THE IMPACT OF BLACK CARBON DEPOSITION ON SNOWPACK AND STREAMFLOW IN THE WASATCH MOUNTAINS IN UTAH: A STUDY USING MODIS ALBEDO DATA, STATISTICAL MODELING AND MACHINE LEARNING

by

Jai Kanth Panthail

A thesis submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Civil and Environmental Engineering

The University of Utah

May 2015

The University of Utah Graduate School

STATEMENT OF THESIS APPROVAL

The thesis of **Jai Kanth Panthail**

has been approved by the following supervisory committee members:

Steven Burian , Chair **03/09/2015**
Date Approved

Michael Barber , Member **03/09/2015**
Date Approved

Simon Brewer , Member **03/09/2015**
Date Approved

and by **Michael Barber** , Chair/Dean of

the Department/College/School of **Civil and Environmental Engineering**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Salt Lake City, located at the base of the Wasatch mountain range in Utah, receives a majority of its potable water from a system of mountain creeks. Snowmelt runoff from mountain watersheds provides the city a clean and relatively inexpensive water supply, and has been a key driver in the city's growth and prosperity. There has been keen interest recently on the possible impact of the deposition of darkening matter, such as dust and black carbon (BC) on the snow, which might lead to a decrease in its 'albedo' or reflective capacity. Such a decrease is expected to result in faster melting of the snow, shifting springtime streamflows to winter. This study aimed to develop a modeling framework to estimate the impact on snowmelt-driven runoff due to various BC deposition scenarios.

An albedo simulation model, Snow, Ice, and Aerosol Radiation (SNICAR) model, was used to understand the evolution of albedo under different BC loadings. An Albedo-Snow Water Equivalent (A-SWE) model was developed using a machine learning technique, 'Random Forests', to quantify the effect on the state of snowpack under various albedo-change scenarios. An Albedo-Snow Water Equivalent-Streamflow (A-SWE-S) model was designed using an advanced statistical modeling technique, 'Generalized Additive Models (GAMs)', to extend the analysis to streamflow variations.

All models were tested and validated using robust k-fold cross-validation. Albedo data were obtained from NASA's MODIS satellite platform. The key results found the snowpack to be depleted 2-3 weeks later with an albedo increase between 5-10% above current conditions, and 1-2 weeks earlier under albedo decrease of 5-10% below current conditions. Future work will involve improving the A-SWE-S model by better accounting for lagged effects, and the use of results from both models in a city-wide systems model

to understand water supply reliability under combined deposition and climate change scenarios.

TABLE OF CONTENTS

Chapters

Appendices

LIST OF FIGURES

Figures

LIST OF TABLES

Tables

ACKNOWLEDGEMENTS

I would like to express my appreciation to my thesis advisory committee—Dr. Steven Burian, Dr. Simon Brewer (Geography), and Prof. Michael Barber. I am particularly thankful to Dr. Steven Burian, my thesis committee chair and master’s adviser, for constantly guiding me during my time here at the University of Utah. He always encouraged me to innovate at research, and that is something I will take forward. I am also extremely thankful to Dr. Simon Brewer for introducing me to the wonderful world of advanced statistics, and for always being patient in helping me tackle the many obstacles I faced.

I am also thankful to my other teachers here at the University, Dr. Christine Pomeroy, Prof. William Johnson (Geology) and Prof. Brian McPherson, from whom I gained the essential knowledge of environmental and water resources engineering.

CHAPTER 1

INTRODUCTION

1.1. Link between air quality and snowmelt runoff

Salt Lake City is located to the southeast of the Great Salt Lake, and is surrounded by the Wasatch Mountains to the east and the Oquirrh Mountains to the southwest. During the winter, the mountain ranges trap pollutants within the Salt Lake Valley. Temperature inversions in the valley, a phenomenon where cold air is trapped below warm air, further cause air pollution to accumulate near the valley floor. Air quality is particularly poor between November and February [Silcox et al., 2012], with the worst inversion episodes occurring during January and February. The pollution is usually characterized by particulate matter smaller than 2.5 microns in diameter (PM 2.5) [Silcox et al., 2012]. It has been found that 57% of wintertime PM 2.5 emissions are contributed by "mobile sources", which mainly consist of automobiles [Utah Division of Air Quality, 2014]. It was also found that there are an average of 4.3 dust events in the Salt Lake Valley each water year, and that most such events occur in April and September [Steenburgh and Massey, 2012].

Air quality data from an EPA site in Salt Lake City (located at Hawthorne Elementary School, 1675 South 600 East) were used to understand the frequency and intensity of air quality episodes. According to EPA standards, the PM2.5 concentrations 2004 had the most number of days with air quality below mandated standards, with 36 days exceeding 35 μg m^{-3}. Appendix A provides the number of days air quality standards was not met in Salt Lake City, along with the number of days a much higher PM2.5 level of 60 µg m^{-3} was exceeded.

Appendix A shows the PM2.5 concentrations recorded by the EPA site for years 2004, 2007 and 2012, respectively. These years were chosen because they represent years of relatively poor, average and good air quality. Appendix A also shows the concentrations of PM2.5 for the water years (1st October-30th September) 2008-09 and 2011-12, respectively. Regardless of the severity of the pollution episodes, it can be observed that they mainly occur during winter and spring. The majority of the severe events occur between late December and mid-March.

The melting of snow is controlled by, among other factors, its *albedo*. Albedo can be defined as a nondimensional, unitless quantity that indicates how well a surface reflects solar energy [National Snow and Ice Data Center, "Thermodynamics: Albedo"]. Albedo varies between 0 and 1, where 0 indicates 'complete absorption' and 1 indicates 'complete reflection'. The albedo of snow ranges from about 0.9 for freshly fallen snow, to about 0.4 for melting snow, and around 0.2 for dirty snow [Hall and Martinec, 1985]. Albedo is an important parameter in the energy balance of the earth [Dobos, 2003], and depends on both the reflective properties of the surface (e.g. snow grain size, liquid water content, dust or impurities, and surface roughness) and on atmospheric parameters (e.g., solar incident angle, cloud characteristics and air turbidity) [Warren, 1982].

Extensive research has documented the effect of deposition of various darkening matter on the albedo of snow, with the greatest focus on desert dust. It has been hypothesized that measured snow albedos at visible wavelengths are significantly lower than pure-snow values due to the presence of dust or soot, and it has been understood that smaller particles are more effective at lowering albedo [Warren and Wiscombe, 1980]. Absorbing impurities such as dust and carbonaceous particles decreases the spectral albedo in the visible wavelengths, from 0.95–0.98 down to as low as 0.30 [Painter et al., 2012]. Also important is the wavelength in which the albedo of snow is measured. In the near infrared (0.7 to 1.5 µm) and shortwave (1.5 to 3 µm) wavelengths, snow grain growth is the most important parameter that decreases albedo. Light absorbing impurities generally decrease the spectral albedo in the visible wavelength (0.4 to 0.7 µm) [Painter et al., 2012].

The impact of darkening matter on snow albedo can be understood in terms of the energy balance of snow. Solar radiation is the primary driver of snow melt in mountainous areas, with the irradiance and albedo of snow being other factors [Oerlemans, 2009; Bales et al., 2009; Painter et al., 2007]. Solar radiation heats the dust and carbonaceous matter, and these particles then heat the surrounding snow grains by conduction. On reaching 0^0C, the snow grains are affected by further radiative forcing and start to melt. As the snow layer starts to melt, the pollutant particles percolate to the lower layer. This decreases the albedo of the lower layer, and accelerates its melt [Painter et al., 2012]. This feedback loop of decreased albedo and accelerated melting due to pollutant deposition is what contributes to the radiative forcing on snow.

One study has shown that desert dust causes snow to melt 1 month earlier in the San Juan Mountains of Colorado [Painter et al., 2007], and various other studies have further confirmed the impact on snowmelt from mountainous watersheds. One study related errors in the National Weather Service - Colorado Basin River Forecast Center (CBRFC)'s streamflow predictions to interannual variability of dust radiative forcing in snow, using data from NASA's MODIS satellite [Bryant et al., 2013]. It was found that each 10 Watt/m^2 of dust forcing during the melt period contributed to a runoff prediction bias of 10.0% ± 1.5% and a 1.5 ± 0.6 day shift in runoff center of mass. The same study also found that 11 years of mean dust-on-snow (DOS) forcing corresponded to an earlier melt of between 25-30 days relative to clean snow. A related study [Deems et al., 2013] found that extreme dust on snow absorbs up to four times as much radiation as moderate dust, and shifts peak snowmelt between 3-6 weeks earlier. The study also found that extreme dust scenarios mean an annual flow reduction of 1% compared to moderate dust, and a reduction of 6% compared to no dust.

The focus of this study is more on black carbon (BC) deposition on snow, as compared to dust-on-snow. BC is known to be the main component of microscopic soot particles produced from the burning of fossil fuels and biomass, and strongly absorbs solar radiation [Hadley and Kirchstetter, 2012]. Soot can be further defined as a combination of mostly BC, organic carbon (OC), metal and sulfate. BC and soot are therefore more indicative of human activities, mainly in urban areas, and can be hypothesized to be a major component of darkening matter deposition in the Wasatch Mountains near Salt Lake City. Research at the Zhadang glacier in China found BC concentrations of 334-473 ppm in snow, contributing to a radiative forcing of between 1.1-8.6 Watt/m^2 [Qu et al., 2014]. A study in Nepal [Yasunari et al., 2010], at the Yala Glacier, found BC concentrations between 26.0-68.2 μg/kg. These concentrations were estimated to cause albedo reductions of 2.0-5.2%, resulting in a decrease of 11.6-33.9% annual runoff of a typical Tibetan glacier.

Numerous studies have found that snowmelt runoff in the United States and elsewhere is being affected considerably by the impact of rising temperatures [Khadka et al., 2014; Stewart et al., 2004; Stone et al., 2002;]. Spring peak runoff is expected to occur much earlier, and total runoff volume is expected to decrease considerably. Although this study is focused on understanding the changes in snowmelt runoff due to black carbon deposition, future efforts are expected to focus on water supply reliability. This will be done using various climate change scenarios for precipitation and temperature.

BC data for the Wasatch are extremely rare, and it is only possible to estimate the concentrations using other deposited particles as proxies. A study of the Wasatch Mountain snow suggested anthropogenic sources for some carbonaceous matter, including emissions from transportation and industrial activities [Reynolds et al., 2014]. The study analyzed DOS samples, and found organic carbon to range from 0.66% to 5.35% at various locations. Many of the sampling locations explored in the above paper are located extremely close to the study watershed used for this particular study, indicating that results from the paper could be possibly used. However, the paper does not mention the sample weight used to calculate the previously mentioned carbon percentages, which made it impossible to estimate the exact concentrations of carbon at the test sites. Personal correspondence with the authors revealed that the sample weight was possibly not documented. PM2.5 is primarily composed of ammonium sulfate, ammonium nitrate, organic carbonaceous matter and elemental carbon [Neil Frank, EPA], indicating that it could be used as a proxy for BC deposition.

1.2. Salt Lake City's water supply system

Salt Lake City in Utah, USA receives a majority of its potable water from snowmelt-fed streams originating in watersheds in the Wasatch Range of mountains. Seven major canyons in the Wasatch Mountains are the primary water sources for the city, encompassing about 200 square miles and draining approximately 152,000 acre-feet of water every year. While access to some of the watersheds is regulated, intense human activity such as skiing and tourism is prevalent in other watersheds [Salt Lake City Department of Public Utilities, March 1999].

Four creeks, Big Cottonwood, Little Cottonwood, Parleys and City Creeks, supply the majority of Salt Lake City's potable water (Environmental Protection Agency, Jan. 2010). The City Creek watershed has a maximum elevation of 9400 feet, is about 12 miles long and has 19.2 square miles of drainage area. Almost completely owned by the government, the watershed is primarily used for recreational activities apart from supplying water to the city. The Big Cottonwood Canyon watershed drains 50 square miles of area, and has elevations ranging from 5000 feet to more than 10,500 feet. The watershed yields more than 51,000 acre-feet of water, making it the highest contributing watershed in the Salt Lake City area. The *Solitude* and *Brighton* ski resorts are located in the Big Cottonwood Canyon, and constitute the largest human activity in the watershed.

Little Cottonwood Canyon consists of elevations ranging from 5200 to 11,200 feet, and drains an area of 27.4 square miles. It has the second highest yield of all the seven watersheds, with an annual yield of more than 46,000 acre-feet [Salt Lake City Watershed Management Plan]. The *Alta* and *Snowbird* ski resorts are located in the canyon, and bring a considerable number of visitors annually.

Parleys Canyon watershed is the largest among the previously mentioned four watersheds, with a total area of about 50 square miles. Elevations in the watershed range from 4700 feet to 9400 feet above mean sea level. The average annual yield exceeds 18,000 acre-feet [Salt Lake City Watershed Management Plan]. The Interstate-80 freeway passes through the lower part of the canyon, and various recreational activities are common in the watershed area. Two reservoirs, Little Dell and Mountain Dell, are located within the watershed and are used to store peak springtime flows for future use. Parleys Canyon was chosen as the watershed of choice for this study because of its proximity to both the city and possibly to another source of particulate pollution, the I-80. It is also impacted by various anthropogenic activities. Another important factor for selecting Parleys was the availability of a continuous streamflow record for the study period (2001-2013). The modeling workflow and techniques used in this study can be easily replicated for the other watersheds in Salt Lake City, or watersheds located elsewhere, if need be and the data for calibration exist.

1.3. Study region

Parleys Canyon lies between Emigration Creek watershed to the north and Mill Creek watershed to the south. Figure 1 shows the watershed delineated using a US Geological Survey (USGS) Digital Elevation Model (DEM) in ArcMap. The watershed consists of three major streams – Parleys Creek, Lamb’s Creek and Alexander Spring Creek. The outlet chosen for watershed delineation is located very close to the streamflow gauging station, and is just upstream of the Mountain Dell Reservoir. It should be noted that there are actually three Parleys Canyon watersheds. The watershed used for this study includes the area draining to Mountain Dell, and does not include the portion to the north (draining to Little Dell Reservoir) and the controlled flow area downstream of Mountain Dell Reservoir.

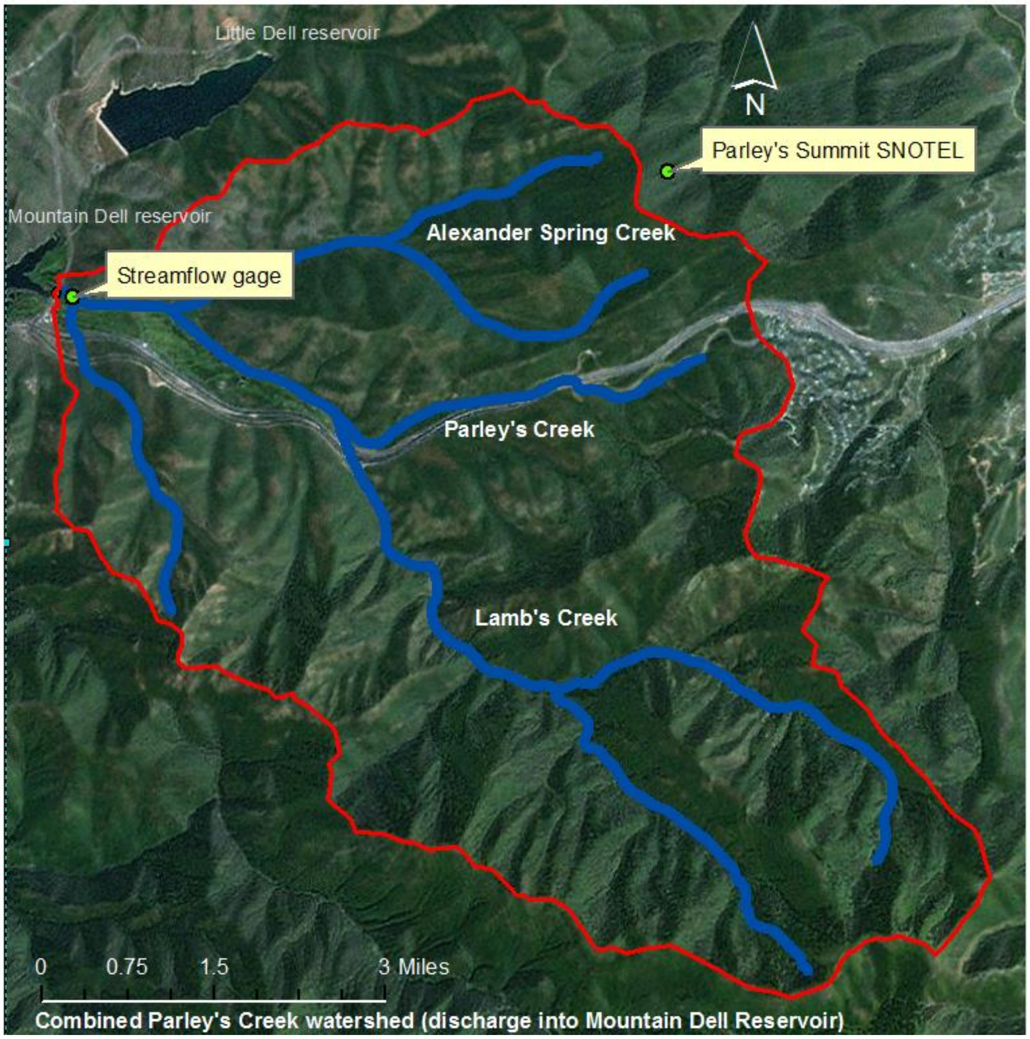

Figure 1: Parleys Creek watershed (shown along with data sources)

CHAPTER 2

METHODS AND DATA

2.1. Snowmelt modeling: An introduction and comparison of methods

Snow is a unique component of the hydrological cycle, because snowpack acts like a reservoir, releasing runoff due to variations in temperature and other factors. Hydrological modeling without snowmelt modeling is usually a combination of the processes of precipitation, infiltration, evapotranspiration (ET), subsurface transport and surface runoff. The inclusion of snowpack increases the effect of lag, which can often be a challenging phenomenon to model. It is therefore important to understand the energy balance of snow, along with the effects of surface topography, vegetation and other factors on melt.

Energy balance processes related to snowmelt are known to include net radiation (‘shortwave’ or solar radiation, and ‘longwave’ or atmospheric radiation), latent and sensible heat transfer, and heat due to precipitation [Anderson, 2006]. Net shortwave radiation depends on solar output and surface albedo, with the albedo varying between 0.9 and 0.4 for snow. Longwave radiation is emitted by the atmosphere and various particles in the atmosphere, and is influenced by the amount of water vapor and air temperature [Anderson, 2006]. Melt occurs when the temperature of the snow surface rises to 0^0C, and runoff from the snowpack occurs when it cannot hold any further melt water in the pore spaces. The melt process can be further accelerated by the absorption of heat from precipitation falling on the snow [USACE, 1998].

This study presented a unique challenge in terms of the choice of modeling framework, especially due to the complex interactions between air quality and hydrology required to be studied. Substantial effort was put into exploring various physically-based hydrological models, including the Snowmelt Runoff Model (SRM), the SNOW-17 snow accumulation and ablation model, Utah Energy Balance (UEB) model, Variable Infiltration Capacity (VIC) model and the Gridded Surface Subsurface Hydrologic Analysis (GSSHA) model. Some of the models (example, SRM and UEB) were designed to be primarily snowmelt models for use in alpine watersheds, whereas others were general hydrological models (VIC and GSSHA) with snowmelt-routines included. The models can also be classified as temperature-index models (example, SRM and Snow-17), with simple melt equations based on air temperature, and energy-balance models (example, UEB), which depend on relationships between incident, absorbed and reflected energies to calculate melt.

It was understood that most snowmelt models did not have a parameter such as ‘snow albedo’ that could be modified to account for particulate deposition, and even if they did, were extremely complicated to set up and use. UEB and GSSHA showed the most promise for use in the study, but it was possible that streamflow variations arising due to errors in model calibration could be confused with variations due to albedo reductions. Hydrological models also have many parameters and state variables that need to be estimated based on watershed characteristics.

2.2. Data-driven methods: Statistical modeling in hydrology

Based on the study of various physically-based hydrological models, other modeling options were explored. The recent surge in the application of data analytics to various fields such as the social sciences, bioinformatics, computer science and business has resulted in the availability of a plethora of data-driven modeling options. Statistical modeling has advanced much beyond basic linear regression and classification to advanced methods such as Generalized Additive Models (GAMs) and Generalized Linear Mixed Models (GLMMs), capable of interpreting nonlinear relationships and lag in time series data. Machine learning techniques such as Recursive Partitioning and Random Forests allow classification of a large amount of data in order to predict the effect on one variable due to changes in another variable.

Many statistical modeling techniques and almost all machine learning (ML) methods are very computationally-intensive, but the availability of easy-to-use and optimized libraries in the *R* statistical language [R Core Team, 2013] make their application relatively straightforward. Robust validation methods such as k-fold cross-validation allow the predictive skill of statistical models to be tested.

2.3. Application of remote sensing data to snowmelt modeling

Snowmelt modeling, like most parts of hydrological modeling, is extremely dependent on accurate measurement of various parameters and state variables for accurate results. Various snow data measurements, like snow water equivalent (SWE), snow albedo, snow depth and snow cover, have historically been made using snow course data. The SWE data used in the study have been obtained from a snow telemetry (SNOTEL) site operated by the United States Natural Resources Conservation Service (NRCS), collected using a snow pillow and transmitted using telemetry [Schaefer and Paetzold, March 2000]. Field measurements, unless automated like SNOTEL, can be cumbersome and expensive. Also at best they represent point data at the station of measurement, and not spatial data about the watershed of interest.

Remote sensing allows hydrologists to understand the spatial and long-term temporal trends in snow properties and behavior, especially in remote alpine watersheds. Data can be collected either from low-flying aircraft, or from satellites sources like

Scanning Multichannel Microwave Radiometer (SMMR), Special Sensor Microwave/Imager (SSM/I), Advanced Microwave Scanning Radiometer - Earth Observing System (AMSR-E) and Moderate Resolution Imaging Spectroradiometer (MODIS) [WMO]. Data are collected in various bands of the electromagnetic spectrum like visible, shortwave, infrared, thermal infrared, microwave and gamma [WMO].

Although there are many advantages with using satellite-sensed data for snowmelt studies, there are also certain challenges to overcome. Such data can sometimes be temporally infrequent, as in the case of LANDSAT, with a temporal resolution of 16 days. Although methods have been developed to distinguish various surfaces in satellite images [Crane and Anderson, 1984; Dozier and Marks, 1987], problems still persist when substantial forest cover, shadows and rocks are present [WMO]. Cloud cover can also be a significant issue, with only certain bands in the spectrum capable of differentiating between clouds and snow. Even with these challenges, remotely-sensed data provide the most convenient and spatially accurate data for understanding the role of snowmelt in hydrology.

Datasets widely used in snow hydrology are the MOD- and MYD- suites of products from the Terra MODIS and Aqua MODIS satellites, respectively. The MOD- and MYD- suites include snow cover data with different spatial resolutions, ranging from 500 m on a sinusoidal projection, to 0.05^0 and 0.25^0 resolution on a geographic Lat/Lon projection. Temporal resolutions range between daily and monthly for various products in the suite [Hall and Riggs, 2007]. This particular study has extensively used data from the Moderate Resolution Imaging Spectroradiometer (MODIS) platform, specifically from the MCD43A3 Albedo Product (MODIS/Terra Albedo Daily L3 Global 500m SIN Grid) dataset [Professor Schaaf's Lab].

Along with trying to answer some important questions on water supply reliability under conditions of pollutant deposition on snow, this study also attempts to determine the applicability of MODIS albedo data for long-term hydrological analysis. Future work is also expected to include other satellite datasets in improving the process of connecting air pollution, contaminant deposition, snow processes and runoff.

2.4. Black carbon (BC) deposition scenarios

Due to exact black carbon (BC) concentrations for the Wasatch not being available, it was not possible to determine albedo impact under contaminant deposition as it was initially planned in the study. The Snow, Ice, and Aerosol Radiation (SNICAR) model (Flanner et al., 2007; Flanner at el., 2009) uses a two-stream radiative transfer solution from Toon et al. (1989) to calculate the albedo of snow for various combinations of deposited pollutants. The SNICAR model allows for the calculation of albedo affected by black carbon deposition, and the SNICAR analysis provided in this study can be used to determine the impact on Wasatch albedo once BC concentrations are determined.

Future work is expected to use these concentrations to formulate scenarios for BC deposition and to understand water system reliability under such circumstances.
The albedo values used in this study, a daily time-series obtained from MODIS satellite data, represent the actual state of albedo in the watershed. That is, the time-series represents albedo under current deposition conditions and can be used as the ‘base case’ for any scenario formulation. Once actual BC concentrations are obtained, for example 500 ppb (parts per billion), SNICAR can then be used to generate the albedo for that concentration of BC, and for concentrations of BC lower than (200 ppb, 300 ppb etc.) and greater than (800 ppb, 1000 ppb etc.) the actual concentration. These albedo values could then be used to generate scenarios based on the percentage change of albedo from the base case.

This study presents an albedo change analysis that is disconnected from the SNICAR model. The SNICAR analysis presented in this paper is only meant to demonstrate the model’s capability in determining albedo under various BC deposition conditions. The statistical and machine learning models described in this paper operate

independently, and are used to track snowpack state change (via SWE) and streamflow variations. The models operate based on theoretical percentage changes in albedo.
In order to understand the impact of varying albedo on snowpack state and streamflow, various albedo scenarios were applied to both models developed. As described earlier, due to lack of black carbon (BC) data for the Wasatch, these scenario represent percentage change in albedo over each year. If continuous (time-series) or frequent BC data becomes available at some point, the input albedo values to the models can be modified appropriately using the SNICAR model results described in this study.

The models were run using ‘percentage-change in albedo’ scenarios of +10%, +5%, -5% and -10%. The albedo time-series was only modified for the months of January, February, March, April, May, November and December each year, as snowpack in the Wasatch is not known to commonly exist in noticeable amounts outside these months. If there is assumed to be no BC in the snow, albedo change scenarios of -10% and -5% would represent deposition of about 1500 ppb and 500 ppb BC, respectively, based on the SNICAR model results. Similarly, the +10% and +5% albedo change scenarios would then represent a change from 1500 ppb and 500 ppb deposition to zero BC deposition. Other change scenarios could be used to understand the effect on SWE due to relatively lower or greater albedo impacts.

2.5. Use of the SNICAR snow albedo model to estimate the sensitivity of broadband snow albedo to various pollutants and varying concentrations

Pollutants in the snow can include BC, dust and volcanic ash. Inputs to the model include type of incident radiation (‘Diffuse’ or ‘Direct’), solar zenith angle (if radiation is direct), snow grain effective radius, snowpack thickness, snowpack density and albedo of underlying ground. Concentrations of uncoated and sulfate-coated BC, dust of various sizes and volcanic ash can be entered to calculate the effects of impurities on snow albedo. Although an online interface for SNICAR exists (http://snow.engin.umich.edu/), a MATLAB script provided by Dr. Mark Flanner (Atmospheric, Oceanic and Space Sciences, University of Michigan) was used for the purpose of this study. Appendix B lists some of the parameter values used.

SNICAR was run using varying uncoated BC concentrations from 0 to 3000 ppb (parts per billion). Figure 2 graphically shows the depletion of snow albedo with increasing BC. Appendix B shows the % change for each 100 ppb increase in BC concentration, both from the previous BC concentration and from zero BC concentration. The albedo calculated by SNICAR for zero BC concentration is 0.8273, which drops to 0.8105 with 100 ppb BC (an approximate 2% reduction in albedo).

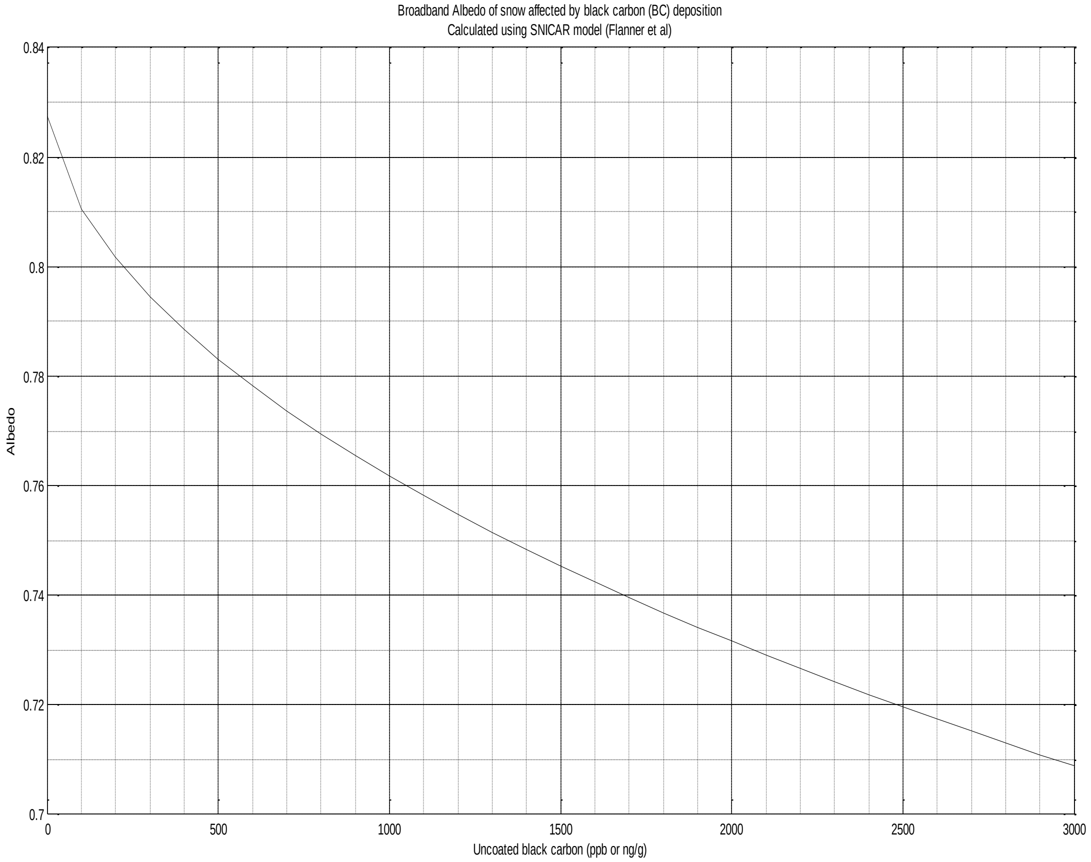

Figure 2: Effect of varying black carbon concentrations on snow albedo

It can be observed that although the albedo keeps decreasing with increasing BC deposition, the rate of albedo depletion slows down. The albedo only decreases between 0.3-0.4 % over the previous BC concentration, once BC exceeds 1500 ppb in the snow. Even then, the percentage of albedo depletion from zero albedo in snow is significant. It can be seen that with 500 ppb BC in snow, albedo is depleted 5.34% from zero BC in snow. The depletion is 7.39% and 9.31%, for 1000 ppb and 1500 ppb, respectively. Snow albedo is reduced by 11.57% and 14.31%, with 2000 ppb and 3000 ppb BC, respectively.

2.6. Description of data

2.6.1. MODIS albedo data: The MCD43A3 dataset

Albedo can be defined as the 'ratio of upwelling to downwelling radiative flux at the surface', with downwelling flux being a sum of a direct component and a diffuse component [Professor Schaaf's Lab]. White-sky albedo is the bihemispherical reflectance under conditions of isotropic illumination, with the angular dependency removed. Black-sky albedo is the directional hemispherical reflectance computed at local solar noon [MODIS-Atmosphere, NASA GCFC]. The white-sky and black-sky albedos allow the actual albedo to be calculated for a number of illumination conditions [Roman et al., 2010], by interpolating as a function of the diffuse skylight [Lewis and Barnsley, 1994].

The MCD43A3 Albedo Product, in the MODIS/Terra Albedo Daily L3 Global 500m Sinusoidal Grid, provides the white-sky and black-sky albedos (at local solar angle) for MODIS bands 1-7 as well as for three broad bands ((0.3-0.7µm, 0.7-5.0µm and 0.3-5.0µm) [Professor Schaaf's Lab]. The shortwave (0.3-5.0µm) broadband domain is the most important for this study, as it primarily characterizes the total energy reflected by the earth's surface [Liang and Walthall, 1999]. Version v006 MCD43A3 data for MODIS h09v04 grid (containing Utah) were obtained from University of Massachusetts, Boston [Professor Schaaf's Lab] as daily HDF (Hierarchical Data Format) files. Figure 3

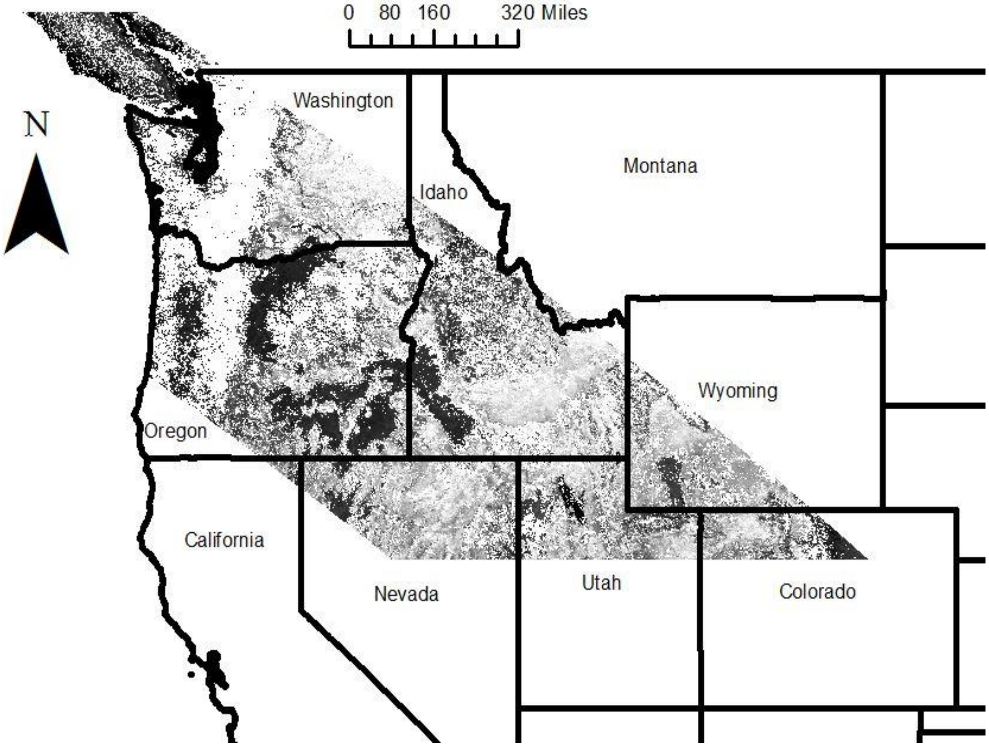

Figure 3: MCD43A3 band 29 (shortwave broadband white-sky albedo) for January 1, 2009 (h09v04 grid)

shows a sample MCD43A3 file for the h09v04 grid.

White-sky and black-sky albedos were extracted from MCD43A3, for all days in 2001-2013. Within each HDF file, shortwave broadband white-sky and black-sky albedos are specified as sub-datasets 29 and 19, respectively. The albedo is scaled down by a factor of 1000, with ‘no value’ pixels indicated by a value of 32766. ArcMap’s Model Builder was used to extract the two sub-datasets, clip each file to the extent of the study watershed and convert to NetCDF format for analysis with *R*. Figure 4 shows the Model Builder schematic used for this process, and Figure 5 shows a sample of white-sky albedo extracted for the study watershed using the Model Builder model. A script written in *R*, utilizing the *ncdf* package [NCDF], was used to extract spatially-averaged (mean) albedo from each clipped NetCDF file. The script is shown in Appendix D.

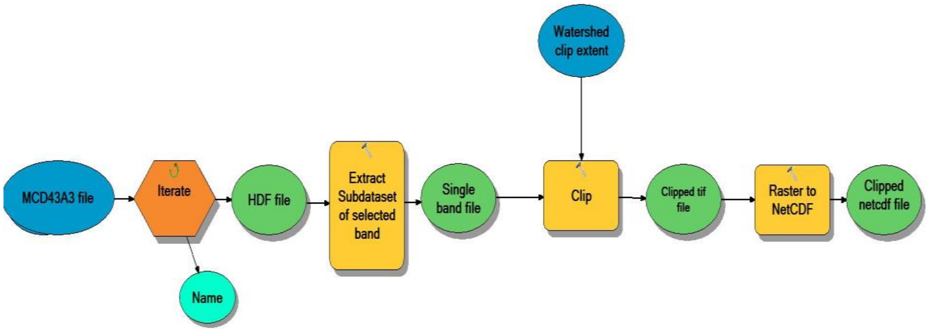

Figure 4: ArcMap ModelBuilder tool used to extract albedo data from HDF files

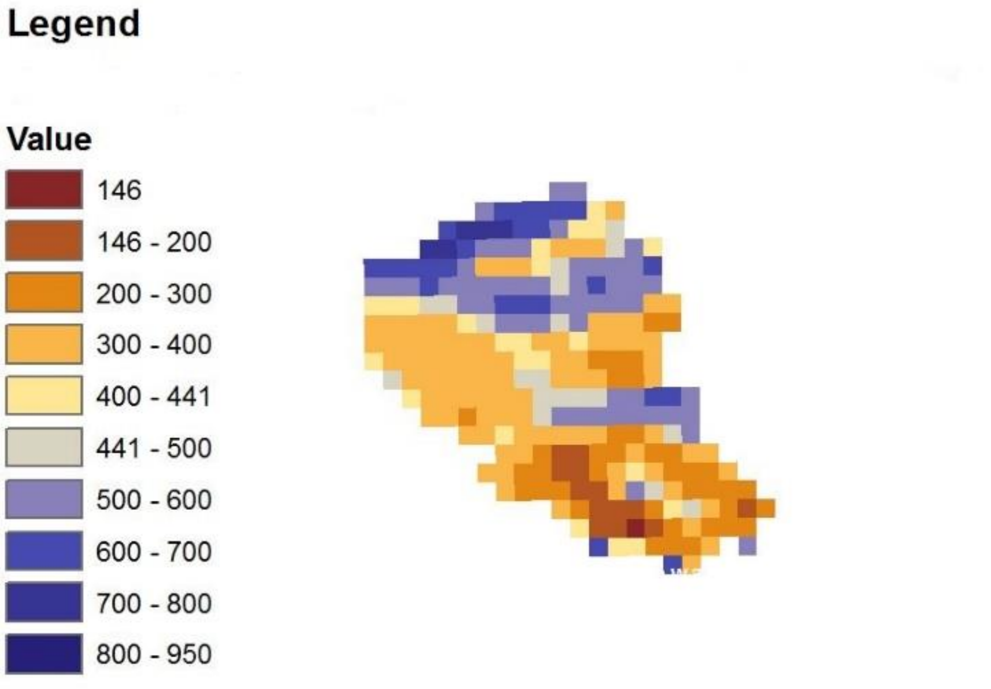

Figure 5: Sample white-sky albedo extracted from MCD43A3 band 29 (January 1, 2009) for Parleys watershed (albedo scaled up by 1000)

An analysis of spatially-averaged albedo extracted from MCD43A3 shows close correlation between white-sky and black-sky albedo, especially during periods of dense snowpack (Figure 6). The Pearson's product-moment correlation test, a measure of linear correlation, on the two albedos provides a correlation value greater than 0.99, with a p-value < 2.2e-16. For the purpose of this study, only white-sky albedo will be considered so as to simplify analysis. Also, the method of spatial averaging is assumed to capture the temporal trend of the albedo and to suffice for modeling. This assumption might be considered relevant in this study due to the fact that all other data used, from the SNOTEL site, EPA site and streamflow gage, are also point-measurement data. Future studies could benefit from integrating spatially varying albedo in the modeling framework, or by understanding the variation between spatially-averaged albedo and point albedo at various locations on the watershed.

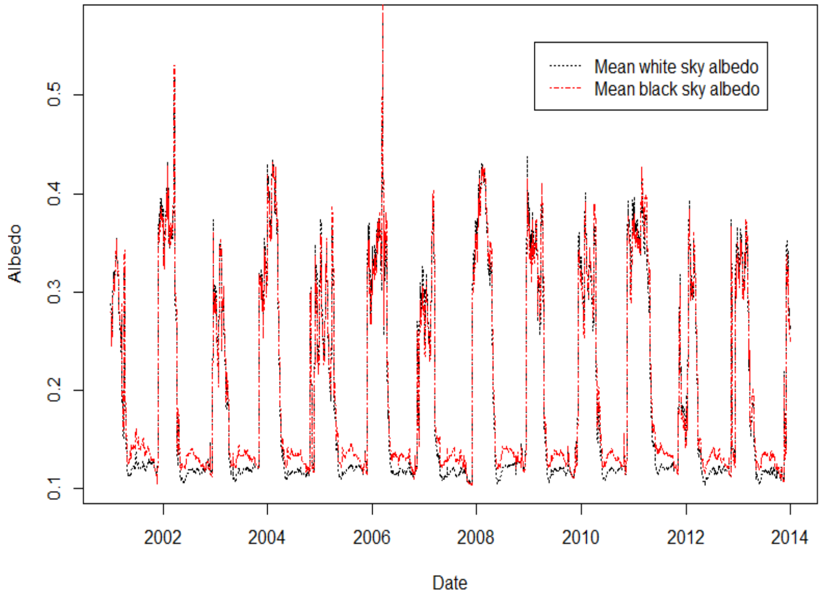

Figure 6: Correlation between mean white-sky and mean black-sky albedo for Parleys Creek (albedo derived from MCD43A3 dataset)

2.6.2 SNOTEL data (precipitation, temperature and snow water equivalent)

SNOTEL is a system of automated sensors that measure snowpack and other related climate data, operated by the Natural Resources Conservation Service (NRCS) of the United States Department of Agriculture (USDA). There are more than 600 SNOTEL (snow telemetry) sites in 13 states. Variables measured by the SNOTEL network include snow depth, soil moisture and temperature, precipitation, wind speed, solar radiation, humidity and atmospheric pressure. For this study, data were obtained from SNOTEL site 684 'Parleys Summit' located at latitude 40 deg 46 min N and longitude 111 deg 38 min W, at an altitude of 7500 feet (2286 meters). Figure 1 shows the location of the SNOTEL site. The station has been reporting since 1978, with a combination of daily and hourly sensors.

Daily accumulated precipitation, average temperature and snow water equivalent data from 2001-2013 were obtained in the form of CSV files from the station's web interface (http://www.wcc.nrcs.usda.gov/nwcc/site?sitenum=684), with one CSV file for each year. An *R* script (Appendix E) was used to automate the import process to *R*, the interpolation of missing data and the conversion of accumulated precipitation to continuous measurements. Appendix C shows the extracted precipitation, temperature and snow water equivalent (SWE) data, respectively.

2.6.3 Streamflow data for Parleys Creek

Salt Lake City Department of Public Utilities (SLCDPU) provided streamflow data for Parleys Creek from 2001-2013, for a gauging station located on Lamb's Creek at Latitude 40.754761 and Longitude 111.708534. Lambs Creek and Alexander Spring Creek combine with Parleys Creek before it drains to the Mountain Dell Reservoir, and the gauging station is located after this merging. Figure 1 shows the location of the gauging station, and Appendix C shows the streamflow data extracted from the files provided by SLCDPU.

Flows in the Parleys Creek watershed are known to exhibit distinct springtime peaks typical of snowmelt-driven systems [Salt Lake City, 1999], with wide variability among year-to-year peak flows. Flows in the water years 2005-2006, 2006-2007 and 2011-2012 are significantly higher compared to other years, driven by greater snowpack and possible mid- and late-winter storms. Appendix C shows discharge in Parleys Creek for the years 2002, 2005 and 2008, respectively.

It was observed that some years, like 2002, have a fairly steep rise towards peak discharge, followed by an equally steep drop towards summer flows. The steep drop possibly indicates that spring temperatures were relatively higher, with no spring-time snowstorms, and this led to rapid snowpack depletion. Other years, like 2005, exhibited less drastic changes, but also some substantial late-spring and early-summer streamflow changes. These changes were possibly driven by storms, and by sudden changes in air temperature. For certain years, for example 2008, it was difficult to accurately assign a pattern to streamflow. Such years exhibited numerous drops and rises in streamflow, possibly driven by severe storms between May and July.

Considering that a data-driven modeling approach is used for this study, it is essential to note trends such as those mentioned above. A physical model is driven by mathematical relationships between various variables, whereas a data-driven approach is based on relationships between trends of various variables.

2.6.4. Air quality data (PM 2.5 and PM 10) from EPA

Black carbon aerosols have been defined as the solid component of PM 2.5, with PM 2.5 being particulate matter (PM) with sizes less than 2.5 micrometers. PM 10, composed of larger particles and usually representative of dust, have sizes less than PM 10. The closest US Environmental Protection Agency (EPA) reporting station to the Parleys Creek watershed is located at Hawthorne Elementary School, 1675 South 600 East, Salt Lake City. There is no station reporting continuous air quality at the watershed location, and the EPA site remains the best data source at this moment.

The station is located at elevation of 4285 feet, at latitude 40.736389 and longitude -111.872222. PM 2.5 data were obtained from the ‘Federal Reference Method (FRM) Network’ (Parameter code 88101), and PM 10 data were obtained from Parameter code 81102 [Utah State DAQ, 2010]. Appendix C shows the PM 2.5 and PM 10 data obtained from EPA’s Hawthorne site, respectively.

2.6.5 The combined dataset: parleys_data

All the variables used for this study were compiled into a single *R* dataframe, parleys_data. Dataframes are the fundamental data structure used within *R*, and contain variables with the same number of rows. They have a class name of “data.frame” within *R*, and each variable in a dataframe is represented by a unique row name [R documentation, data.frame {base}].

In the dataframe, average air temperature is represented by the variable ‘tavg’, precipitation by ‘precip’, streamflow by ‘flow’, mean white-sky albedo by ‘mean_wsa’, mean black-sky albedo by ‘mean_bsa’, PM 10 by ‘pm10’, PM 2.5 by ‘pm2.5’, snow water equivalent by ‘wteq’ and dates by ‘date’. Table 1 shows the summary statistics of various continuous, time-series variables in parleys_data.

Various other temporal variables were also created, in addition to the time-series variables. These include the current ‘timestep’ (from 1 to 4748), ‘month_number’ (1-12), ‘year’ (2001-13), ‘day_ofyear’ (1-365/366) and ‘day_number’ (day of the month). These variables were used to include the temporal trends in the statistical models. Certain additional variables were added to parleys_data while creating the models, and these are separately described in the sections about the models.

The Pearson product-moment test, using the *R* function cor() [Becker et al., 1988], can be used to understand the correlation between variables. The correlation test provides the *Pearson correlation coefficient*, which falls between +1 and -1 inclusive. A value of 1 indicates total positive correlation, 0 indicates no correlation and -1 is total negative correlation. Figure 7 shows the correlation values generated for various variable pairs in parleys_data, along with a matrix of scatterplots generated using the ggpairs() function [R documentation, GGally: Extension to ggplot2]. Table 1 shows summary statistics for various variables in parleys_data.

PM2.5 and PM10 are observed to be positively correlated with a Pearson coefficient of +0.76, indicating that it might be possible to use only one of them while modeling the air quality link with albedo. PM2.5 is better correlated (+0.32) with mean white-sky albedo (mean_wsa) than PM10 is with the albedo (+0.18). The albedo (mean_wsa) is negatively correlated (-0.67) with average temperature (tavg), which is expected considering that albedo decreases with increase in temperature and subsequent melting. The snow water equivalent (wteq) is negatively correlated with temperature (-0.53), since SWE decreases with increase in temperature. SWE is also strongly correlated with the albedos (+0.76 and +0.73), which points to the fact that air quality and snowpack properties follow the same temporal trend. This indicates that adding a temporal trend to any model relating both might increase prediction power significantly. Most other variable pairs do not have significant coefficients, either due to nonlinear relationships or high lag.

2.7. Statistical modeling techniques

Machine learning is driven by large amounts of data and algorithms, whereas statistical modeling is driven by assuming a model for the data [Breiman, 2001]. One model used in this study, the Albedo-SWE model, was built using the Random Forests machine learning technique. Another model, the Albedo-SWE-Streamflow model, was built using the Generalized Additive Models (GAMs) statistical modeling framework.

Linear regression, in statistics, is a method of modeling the relationship between a dependent variable and one or more predictor or explanatory variables. In case a single

Table 1: Summary statistics of variables in parleys_data

Statistic/ Variable	**precip**	**flow**	**tavg**	**pm10**	**pm2.5**	**mean _wsa**	**mean _ bsa**	**wteq**
Min	0	0.5	-19.2	2	0	0.104	0.104	-0.050
Max	2.2	134.82	26.5	360	94.2	0.592	0.573	24.45
Mean	0.089	8.119	5.78	25.67	10.737	0.2	0.194	4.177
Median	0	4.64	5.1	21	7.1	0.139	0.124	0.25
Standard Devation	0.203	11.327	9.063	18.844	11.099	0.096	0.102	5.802

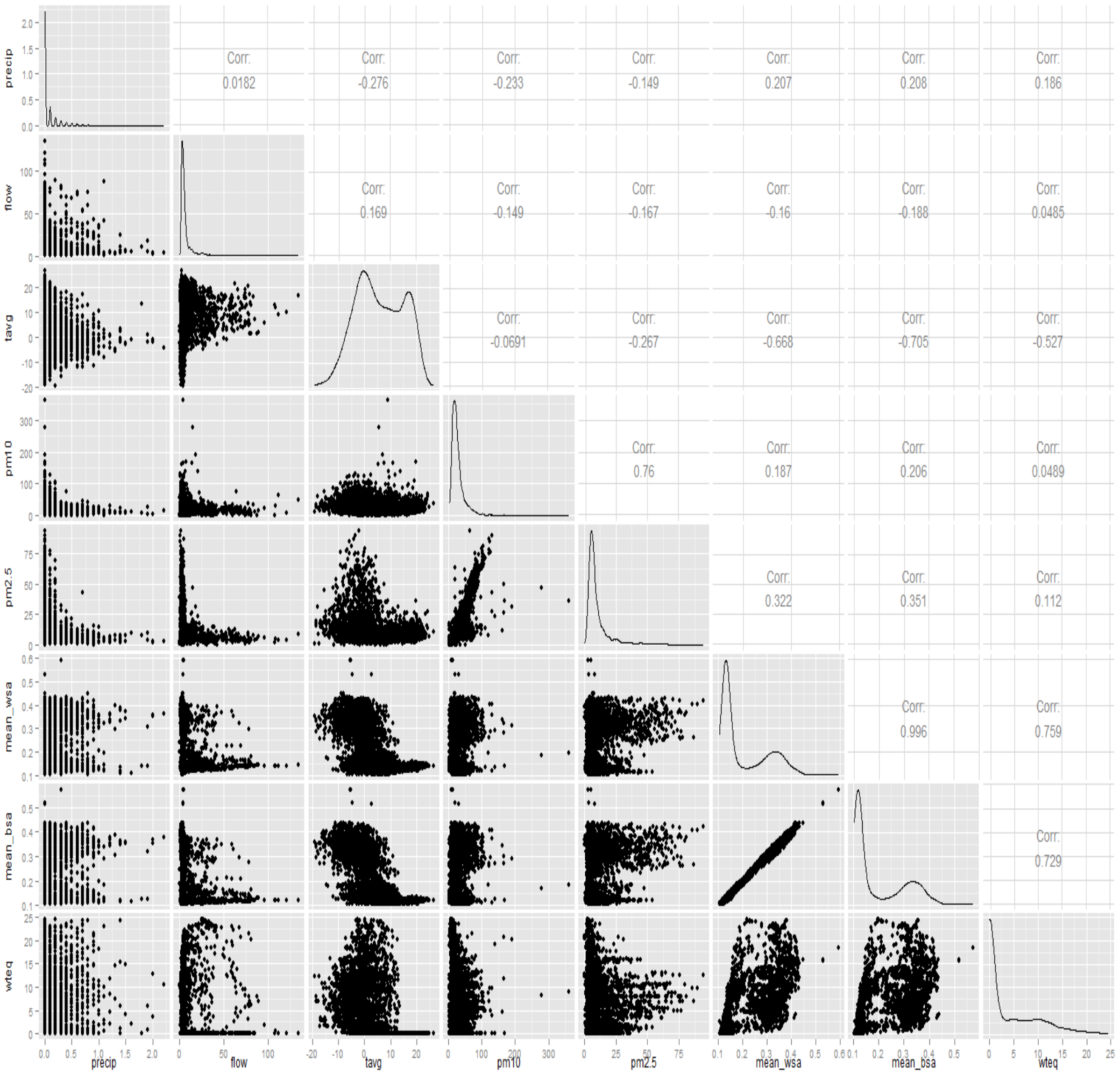

Figure 7: Correlation scatterplot for parleys_data

predictor variable is used, it is referred to as *simple linear regression.* Use of multiple predictor variables is termed *multiple linear regression*. Such models are usually fitted using a least squares approach (*ordinary linear regression*) or a maximum-likelihood estimation approach.

Simple Linear Regression:

$$\boldsymbol{y = bx + \epsilon}$$

Multiple Linear Regression:

$$\boldsymbol{y_i = b_0 x_{i0} + b_1 x_{i1} + b_2 x_{i2} \ldots\ldots + \epsilon}$$

Where;

x=predictor variable(s)

y=predicted variable

b=coefficient of relationship (slope)

ϵ=model error

A further extension of ordinary linear regression is the generalized linear model (GLM), which allows for the use of predicted (or dependent) variables that are necessarily not normally distributed. Such models incorporate other distributions, usually of the exponential family, through a link function [Clark]. The link function links the mean of the predicted variable to the predictors, and performs internal transformation to linearize the relationship between variables. The basic structure of a GLM is shown below.

$$\boldsymbol{E(Y) = g^{-1}(\eta)}$$

Where;

E(Y)=expected value of predicted variable Y, generated from a distribution

η=linear predictor

g=link function that relates linear model to predicted variable, Y

Generalized Additive Models (GAMs) are another approach to incorporate nonlinear predictors into the modeling framework, while retaining the ability to model non-normal dependent variables. GAMs use smooth functions of the predictor variables to determine the predicted variable. Nonparametric methods are used to fit individual functions to each predictor, usually in the form of a spline or loess of some form. GAMs are extremely useful when there exists a very complex relationship between the predictor and predicted variables, which cannot be fitted even using GLMs. The R package mgcv, using the function gam, fits a generalized additive model to data with the Generalized Cross Validation (GCV) method [R gam {mgcv} documentation]. The function allows the user to specify the family to be used for the distribution and link, along with other function parameters. The general form of a GAM is given below.

Example GAM:

$$\boldsymbol{g}\big(\boldsymbol{E}(\boldsymbol{Y})\big) = \boldsymbol{b_0} + \boldsymbol{f}(\boldsymbol{x_1}) + \boldsymbol{k}(\boldsymbol{x_2})$$

Where;

E(Y)=expected value of predicted variable Y, generated from a distribution

g=link function that relates predictors to predicted variable, Y

x=predictor variable(s)

f, k=smoothing functions

Model selection and validation allows the modeler to choose optimal parameters and values, such as the number of predictor variables and model fitting options, in order to get the best possible model that generalizes to new data. It is often useful to compare the fit and predictive skill of a model, while minimizing model complexity and model run time. In many cases, a model fits the training data very well, but is not able to predict using new data-a phenomenon referred to as ‘over-fitting’. Measures of model prediction error are only useful when combined with a method to test the model using new data.

In this study, *k-fold cross-validation* was used as the model validation and parameter selection method. This method creates k partitions of the dataset, and k-1 partitions are used for training the model in each iteration. The remaining one partition is used to test the model prediction. This process of iteration can be used to calculate the *root mean squared error of prediction (RMSEP)* and the *r-squared of prediction (r^2p)* of model prediction accuracy, of each combination of model parameters. The best combination is the one that either minimizes the RMSEP or maximizes the r^2p, with the RMSEP a better judge of model prediction accuracy than the r^2p [Cornell University, 2012].

$$\boldsymbol{RMSEP_p = \sqrt{\frac{\sum_{i=i}^{n}\left(y_i - y_p\right)^2}{n}}}$$

Where;

y=actual (real) value of predicted variable (in test set)

y_p=modeled value of predicted variable

$$\boldsymbol{r^2p = 1 - SS\frac{res}{SS_{tot}}}$$

Where;

$$SS_{res} = \sum_i (y_i - f_i)^2 = residual\ sum\ of\ squares$$

$$S_{tot} = \sum_i (y_i - \bar{y})^2 = total\ sum\ of\ squares\ \propto variance\ of\ the\ data$$

f=modeled value of predicted variable

y=actual (real) value of predicted variable

$\bar{y}$=mean of actual data

2.8. Machine learning, trees and random forests

Machine learning is a field of study and development of algorithms that learn from data [Kovahi, 1998] to build models that can predict. Designed to make decisions based on available data [Simon, 2013], machine learning is both related to and different from statistical modeling. It is extremely useful when there are complex interactions between the predictor variables, and when the relationship between predictor and predicted variables is nonlinear. Machine learning is widely used in computer science, especially in the fields of computer vision, pattern recognition, artificial intelligence and data mining. Machine learning is an algorithmic modeling approach, differing from statistical modeling because it begins with a black-box assumption of no known relationship between predictor(s) and predicted variables [Breiman, 2001].

Machine learning methods are usually classified as ‘supervised learning’, ‘unsupervised learning’ and ‘reinforcement learning’ [Russell et al., 2003]. Supervised learning refers to models that develop a function based on some learning data [Mohri et al., 2012], in order to make predictions on or classify new data. In unsupervised learning, a learning algorithm finds patterns directly in input data. The method used in this study, Random Forests, is an ensemble supervised learning technique. Ensemble methods use multiple learning algorithms to make decisions [Polikar, 2006]. Such models allow for the exploration of model uncertainty, and account for the fact that many weak models together can be more robust than a single overfitted model.

Decision trees is a machine learning technique that uses tree-like models relating predictors and predicted variables. Decision trees can be used both for classification and regression, and a combination of both is usually referred to as Classification And Regression Tree (CART) analysis [Breiman et al., 1984]. A method termed ‘recursive partitioning’ is usually used to split the input dataset into the tree’s branches and nodes, based on the input data and a purity measure [Strobl et al., 2015]. The algorithm progressively splits the independent variable, with the purity of the node calculated at each split, and the split with the highest purity kept. The purity measure indicates the homogeneity of the data under each node, and is calculated using mean squared error. The *R* language [R Core Team, 2013] has various packages for decision tree modeling: the rpart package for CART analysis, the party package for nonparametric regression trees and the randomForest package for Random Forests.

Random Forests [Breiman, 2001] are designed to build multiple decision trees from the training dataset, with each tree constructed using a different bootstrap sample, and using the mean of those trees to make predictions. A common issue with standard tree methods is overfitting, in which random relationships result in noisy predictions. Random Forests are specifically designed to avoid this problem, which was also an issue observed during the GAM model formulation in this study. Random Forests are implemented in *R* using the randomForest package [Liaw and Wiener, 2002], which provide methods for creating, comparing, modifying and predicting using random forests. Unlike in regression-based models, which use p-values to rank the importance of predictor variables, random forests generally use a **variable importance** measure [Breiman, 2001]. The randomForest package contains a very useful method (varImpPlot) to plot a dotchart of the variable importance as measured by the forest. Unlike simpler tree-based methods, due to the large number of component trees and with each tree having a slightly different outcome, it can be difficult to visually interpret the node and branches in random forests. Therefore, the variable importance plot was extensively used while formulating the A-SWE model to decide which variables were important to describe the relationship being modeled, and which variables could be left out.

2.9. Time series models

Regardless of the technique used (statistical, machine learning or any other), there are multiple challenges in modeling time-series data. Each of these challenges can sometimes be tackled using multiple techniques, each with its own merits and demerits. One major issue with time-series data is that they are often auto-correlated, where the value of a variable at a time step is related to its value at one or more preceding time steps. A common method to include the auto-correlation effect in the model is to add lagged variables as predictor variables. This allows the model to predict the dependent variable from both the predictor and from a lagged version of the predictor.

The above method depends on the assumption that the lag itself is ‘stationary’ over time, referring to the fact that its means and variances do not change over time.

Nonstationarity can be a bigger challenge to handle, as was observed while building one model to predict streamflow for this study ('3.2. A-SWE-S MODEL'). Since streamflow is a function of various hydrological processes, sometimes driven by reservoir-like storage effects, its lags are not stationary. In such cases, simply adding a lagged version of the predictor might not improve prediction significantly.

CHAPTER 3

MODEL FORMULATION AND VALIDATION

3.1. A-SWE model

3.1.1 Model formulation

The A-SWE model was initially built using the Generalized Additive Model (GAM) framework, which allows relationships between nonlinear data to be modeled using smoothing functions. The model was built using average air temperature (tavg, in degrees Celsius) and spatially averaged mean white-sky albedo (mean_wsa) as predictors, with the temporal pattern of snowpack represented by a 'day of the year' component. The GAM performed satisfactorily in terms of model fit, with r2p (R squared of prediction) values between 0.49 and 0.87 during k-fold cross-validation using 12 years of training data and 1 year of validation data, and explained 85% of the variance in SWE (wteq). But the model RMSEP (root mean squared error of prediction) was very high, with prediction errors of up to 4.3 inches for some years. Most importantly, the model was not able to suitably predict the SWE in the transition between snowpack and no-snowpack periods. Inclusion of lagged terms for albedo did not significantly improve model performance.

Based on the hypothesis that autocorrelation and significant lag was reducing model prediction capability, a Random Forest approach was attempted. A model was built using the randomForest package [Liaw and Wiener, 2002], using mean_wsa to predict wteq. The yearly albedo cycle was included using the day_ofyear variable. After observing that the model was not able to track the albedo transition period and peak albedo perfectly, two new variables were added to the model formulation. The max_mean_wsa variable contains the maximum mean_wsa for the calender year, and day_sincemaxwsa represents the number of days since the previous peak mean_wsa.

Further model runs indicated that adding lagged terms for the predictor might improve the model, and this was confirmed by the Partial Autocorrelation Function (PACF) plot [R Documentation acf{stats}] for mean_wsa (shown in Figure 8). The PACF explains the linear dependence of an element in the series with a previous element in the same series, and the ensuing lagged effect, corrected for correlation across shorter lags. It can help explain the amount of lag to be added to the model. Based on the PACF plot, the first and second lagged terms for mean_wsa (lag1_mean_wsa and lag2_mean_wsa) were added to the model. These terms lag the mean_wsa time-series by one day and two days, respectively. Appendix F contains the code used to automate the calculation of these additional variables. All the variables used in the A-SWE model are shown in Table 2. The table also describes the physical meaning of all variables.

The final model formulation for the A-SWE model is as below:

```
swe.rf<-randomForest(wteq ~ day_ofyear + mean_wsa + lag1_mean_wsa +
lag2_mean_wsa + max_mean_wsa + day_sincemaxwsa, ntree=1000, mtry=3,
data=parleys_data3)
```

Initially the model was built using 500 trees, but the number of trees was increased to 1000 to improve model performance. Further increasing the number of trees did not decrease mean squared error, which was constant at around 0.19. Climatic variables tavg and precip were added to the model formulation, but were removed when it was observed that they were decreasing model accuracy by increasing mean squared error to about 0.32.

The percentage of variance explained by the randomForest model can be obtained using the print() command, as shown below. It can be seen that the model explains more than 99% of the variance in the predicted variable, using 3 variables at a time to create

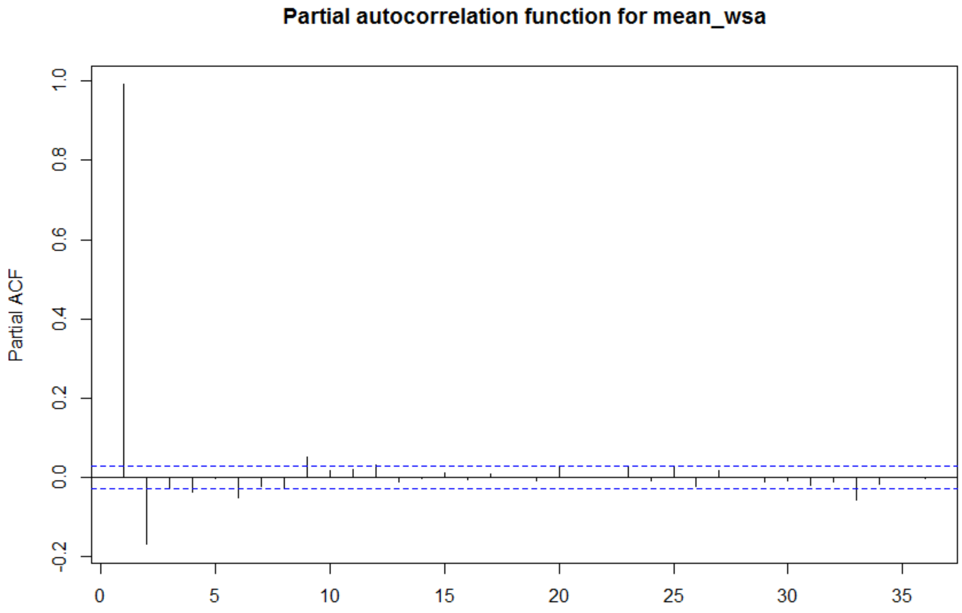

Figure 8: PACF for mean_wsa

Table 2: Time series variables used in the A-SWE model

Variable	Description	Physical meaning
Wteq	Snow water equivalent (inches)	Predicted variable
day_ofyear	Day of the year (1-365/366)	Represents the seasonal cycle of the snowpack; acts as a proxy for average seasonal temperature
mean_wsa	Mean white-sky albedo (0-1)	Main predictor variable
lag1_mean_wsa	Lagged mean white-sky albedo (lag=1 day, 0-1)	Connects the effect of albedo lagged by 1 day on current snowpack state
lag2_mean_wsa	Lagged mean white-sky albedo (lag=2 days, 0-1)	Connects the effect of albedo lagged by 2 days on current snowpack state
max_mean_wsa	Maximum mean white-sky albedo in the calendar year (0-1)	Represents inter-annual variability in snowpack albedo
day_sincemaxwsa	Days since the mean white- sky albedo peaked in the previous year	Represents temporal trend of interannual variability in snowpack albedo

splits in the trees. A variable importance plot shown in Figure 9, generated using the varImpPlot() command, visually describes the importance of each variable in increasing node purity. It can be observed from the plot that the most important variable for increasing node purity (roughly, decreasing the mean squared error) is day_ofyear, indicating that the temporal component is essential to model snow processes. The mean white-sky albedo (mean_wsa) and its lags are also important, and they are followed by the other variables that slightly improve model accuracy. The variable importance plot is generally used as a method to select variables for deletion to reduce model run time and complexity, but all variables were left in since the model took only about 30 seconds to create.

```
> print(swe.rf)
Call: randomForest(formula = wteq ~ day_ofyear + mean_wsa + lag1_mean_wsa +
lag2_mean_wsa + max_mean_wsa + day_sincemaxwsa, data = parleys_data3,ntree =
1000, mtry = 3)
            Type of random forest: regression
```

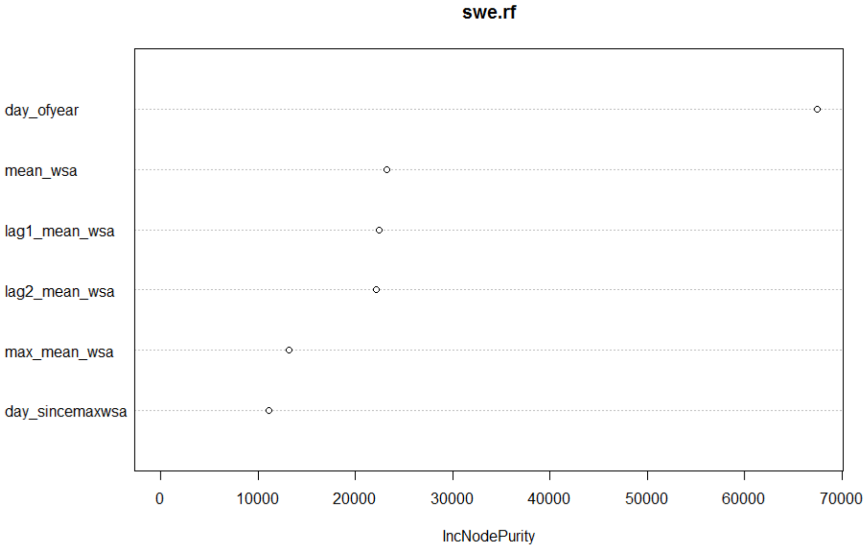

Figure 9: Variable importance plot for the A-SWE random forest (swe.rf)

```
                     Number of trees: 1000
No. of variables tried at each split: 3
          Mean of squared residuals: 0.19478
                    % Var explained: 99.42
```

3.1.2 Model validation

Model validation by k-fold cross-validation was performed to understand the year-to-year model prediction accuracy. An R script (shown in Appendix F) was used to create the model using 12 years of data and use the remaining 1 year of data to calculate SWE (wteq). The code contains a loop to perform this subsampling process for all 13 years. Figure 10 shows the results of the k-fold cross-validation, with predicted SWE in red and actual SWE in black.

It can be seen that the model predicts SWE remarkably well for certain years like 2003, 2008 and 2009. Other years like 2004 and 2007 are slightly over-predicted, and years like 2010 and 2011 are slightly under-predicted. Some years like 2002, 2005, 2006, 2012 and 2013 are not predicted very well, with the model unable to track abrupt variations in albedo due to winter storms and sudden melt events. The year 2001 is a special case, as the model predicts SWE even in the summer months. This can be explained by the fact that due to 2001 being the first year in the dataset, the day_sincemaxwsa variable could only be approximately calculated since there were no data for the melt event in the year 2000. This was expected to confuse the model and generate SWE in the summer months, a phenomenon not observed for the other years.

Overall the A-SWE random forest model performs well for years that follow the general snowpack development and depletion temporal trend, and does not for years in which the SWE fluctuates due to other factors. Even with this issue, the model is able to capture the trends in SWE changes fairly well. Future efforts are planned at improving model accuracy in years that do not follow general snowpack depletion trends.

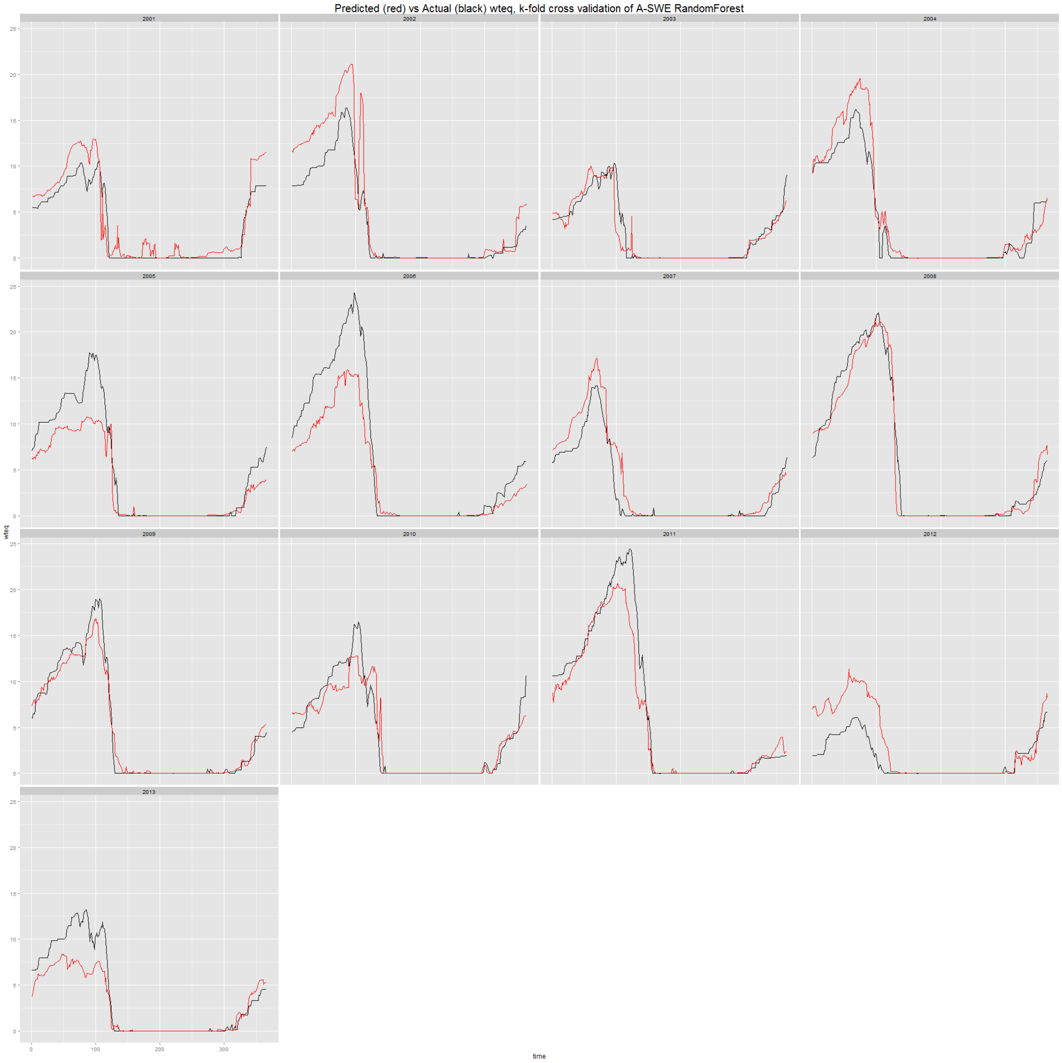

Figure 10: k-fold cross-validation of the A-SWE model

3.2. A-SWE-S model

3.2.1. Model formulation

The A-SWE-S (Albedo-Snow Water Equivalent-Streamflow) model was earlier formulated as the SWE-S model, to use output snow water equivalent from the SWE-S model to predict streamflow. The rather complicated lagged and auto-correlated relationship between SWE and streamflow resulted in extremely low prediction accuracies with the Random Forest approach, and focus was shifted to the Generalized

Additive Model (GAM). GAMs allow for nonlinear predictors to be included using ‘smooth functions’ linked by a ‘link function’ to the predicted variable, and are described in the section ‘2.7 Statistical modeling techniques’.

Some additional variables were obtained for the A-SWE-S model formulation, to better simulate the hydrological processes of precipitation, snowmelt, infiltration, base flow, evapotranspiration and runoff. Since data for these variables were only available from the year 2004, the model was formulated, validated and implemented using data from 2004-2013. These additional variables included soil moisture (sms) at 2, 8 and 20 inches depth, soil temperature (sto) at 2, 8 and 20 inches depth, and snow depth (snwd, in inches), and were all obtained from the Parleys Summit SNOTEL site. The mean white-sky albedo, in its log10 form, was added to the model to allow streamflow prediction under albedo change scenarios. The albedo variable was also added to allow predictions independent of the SWE-S model. The AIC() function, which calculates the ‘Akaike Information Criterion’ (Sakamoto et al., 1986), was used to determine the final model formulation. AIC allows the modeler to choose the best combination of variables by comparing AICs of various model formulations, with the lowest AIC being best.

Three other variables, precip_memory10, precip_memory15 and precip_memory30, were created using the available continuous precipitation data. These variables represent ‘precipitation memory’, and are essentially precipitation accumulation for 10, 15 and 30 days respectively, after which they reset to zero. They are designed to simulate the effects of lagged runoff in the watershed. Also included is ‘precip_accum’, which represents the accumulated precipitation reset to zero at the start of each water year.

The final A-SWE-S model formulation is shown below, and the smooth functions and number of knots used for each variable are described in Table 3. The table also describes the physical significance of each variable. A summary() call showed that the GAM explained 93.7% of the variance in the predicted variable. The R code used to

Table 3: Time series variables used in the A-SWE-S model

Variable	Description	Smooth function	Number of knots	Physical meaning
lflow (predicted)	Log10 of flow	-	-	Predicted variable
day_ofyear	Day of the year (1-365/366)	Cubic regression spline (shrinkage version)	100	Represents the seasonal cycle of streamflow
precip_accum	Accumulated precipitation (inches)	Cyclic cubic regression spline	100	Represents long-term trends in precipitation over the water year
day_sincemelt	Days since the SWE peaked in the previous year	Cyclic cubic regression spline	100	Represents temporal trend of interannual variability in snowpack depletion
sms20	Soil moisture % at 20”	Cubic spline basis	150	Represents vadose zone water content
sto20	Soil temperature at 20” (DegC)	Thin plate regression spline	150	Represents ground heat flux
wteq	Snow water equivalent (inches)	Thin plate regression spline	150	Predictor variable
sms2	Soil moisture % at 2”	Cubic spline basis	150	Represents vadose zone water
sms8	Soil moisture % at 8”	Cubic spline basis	150	Represents vadose zone water content
log10mean_wsa	Log10 of mean white-sky albedo	Cubic spline basis	100	Predictor variable

Table 3 (Continued):

Variable	Description	Smooth function	Number of knots	Physical meaning
precip_memory15	Precipitation accumulation over each 15 day period	Thin plate regression spline	40	Represents the short term trends in precipitation; represents to ET and ponding
precip_memory10	Precipitation accumulation over each 10 day period	Thin plate regression spline	40	Represents the short term trends in precipitation; represents to ET and ponding
precip_memory30	Precipitation accumulation over each 30 day period	Thin plate regression spline	40	Represents the short term trends in precipitation; represents to ET and ponding
Snwd	Snow depth (inches)	Thin plate regression spline	50	Additional snowpack prediction variable

generate various additional variables for the model, along with other model code and output, is given in Appendix H.

```
flow.gam =
gam(lflow~s(day_ofyear,bs='cc',k=100)+s(precip_accum,bs='cc',k=100)+s(day_sincemel
t,bs='cc',k=100)+s(sms20,bs='cr',k=150)+s(sto20,k=150)+s(wteq,k=150)+s(sms8,bs='cr',
k=150)+s(sms2,bs='cr',k=150)+s(log10mean_wsa,bs='cr',k=100)+s(precip_memory15,k=
40)+s(precip_memory10,k=40)+s(precip_memory30,k=40)+s(snwd,k=50),data=parleys_
data3,family=gaussian)
```

3.2.2. Model validation

The A-SWE-S model was validated for the 10 years of data available, using k-fold cross-validation. Appendix H contains the code used to perform the validation, and Figure 11 shows the results. It can be observed that the model is able to predict the general trend of streamflow each year with fairly good accuracy, except for the first year of record (2004). Peaks and drops in the actual streamflow are reflected in the modeled streamflow, with a certain amount of lag. This lag is especially pronounced in the year 2011. Future model improvement efforts are planned at further reducing the effect of lag and autocorrelation, which is a frequent challenge to statistical modeling of complex time-series data. Unlike the A-SWE Random Forest, due to insufficient accuracy with tracking the end of spring streamflow, the A-SWE-S GAM will only be used to model the effect of albedo change on peak runoff for this study. This can be used in conjunction with the A-SWE model results to understand dust deposition impacts on snowpack-driven streamflow.

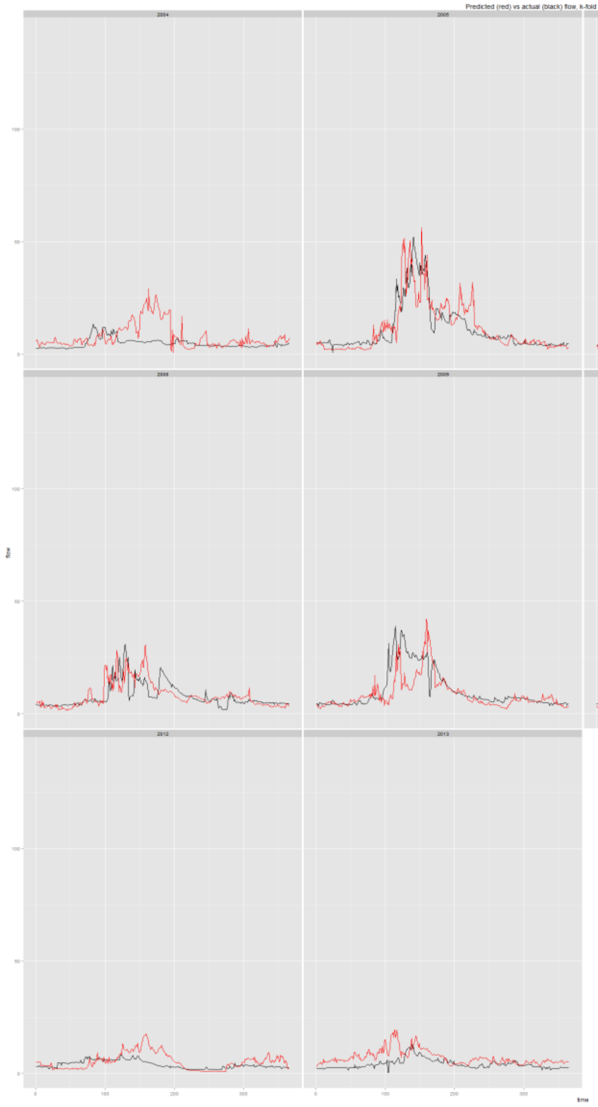

Figure 11: k-fold cross-validation of the A-SWE-S model

CHAPTER 4

RESULTS

Figure 12 diagrammatically describes the modeling process used for this study. Table 4 describes the albedo change impacts on SWE. Figures 13, 14, 15 and 16 show plots of predicted SWE, for various scenarios, of years 2002, 2005, 2008 and 2012. Appendix G contains plots for all the other years. It was observed that SWE under 1.1x (+10%) albedo change reduces to zero on an average 3-4 weeks later than actual SWE. The delay in SWE reaching zero, representing the end of the snowpack, was as high as 6 weeks in the year 2005 for 1.1x albedo, compared to actual albedo. SWE under 1.05x (+5%) albedo change was seen to reach zero on an average 2-3 weeks later compared to actual albedo. Under albedo decrease conditions, representing an increase in BC deposition, it was seen that SWE under 0.95x (-5%) and 0.90x (-10%) scenarios was closer to actual SWE for most years. Under such scenarios, the snowpack was either depleted on an average between 1-2 weeks earlier compared to actual conditions or matched the actual snowpack depletion time. Although there were years in which depletion under the 0.90x scenario was earlier than under the 0.95x scenario by a few days, this trend was not consistent. This indicates that beyond a certain amount of snowpack depletion, any further reduction in albedo might not have a great impact on SWE due to the low surface area of snow present.

Since the A-SWE-S model was not able to track the end of the melt season accurately and suffered from errors due to lagged effects, it was only used to understand if peak discharge varied due to albedo change. The model was used to predict flow under

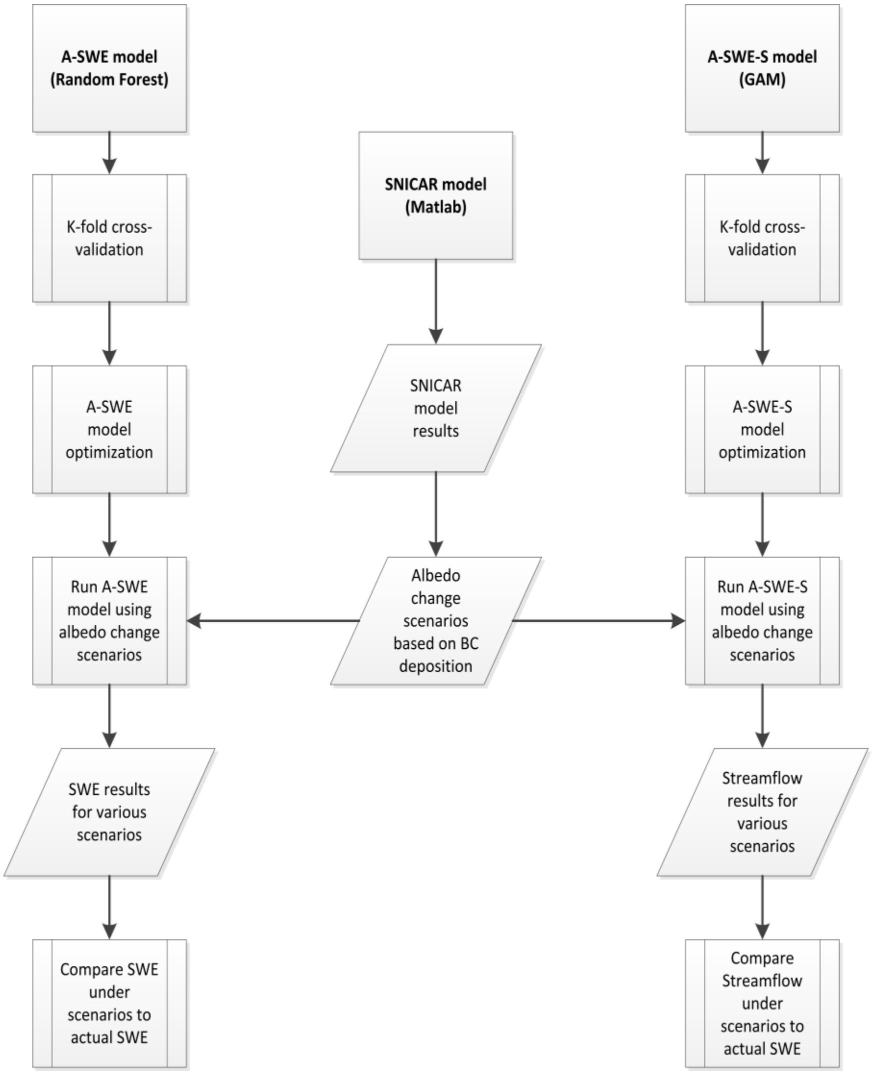
A-SWE model (Random Forest)
K-fold cross-validation
A-SWE model optimization
Run A-SWE model using albedo change scenarios
SWE results for various scenarios
Compare SWE under scenarios to actual SWE
SNICAR model (Matlab)
SNICAR model results
Albedo change scenarios based on BC deposition
A-SWE-S model (GAM)
K-fold cross-validation
A-SWE-S model optimization
Run A-SWE-S model using albedo change scenarios
Streamflow results for various scenarios
Compare Streamflow under scenarios to actual SWE

Figure 12: Modeling schematic

Table 4: Albedo change impact on the snowpack

Albedo Scenario	**Deposition of black carbon**	**Change from ‘Base’ scenario**	**Impact on SWE depletion**
1.10x	-1500 ppb	1500 ppb deposition to zero deposition	3-4 weeks later
1.05x	-500 ppb	500 ppb deposition to zero deposition	2-3 weeks later
Base (1x)	-	-	-
0.95x	+500 ppb	Zero deposition to 500 ppb deposition	1-2 weeks earlier
0.90x	+1500 ppb	Zero deposition to 1500 ppb deposition	1-2 weeks earlier

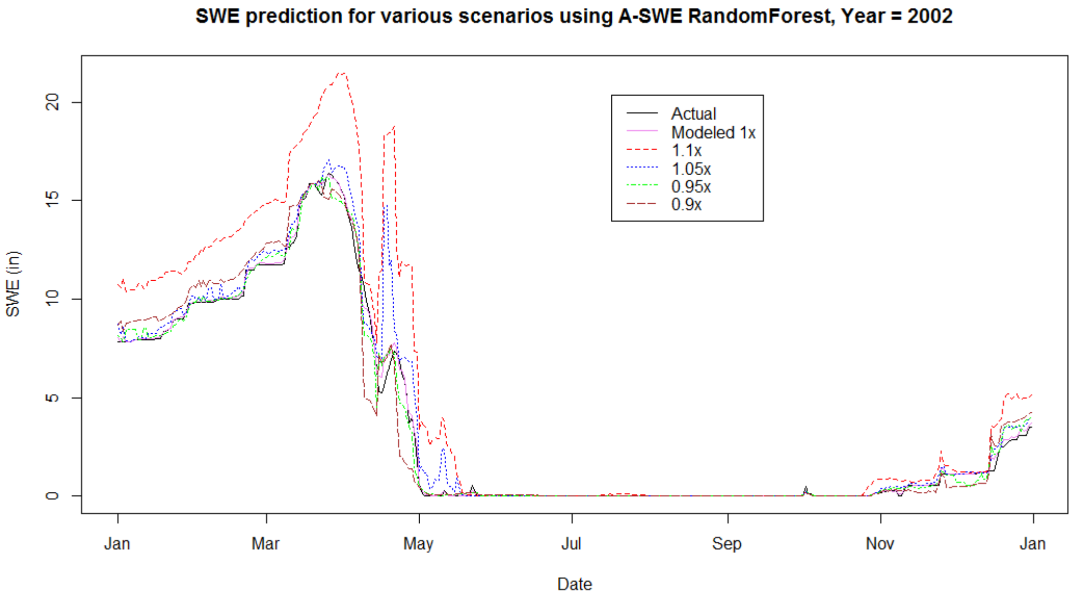

Figure 13: Predicted SWE for year 2002

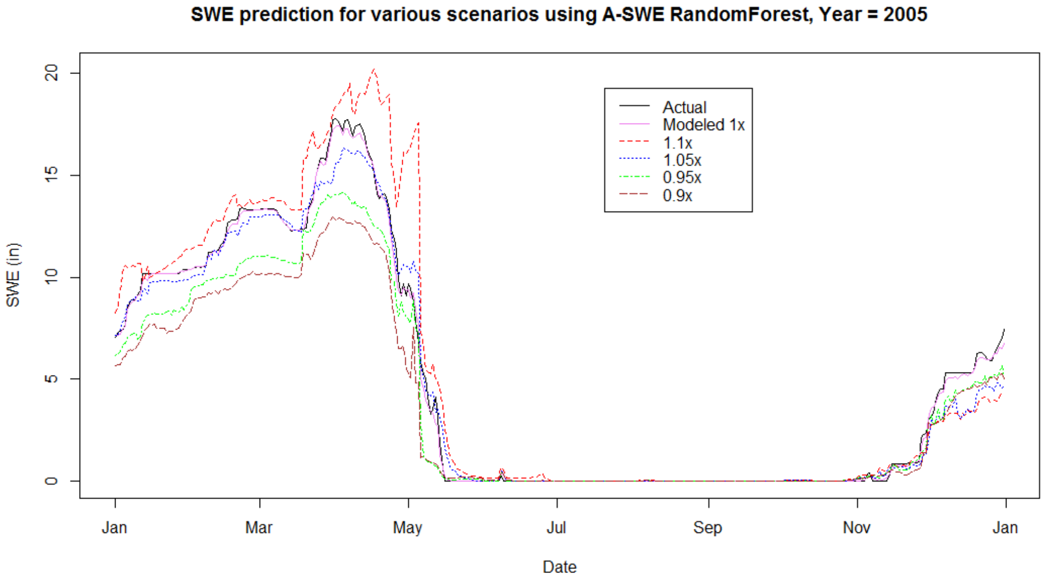

Figure 14: Predicted SWE for year 2005

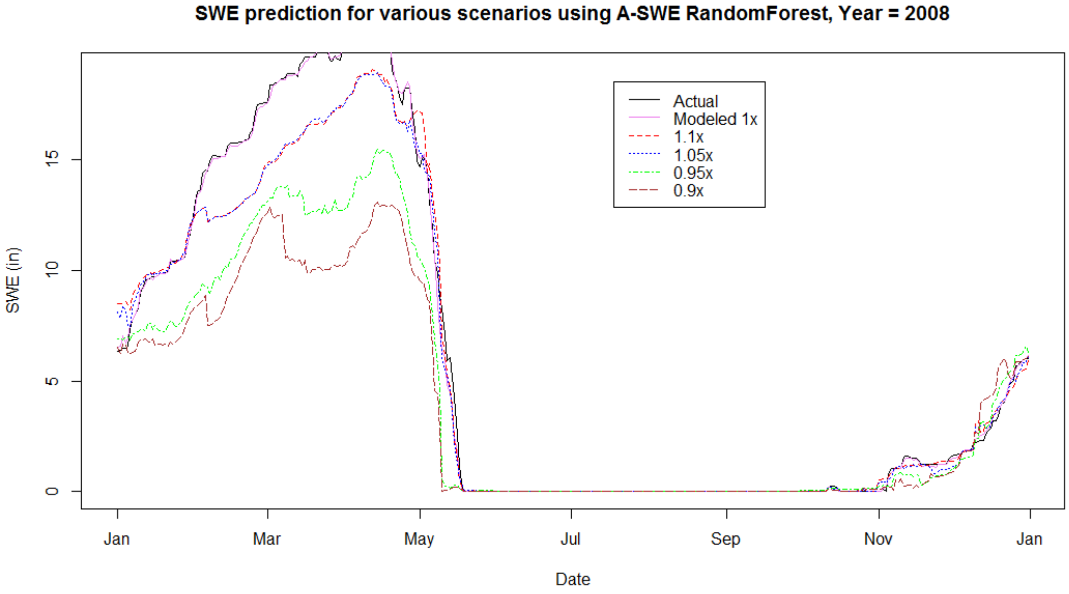

Figure 15: Predicted SWE for year 2008

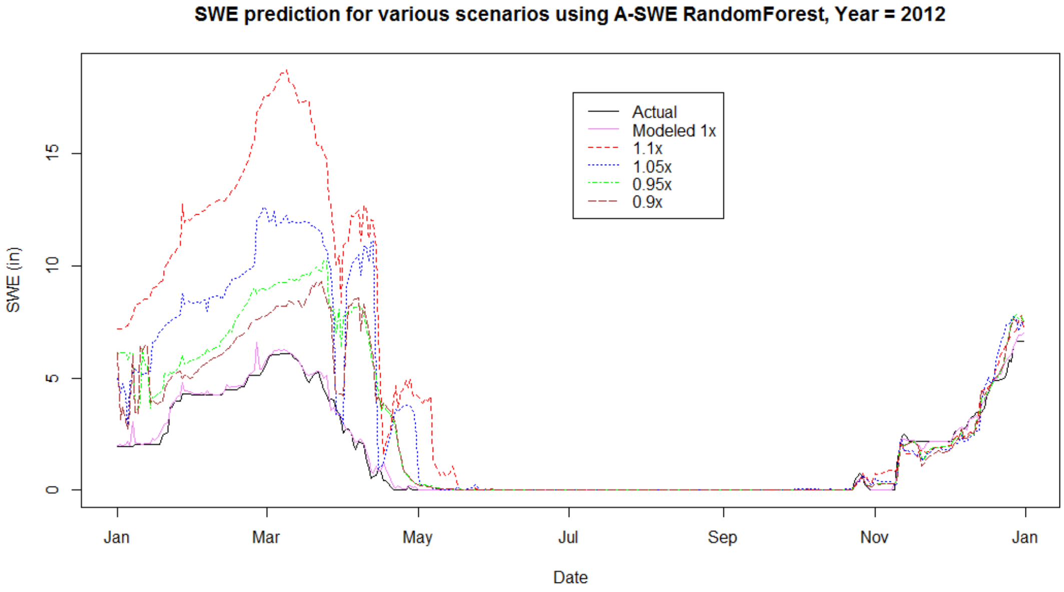

Figure 16: Predicted SWE for year 2012

1.1x, 1.05x, 1x, 0.95x and 0.90x mean white-sky albedo (mean_wsa) conditions. The change in sum of predicted streamflows (Σflow) was calculated to vary between +218 cfs and -71 cfs for various years, relative to the actual streamflow predicted (1x).

Figures 17 shows the histograms of 1.1x and 1x scenarios obtained by this analysis for year 2005, and both plots can visually be interpreted as being similar. A similar phenomenon was observed for the other years, and was confirmed using a two-sample t-test. Therefore, the variations in flow due to albedo variations are either too minimal to be statistically significant, or the model is not able to predict flow under such variations.

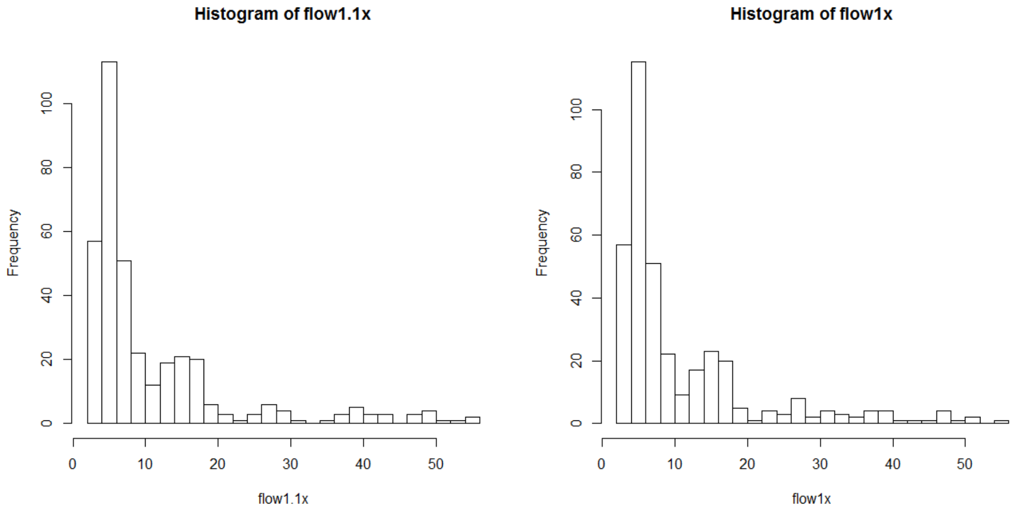

Figure 17: Histograms of flow results from A-SWE-S model (1.1x and 1x albedo)

CHAPTER 5

CONCLUSIONS

This study, unique in its combined analysis of air quality and hydrology, resulted in many useful conclusions about the impact of albedo change on snowpack state and streamflow in the Wasatch. A very important finding was that snowpack was found to be depleted 2-4 weeks later under decreased black carbon deposition, and 1-2 weeks earlier due to elevated black carbon deposition, compared to respective base conditions. The Parleys Creek watershed is only one among four major watersheds that supply potable water to Salt Lake City, and the modeling techniques used in this study can be applied to the other watersheds. Results from this study are planned to be used to drive a systems model of the city's water supply, to understand the impact of black carbon deposition on water system reliability. This analysis will also be conducted under climate change scenarios, to quantify dual deposition and climate impacts. The study also explored the applications of long-term MODIS albedo datasets in hydrology, and future efforts will look at furthering the application of satellite remote sensing datasets in research involving contaminant deposition on snow. Future work is targeted at improving the prediction models, especially the flow prediction model. The models, unique in their aspect of application of advanced statistical and machine learning techniques, can be further improved by better accounting for lagged and autocorrelation. Also the models, constrained under the variable range used to build them, could be tested and improved using data with greater variability or random sampling. With the availability of reliable BC data in the future, it might be possible to analyze flow variation trends under actual deposition conditions.

APPENDIX A

ANALYSIS OF AIR QUALITY IN SALT LAKE CITY

Table A1: PM2.5 exceedance in Salt Lake City from 2001-13

Year	**Number of days PM2.5 > 35 μg m^{-3}**	**Number of days PM2.5 > 60 μg m^{-3}**
2001	25	9
2002	26	6
2003	5	0
2004	36	12
2005	22	1
2006	10	0
2007	18	7
2008	10	1
2009	16	4
2010	15	4
2011	9	2
2012	0	-
2013	35	5

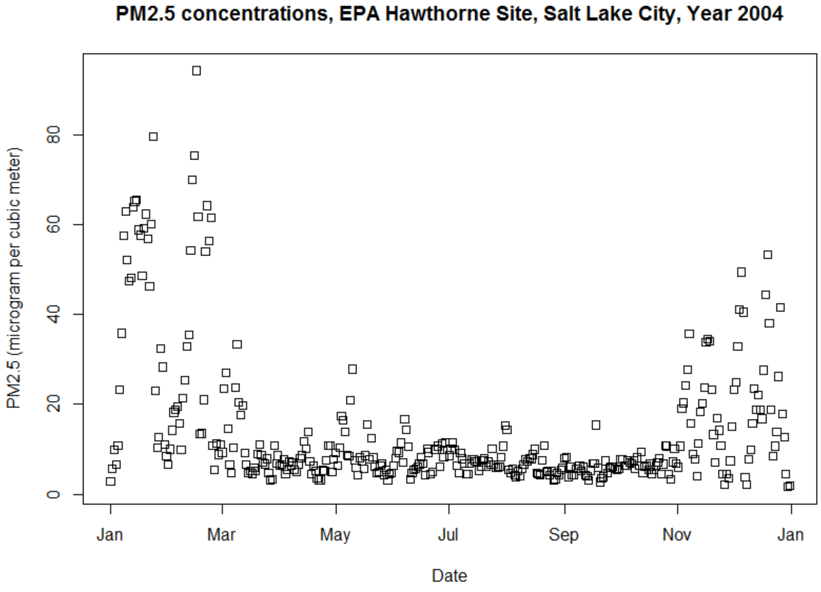

Figure A1: PM2.5 in 2004

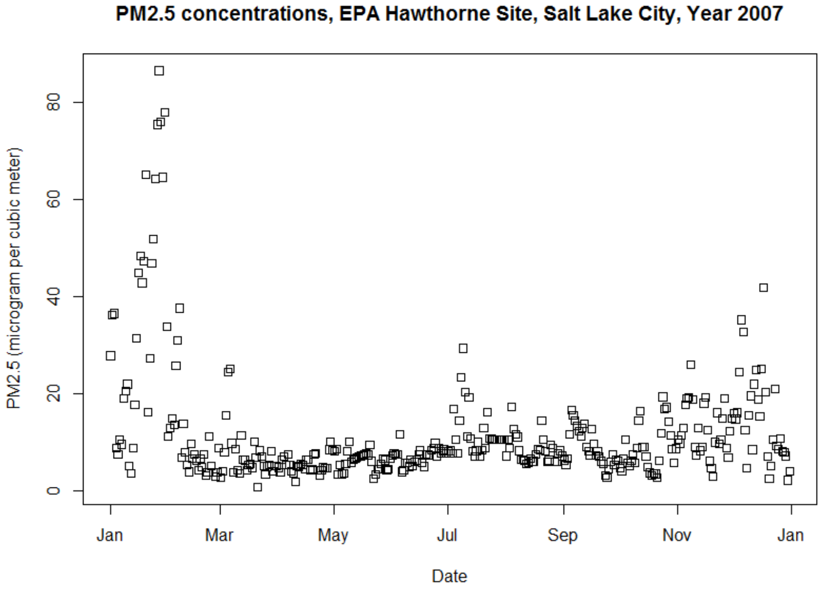

Figure A2: PM2.5 in 2007

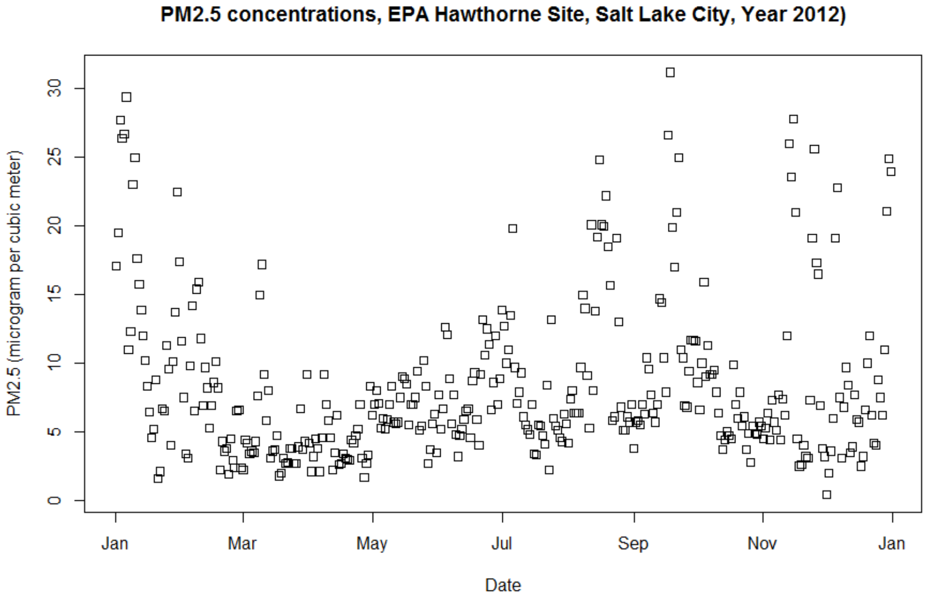

Figure A3: PM2.5 in 2012

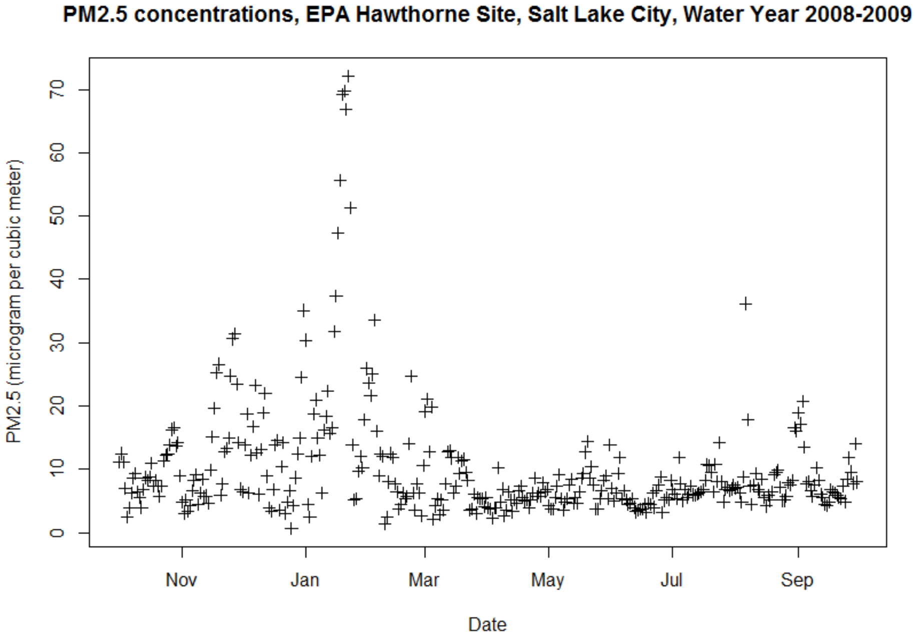

Figure A4: PM2.5 in water year 2008-09

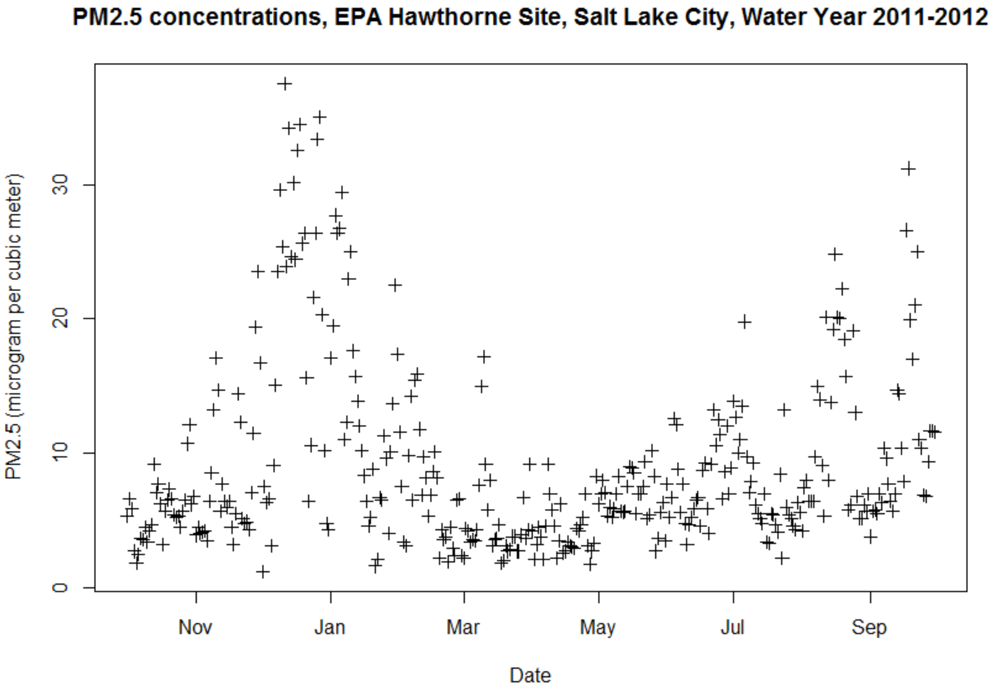

Figure A5: PM2.5 in water year 2011-12

APPENDIX B

SNICAR ANALYSIS

Table B1: Parameters used for SNICAR analysis

Parameter	**Value/Selection used**
Type of incident radiation	Direct-beam incident flux
Two-stream approximation type	Hemispheric Mean
Broadband albedo of underlying surface	0.25
Cosine of solar zenith angle for direct-beam	0.5
Number of snow layers	2
Snow layer(s) thickness	0.02, 9.98
Snow density of each layer	150 kg/m^3
Snow effective grain size for each layer	100 microns

Table B2: Change in snow albedo due to varying black carbon concentrations

BC conc **(ppb or ng/g)**	**Albedo**	**Change over previous BC concentration**	**% Change over previous BC concentration**	**Change from zero BC**	**% Change from zero BC**
0	0.8273	0	0	0	0
100	0.8105	-0.017	-2.03	-0.017	-2.03
200	0.8016	-0.009	-1.10	-0.026	-3.11
300	0.7946	-0.007	-0.87	-0.033	-3.95
400	0.7885	-0.006	-0.77	-0.039	-4.69
500	0.7831	-0.005	-0.68	-0.044	-5.34
600	0.7782	-0.005	-0.63	-0.049	-5.93
700	0.7737	-0.004	-0.58	-0.054	-6.48
800	0.7695	-0.004	-0.54	-0.058	-6.99
900	0.7655	-0.004	-0.52	-0.062	-7.47
1000	0.7617	-0.004	-0.50	-0.066	-7.93
1100	0.7582	-0.004	-0.46	-0.069	-8.35
1200	0.7548	-0.003	-0.45	-0.073	-8.76
1300	0.7515	-0.003	-0.44	-0.076	-9.16
1400	0.7484	-0.003	-0.41	-0.079	-9.54
1500	0.7453	-0.003	-0.41	-0.082	-9.91
1600	0.7424	-0.003	-0.39	-0.085	-10.26
1700	0.7396	-0.003	-0.38	-0.088	-10.60
1800	0.7369	-0.003	-0.37	-0.090	-10.93
1900	0.7342	-0.003	-0.37	-0.093	-11.25
2000	0.7316	-0.003	-0.35	-0.096	-11.57
2100	0.7291	-0.003	-0.34	-0.098	-11.87

Table B2 (Continued):

BC conc **(ppb or ng/g)**	**Albedo**	**Change over previous BC concentration**	**% Change over previous BC concentration**	**Change from zero BC**	**% Change from zero BC**
2200	0.7266	-0.002	-0.34	-0.101	-12.17
2300	0.7243	-0.002	-0.32	-0.103	-12.45
2400	0.7219	-0.002	-0.33	-0.105	-12.74
2500	0.7196	-0.002	-0.32	-0.108	-13.02
2600	0.7174	-0.002	-0.31	-0.110	-13.28
2700	0.7152	-0.002	-0.31	-0.112	-13.55
2800	0.7131	-0.002	-0.29	-0.114	-13.80
2900	0.711	-0.002	-0.29	-0.116	-14.06
3000	0.7089	-0.002	-0.30	-0.118	-14.31

APPENDIX C

ADDITIONAL GRAPHICS OF VARIOUS DATASETS

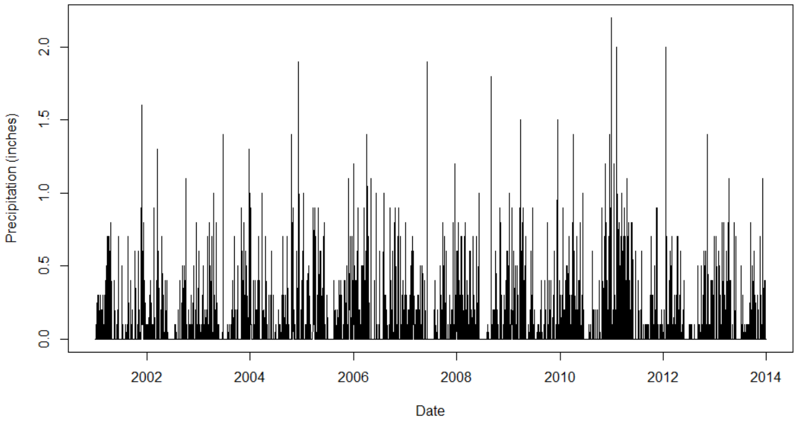

Figure C1: Daily precipitation data from Parleys Summit SNOTEL

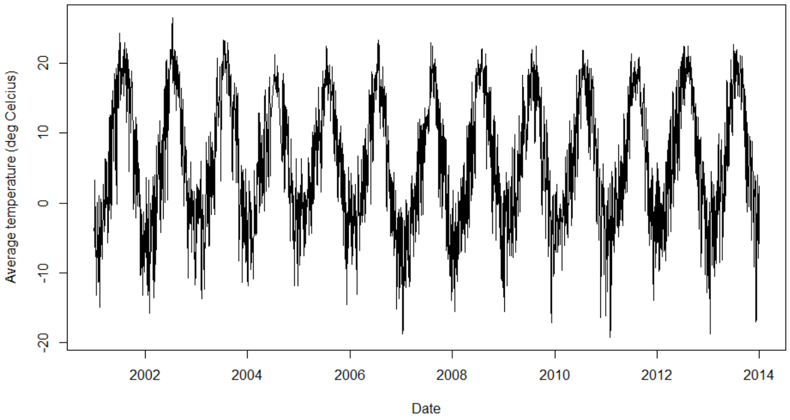

Figure C2: Daily average temperature data from Parleys Summit SNOTEL

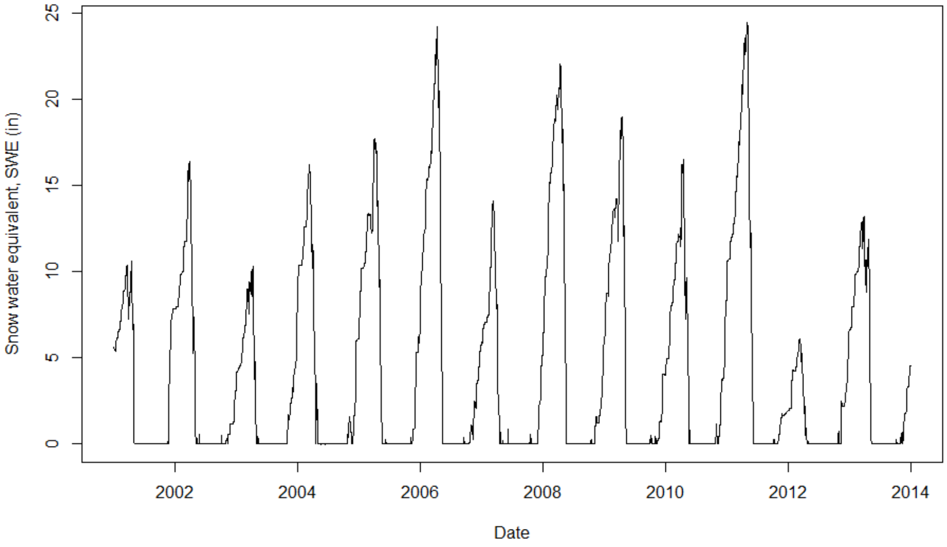

Figure C3: Daily SWE data from Parleys Summit SNOTEL

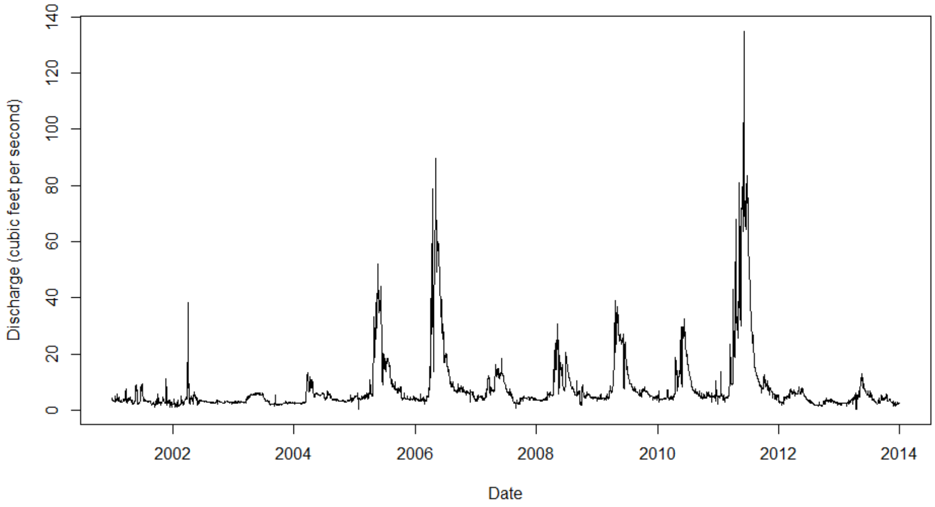

Figure C4: Daily streamflow (discharge) data for Parleys Creek

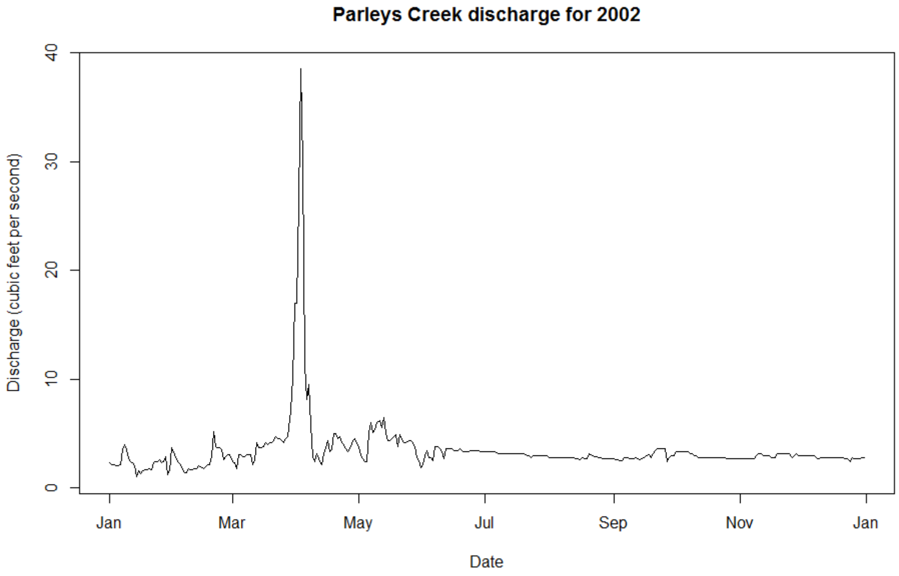

Figure C5: Discharge for 2002

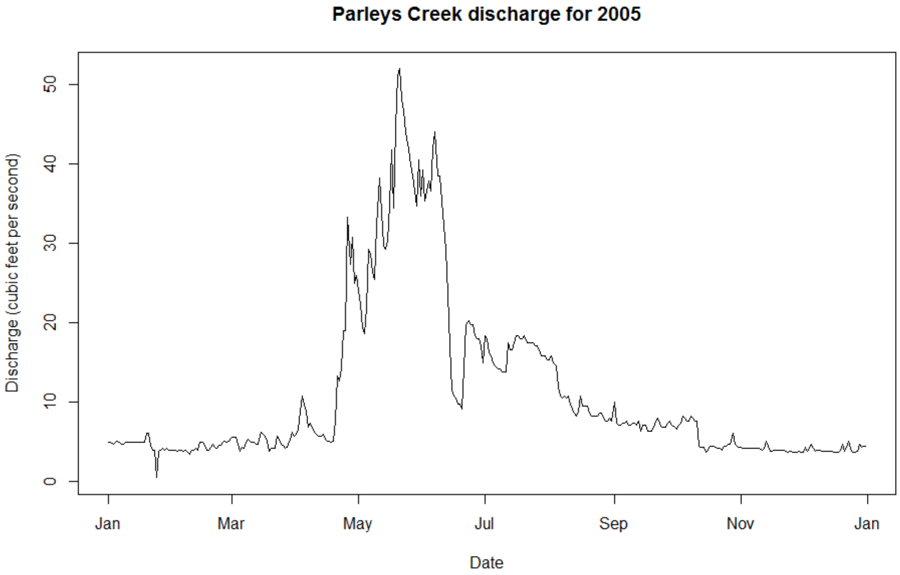

Figure C6: Discharge for 2005

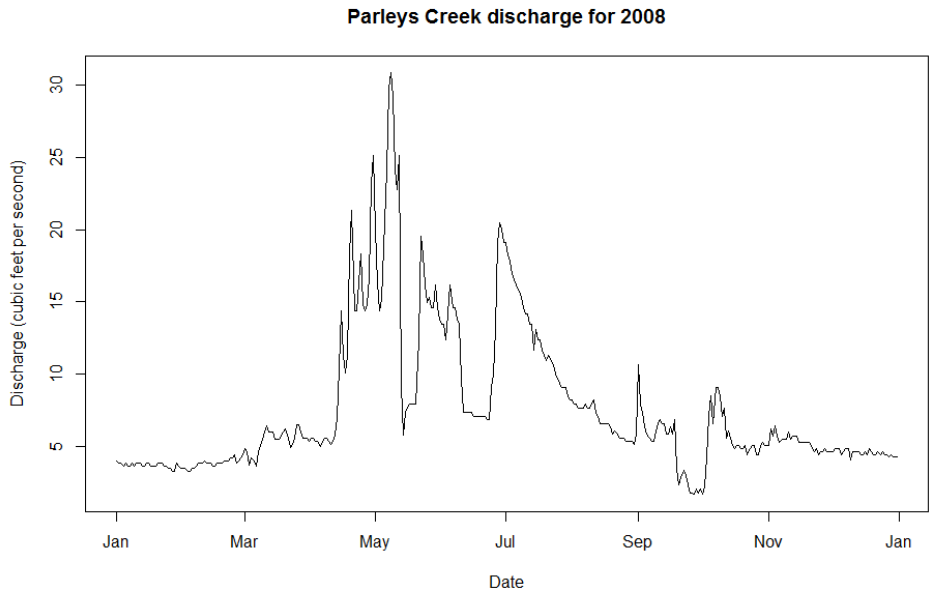

Figure C7: Discharge for 2008

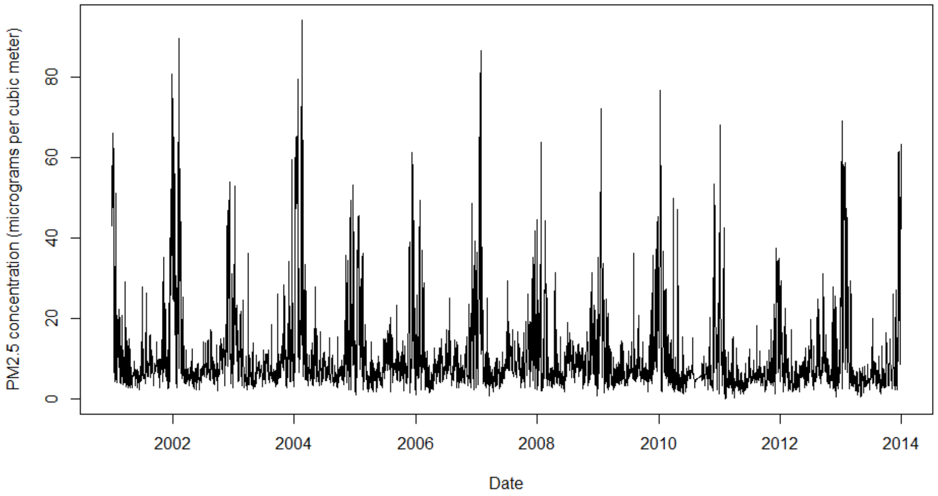

Figure C8: Daily PM2.5 data from EPA's Hawthorne Site

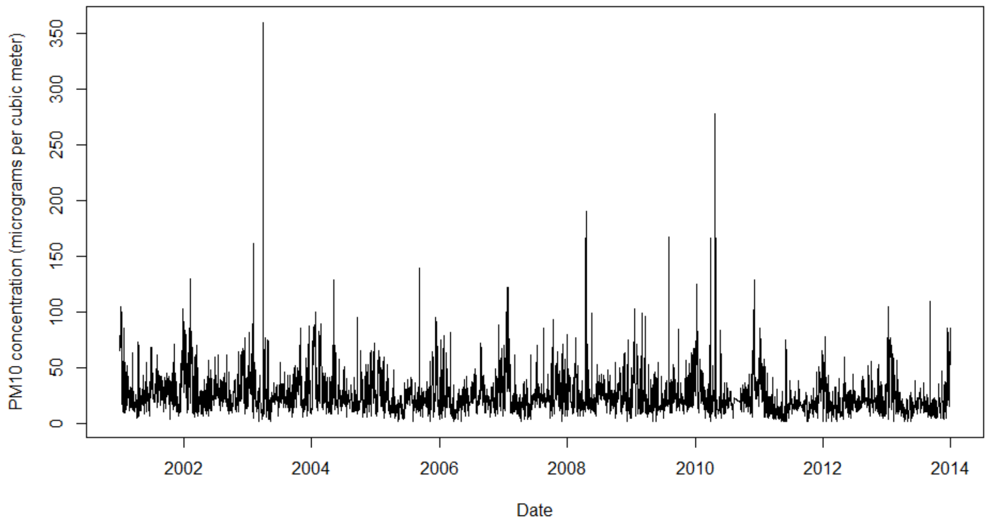

Figure C9: Daily PM10 data from EPA's Hawthorne Site

APPENDIX D

R SCRIPT USED FOR EXTRACTION OF MEAN ALBEDOS

```
#Load required libraries;install them prior to running code
library(ncdf)
library(pROC)
#........................
#MODIS NetCDF file statistics calculator updated with clock,save folder select and
better file name appending.
# ***Doesn't remove NAs and Infs from CSV file. Does not plot***
#........................
#Script calculates statistics (max,min, mean and sd) for MODIS satellite image files files
in NetCDF3 format. Use ArcGIS to convert MODIS Tif files to
#NetCDF if required. Check consistency of NetCDF files before using script. The script
creates a data frame containing julian year, julian day of the year,
#actual dates, mins, maxes, means and sds for all the NetCDF files. It also creates a CSV
from the data frame in a sub-folder of the folder-
#containing the netcdf files.
#(Use Panoply NetCDF reader or the ncdf library in R to check NetCDF file before
running script. Check its variable names and other data before using script)
#...............................
#Edit below before running script
folder_path<-("F:/Research/Statistical modeling/MCD43A3 Parleys NetCDF/Band29
SW WSA Parleys/clipped_netcdf") #Set to file path containing only .nc netcdf3 files
variable='Albedo' #Edit to the name of the variable in netcdf files to calculate statistics for
savefile_name<-'MCD43A3 Albedo Band29 WSA 2000-13 Parleys Creek.csv' #Edit to change CSV save file name
savefolder_name<-'Statistics-MCD43A3 Albedo Band29 WSA 2000-13 Parleys Creek' #Edit savefolder_name to change name of folder that will contain the CSV file and plots
# plotname_append<-'_WSA_band29_' #Edit to change what will be appended to the plots generated by script
# title_append='2000-13 Parleys Creek, MODIS Band 29 Shortwave White Sky Albedo' #Edit this variable as required; this is used as subtitle for each plot
#.........................................................
ptm <- proc.time() #Start clock
setwd(folder_path)
files <- list.files(path=folder_path, pattern=".nc", all.files=T, full.names=F, no.. = T) #creates list of netcdf files in folder
number<-length(files) #number of netcdf files in folder
years <- rep(NA, number) #creates vector for storing years
days <- rep(NA, number) #creates vector for storing julian days
actual_dates <- vector() #creates vector for storing actual date in YYYY-MM-DD format
mins <- rep(NA, number) #creates vector for storing minimum values
maxes <- rep(NA, number) #creates vector for storing maximum values
means <- rep(NA, number) #creates vector for storing mean values
sds <- rep(NA, number) #creates vector for storing standard deviation values
count<-0 #Initialize variable to count number of files processed
for(r in 1:number) #loop over all netcdf files in folder
{
  ncin<-open.ncdf(files[r]) #opens netcdf file
    julian_date<-substr(files[r],10,16) #extracts julian date YYYYDDD
  actual_date<-as.POSIXct(julian_date, format="%Y %j") #converts julian date to
YYYY-MM-DD format
  current_year<-substr(files[r],10,13) #reads current file's year. The two numeric values in
substr represent starting and ending position in file name for year extraction.
  current_day<-substr(files[r],14,16) #reads current file's julian day. The two numeric
values in substr represent starting and ending position in file name for julian day
extraction.
  x<-get.var.ncdf(ncin,variable) #reads in current file into x; the second parameter
represents variable name to be read from NetCDF
  x[x<0] <- NA #converts all negative values to NA-this is required because NA values
are automatically converted to -128 when the netcdf file is read into R
   min_val<-min(x,na.rm=TRUE) #calculates minimum value from file, ignoring NA
values
  max_val<-max(x,na.rm=TRUE) #calculates maximum value from file, ignoring NA
values
  mean_val<-mean(x,na.rm=TRUE) #calculates mean value from file, ignoring NA
values
  sd_val<-sd(x,na.rm=TRUE) #calculates sd value from file, ignoring NA values
close.ncdf(ncin) #Closes netcdf file to prevent memory overload
  years[r]<-current_year #saves current file's year into vector
  days[r]<-current_day #saves current file's julian day into vector
  actual_date<-format(actual_date,format="%Y-%m-%d") #converts POSIXct class
object dates to characters for insertion into vector
  actual_dates<-c(actual_dates,actual_date) #saves current date into vector
    mins[r]<-round(min_val*0.001,digits=3) #saves current mimimum values into vector,
rounding to specified number of digits
  maxes[r]<-round(max_val*0.001,digits=3) #saves current maximum values into vector, rounding to specified number of digits
  means[r]<-round(mean_val*0.001,digits=3) #saves current mean values into vector, rounding to specified number of digits
  sds[r]<-round(sd_val*0.001,digits=3) #saves current sd values into vector, rounding to specified number of digits
  count<-count+1 #Increment count of files processed
}
x<-paste(getwd(),"/",as.character(savefolder_name),sep="") #Create folder path as subfolder of working directory to save output
dir.create(x) #Create subfolder to save output
#Create data frame containing extracted statistics and save to CSV file
albedo_values=data.frame(years,days,actual_dates,mins,maxes,means,sds) #creates data frame from vectors
albedo_values$actual_dates<-as.POSIXct(albedo_values$actual_dates,format="%Y-%m-%d") #changes format of actual_dates in dataframe to POSIXct
write.csv(albedo_values, file=paste(x,"/",savefile_name,sep="")) #writes CSV file from dataframe
print(paste0("Total number of files processed: ", count)) #Displays number of files processed
# Stop the clock and display elapsed time
proc.time() – ptm
```

APPENDIX E

R SCRIPT USED FOR DATA EXTRACTION FROM CSV FILES

```
require(zoo)
#""Read multiple CSVs in a folder and extract the data in them to a single dataframe""#
#Code by Jai K. Panthail#
#Code allows to specify how many lines to skip before reading
#Code allows to specify which columns to read
#Code allows to specify final column names in saved dataframe
#To edit:
#folder_path: path to folder containing CSV files
#number_columns: number of columns in the CSV file
#pos_import:the position of columns to import; for example: '2' indicates to import second column
#skip:the number of rows to skip before starting to read data
#col_names:the vector containing the new names to be given to the imported column (optional:required for post-processing below)
#################"Edit everything below"#################################
folder_path<-("F:/Research/Statistical modeling-ATPS/Temperature data")
number_columns<-7
pos_import=c(2,7)
skip_count=2
col_names<-c('date','tavg')
```

```
df <- data.frame()
##########################"Main
Code"####################################
import_true<-rep("NULL",number_columns)
for (r in 1:length(pos_import)){
  import_true[pos_import[r]]=NA
}
setwd(folder_path)
files <- list.files(path=folder_path, pattern=".csv", all.files=T, full.names=F, no.. = T)
#creates list of csv files in folder
number<-length(files)
for (i in 1:number){
  x<-read.csv(files[i],skip=skip_count,header=T,colClasses=import_true)
  df<-rbind(df,x)
}
#######
#Post processing: edit as per requirement. This is usually to change column names, to
convert date to as.Date(), to remove NA values etc
colnames(df)<-col_names #Change column names to those contained in col_names
df$date<-as.Date(df$date) #Change date format to as.Date
df$tavg[df$tavg==-99.9]<-NA #Convert missing values (-99.9, this case) to NA
df$tavg=na.approx(df$tavg,na.rm=F) #Interpolate NA values using 'zoo' package
parleys_tavg=df #Save dataframe created from previous steps to final dataframe
###################################################################
```

APPENDIX F

A-SWE MODEL FORMULATION AND K-FOLD CROSS-VALIDATION

```
##Create variables (day_sincemaxwsa and max_mean_wsa)##
require(nnet)
#Calculate day of max mean_wsa each year#
i=which.is.max(subset(parleys_data,year==2001)$mean_wsa)
j=which.is.max(subset(parleys_data,year==2002)$mean_wsa)
k=which.is.max(subset(parleys_data,year==2003)$mean_wsa)
l=which.is.max(subset(parleys_data,year==2004)$mean_wsa)
m=which.is.max(subset(parleys_data,year==2005)$mean_wsa)
n=which.is.max(subset(parleys_data,year==2006)$mean_wsa)
o=which.is.max(subset(parleys_data,year==2007)$mean_wsa)
p=which.is.max(subset(parleys_data,year==2008)$mean_wsa)
q=which.is.max(subset(parleys_data,year==2009)$mean_wsa)
r=which.is.max(subset(parleys_data,year==2010)$mean_wsa)
s=which.is.max(subset(parleys_data,year==2011)$mean_wsa)
t=which.is.max(subset(parleys_data,year==2012)$mean_wsa)
u=which.is.max(subset(parleys_data,year==2013)$mean_wsa)
#Creating vector containing days since max mean_wsa each year
day_sincemaxwsa=c(seq(0,339),seq(0,365-i+j),seq(0,365-j+k),seq(0,365-k+l),seq(0,365-
l+m),seq(0,366-m+n),seq(0,365-n+o),seq(0,365-o+p),seq(0,365-p+q),seq(0,366-
q+r),seq(0,365-r+s),seq(0,365-s+t),seq(0,365-t+u),seq(0,365-u))
```

```
parleys_data$day_sincemaxwsa=day_sincemaxwsa
#Calculation of max albedo each year#
a=max(subset(parleys_data,year==2001)$mean_wsa)
b=max(subset(parleys_data,year==2002)$mean_wsa)
c=max(subset(parleys_data,year==2003)$mean_wsa)
d=max(subset(parleys_data,year==2004)$mean_wsa)
e=max(subset(parleys_data,year==2005)$mean_wsa)
f=max(subset(parleys_data,year==2006)$mean_wsa)
g=max(subset(parleys_data,year==2007)$mean_wsa)
h=max(subset(parleys_data,year==2008)$mean_wsa)
i=max(subset(parleys_data,year==2009)$mean_wsa)
j=max(subset(parleys_data,year==2010)$mean_wsa)
k=max(subset(parleys_data,year==2011)$mean_wsa)
l=max(subset(parleys_data,year==2012)$mean_wsa)
m=max(subset(parleys_data,year==2013)$mean_wsa)
#Creating vector containing max mean_wsa for each year
max_mean_wsa=c(rep(a,365),rep(b,365),rep(c,365),rep(d,366),rep(e,365),rep(f,365),rep(g,365),rep(h,366),rep(i,365),rep(j,365),rep(k,365),rep(l,366),rep(m,365))
parleys_data$max_mean_wsa=max_mean_wsa
#Create new dataframe (copy of parleys_data) and additional variables
parleys_data2 = parleys_data
parleys_data2$lag1_mean_wsa = lag(parleys_data2$mean_wsa,k=1)
parleys_data2$lag2_mean_wsa = lag(parleys_data2$mean_wsa,k=2)
#Create new dataframe from parleys_data2, removing NA values
valID = which(complete.cases(parleys_data2))
parleys_data3 = parleys_data2[valID,]
#Find individual years
```

```
yrs = unique(parleys_data3$year)
nyrs = length(yrs)
#Create arrays to store rmsep and r2p values
rmsep_array22<-array(NA,nyrs)
r2p_array22<-array(NA,nyrs)
#Variables for ggplot2 plotting
dfyear = NULL
dftime = NULL
dfwteq = NULL
dfpwteq = NULL
#Load randomForest package
require(randomForest)
#k-fold cross-validation using years
for (y in 1:length(yrs)){
  yrID = which(parleys_data3$year==yrs[y])
  swe.train = parleys_data3[-yrID,]
  swe.test = parleys_data3[yrID,]
  swe.xval<-randomForest(wteq ~ day_ofyear + mean_wsa + lag1_mean_wsa +
lag2_mean_wsa + max_mean_wsa + day_sincemaxwsa, ntree=1000, mtry=3,
data=swe.train)
  swe.pred = predict(swe.xval, newdata=swe.test)
  rmsep_array22[y] = sqrt(mean((swe.pred - swe.test$wteq)^2))
  r2p_array22[y] = summary(lm(swe.pred ~ swe.test$wteq))$r.squared
  dfyear = c(dfyear, rep(yrs[y], length(swe.pred)))
  dftime = c(dftime, swe.test$day_ofyear)
  dfwteq = c(dfwteq, swe.test$wteq)
  dfpwteq = c(dfpwteq, swe.pred)
}
#ggplot2 plotting
require(ggplot2)
png("parleys_wsa_wteq_rf_kfold.png",width = 2000, height = 2000)
mydf = data.frame(time=dftime, year=dfyear,
                wteq=dfwteq, predwteq=dfpwteq)
x = ggplot(mydf, aes(x=time, y=wteq)) + geom_line()
x = x + geom_line(aes(x=time, y=predwteq), color="red")
x = x + facet_wrap(~ year)
x = x + ggtitle("Predicted (red) vs Actual (black) wteq, k-fold cross validation of A-SWE RandomForest")
x = x + theme(plot.title=element_text(size=20))
print(x)
dev.off()
```

APPENDIX G

SWE PREDICTED USING THE A-SWE RANDOM FOREST MODEL

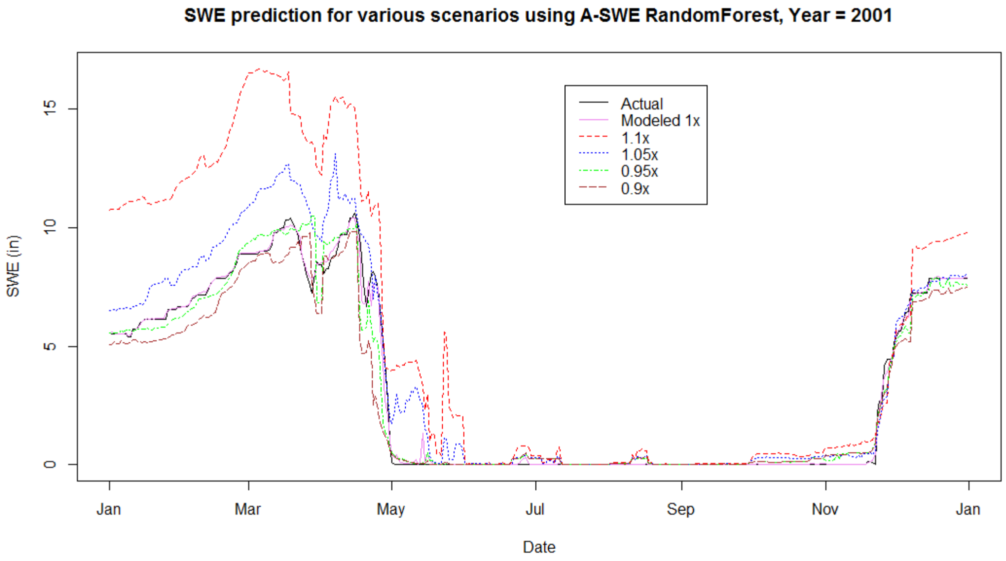

Figure G1: SWE prediction for 2001

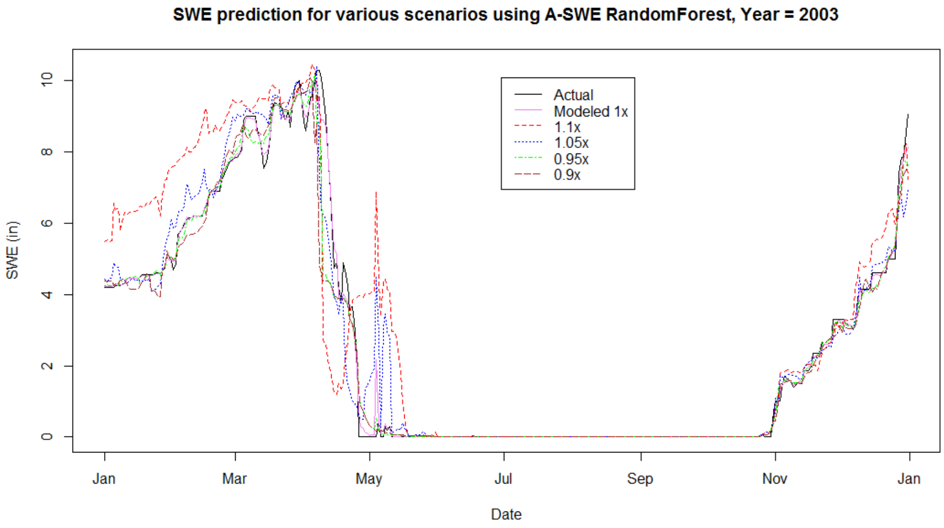

Figure G2: SWE prediction for 2003

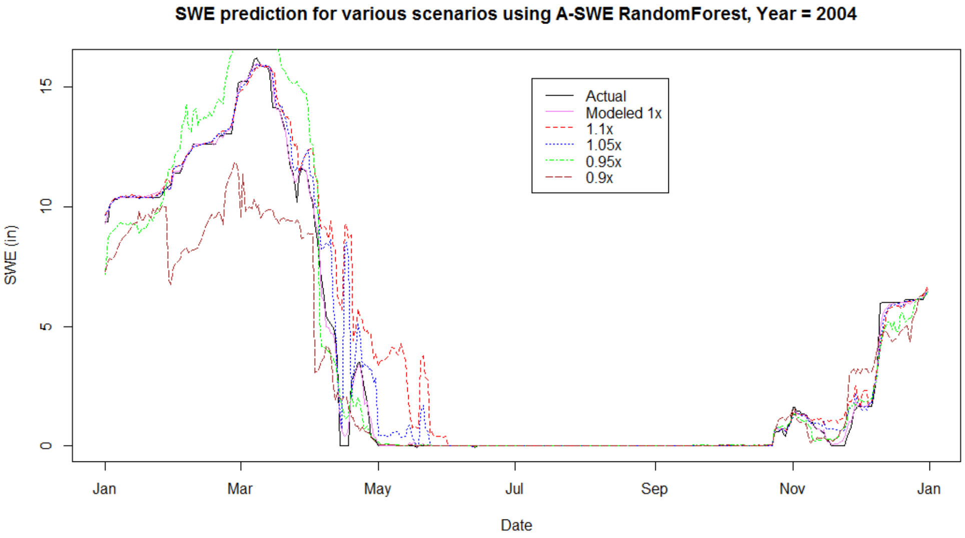

Figure G3: SWE prediction for 2004

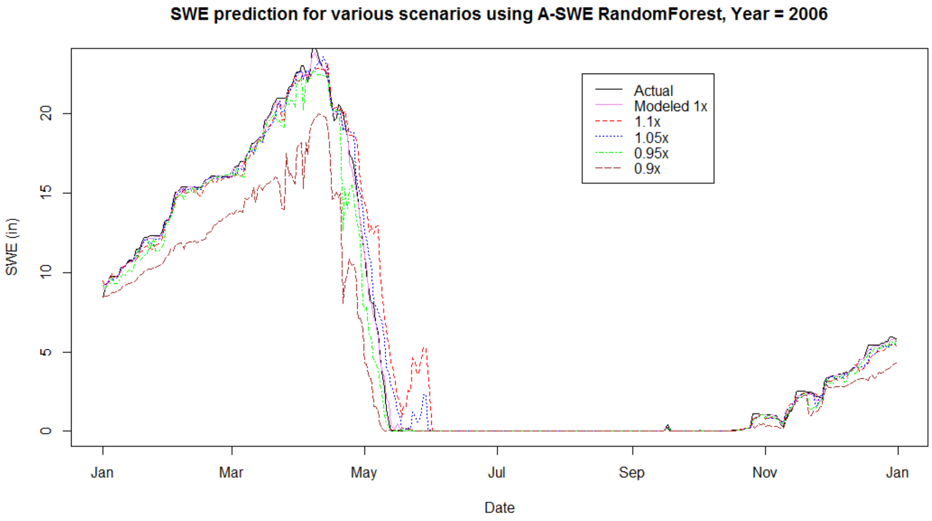

Figure G4: SWE prediction for 2006

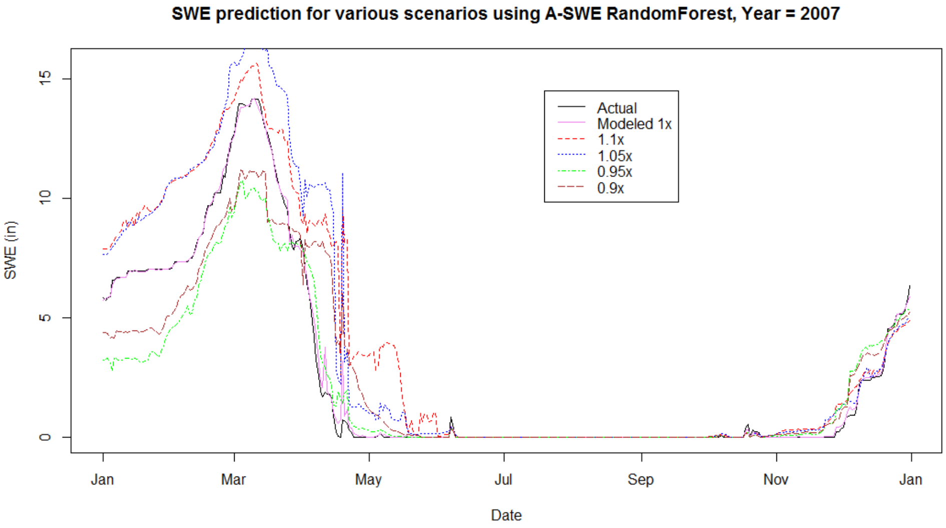

Figure G5: SWE prediction for 2007

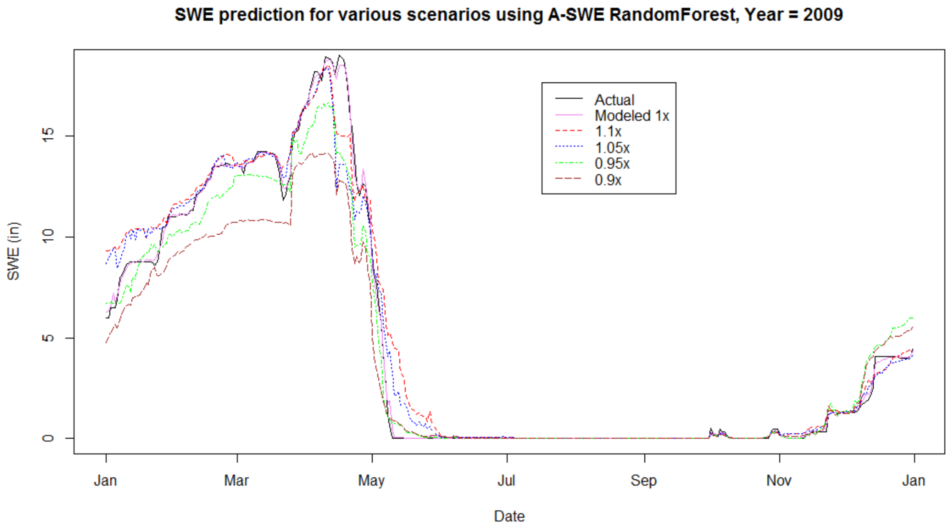

Figure G6: SWE prediction for 2009

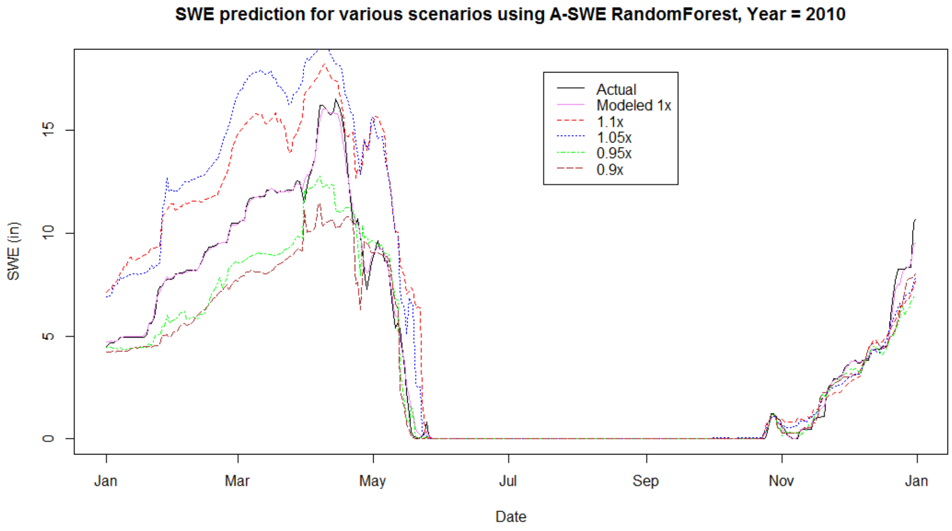

Figure G7: SWE prediction for 2010

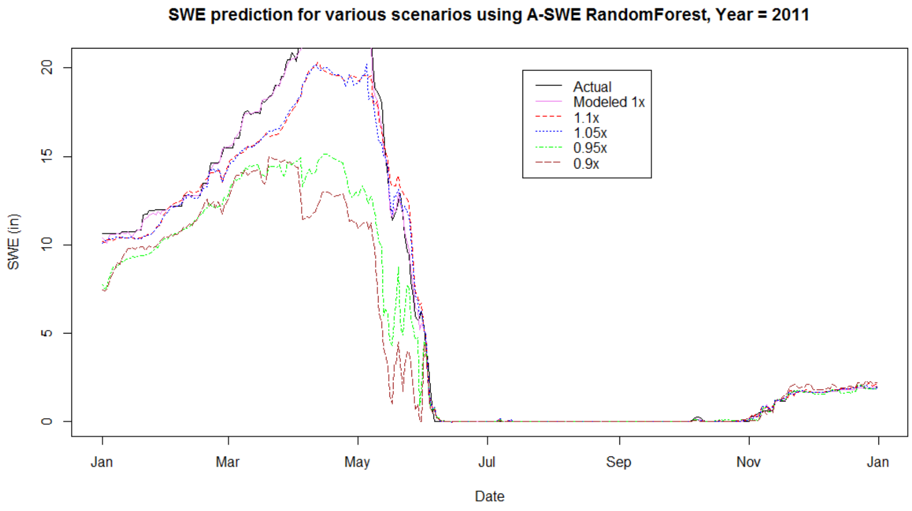

Figure G8: SWE prediction for 2011

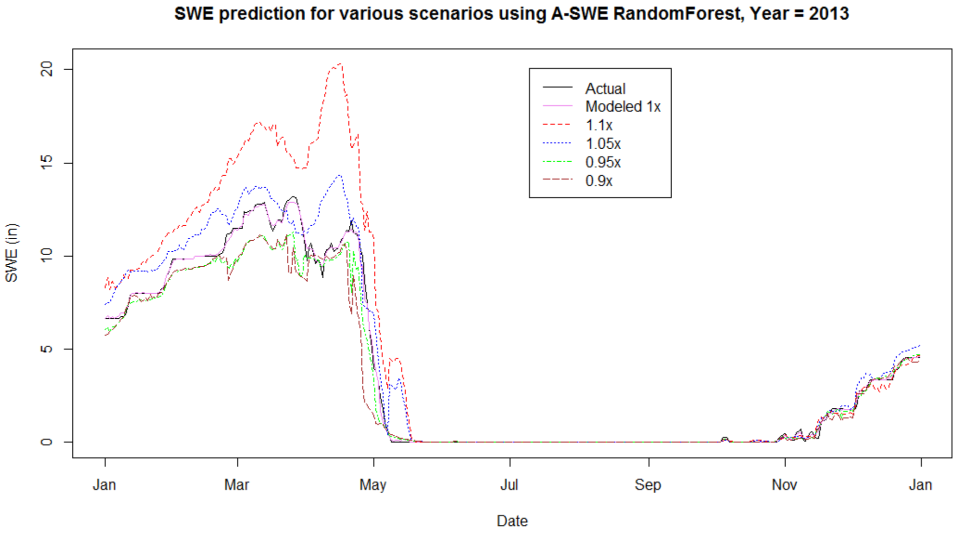

Figure G9: SWE prediction for 2013

APPENDIX H

A-SWE-S MODEL FORMULATION AND K-FOLD CROSS-VALIDATION

```
########################################
##Create variables for use in modeling##
require(nnet)
#Calculate day of max wteq each year
l=which.is.max(subset(parleys_data,year==2004)$wteq)
m=which.is.max(subset(parleys_data,year==2005)$wteq)
n=which.is.max(subset(parleys_data,year==2006)$wteq)
o=which.is.max(subset(parleys_data,year==2007)$wteq)
p=which.is.max(subset(parleys_data,year==2008)$wteq)
q=which.is.max(subset(parleys_data,year==2009)$wteq)
r=which.is.max(subset(parleys_data,year==2010)$wteq)
s=which.is.max(subset(parleys_data,year==2011)$wteq)
t=which.is.max(subset(parleys_data,year==2012)$wteq)
u=which.is.max(subset(parleys_data,year==2013)$wteq)
#Creating vector containing days since melt started (day after max_wteq) each year
day_sincemelt=c(seq(0,365-l+m),seq(0,366-m+n),seq(0,365-n+o),seq(0,365-o+p),seq(0,365-p+q),seq(0,366-q+r),seq(0,365-r+s),seq(0,365-s+t),seq(0,365-t+u),seq(0,365-u))
day_sincemelt=c(seq(311,310+3653-length(day_sincemelt)),day_sincemelt)
```

```
parleys_data$day_sincemelt=day_sincemelt
#Calculation of Max Wteq each year
d=max(subset(parleys_data,year==2004)$wteq)
e=max(subset(parleys_data,year==2005)$wteq)
f=max(subset(parleys_data,year==2006)$wteq)
g=max(subset(parleys_data,year==2007)$wteq)
h=max(subset(parleys_data,year==2008)$wteq)
i=max(subset(parleys_data,year==2009)$wteq)
j=max(subset(parleys_data,year==2010)$wteq)
k=max(subset(parleys_data,year==2011)$wteq)
l=max(subset(parleys_data,year==2012)$wteq)
m=max(subset(parleys_data,year==2013)$wteq)
#Creating vector containing max WTEQ for each year
max_wteq=c(rep(d,366),rep(e,365),rep(f,365),rep(g,365),rep(h,366),rep(i,365),rep(j,365),
rep(k,365),rep(l,366),rep(m,365))
parleys_data$max_wteq=max_wteq
#Create copy of parleys_data dataframe
parleys_data2 = parleys_data
#5 day precip memory
precip_memory5=rep(NA,nrow(parleys_data))
precip_memory5[1]=0
c=0
for (i in 1:nrow(parleys_data)){
  #print(i)
  #print(c)
  c=c+1
  if(c==5){
    precip_memory5[i]=0
    c=0
  }
  precip_memory5[i+1]=parleys_data$precip[i+1]+precip_memory5[i]
}
precip_memory5=precip_memory5[1:nrow(parleys_data)]
#10 day precip memory
precip_memory10=rep(NA,nrow(parleys_data))
precip_memory10[1]=0
c=0
for (i in 1:nrow(parleys_data)){
  #print(i)
  #print(c)
  c=c+1
  if(c==10){
    precip_memory10[i]=0
    c=0
  }
  precip_memory10[i+1]=parleys_data$precip[i+1]+precip_memory10[i]
}
precip_memory10=precip_memory10[1:nrow(parleys_data)]
#15 day precip memory
precip_memory15=rep(NA,nrow(parleys_data))
precip_memory15[1]=0
c=0
for (i in 1:nrow(parleys_data)){
  #print(i)
  #print(c)
  c=c+1
  if(c==15){
    precip_memory15[i]=0
    c=0
  }
  precip_memory15[i+1]=parleys_data$precip[i+1]+precip_memory15[i]
}
precip_memory15=precip_memory15[1:nrow(parleys_data)]
#30 day precip memory
precip_memory30=rep(NA,nrow(parleys_data))
precip_memory30[1]=0
c=0
for (i in 1:nrow(parleys_data)){
  #print(i)
  #print(c)
  c=c+1
  if(c==30){
    precip_memory30[i]=0
    c=0
  }
  precip_memory30[i+1]=parleys_data$precip[i+1]+precip_memory30[i]
}
precip_memory30=precip_memory30[1:nrow(parleys_data)]
#Add precip memory terms to dataset
parleys_data2$precip_memory5=precip_memory5
parleys_data2$precip_memory10=precip_memory10
```

```
parleys_data2$precip_memory15=precip_memory15
parleys_data2$precip_memory30=precip_memory30
#Create additional variables#
parleys_data2$lflow = log10(parleys_data2$flow)
parleys_data2$log10mean_wsa=log10(parleys_data2$mean_wsa)
parleys_data3=parleys_data2
####k-fold cross-validation###
require(mgcv)
require(ggplot2)
#Variables for plotting
dfyear = NULL
dftime = NULL
dfflow = NULL
dfpflow = NULL
#Find all years in data set and number of years
allyrs = unique(parleys_data3$year)
nyrs = length(yrs)
#Create knots of various variables in dataset
day_knots=100
sincemelt_knots=100
precip_knots=100
wteq_knots=150
sms2_knots=150
sms8_knots=150
sms20_knots=150
sto20_knots=150
alb_knots=100
```

```
precip_memory10_knots=40
precip_memory15_knots=40
precip_memory30_knots=40
snwd_knots=50
#Create arrays to hold rmsep and r2p values for various years
rmsep_array50<-array(NA,nyrs)
r2p_array50<-array(NA,nyrs)
#k-fold
for (h in 1:nyrs){
  print(paste("Year",allyrs[h],h))
  yearID = which(parleys_data3$year==allyrs[h])
  flow.train = parleys_data3[-yearID,]
  flow.test = parleys_data3[yearID,]
  flow.xval = gam(lflow~s(day_ofyear,bs='cc',k=day_knots)+s(precip_accum,bs='cc',k=precip_knots)+s(day_sincemelt,bs='cc',k=sincemelt_knots)+s(sms20,bs='cr',k=sms20_knots)+s(sto20,k=sto20_knots)+s(wteq,k=wteq_knots)+s(sms8,bs='cr',k=sms8_knots)+s(sms2,bs='cr',k=sms2_knots)+s(log10mean_wsa,bs='cr',k=alb_knots)+s(precip_memory15,k=precip_memory15_knots)+s(precip_memory10,k=precip_memory10_knots)+s(precip_memory30,k=precip_memory30_knots)+s(snwd,k=snwd_knots),data=flow.train,family=gaussian)
  flow.pred = predict(flow.xval, flow.test)
  rmsep_array50[h] = 10**(sqrt(mean((flow.test$lflow - flow.pred)^2, na.rm=TRUE)))
  r2p_array50[h] = summary(lm(10**(flow.pred) ~ flow.test$flow))$r.squared
  dfyear = c(dfyear, rep(allyrs[h], length(flow.pred)))
  dftime = c(dftime, flow.test$day_ofyear)
  dfflow = c(dfflow, flow.test$flow)
  dfpflow = c(dfpflow, flow.pred)
}
 min_rmsep=min(rmsep_array50)
max_rmsep=max(rmsep_array50)
max_r2p=max(r2p_array50)
min_r2p=min(r2p_array50)
optim_r2p<-which(r2p_array50 == max(r2p_array50), arr.ind = TRUE)
optim_rmsep<-which(rmsep_array50 == min(rmsep_array50), arr.ind = TRUE)
min_rmsep #Best rmsep
max_rmsep #Worst rmsep
max_r2p #Best r2p
min_r2p #Worst r2p
optim_r2p #Find combination with maximum r2p (Best model fit combination)
optim_rmsep #Find combination with minimum rmsep (Combination with least error of prediction)
#Plotting using ggplot2
png("parleys_wteq_flow_gam_kfold.png",width=2400,height=2400)
mydf = data.frame(time=dftime, year=dfyear,
                 flow=dfflow, predflow=10**dfpflow)
x = ggplot(mydf, aes(x=time, y=flow)) + geom_line()
x = x + geom_line(aes(x=time, y=predflow), color="red")
x = x + facet_wrap(~ year)
x = x + ggtitle("Predicted (red) vs actual (black) flow, k-fold cross-validation of A-SWE-S GAM")
print(x)
dev.off()
> summary(flow.xval)
Family: gaussian
```

```
Link function: identity

Formula:

lflow ~ s(day_ofyear, bs = "cc", k = 100) + s(precip_accum, bs = "cc",
    k = 100) + s(day_sincemelt, bs = "cc", k = 100) + s(sms20,
    bs = "cr", k = 150) + s(sto20, k = 150) + s(wteq, k = 150) +
    s(sms8, bs = "cr", k = 150) + s(sms2, bs = "cr", k = 150) +
    s(log10mean_wsa, bs = "cr", k = 100) + s(precip_memory15,
    k = 40) + s(precip_memory10, k = 40) + s(precip_memory30,
    k = 40) + s(snwd, k = 50)


Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.806713   0.001533   526.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Approximate significance of smooth terms:
                     edf Ref.df      F  p-value
s(day_ofyear)       68.74  98.00 12.610  < 2e-16 ***
s(precip_accum)     92.20  98.00 47.687  < 2e-16 ***
s(day_sincemelt)    84.83  98.00 10.777  < 2e-16 ***
s(sms20)            38.87  48.43  6.987  < 2e-16 ***
s(sto20)            48.34  59.67  3.948  < 2e-16 ***
s(wteq)            105.37 122.97  3.569  < 2e-16 ***
s(sms8)             33.27  41.59  6.798  < 2e-16 ***
s(sms2)             19.85  24.94  8.259  < 2e-16 ***
s(log10mean_wsa)    39.98  48.14  3.393 5.36e-14 ***
s(precip_memory15)  26.56  30.86  2.323 5.04e-05 ***
```

```
s(precip_memory10)  23.24  27.22  2.331 0.000115 ***

s(precip_memory30)  34.32  37.14  7.374  < 2e-16 ***

s(snwd)             28.22  34.01  3.658 6.04e-12 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.924   Deviance explained = 93.7%

GCV = 0.010423  Scale est. = 0.0085834  n = 3653
```

REFERENCES

1. Anderson, E.A., 2006. Snow Accumulation and Ablation model–SNOW-17. http://www.nws.noaa.gov/oh/hrl/nwsrfs/users_manual/part2/_pdf/22snow17.pdf

2. Bales, R.C., Q. Guo, D. Shen, J.R. McConnell, G. Du, J.F. Burkhart, V.B. Spikes, E. Hanna, and J. Cappelen, 2009. Annual Accumulation for Greenland Updated Using Ice Core Data Developed during 2000–2006 and Analysis of Daily Coastal Meteorological Data. *Journal of Geophysical Research: Atmospheres* 114:n/a–n/a.

3. Becker, R.A., J.M. Chambers, and A.R. Wilks, 1988. *The New S Language*. Wadsworth & Brooks, 1988 1. Pacific Grove, CA.

4. Breiman, L., 2001a. Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author). *Statistical Science* 16.3 (2001):199–231.

5. Breiman, L., 2001b. Random Forests. *Machine Learning* 45:5–32.

6. Breiman, L., J. Friedman, C.J. Stone, and R.A. Olshen, 1984. *Classification and Regression Trees*. CRC Press. Boca Raton, FL.

7. Bryant, A.C., T.H. Painter, J.S. Deems, and S.M. Bender, 2013. Impact of Dust Radiative Forcing in Snow on Accuracy of Operational Runoff Prediction in the Upper Colorado River Basin. *Geophysical Research Letters* 40:3945–3949.

8. Cornell University. Assessing the Fit of Regression Models. Cornell Statistical Consulting Unit. http://www.cscu.cornell.edu/news/statnews/stnews68.pdf. Accessed 11/20/2014.

9. Crane, R.G. and M.R. Anderson, 1984. Satellite Discrimination of Snow/cloud Surfaces. *International Journal of Remote Sensing* 5:213–223.

10. Deems, J.S., T.H. Painter, J.J. Barsugli, J. Belnap, and B. Udall, 2013. Combined Impacts of Current and Future Dust Deposition and Regional Warming on Colorado River Basin Snow Dynamics and Hydrology. *Hydrol. Earth Syst. Sci.* 17:4401–4413.

11. Dobos, E., 2005. Albedo. *Encyclopedia of Soil Science, Second Edition*. CRC Press. Boca Raton, FL.

12. Dozier, J. and D. Marks, 1987. Snow Mapping and Classification from Landsat Thematic Mapper Data. *Annals of Glaciology* 9:97–103.

13. E.P.A. 2010. Salt Lake City Works with Stakeholders to Protect Its Water Supply. Office of Water (4606M), 816F10048.

14. Flanner, M.G., K.M. Shell, M. Barlage, D.K. Perovich, and M.A. Tschudi, 2011. Radiative Forcing and Albedo Feedback from the Northern Hemisphere Cryosphere between 1979 and 2008. *Nature Geoscience* 4:151–155.

15. Flanner, M.G., C.S. Zender, P.G. Hess, N.M. Mahowald, T.H. Painter, V. Ramanathan, and P.J. Rasch, 2009. Springtime Warming and Reduced Snow Cover from Carbonaceous Particles. *Atmospheric Chemistry and Physics* 9:2481–2497.

16. Flanner, M.G., C.S. Zender, J.T. Randerson, and P.J. Rasch, 2007. Present-day Climate Forcing and Response from Black Carbon in Snow. *Journal of Geophysical Research: Atmospheres* (1984–2012) 112.

17. Hadley, O.L. and T.W. Kirchstetter, 2012. Black-Carbon Reduction of Snow Albedo. *Nature Climate Change* 2:437–440.

18. Hall, D.K. and J. Martinec, 1986. *Remote Sensing of Ice and Snow*. Chapman and Hall Ltd., London.

19. Hall, D.K. and G.A. Riggs, 2007. Accuracy Assessment of the MODIS Snow Products. *Hydrological Processes* 21:1534–1547.

20. Khadka, D., M.S. Babel, S. Shrestha, and N.K. Tripathi, 2014. Climate Change Impact on Glacier and Snow Melt and Runoff in Tamakoshi Basin in the Hindu Kush Himalayan (HKH) Region. *Journal of Hydrology* 511:49–60.

21. Kohavi, R. and F. Provost, 1998. Glossary of Terms. *Machine Learning* 30:271–274.

22. Lewis, P. and M.J. Barnsley, 1994. Influence of the Sky Radiance Distribution on Various Formulations of the Earth Surface Albedo. *6th International Symposium on Physical Measurements and Signatures in Remote Sensing (ISPRS)*. CNES Val d’Isere, France, pp. 707–715.

23. Liang, S., A. Strahler, and C. Walthall, 1998. Retrieval of Land Surface Albedo from Satellite Observations: A Simulation Study. *Geoscience and Remote Sensing Symposium Proceedings, 1998 (IGARSS’98)*. 1998 IEEE International. IEEE, pp. 1286–1288.

24. Liaw, A. and M. Wiener, 2002. Classification and Regression by randomForest. *R News* 2:18–22.

25. Mohri, M., A. Rostamizadeh, and A. Talwalkar, 2012. Foundations of Machine Learning. MIT Press. Cambridge, MA.

26. NASA GCFC, MODIS-Atmosphere: Frequently Asked Questions. http://modis-atmos.gsfc.nasa.gov/ALBEDO/faq.html. Accessed 11/20/2014.

27. National Snow and Ice Data Center. Thermodynamics: Albedo. http://nsidc.org/cryosphere/seaice/processes/albedo.html. Accessed 09/20/2014.

28. NCDF. Ncdf: Interface to Unidata netCDF Data Files. http://cran.r-project.org/web/packages/ncdf/index.html. Accessed 09/20/2014.

29. Neil Frank [EPA]. The Chemical Composition of PM2.5 to Support PM Implementation. Presentation at *EPA State / Local / Tribal Training Workshop: PM 2.5 Final Rule Implementation and 2006 PM 2.5, Designation Process. AQAG/AQAD USEPA.*

30. Oerlemans, J., R.H. Giesen, and M.R. Van den Broeke, 2009. Retreating Alpine Glaciers: Increased Melt Rates due to Accumulation of Dust (Vadret Da Morteratsch, Switzerland). *Journal of Glaciology* 55:729–736.

31. Painter, T.H., A.P. Barrett, C.C. Landry, J.C. Neff, M.P. Cassidy, C.R. Lawrence, K.E. McBride, and G.L. Farmer, 2007. Impact of Disturbed Desert Soils on Duration of Mountain Snow Cover. *Geophysical Research Letters* 34.12 (2007).

32. Painter, T.H., S.M. Skiles, J.S. Deems, A.C. Bryant, and C.C. Landry, 2012. Dust Radiative Forcing in Snow of the Upper Colorado River Basin: 1. A 6 Year Record of Energy Balance, Radiation, and Dust Concentrations. *Water Resources Research* 48.7 (2012).

33. Polikar, R., 2006. Ensemble Based Systems in Decision Making. Circuits and Systems Magazine. *IEEE* 6:21–45.

34. Professor Crystal Schaaf's Lab. MCD43A3 Albedo Product. http://www.umb.edu/spectralmass/terra_aqua_modis/v006/mcd43a3_albedo_product. Accessed 11/20/2014.

35. Qu, B., J. Ming, S.-C. Kang, G.-S. Zhang, Y.-W. Li, C.-D. Li, S.-Y. Zhao, Z.-M. Ji, and J.-J. Cao, 2014. The Decreasing Albedo of the Zhadang Glacier on Western Nyainqentanglha and the Role of Light-Absorbing Impurities. *Atmospheric Chemistry and Physics* 14:11117–11128.

36. R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org/. Accessed 09/20/2014.

37. R documentation. GGally: Extension to ggplot2. http://cran.r-project.org/web/packages/GGally/index.html. Accessed 09/20/2014.

38. R Documentation acf{stats}. Auto- and Cross- Covariance and -Correlation Function Estimation. https://stat.ethz.ch/R-manual/R-patched/ library/stats/html/ acf.html. Accessed 09/20/2014.

39. R documentation. Data.frame {base}, Data Frames. https://stat.ethz.ch/R-manual/R-devel/library/base/html/data.frame.html. Accessed 09/20/2014.

40. R gam {mgcv} documentation. Generalized Additive Models with Integrated Smoothness Estimation. http://stat.ethz.ch/R-manual/R-patched/library/mgcv/html/gam.html. Accessed 09/20/2014.

41. Reynolds, R.L., H.L. Goldstein, B.M. Moskowitz, A.C. Bryant, S.M. Skiles, R.F. Kokaly, C.B. Flagg, K. Yauk, T. Berquó, and G. Breit, 2014. Composition of Dust Deposited to Snow Cover in the Wasatch Range (Utah, USA): Controls on

Radiative Properties of Snow Cover and Comparison to Some Dust-Source Sediments. *Aeolian Research* 15:73–90.

42. Román, M.O., C.B. Schaaf, P. Lewis, F. Gao, G.P. Anderson, J.L. Privette, A.H. Strahler, C.E. Woodcock, and M. Barnsley, 2010. Assessing the Coupling between Surface Albedo Derived from MODIS and the Fraction of Diffuse Skylight over Spatially-Characterized Landscapes. *Remote Sensing of Environment* 114:738–760.

43. Russell, S., P. Norvig, and A. Intelligence, 1995. A Modern Approach. Artificial Intelligence. Prentice-Hall. Egnlewood Cliffs 25.

44. Sakamoto, Y., M. Ishiguro, and G. Kitagawa, 1986. Akaike Information Criterion Statistics. D. Reidel. Dordrecht, The Netherlands.

45. Salt Lake City Department of Public Utilities. Salt Lake City Watershed Management Plan. http://www.townofalta.com/pdf/SLC_Watershed_Management_Plan.pdf. Accessed 09/20/2014.

46. Schaefer, G.L. and R.F. Paetzold, 2001. SNOTEL (SNOwpack TELemetry) and SCAN (soil Climate Analysis Network). *Proc. Intl. Workshop on Automated Wea. Stations for Appl. in Agr. and Water Resour. Mgmt.*

47. Silcox, G.D., K.E. Kelly, E.T. Crosman, C.D. Whiteman, and B.L. Allen, 2012. Wintertime PM 2.5 Concentrations during Persistent, Multi-Day Cold-Air Pools in a Mountain Valley. *Atmospheric Environment* 46:17–24.

48. Simon, P., 2013. Too Big to Ignore: The Business Case for Big Data. John Wiley & Sons. Hoboken, New Jersey

49. Steenburgh, W.J., J.D. Massey, and T.H. Painter, 2012. Episodic Dust Events of Utah’s Wasatch Front and Adjoining Region. *Journal of Applied Meteorology and Climatology* 51:1654–1669.

50. Stewart, I.T., D.R. Cayan, and M.D. Dettinger, 2004. Changes in Snowmelt Runoff Timing in Western North America under Abusiness as Usual’climate Change Scenario. *Climatic Change* 62:217–232.

51. Stone, R.S., E.G. Dutton, J.M. Harris, and D. Longenecker, 2002. Earlier Spring Snowmelt in Northern Alaska as an Indicator of Climate Change. *Journal of Geophysical Research: Atmospheres* (1984–2012) 107:ACL–10.

52. Strobl, C., J. Malley, and G. Tutz, 2009. An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychological Methods* 14:323.

53. Toon, O.B., C.P. McKay, T.P. Ackerman, and K. Santhanam, 1989. Rapid Calculation of Radiative Heating Rates and Photodissociation Rates in Inhomogeneous Multiple Scattering Atmospheres. *Journal of Geophysical Research: Atmospheres* 94:16287–16301.

54. U.S. Army Corps of Engineers (USACE). EM 1110-2-1406, CECW-EH, Runoff from Snowmelt. Department of the Army, March 31, 1998. Washington, DC.

55. Utah Division of Air Quality. Emission Sources of Winter PM 2.5. http://www.deq.utah.gov/FactSheets/fspages/sources.htm. Accessed 09/20/2014.

56. Utah Division of Air Quality, 2010. Utah 2010 Air Monitoring Network Plan. http://www.epa.gov/ttnamti1/files/networkplans/UTPlan2010.pdf. Accessed 09/20/2014.

57. Warren, S.G., 1982. Ice and Climate Modeling: An Editorial Essay. *Climatic Change* 4:329–340.

58. Warren, S.G. and W.J. Wiscombe, 1980. A Model for the Spectral Albedo of Snow. II: Snow Containing Atmospheric Aerosols. *Journal of the Atmospheric Sciences* 37:2734–2745.

59. WMO (World Meteorological Organization). Review on Remote Sensing of the Snow Cover and on Methods of Mapping Snow. http://www.wmo.int/pages/prog/hwrp/chy/chy14/documents/ms/remote_sensing_snow_cover_methods_mapping_snow.pdf. Accessed 09/20/2014.

60. Yasunari, T.J., P. Bonasoni, P. Laj, K. Fujita, E. Vuillermoz, A. Marinoni, P. Cristofanelli, R. Duchi, G. Tartari, and K.-M. Lau, 2010. Estimated Impact of Black Carbon Deposition during Pre-Monsoon Season from Nepal Climate Observatory–Pyramid Data and Snow Albedo Changes over Himalayan Glaciers. *Atmospheric Chemistry and Physics* 10:6603–6615.