5 \times \sin(x_1) - 2 \times \cos(x_1) + \sin(x_2) - 1.5 \cos(x_2) )</td>
<td>( b_2 = 1.5 \times \sin(x_1) - \cos(x_1) + 2 \times \sin(x_2) - 0.5 \cos(x_2) )</td>
</tr>
<tr>
<td>FON</td>
<td>3</td>
<td>([-\pi, \pi])</td>
<td>( f_1(x) = 1 - \exp(-\sum_{i=1}^{3} (x_i - \frac{1}{\sqrt{3}})^2 ) )</td>
<td>( f_2(x) = 1 - \exp(-\sum_{i=1}^{3} (x_i + \frac{1}{\sqrt{3}})^2 ) )</td>
<td>( x_i \in [0, 1] )</td>
<td>[Deb, 2001] non-convex</td>
</tr>
<tr>
<td>KUR</td>
<td>3</td>
<td>([-5, 5])</td>
<td>( f_1(x) = \sum_{i=1}^{n-1} (-10 \exp(-0.2 \sqrt{x_i^2 + x_{i+1}^2})) )</td>
<td>( f_2(x) = \sum_{i=1}^{n} (</td>
<td>x_i</td>
<td>^{0.8} + 5 \sin x_i^3) )</td>
</tr>
<tr>
<td>ZDT1</td>
<td>30</td>
<td>([0,1])</td>
<td>( f_1(x) = x_i )</td>
<td>( f_2(x) = g(x) \left[ 1 - \sqrt[3]{ \frac{1}{g(x)} } \right] )</td>
<td>( g(x) = 1 + 9 \left( \sum_{i=2}^{n} x_i \right) / (n-1) )</td>
<td>( x_i \in [0, 1] )</td>
</tr>
<tr>
<td>ZDT2</td>
<td>30</td>
<td>([0,1])</td>
<td>( f_1(x) = x_i )</td>
<td>( f_2(x) = g(x) \left[ 1 - \left( \frac{x_i}{g(x)} \right)^2 \right] )</td>
<td>( g(x) = 1 + 9 \left( \sum_{i=2}^{n} x_i \right) / (n-1) )</td>
<td>( x_i \in [0, 1] )</td>
</tr>
</tbody>
</table>
### Table 3-2 Settings for each test function

<table>
<thead>
<tr>
<th>Problem</th>
<th>( n )</th>
<th>Number of complexes</th>
<th>Number of members in a complex</th>
<th>Number of members in the population</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCH</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>POL</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>FON</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>KUR</td>
<td>3</td>
<td>4</td>
<td>31</td>
<td>124</td>
</tr>
<tr>
<td>ZDT1</td>
<td>30</td>
<td>4</td>
<td>31</td>
<td>124</td>
</tr>
<tr>
<td>ZDT2</td>
<td>30</td>
<td>4</td>
<td>31</td>
<td>124</td>
</tr>
<tr>
<td>ZDT3</td>
<td>30</td>
<td>4</td>
<td>31</td>
<td>124</td>
</tr>
<tr>
<td>ZDT4</td>
<td>10</td>
<td>4</td>
<td>31</td>
<td>124</td>
</tr>
</tbody>
</table>
The key settings for each algorithm are listed below: For the MODE, the crossover constant is set to 0.9 and the scalar factor is 0.45; For the MOGA, the crossover probability is 0.5, and the mutation probability is 0.1; For MOSA, the cooling factor is set to 0.87; reheating temperature is set to 5. For the MOPSO, the cognitive learning factor is 2, the social learning factor is 2, the movement velocity is 0.5, the inertial constant is 0.5, and the maximum number of individuals in each particle is set to 5. All test runs are set a maximum of 50,000 function evaluations.

3.3.2 Comparison of Different Multi-Objective Optimization Algorithms

The final non-dominated solutions derived with MOSPD, MOCOM, MODE, MOGA, MOSA, and MOPSO, along with simulated global Pareto front are shown in Figure 3-4. For MOSPD and MOCOM, the population at selected number of function evaluation during the evolution on the test function SCH, POL, FON, and KUR are shown in Figure 3-5. And the evolution processes for MOSPD and MOCOM on ZDT1, ZDT2, ZDT3, and ZDT4 are presented in Figure 3-6.
Figure 3-4 Test results of MODE, MOGA, MOSA, MOPSO, MOSPD, and MOCOM
Figure 3-5 Evolution process of MOCOM and MOSPD on the SCH function (a) and (b), the POL function (c) and (d), the FON function (e) and (f), and the KUR function (g) and (h)
Figure 3-6 Evolution process of MOCOM and MOSPD on ZDT1 function (a) and (b), ZDT2 function (c) and (d), ZDT3 function (e) and (f), and ZDT4 function (g) and (h)
To evaluate the performance of a multi-objective optimization algorithm, Deb [2001] suggested at least two metrics should be used. One metric measures how close is the non-dominated solution set towards the global Pareto front, and the other evaluates the spreading extent of the non-dominated solution set along the global Pareto front. Therefore, the diversity metric $\Delta$ and generational distance (GD), which are two well-established performance-measurement indices according to Zitzler and Thiele [1998], [1999] and Deb [2001], are chosen in this research.

The range of GD is greater or equal to zero, and $\Delta$ lies within the range of $[0,1]$. A smaller value of GD indicates that the non-dominated solutions have better convergence towards the global Pareto front. Similarly, when the $\Delta$ value is closer to zero, a more diverse spread of the non-dominated solutions along the global Pareto front is denoted. Table 3-3 and Table 3-4 list the spread metric $\Delta$ and GD values for each of the algorithms on the six test problems.

### Table 3-3 Diversity metric for MODE, MOGA, MOSA, MOPSO, MOCOM and MOSPD on the test functions

<table>
<thead>
<tr>
<th>Function\Algorithm</th>
<th>MODE</th>
<th>MOGA</th>
<th>MOSA</th>
<th>MOPSO</th>
<th>MOCOM</th>
<th>MOSPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCH</td>
<td>0.1445</td>
<td>0.3151</td>
<td>0.157</td>
<td>0.5023</td>
<td>0.2228</td>
<td>0.129</td>
</tr>
<tr>
<td>POL</td>
<td><strong>0.0333</strong></td>
<td>0.2241</td>
<td>0.0487</td>
<td>0.1313</td>
<td>0.3071</td>
<td>0.2061</td>
</tr>
<tr>
<td>FON</td>
<td>0.2203</td>
<td>0.2374</td>
<td>0.1036</td>
<td><strong>0.038</strong></td>
<td>0.2637</td>
<td>0.1533</td>
</tr>
<tr>
<td>KUR</td>
<td>0.0869</td>
<td>0.0418</td>
<td>0.0465</td>
<td><strong>0.0405</strong></td>
<td>0.6403</td>
<td>0.0456</td>
</tr>
<tr>
<td>ZDT1</td>
<td>0.5842</td>
<td>0.5835</td>
<td>0.5769</td>
<td>0.6656</td>
<td>0.7708</td>
<td><strong>0.2344</strong></td>
</tr>
<tr>
<td>ZDT2</td>
<td>0.6147</td>
<td>0.7265</td>
<td>0.7562</td>
<td>0.8508</td>
<td>0.8364</td>
<td><strong>0.3484</strong></td>
</tr>
<tr>
<td>ZDT3</td>
<td>0.509</td>
<td>0.519</td>
<td>0.5281</td>
<td>0.643</td>
<td>0.7832</td>
<td><strong>0.126</strong></td>
</tr>
<tr>
<td>ZDT4</td>
<td>0.5339</td>
<td>0.5624</td>
<td>0.7555</td>
<td>0.6061</td>
<td>0.6552</td>
<td><strong>0.3293</strong></td>
</tr>
</tbody>
</table>
Table 3-4 Convergence metric GD for MODE, MOGA, MOSA, MOSPD, MOCOM, MOCOM

and MOSPD on the test functions

<table>
<thead>
<tr>
<th>Function</th>
<th>MODE</th>
<th>MOGA</th>
<th>MOSA</th>
<th>MOPSO</th>
<th>MOCOM</th>
<th>MOSPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCH</td>
<td>0.0048</td>
<td>0.0049</td>
<td>0.0057</td>
<td>5.8157</td>
<td><strong>0.0045</strong></td>
<td>0.0046</td>
</tr>
<tr>
<td>POL</td>
<td>0.1342</td>
<td>1.1736</td>
<td>0.189</td>
<td>0.1054</td>
<td>0.8252</td>
<td><strong>0.038</strong></td>
</tr>
<tr>
<td>FON</td>
<td>0.0071</td>
<td>0.0112</td>
<td>0.0084</td>
<td>0.0106</td>
<td>0.0058</td>
<td><strong>0.005</strong></td>
</tr>
<tr>
<td>KUR</td>
<td>0.0111</td>
<td><strong>0.0103</strong></td>
<td>0.0152</td>
<td>0.0263</td>
<td>0.0299</td>
<td>0.0133</td>
</tr>
<tr>
<td>ZDT1</td>
<td>0.1451</td>
<td>0.2836</td>
<td>0.1047</td>
<td>0.3573</td>
<td>0.264</td>
<td><strong>0.0056</strong></td>
</tr>
<tr>
<td>ZDT2</td>
<td>0.1777</td>
<td>0.6395</td>
<td>0.0986</td>
<td>0.7097</td>
<td>0.6007</td>
<td><strong>0.0122</strong></td>
</tr>
<tr>
<td>ZDT3</td>
<td>0.0892</td>
<td>0.2608</td>
<td>0.0672</td>
<td>0.3543</td>
<td>0.0214</td>
<td><strong>0.016</strong></td>
</tr>
<tr>
<td>ZDT4</td>
<td>0.9199</td>
<td>12.8908</td>
<td>0.1376</td>
<td>15.695</td>
<td>10.3852</td>
<td><strong>0.014</strong></td>
</tr>
</tbody>
</table>

3.4 Discussion

According to the test results, MOSPD shows better capability of expanding non-dominated solutions along the Pareto front. According to Table 3-3, for low-dimension problems (SCH, POL, FON, and KUR), in which the number of decision variable ranges from 1 to 3, MOSPD demonstrates a comparable Δ value compared to the other four algorithms (MODE, MOGA, MOSA and MOPSO), while it consistently shows a better Δ value on all of the higher dimension problems (ZDT1-4 with 10 to 30 decision variables). Note that for all tests, MOSPD has a smaller Δ value when compared to MOCOM. Based on the result of limited sensitivity test shown in this study, MOSPD has exhibited superior performance over MOCOM in all cases and better diversity measurements over the other four algorithms, especially for high dimension problems. In some of the test functions (Figure 3-4 (d-g)), the non-dominated solutions with MOCOM tend to cluster in a fairly small region in the
objective space, while the solutions with MOSPD have a likely uniformed spread. As the Table 3-4 shows, excluding the SCH and KUR cases, the smaller values of GD indicate that the MOSPD algorithm is also able to generate more converged non-dominated solutions than those generated from other algorithms. However, the GD values for SCH and KUR with MOSPD are still very competitive when compared to others. For the functions POL, ZDT1, ZDT2, and ZDT4 (Figure 3-4 (b), (e), (f), and (h)), MODE, MOGA, MOSA, MOPSO, and MOCOM fail to discover the global Pareto front in the objective function space and the search stops at a local optimal with higher values for both of the objectives. In contrast, the MOSPD is able to escape from the local attractions and reach the objective function values that are very close to the global optimums.

When compared with the MOCOM and MOSPD evolution processes (Figure 3-5 and Figure 3-6), MOCOM fails to maintain the population diversity during the evolution (Figure 3-5(g), Figure 3-6(a, c, e, and g)). Although MOCOM converges more quickly toward lower objective function values for some cases during the evolution (Figure 3-6 (a) and (c)), the expansion of the population is poor. For most of the cases, the population evolution of MOCOM eventually stops with higher objective function values due to the large local minimum attraction, while the population evolution of MOSPD in the same tests (Figure 3-5 (h), Figure 3-6 (b, d, f, and h)), are likely to maintain the population diversity, and the final population is able to reach a more expanded location with lower objective function values.
The improved performance of MOSPD is due to the two newly introduced modules. In the original MOCOM, the evolution process is guaranteed to be competitive based on the selection criteria that the possibility that “better” parents contribute to the generation of offspring is higher than that of “worse” parents. However, this strategy does not guarantee that the offspring can uniformly locate along a certain Pareto front. In MOCOM, according to the identical Pareto ranking, the parents, which can contribute to generating an offspring with a better location that is towards a sparse surrounding area, are equally treated as the parents that are only able to produce the offspring with a worse location. The new strategy used in MOSPD remedies the equal treatment criteria by adjusting the selection possibility using the crowding distance measurement. The crowding distance-based possibility selection strategy ensures that the parents that produce offspring with a better location are assigned with a higher chance to be selected. This is a more robust strategy because more diversified offspring help to form the final non-dominated solutions towards a uniform distribution along the global Pareto front. Similarly, for some of the cases (Figure 3-5(d)-(f)), the original MOCOM cannot escape the local attractions due to the fact that the decision variables in certain dimensions happen to be the same or very close values, which causes the population to lose the ability to continue searching decision variable space in this dimension. Eventually, the members with lost dimensions will be stuck at the local Pareto optimal with higher values on both objectives than the global Pareto optimal. MOSPD overcomes this problem by repeatedly restoring the lost dimensions and preserving the population’s vitality of searching larger parameter spaces.
3.5 Conclusion

In this Chapter, a new multi-objective global optimization algorithm, termed the Multi-Objective Shuffled Complex Evolution with Principal Component Analysis and Crowding Distance Operator (MOSPD), is developed. Extensive numbers of comparison studies have been carried out. The newly developed MOSPD algorithm is compared with original MOCOM, MODE, MOGA, MOSA, and MOPSO algorithm on 8 various standard test functions. It is found that the MOSPD algorithm has a generally superior performance over other algorithms in reaching better values of the diversity and convergence measures.

The newly developed MOSPD algorithm combines (1) the strengths of the MOCOM algorithm [Yapo et al., 1998]; (2) the concept of the crowding distance-based offspring selection probability strategy [Deb, 2001]; and (3) the tool of PCA [Hotelling, 1933; 1936] that restores and maintains the population diversity during searching. According to the test results, MOSPD is a more efficient and effective algorithm over MOCOM regarding convergence and diversity of the non-dominated solutions. In the next Chapter, the improved algorithm (MOSPD) is implemented into optimizing a realistic reservoir operation problem for the Oroville-Thermalito Complex (OTC) for the California’s State Water Project. The diversified and converged non-dominated solutions developed by MOSPD serve as OTC’s daily operation alternatives in response to various dry, wet and normal water supply conditions. The focus of applying MOSPD to OTC’s cascade reservoir model is to provide decision maker with optimal operation strategies for water supply and hydropower generation. In the model formulation stage, two non-linear aspects involved in the Oroville-Thermalito Complex are
specifically addressed. And the performances of MOSPD and the original MOCOM algorithm on optimizing the mentioned reservoir model are further compared in the next Chapter.
4. Development of Non-linear Water and Hydropower Models

4.1 Introduction

4.1.1 The Oroville-Thermalito Complex and State Water Project

The Oroville-Thermalito Complex (OTC) is a group of reservoirs, structures, and facilities located in and around the city of Oroville in Butte County, California. The OTC is one of the most important freshwater supply for the California’s State Water Project (SWP) (Figure 4-1). The main functionalities of OTC are to supply fresh water and generate hydro-electricity for the California’s SWP. The SWP (Figure 4-1) is the nation’s largest state-built water and power development and conveyance system [DWR, 2013c] operated by the California Department of Water Resources (DWR), with its main purpose to store and supply water from the precipitation-concentrated northern area of to the water-scarce central and southern regions. Currently, SWP has conveyed an annual average of $3.577 \times 10^9$ cubic meters ($m^3$) of water with the potential of providing $5.181 \times 10^9$ cubic meters ($m^3$) designated water allocation to its users. The main challenge for the future SWP project lies in its ability to make efficient water release to meet increasing water demands. As DWR [1993] projected, a net, statewide water-supply deficit of $4.07 \times 10^9$ to $5.181 \times 10^9$ $m^3$ by 2020 is expected, which implies that SWP is under a significantly severe burden to supply water under potential drought condition. For example, the recent California drought began in 2011 and dramatically affected California’s water supply and hydropower generation. Calendar year 2013 closed as the driest year in recorded history for many areas of California, and current conditions suggest no change in sight for 2014. On January 31, 2014, the DWR announced several actions to protect the health and safety of Californians from the effects of more severe water shortages. Those actions include lowering the anticipated allocation of water to customers of the SWP from 5% to zero, which marks the first zero
allocation announcements for all customers of the SWP in its 54-year history. In order to better meet the water shortage and response to varying water-supply conditions, a more efficient operation of SWP's water storage facilities is required so that the SWP's water supply can be more stable and sustainable.

Figure 4-1 SWP and OTC (Courtesy of the California Legislative Analyst's Office and the California DWR)

4.1.2 The Oroville-Thermalito Complex

The Oroville-Thermalito Complex (OTC) (Figure 4-1), which is located in northern California in the foothill of the Sierra Nevada Mountains, consists of several reservoirs and hydropower plants in and around the city of Oroville in Butte County. As the most vital fresh head-water supply source and power development for SWP, OTC delivers about $4.317 \times 10^9 \text{ m}^3$ of water at maximum capacity and generates more than 2.8 billion kilowatt-hours of power annually. When water is needed in SWP, the OTC releases water into the Feather River through the Oroville Dam, Thermalito Diversion Dam, and the Thermalito Afterbay. The released water travels down the river to where the river converges...
with the Sacramento River and continues to be pumped or diverted to southern and central California for various demands. These processes are carried out by jointly operating OTC's 10 major facilities: (1) three cascade reservoirs: Lake Oroville, the Thermalito Forebay, and the Thermalito Afterbay; (2) three hydro-electrical powerplants: the Hyatt Powerplant, the Thermalito Diversion Powerplant, and the Thermalito Pumping- Generating Plant; and (3) five other facilities, including the Thermalito Diversion Dam, the Lake Oroville Dam, the Feather River Fish Hatchery, the Fish Barrier Pool, and the Thermalito Power Canal.

To increase the water's sustainability in OTC, the accumulated daily-storage volume during one month was considered as the first objective. This objective indicates the regional "water supply loaning" availability and capability to consolidate storages in response to emergent drought conditions. The concept of "water supply loaning" is originally introduced by Kelly [1986], which is a special reservoir operation scheme during drought conditions. According to Kelly [1986], “water supply loaning” is recognized as an efficient special reservoir operation during drought conditions, which temporarily transfers large amounts of water from one water-supply system to another water-scarce system in a short period of time (days or weeks) to mitigate the drought condition. This action requires the loading system to have sufficient and continuous water storage so that the transferring of water does not jeopardize the loading system’s engineering constraints and demand constraints. Historically, the “water supply loaning” operation has been carried out between SWP and California's Central Valley Project (CVP) during the California drought from 1976-1977. In August 1977 during which \(9.25 \times 10^6 \, m^3\) of water has been transferred from the OTC and the SWP’s reservoirs to the
San Luis Reservoir to temporarily meet the CVP irrigation demand. This action is completed in 4 days. During that time the SWP stores relatively high levels of water storage in its reservoir system, while the CVP is facing a forecasted water shortage caused by the 1976-1977 California droughts. Consolidation of water storage is another drought-response operation scheme that merges the water storage of several reservoirs into one as quickly as possible to decrease the evaporation, seepage and in-stream losses. The consolidation of water storage is extended through the Orland Project, a project within the CVP, during the same period of 1976-1977. A necessity of these two special operations is that water must to be transferred from the region with relatively higher storage volumes in a short period of time (days or weeks) to the region with low storage volumes. Therefore, the accumulated daily storage volume during a particular month is able to measure the capability to quickly enable these special operations.

In order to increase OTC's power-supply stability, the second operation objective is set to be the net electricity generation during a given month. California's SWP is the largest single electricity user in the state, accounting for 3% of all electricity consumed statewide [DWR, 2013d]. Most of the energy is used to deliver water to the southern California region, where pumping 1 m³ water through the Tehachapi Mountain consumes about 2.432 kWh. Annually, about two-thirds of SWP's power comes from its hydro-electrical plants, and the remainder is supplemented by the coal-fired Reid Gardner plant in Nevada and power purchases and exchange programs [DWR, 2011; 2012; 2013b]. Thus, increasing OTC's total net hydro-power generation could help SWP build its electrical power development towards cleaner and more self-sufficient levels. Furthermore, considering the state and
federal legislators have already been establishing laws and special teams focusing the water-energy problems, the study of water-energy nexus [CEC, 2005b; Cohen et al., 2004; DWR, 2013d; 2006; Wilkinson, 2000] and how to optimize these two objectives through reservoir system will provide useful information regarding how to manage available water and energy resources in California.

The two objectives considered in the OTC reservoir problem are based on the functionalities and the purposes of the reservoirs in the complex. In other literature, there are significant ways to include hydropower generation in reservoir operation, such as considering short-term hydro scheduling, annual production, and trade-offs with agricultural water uses [Cheng et al., 2014; Gil et al., 2003; Hassanzadeh et al., 2014; Li et al., 2013] In these ways, hydropower generation is either considered as a single objective in complex reservoir systems or measured by economic value to other water uses. However, the settings of these two objectives are derived from the realistic functionalities of the OTC’s reservoir system, as well as the demands and requirements from the operating agencies. The total storage in the complex indicates the water supply potential for California’s SWP, while the hydropower generation ensures that the water released from this complex will be delivered through the whole California’s water distribution system. Again, the selected two objectives are essential to the operation of OTC and SWP.

Besides the two major purposes (water supply and power generation), current operation in OTC has other goals, such as flood control, temperature control for fish habitat, and water-quality monitoring. The priorities for these goals vary from month to month. First, OTC’s facilities are operated under
flood-control requirements specified by the U.S. Army Corps of Engineers (USACE), which requires that the OTC's reservoirs to keep a flood-control space during the peak flood season (October 15–March 31). The required flood-control space varies between $462.55 \times 10^6$ and $925.11 \times 10^6$ cubic meters, depending on the accumulated precipitation parameter prescribed in the flood-control manual [DWR, 2013a]. In addition, OTC is required to maintain a specific outlet temperature for salmon and steelhead trout spawning conditions during the whole year. A temperature range of plus or minus 4°F is allowed [DWR, 2006] only from April through November. Moreover, water-quality standards are designed to meet several water quality objectives such as salinity, Delta outflow, river flows, and export limits [DWR, 2006].

4.2 Model Development

4.2.1 Mathematical Formulation

In Figure 4-2, the facilities and key infrastructures for the OCT’s reservoir system are presented. In order to mathematically formulate OTC’s reservoir system, the main components are described separately. Reservoirs, dams and power plants are represented by node numbers 1-6. Water flows are illustrated by blue arrows, and the interactions between power facilities and electricity grids are shown with red arrows. OTC has two major outlets to the Feather River. One is located at the Thermalito Diversion Dam, which belongs to the Thermalito Forebay area, and the other is connected to the Thermalito Afterbay. Other water flows are upstream inflows and the regulated daily flows, such as deliveries to nearby cities, power flows, and pumpback flows. In this system, the Feather River daily releases were considered to be optimized because of the regulations as mentioned above.
The two main types of constraints that relate to mean sea-level water elevation are included in Figure 4-2 and are the reservoir water-elevation constraints and the normal static water head constraints of power plants \cite{DWR2013c}. The reservoir water-elevation constraints are described as the upper and lower bounds, which represent the reservoirs’ physical water capacities based on the engineering design of the dams. The normal static head upper and lower bounds ensure the safety operation requirements for the power plant's generator and other components. The OTC’s problem can be expressed as:

Optimizing decision variables $U_t$ in order to
\[
\begin{align*}
\text{Max} & \left\{ F_1 = \sum_{m=1}^{M} \sum_{t=1}^{T} S_t, \\
F_2 = \sum_{m=1}^{M} \sum_{t=1}^{T} (G_t - P_t) \cdot \Delta t = \sum_{m=1}^{M} \sum_{t=1}^{T} 9.81 \cdot (\eta \cdot Q - \mu \cdot Q') \cdot H_t \cdot \Delta t \right. \\
\end{align*}
\]

which is subject to:

\[
\begin{align*}
S_{t+1} &= S_t + I_t - O_t \\ 
O_t &= Q_t + U_t + E_t + D_t \\ 
H_t &= h_{upper} - h_{lower} = f(S_{upper,t}) - g(S_{lower,t}) \\ 
S_{t,\min} &\leq S_t \leq S_{t,\max} \\ 
Q_{t,\min} &\leq Q_t \leq Q_{t,\max} \\ 
D_{t,\min} &\leq D_t \leq D_{t,\max} \\ 
G_{t,\min} &\leq G_t \leq G_{t,\max} \\ 
P_{t,\min} &\leq P_t \leq P_{t,\max} \\ 
H_{t,\min} &\leq H_t \leq H_{t,\max} 
\end{align*}
\]

for \( t=1,2,\ldots,T, \ n=1,2,\ldots,N, \) and \( m=1,2,\ldots,M. \)

In Eqs. (4-1)-(4-10), \( T \) is the number of days in a month; \( N \) is the number of storage facilities; \( M \) is the number of power facilities; \( S_t \) is the storage at time step \( t \); \( I_t \) is the reservoir inflow at time step \( t \); \( O_t \) is the reservoir outflow which consists of regulated power discharge term \( Q_t \), Feather River release term \( U_t \), evaporation and other losses \( E_t \), and water deliveries to local urban areas \( D_t \) at each time step \( t \); \( G_t \) is the power-generation capacity at time step \( t \); \( P_t \) is the pump-back electricity capacity at time step \( t \); \( \eta \) and \( \mu \) are the electricity generation and pumping efficiency, respectively; \( Q'_t \) is the regulated pump-back flow; and \( H_t \) is the normal static water-head difference.
between the upper and lower reservoirs. In Eqs. (4-5)-(4-10), lower bounds of the (upper bounds) constraints are noted with “min” ("max") as subscripts.

4.2.2 Non-linearity

Based on the mathematical formulation, especially Eq. (4-4), the normal static water-head difference is determined by the upper and lower reservoir water elevation $h_{upper}$ and $h_{lower}$, each of which is a function of the reservoir's storage volume. The relationship between reservoir's storage volume and elevation is non-linear due to the irregular shape of reservoir topography. Here, the function $f$ and $g$ are used to represent this relationship.

In Eqs. (4-1)-(4-10), there are two crucial non-linear aspects that could influence decision maker's choice and the optimized solution. First, according to Eq. (4-1), the net electricity generation changes non-linearly with the discharge. The electricity generation is calculated by the discharge $Q_t$ multiplied by the water-level difference $H_t$. More discharge generates more hydro-electricity and also results in less storage volume in a given reservoir. The less storage volume leads to a lower water level, according to the non-linear functions $f$ and $g$. Eventually, the increasing discharge also decreases the net electricity generation amount. Second, in this process, reservoir topography functions $f$ and $g$ play an important role in determining the actual changing value of normal water static head difference $H_t$. In the model framework, the time step is chosen to be daily for the mid-range (month or seasonal) optimal reservoir operation. By using a shorter time step, such as hourly, the formulation is more realistic to the dynamic hydropower generating and pumping
mechanics, because the turbine efficiency curve and other dynamic process are able to be included [Diniz et al., 2007]. However, the hourly operation requires quick opening and closing the releasing of the releasing vaults or gates, and the failure to ensure this could cause operation difficulties and quicken the aging of facilities. Moreover, with regard to the optimal reservoir water-supply objective, the operation on a daily scale is a relatively good tuning scale for mid-range optimal water-supply planning. Therefore, daily-scale is used as the time step, which also follows the similar manner that used by Li et al. [2013] to estimate hydropower generation and pumping in cascade reservoir modeling framework.

To present the relationship between storage volume and water elevation, a piece-wise storage-elevation curve is typically used. As mentioned earlier, Bayón et al. [2009] argued that linear approximation of the reservoir storage-elevation (S-E) curve could result in serious errors in calculating hydro power. Here, a non-linear approach is used, in which the S-E curves for Oroville Lake, Thermalito Forebay, and Thermalito Afterbay are fitted with 8-, 13- and 14-order polynomial functions (Figure 4-3 (a)-(c)). The orders of fitting functions are chosen based on error variability between observations and fitted values. As shown in Figure 4-3(d)-(f), with the increasing polynomial orders, the sum of the squared residual is decreasing. The polynomial function can generate a near-perfect fit by increasing the number of orders. However, higher order of polynomial could result in requirements for more complex computation and difficulties for optimization algorithm to find local minimums. The order of fitting function is based on the criterion that the sum of squared residual is relatively small, but it does not decrease dramatically with increasing orders. According to
Figure 4-3, the discovered polynomial function orders are 8, 13, and 14 for the Oroville Lake, the Thermalito Forebay and the Thermalito Afterbay, respectively. The sample S-E measurements are retrieved from the SWP construction reports [DWR, 1974] and the monthly reports published by SWP operations control office.

Figure 4-3 (a) Storage-elevation curve of Lake Oroville; (b) Storage-elevation curve of Thermalito Forebay including power canal and pool; (c) Storage-elevation curve of Thermalito Afterbay. Sum of squared residuals for (d) Lake Oroville, (e) Thermalito Forebay including power canal and pool, and (f) Thermalito Afterbay.
4.3 Results

In this section, the newly developed MOSPD algorithm in Chapter 3 will be implemented into the OTC's cascade reservoir model in order to generate operational alternatives based on different water supply conditions. Specifically, the differences between two extreme solutions, in which one maximizes storage and one maximizes net electricity generation are analyzed. Their potential benefits in responding to these wet, average, and dry conditions are discussed as well.

4.3.1 Settings

In the OTC's problem, the simulation with both the MOSPD and MOCOM algorithms with identical settings is conducted:

(1) The tunable parameters are the Thermalito Forebay and Afterbay daily Feather River release amounts and the remaining internal flow amounts within the OTC system are set to the realistic operation values. As mentioned in Section 4.1, there is a constraint on the monthly total release amount for the Thermalito Forebay and the Afterbay. Thus, the tunable parameters have a dimension of $2 \times (\text{number of days in one month} - 1)$. Other constraints are (1) the storage capacity constraints for the upper and lower limits as shown in Eq. (4-5); (2) power generation capacity constraints as shown in Eq. (4-8); (3) pump-back electricity capacity constraints as shown in Eq. (4-9); and (4-4) the static water-head constraints represented in Eq. (4-10).

(2) The initial conditions (the Feather River releases) are derived using the data-driven model introduced in Chapter 2 with daily resolution.
(3) The boundary-handling method (referred to as the reflecting method) is intended to reset any newly generated offspring that violates its respective constraint. During the evolution, the boundary acts as a mirror and reflects the projection of the displacement. Then, the displacement adjusts the offspring’s location in the parameter space.

(4) Objective functions are the daily storage volume total and net electricity generation as shown in Eq. (4-1). As mentioned in Section 2 of this Chapter, the first objective is an important factor for initiating special operations during drought conditions, and the second objective supplements the energy shortage for transferring and pumping water in the SWP.

(5) The simulations are carried out for the period of April-June, which is snow-melt season for the Sierra Nevada Mountain. Three different years (1998, 2000 and 2001) are included, because these three years are officially recognized as the typical wet, average and dry year, respectively, according to the SWP water supply office.

4.3.2 Uncertainty of Model

To demonstrate the accuracy of the joint model, a model sensitivity analysis is carried out. The sensitivity study compares the objective function value between the real operations scenarios, and the model calculated scenarios with randomly sampled initial parameters for April, May, and June in 1998, 2000, and 2001. The comparison result is shown in Figure 4-4. The colored solid circles represent the real operation scenarios for each month. The hollow star symbols are the objective function values for 25 independent initial parameter sets. For better illustration, the symbols for the model simulated values for each month are all plotted with the same color (black). Nevertheless, the model simulated points of each month are closely clustered around the corresponding real operation.
points. The results indicate the model is able to give reasonably accurate simulations with randomly sampled parameters with the initial settings mentioned above.

Figure 4-4 The comparison of the objective function values between the real operation scenarios and 25 independent runs of the model generated scenarios using the randomly sampled initial parameters for April, May June in 1998, 2000, and 2001

4.3.3 Simulation Results for Dry, Wet, and Normal Years

To apply the proposed optimization scheme to the model, the population size is set to be 128 with 64 individuals in each complex, and the maximum of function evaluation to 10,000 as one of the stopping criteria. The simulation results for the accumulated daily storage and net electricity generation in each month of OTC are shown in Figure 4-5, in which each of the solution points represents one feasible Feather River release strategy during a given month. Different S-E
curve-fitting methods (polynomial, piecewise linearization and successive parabolic interpolation) are also compared. The comparison results for different months using MOSPD are shown in Figure 4-6. Last, the non-dominated solution set for May-2000 is presented in Figure 4-7 (a)-(b). Among the non-dominated solutions, each single point represents a time-series of daily release strategies. To better illustrate the optimized daily operation, two extreme points (solutions) are further demonstrated in Figure 4-5 and Figure 4-7 (c)-(d). One of the solutions maximizes total daily storage volume and is termed the storage optimal operation. The other one maximizes the total net electricity generation and is titled the electricity optimal operation. Their observed daily storage volume is also plotted in Figure 4-7 (c)-(d).

Figure 4-5 Simulation results for MOCOM and MOSPD for April, May, and June in 1998, 2000, and 2001

96
Figure 4-6 Comparison of different S-E curve-fitting methods for April, May, and June in 1998, 2000, and 2001
4.4 Discussion

According to Figure 4-5, both MOCOM and MOSPD are able to generate daily operation strategies in the feasible space with two differences: (1) the non-dominated solutions with MOSPD are located towards higher objective function values and (2) the generated solutions from MOSPD are more uniformly and diversely distributed along the Pareto front. These differences imply that MOSPD is capable of generating better non-dominated solutions in the OTC’s problem, which means that, compared to original MOCOM algorithm, the MOSPD algorithm will be a better tool supporting
reservoir releases decision making. The simulation results from Figure 4-5 also show that, when a monthly total release volume is regulated by DWR, an efficient daily-release adjustment can still benefit the system's output regarding the potential total daily storage and net electricity generation during the month.

The S-E curve-fitting methods comparison results (Figure 4-6) show that the polynomial fitting is a better method for reducing the residuals. The results generated using the polynomial fitting are closer to the assumed “true” (successive parabolic interpolation) compared to that using piecewise linearization. Here, the successive parabolic interpolation is used as a reference because the true Pareto front of the realistic system is unknown for most of the case. By successive fitting of the S-E measurement points with the parabolic function, the global Pareto front is approximated by the “near real” solution front. The distance, which measures how close one solution front is to the “near real” solution front, represents the errors caused by different fitting methods. The closer of the two fronts indicates a better explanation of the “near real” situation and vice versa.

Even though the daily releases are similar among different strategies (Figure 4-7 (a)-(b)), the daily storage volumes could be dramatically different (Figure 4-7 (c)-(d)). Therefore, it is can be inferred that even small changes or adjustments are the crucial factors to influence the entire system regarding the storage volume and other targets. This inference is in agreement with the consultation with DWR staff saying that the Feather River releases are very important operating objectives with regard to the management of OTC’s facilities and SWP. The reason for the large change in storage volume is that the daily release contributes a daily carryover of storage, which forces the storage to either drain or
fill the reservoirs. The carryover of storage accumulates so quickly that in several days the level of reservoir storage reaches its maximum, such as that shown in Figure 4-7 (c) on the 8th, 13th and 19th days. Similarly, the carryover storage forces the storage volume to reach the lower bounds of the reservoirs on the 21th and 26th day in the optimal-electricity operation (Figure 4-7(d)). Larger static water-head difference arises, and more electrical power is generated, when the water level reaches the lower boundary on these two days.

Moreover, the optimized solutions are able to provide better daily operation alternatives in response to dry and wet water supply conditions. If drought conditions are foreseen in the near future, the implements of storage-dominated operation will increase the resiliency of the system in response to the needs of potential water loan or special drought operation. The reservoirs in the system are holding a higher storage volume by reducing the release amount for certain days but supplementing on other days. The higher short-term storage volume is able to execute emergent drought-response operations, such as water supply loaning and storage consolidation. Both these special operations require high-storage volume of one region so that the short-term (days or weeks) water transfer does not drain the water lenders. If the water-supply condition is above normal and no shortages are expected, electricity generation-dominated solutions result in increased hydropower generation in order for the SWP to pump and deliver the water to its users. The increased clean-energy production from hydropower sectors also helps to mitigate the greenhouse gas emissions because the power supply from coal-fired and other forms of energy is replaced by the extra hydropower generation. For the average water supply condition, whether one objective surpasses the other one depends on
decision maker's preference and consideration, which are difficult to be generalized at this point. Nevertheless, the compromised solutions are recommended because both strategies for drought mitigation and increased hydropower generation become important.

4.5 Conclusion

In this chapter, the aim is to enhance the capability and strengthen the application of the newly developed MOSPD algorithm on one of the major cascade reservoir system in California, termed the Oroville-Thermalito Complex (OTC). To address the realistic concerns on daily reservoir operation of OTC, a non-linear optimization model is built based on the realistic reservoir system configuration, engineering constraints and decision makers’ goals under allowable simplifications and assumptions. The impact of the simplifications and assumptions on reservoir topography are analyzed by comparing different curve-fitting methods and corresponding errors. Different from the traditional procedure that separately considers algorithm development and modelling of the real-world problem, the gap between theoretical development and real world implementation is narrowed by simultaneously improving the algorithm and model for the OTC’s reservoir management. The studies demonstrated in both Chapter 3 and 4 provide an integrated platform to exhibit choices in a more transparent and clear format to decision makers in OTC. In detail, OTC’s reservoir operation problem is studied, in which reservoir topography and hydro power generations are explicitly formulated, which dramatically reduces the errors when estimating the S-E relationship. Different curve-fitting methods are compared. The goal is to make the problem formulation as realistic as possible so that
decision makers gain more confidence about the simplifications and assumptions in modelling the real world problem.

Based on the test results demonstrated in this chapter, the following conclusions can be made:

(1) Test results further show that the newly developed MOSPD algorithm is able to generate better non-dominated solutions, based on the diversity and the convergence performance criteria, as compared to the MOCOM algorithm. Different from the comparison results in Chapter 3, which are on human-designed test functions, a real-world application of the newly developed algorithm is demonstrated. Both the diversity and convergence criteria are important in the decision making process allowing water managers to choose the most appropriate reservoir-operation options. The improvements are due to the effectiveness of two distinct enhancements modules, namely, the “possibility-adjustment” and the “dimension monitoring and restoring” modules.

(2) The comparison among the optimal orders of polynomial fitting, linearization and successive parabola fitting helps us to understand the impact of simplifications and assumptions in the way the real reservoir topography is mathematically represented. The results again confirm the claim by Labadie [2004] that non-linear challenges in optimal reservoir-system management should be addressed directly by non-linear programming, as well as the conclusion by Bayón et al. [2009] that linear simplification of the storage-elevation (S-E) curve can produce serious errors.

(3) The optimal solutions derived by the proposed algorithm (MOSPD) are able to provide operation alternatives in response to different water supply conditions, as well as various preferences from decision makers. For the case study provided in this paper, the following overarching recommendations emerge.
i. During dry conditions, the storage maximizing solutions are recommended in order to better respond to any special operating scheme triggered by drought.

ii. During wet conditions, the electricity maximizing solutions are recommended in order to mitigate power shortages and allow production of more clean energy.

iii. The compromised solutions (in the middle ranges of the Pareto front) might be preferred by decision makers, based on their consensus preferences.

Finally, it is believed that the proposed approach, which combines the capabilities of advanced multi-objective optimization algorithms with more realistic (i.e., considering the nonlinearity and complexity) formulations of the system, can provide decision makers with the better picture of the range of options to choose from.

Regretfully, there are still several other non-linear aspects in modelling the OTC’s problem (i.e., water rights [DWR, 2013b], environmental requirements [DWR, 2006], and engineering-optimal design including heterogeneous hydropower units [Li et al., 2013], and non-stable short-term turbine efficiency influence [Diniz et al., 2007], which are not fully considered in this study. These issues currently are either simplified or omitted for further study.. In addition, more interactions between decision makers and algorithm developers are needed in order to allow for better and more realistic formulation of the real-world problem and greater appreciation by the algorithm developers about the complexity of issues facing decision makers. Other potential future work involves the adaptive
changes of the constraints in the optimization process to obtain better Pareto optima [Piscopo et al., 2014].
5. Summary and Recommendation

5.1 Dissertation Summary

The focuses of this dissertation are on (1) developing an data-driven reservoir simulation model and applying statistical approaches on 9 major reservoirs in California; (2) enhancing a global shuffled complex global optimization algorithm and comparing it with multiple cutting-edge algorithms on human-designed test functions; and (3) building a joint water-energy cascade reservoir management model which is capable of addressing some non-linearities for the head-water sources, termed Oroville-Thermalito Complex, in the California’s State Water Project.

The ultimate goal of my studies is to optimize the water supply resilience and hydropower production through the California’s reservoir system and close the “gap” between theoretical approaches and realistic application. To reach this goal, Chapter 2, 3 and 4 are devoted to improve the understanding of general reservoir operation and human’s interference, to provide more powerful heuristic search tool for discovering optimal reservoir operation strategies in a multiple objective context, and to develop physical-based, water-energy joint reservoir operation model with the capability of incorporating realistic non-linearities into decision making, respectively.

5.2 Major Findings and Contributions

The major contribution of this dissertation is the provision of computer-based framework, including advanced analytical approaches, decision making support tools, and modeling platforms to systematically improve reservoir operation regarding water-energy related objectives. Each chapter is
also devoted to partially address individual challenge of how to close the “gap” between theoretical development and real-world application. Specific findings and contributions of this dissertation are listed as follow:

(1) A newly developed data-driven reservoir simulation model is introduced in Chapter 2. Besides traditional hydrological information, such as precipitation, inflows, snow melt, and evaporation that influence reservoir decision making, more types of information, recently being recognized as important factors to reservoir operation, are able to be taken into the simulation framework, i.e. policy/regulation, snow depth, downstream river stages, etc. The flexibility of the proposed model will allow a broader use by communities, and implemented in any reservoir worldwide.

(2) The mechanism and structure of proposed model in Chapter 2 remains simple and computational efficient. The tree-like structure and simple Boolean logic embedded in the CART algorithm provide certain levels of transparency that both model developer and reservoir operator could design proper decision variables and constraints together. Such efforts will be crucial in closing the “gap” between theoretical approaches and real world application.

(3) The proposed model in Chapter 2 is able to provide an accurate estimation on the human controlled outflows. The outflow simulation allows not only a better understanding of human’s interference on natural process, but also by itself stands for a more relevant information than reservoir inflows for downstream water users to make proper water management plans.

(4) The operation patterns for 9 major reservoirs in California are analyzed by proposed approaches in Chapter 2. Without any prior information, model is able to automatically discover that the operation in Oroville Lake, Shasta Lake and Trinity Lake are highly correlated to the changes of
policy and regulation. The influences of policy and regulation could exceed average 30% of that from other operational factors. In addition, the operation in reservoirs which located in low elevation areas or near demand areas, such as the Folsom Lake and New Melones Reservoir, are dominated by reservoir inflow amount with an average contribution percentage over 80%.

(5) A newly developed Multi-Objective Shuffled Complex Evolution Global Optimization with Principal Component Analysis and Crowding Distance (MOSPD) is introduced in Chapter 3, in which two distinct enhancement modules are added to the original Multi-Objective Shuffled Complex Evolution Global Optimization algorithm (MOCOM) in order of improving the diversity and convergence performances of non-dominated solutions.

(6) The proposed MOSPD algorithm in Chapter 3 is tested on 8 human-designed test functions and compared with multiple existing multi-objective heuristic search algorithms, namely, the Multi-Objective Differential Evolution (MODE), the Multi-Objective Genetic Algorithm (MOGA), the Multi-Objective Simulated Annealing (MOSA), and the Multi-Objective Particle Swarm Optimization (MOPSO).

(7) With regard to the diversity and convergence measures, the MOSPD shows the best performance in general and the improvement from original MOCOM is considerable. The enhanced tool developed is universally adaptable in various fields, not limited to reservoir operation.

(8) In Chapter 4, a physical-based cascade reservoir operation model is built for the SWP head-water region, the Oroville-Thermalito Complex. The novelty of the proposed model lies in that two non-linear aspects, namely, the hydropower production and reservoir topography, are incorporated into the model development.
(9) The capability of newly developed MOSPD algorithm is further demonstrated through the application to the cascade reservoir model introduced in Chapter 4. Results show that there is a great potential to improve both reservoir water storage and hydropower generation by altering the daily releases within the engineering constraints.

(10) Using both the MOSPD algorithm and OTC’s reservoir model, multiple operational strategies and alternatives are developed for the run-off season under different dry/wet scenarios. The optimal non-dominated solution allows provides local reservoir operators to analyze the trade-off between water supply and hydropower production, which is also valuable for policy makers to mitigate conflicting interests and choose the most beneficial operational strategies to reach better sustainability for water and energy.

5.3 Future Research Recommendations

As discussed in Chapter 2, 3 and 4, there are still many limitations of the studies conducted in this dissertation. However, the approaches, tools, and modelling frameworks presented in this dissertation are all universally adaptable to other problems under the topic of reservoir operation, optimization algorithm development and reservoir modelling with consideration of non-linearity intrusion. Here, some future research recommendations and promising technologies for further improving the existing tools, models, and approaches are suggested.

(1) Currently, the decision tree algorithms might not successfully predict the dynamic features of the data. In other words, if the value of a predictor is out of the training dataset or the features are out
of the “box” of training sets, the predicted target variable, at the most ideal case, will only have the extreme value of that within the training dataset. The regression trees might not work in predicting the black swan events. Recommendation is made to merge the idea of the Bayesian Process Model with the decision tree algorithms so that the “out-of-box” issues could be properly addressed.

(2) Many performance measures are available besides the diversity and convergence indices that used in this dissertation. Nevertheless, diversity and convergence are the two most essential and popular measures in comparing multi-objective algorithms. Furthermore, the current development of multi-objective optimization algorithm still complies with the famous “no-free-lunch” theory. Eventually, the trading-off among algorithms could be the mitigation of one aspect of algorithm performance with another, as well as the computational efficiency. With the rapid development of high performance computing platform and cloud computing techniques, the limitation of computation resources could be relaxed. If the consideration of computational efficiency is trivial, how to integrate multiple aspects of algorithm performance and provide users with the effective but simple algorithms will be the highest concerns. In the author’s perspective, hybrid optimization algorithm and multi-algorithm ensemble approach will become more and more important in providing such usefulness.

(3) To close the “gap” between theoretical approaches and real world application in reservoir operation not only requires a better knowledge of the complex dynamics of natural processes, proper tools and modelling frameworks, but also needs the involvement from decision makers, communities, and users into the development process of any model and tool. Author believes that
no “silver bullet” is able to easily close the “gap”. Therefore, more real world application cases over different areas worldwide should be carried out to repeatedly test, verify, and improve the proposed approaches, tools, and models in this dissertation so that the existing knowledge and know-how can be bridged and transferred to border industries and communities.
REFERENCES:

Water Supply Overview.


Ackerman, F., and J. Fisher (2013), Is there a water–energy nexus in electricity generation?

optimization (HBMO) algorithm for optimal reservoir operation, Journal of the
Franklin Institute, 344(5), 452-462.

method for prediction, Technometrics, 16(1), 125-127.

Araújo, M. B., and M. New (2007), Ensemble forecasting of species distributions, Trends in
ecology & evolution, 22(1), 42-47.

Arlot, S., and A. Celisse (2010), A survey of cross-validation procedures for model selection,
Statistics surveys, 4, 40-79.

Azadnia, A., and B. Zahraie (2010), Application of Multi-Objective Particle Swarm
Optimization in Operation Management of Reservoirs with Sedimentation Problems, in
World Environmental and Water Resources Congress 2010, edited, pp. 2260-2268,
American Society of Civil Engineers.


DOE (2014), The Water-Energy Nexus: Challenges and Opportunities.


DWR (2009), California Water Plan Update 2009.


DWR (2013a), Feather River Regional Flood Management Plan.


Pearson, K. (1901), LIII. On lines and planes of closest fit to systems of points in space, 

*Philosophical Magazine Series 6*, 2(11), 559-572.


Schaffer, J. D. (1984), Some experiments in machine learning using vector evaluated genetic algorithms (artificial intelligence, optimization, adaptation, pattern recognition), 166 pp, Vanderbilt University.


Timofeev, R. (2004), Classification and regression trees (CART) theory and applications.

Tsai, C.-C., M.-C. Lu, and C.-C. Wei (2012), Decision Tree–Based Classifier Combined with Neural-Based Predictor for Water-Stage Forecasts in a River Basin During Typhoons: A Case Study in Taiwan, *Environmental engineering science, 29*(2), 108-116.


USACE (1990), Modifying Reservoir Operations To Improve Capabilities for Meeting Water Supply Needs During Drought Rep., Institution for Water Resources Corps of Engineers.

USBR (2004), Long-Term Central Valley Project Operations Criteria and Plan CVP-OCAP


