Snow Depth Estimation, Structure, Prediction, and Hydrologic Modeling at the Kilometer Scale in the Colorado Rocky Mountains.

by

James L. McCreight

B.A., University of Montana, 2000

M.S., University of Colorado, 2003

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Civil, Environmental, and Architectural Engineering

2010

This thesis entitled:
Snow Depth Estimation, Structure, Prediction, and Hydrologic Modeling at the Kilometer Scale in the Colorado Rocky Mountains.
written by James L. McCreight
has been approved for the Department of Civil, Environmental, and Architectural Engineering

Rajagopalan Balaji

Prof. Thomas Chase

Dr. Lauren Hay

Dr. Hans-Peter Marshall

Prof. Harihar Rajaram

Dr. Andrew Slater

Date _____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

McCreight, James L. (Ph.D., Hydrology, Water Resources, and Environmental Fluid Mechanics)

Snow Depth Estimation, Structure, Prediction, and Hydrologic Modeling at the Kilometer Scale in the Colorado Rocky Mountains.

Thesis directed by Prof. Rajagopalan Balaji

Research focuses on observing and predicting spatial distribution of snow depth at the kilometer scale. Observation of spatial snow depth distribution is considered by its estimation from random, sparse observations and important factors affecting this estimation. Predicting spatial distribution of both snow depth and melt rates begins from simple hypothesis wherein the spatial distribution of snow depth is structured by the spatial distribution of controlling variables. Predictions made by this structured view are evaluated in spatial modeling of peak-accumulation snow depth and applied to spatial distribution of a point-scale, temperature-index model of snowmelt runoff using minimal parameter complexity.

High-resolution light detection and ranging (LiDAR) measurements provide a rich backdrop for understanding estimation from sparse observations and developing our structured view of snow distribution. The data are used to illuminate the effects of sample size on estimation skill, the uncertainty in estimation due to random sampling, the effect of model resolution on estimation skill, and the difference between cross-validated skill and skill based on the entire distribution. None of these topics have previously been explored in the literature. The effect of predictor quality is also investigated. LiDAR derived predictors are compared to readily available predictors downloaded from the internet.

Hierarchical cluster analysis is used to decompose spatial non-stationarity of snow depth and results match qualitative understanding of the spatial distribution of physical controls. The same methodology is then used to decompose spatial non-stationarity of physical controls and infer patterns of snow depth distribution independent of observations. Even when using readily-available predictors, predicted patterns require at least 100-200 observations to be matched by standard

estimation methods.

Predicted patterns are then applied to formulate a parameterized spatial distribution of a 1-dimensional, temperature-index model to account for heterogeneity of both snow accumulation and melt. Our new method introduces fewer or comparable parameters as the current subgrid distribution, the areal depletion curve. Given highly uncertain parameter selection in practical application, we demonstrate that our more physically intuitive method virtually always results in significant improvement in simulated streamflow timing when compared to the depletion curve method.

Dedication

To my family, old and new.

I never could have done this without your support.

I love you all.

Acknowledgements

This research was supported by NASA headquarters under the NASA Earth and Space Science Fellowship Program - Grant NNX07AO40H. Thanks to the reviewers of my proposal, even if this work is rather different.

Thanks to my friend and body work expert, Dave Moore, for fixing my shoulders, neck, back, and arms throughout this marathon of coding and typing. Thanks to Chappell, for keeping me tethered to what is really important, for feeding me, for putting up with this whole process, and, not the least, for distracting me always.

Thanks to Adam Winstral for help with his maximum upwind slope code. Thanks to Jeff Deems for assistance at many steps along the way. Thanks to Chris Landry (a.k.a the pit viper) at the Center for Snow and Avalanche Studies for his excellent observational record at Senator Beck Basin and his insights into the data. Thanks to Tom Painter and Andy Barrett for snow course observations at Senator Beck Basin and discussion. Thanks to Andy Gleason for zombie pets, fish tacos, a great attitude, love of the mountains and of snow science. I instructed the printers to make this page yellow. Thanks to H.P. Marshall for his tireless enthusiasm for snow and science, discussions on the banks of the Beas, the opportunity to run radar and snow probes in Senator Beck Basin, and for all his help and encouragement. Thanks to Drew Slater for pointing me in the right direction so many times, for being very practical, and for many years of useful discussion, assistance, and encouragement. Thanks to Balaji, for his encouragement, enthusiasm, creativity, and the freedom to follow my own path.

Finally, thanks to the snow, the water, the winds, the sun, and the mountains, for making this all possible and non-trivial in the first place.

Contents

Chapter

1 Introduction 1

2 CLPX LiDAR Sites 5

2.1 CLPX overview 5

2.2 High-resolution LiDAR data 5

2.3 Low-resolution data 9

2.4 Derived predictor variables 11

2.5 Overview of sites, F and R groups 15

3 Estimation from Sparse Observations 18

3.1 Introduction 18

3.2 Background 21

3.3 Experimental design and methodology 24

3.4 Skill as a function of observation density and model resolution 28

3.4.1 Percent error 28

3.4.2 RMSE 30

3.4.3 R^2 33

3.5 Skill of downloaded and LiDAR-measured predictors 36

3.5.1 PE 36

3.5.2 RMSE 39

3.5.3 R^2 41

3.6 Sampling design 44

3.7 Discussion 49

3.8 Conclusion 51

4 Snow Depth Spatial Structure 53

4.1 Introduction 53

4.2 Background 54

4.3 Methodology 57

4.3.1 Hierarchical k-means clustering 57

4.3.2 Snow depth structure and clustering 58

4.3.3 Predictor non-stationarity and clustering: cluster prediction 59

4.3.4 Binary regression trees and R-squared 60

4.4 Site snow depth non-stationarity 61

4.4.1 FA 63

4.4.2 FF 64

4.4.3 FS 65

4.4.4 RB 66

4.4.5 RS 67

4.4.6 RW 67

4.5 Predictor non-stationarity and prediction of snow depth non-stationarity 69

4.5.1 Application to spatial modeling 72

4.6 Conclusions 78

5 Spatial Structure and Hydrologic Modeling 80

5.1 Introduction 80

5.2 Background 81

5.3 Data: Senator Beck basin 84

5.4 Methodology 87
5.4.1 Experimental design 87
5.4.2 Snow-17 model and calibration at SASP 90
5.4.3 Parameter-region modeling 92
5.4.4 Runoff model 98
5.4.5 Parameter sampling 99
5.5 Results 101
5.5.1 Spatially distributed melt rates and the ADC 103
5.5.2 Cluster model parameters and streamflow simulation 105
5.5.3 Three model deficiencies 111
5.6 Conclusions 114

6 Summary, Conclusions, & Directions 116

Bibliography 121

Appendix

A CLPX LiDAR: Details 125
A.1 LiDAR and in situ measurements 125
A.2 Power-law variogram 126
A.3 Nearest neighbor regridding 126
A.4 Derivation of vegetation height from NLCD2001 class 126

B Estimation from Sparse Observations: Details 127
B.1 Algorithm outlines 127
B.2 Models and fitting 128
B.2.1 Binary regression trees 128

B.2.2 Bagging trees 130

B.2.3 LM - linear regression model 131

Tables

Table

2.1 Observed and derived statistics of snow depth and physiographic variables at each site. 16

3.1 Summary of results from previous studies of snow depth estimation using binary regression trees (BRT) at kilometer scales ($< 20km^2$). Predictor sources listed are United States Geological Survey (USGS) 7.5 minute maps, National Elevation Data set (NED), Shuttle Radar Topographic Mission (SRTM), and the 2001 National LandCover Data set (NLCD2001). 23

3.2 Apparent and true R^2 for CLPX LiDAR measurements using identical sampling design to the in situ measurements. Results for downloaded and LiDAR-measured predictor sets. 46

3.3 Comparison of snow conditions to Erxleben et al. (2002). Depths and standard deviations in meters. 49

3.4 Comparison of RMSE (m) and R^2 results to Erxleben et al. (2002). 50

4.1 Statistics of independent variables within cluster nodes at each site as shown in figure 4.1. All are mean values except for aspect which is the mode over 10 degree bins and measured in degrees CCW from east. Snow depth and vegetation density are measured in meters. Potential accumulated solar radiation is measured in $Wh/m^2/day$ and maximum upwind slope in degrees from horizontal. 68

4.2 R^2 for three methods of predictor selection at each site for LiDAR-measured and downloaded predictor variable sets. Variables are elevation (ele), vegetation height (veg), vegetation density (vegdense), aspect (asp), slope (slp), potential accumulated solar radiation (sun), maximum upwind slope (sx), and northness (n). The ''Shuffle .95'' column gives a 95% confidence limit on the maximum amount of R^2 resulting from autocorrelation. Percent of BRT skill is calculated at 1600 observations and compared for identical predictor sets. 77

5.1 Table of calibrated snow-17 model parameters at SASP in 2005 and 2006. See Anderson (2006) for details on the parameters. 91

5.2 Predictor variable statistics in each SBB cluster region shown in the middle panel of figure 5.6. All are mean values except aspect which is the mode over 10 degree bins. Potential accumulated solar radiation is in units of $Wh/m^2/day$, maximum upwind slope (Sx) is in units of degrees from horizontal, vegetation density and elevation are in units of meters. 95

Figures

Figure

2.1 Location of sites in Colorado, USA. Latitude and longitude shown on figure border. 6

2.2 Overview of the separate areas containing the F and R sites. Contours 25m, bold contours 100m. 10

2.3 High-resolution elevation, vegetation, and snow depth at each site. Site dimensions shown on the first plot, all other sites are the same dimension. Contours shown at 25m. Location of weather stations indicated with a bold x in each. Location of each site can be seen in both figures 2.1 and 2.2, including surrounding terrain and distances between sites in the latter. 12

2.4 NLCD2001 landcover classification at each site. Developed areas at FS and RS are roads. 13

2.5 Ten meter wind speeds and directions measured at each site's weather station from November 1, 2002, through April 8, 2003. Directions summarized are for only wind speeds of at least $5ms^{-1}$. 14

2.6 Snow depth histograms at all sites, the F sites, the R sites, and all individual sites. Solid line is high-resolution and dotted is low-resolution snow depth. Bin size is 15cm. 16

3.1 Snow depth sampling design from Erxleben et al. (2002). 20

3.2 Uncertainty (90% confidence) in percent error resulting from random selection of observations as a function of observation count and model resolution. Median percent error is not shown because it is essentially zero at all but the lowest observation densities. Median precent error is shown for 30m resolution in figure 3.5. 27

3.3 Median RMSE and its uncertainty (90% confidence) due to random selection of observations as function of observation count and resolution. 31

3.4 R^2 and its uncertainty (90% confidence) due to random selection of observations as a function of observation density and model resolution. 34

3.5 Percent error and its uncertainty (90% confidence) as a function of observation density at 30m resolution. 38

3.6 RMSE and its uncertainty (90% confidence) as a function of observation density at 30m resolution. 40

3.7 R^2 and its uncertainty (90% confidence) as a function of observation density at 30m resolution. 43

3.8 Location of ground-based (in situ), manual measurements from CLPX at each site. . 45

3.9 Comparison of R^2 from the CLPX sampling design and those from the random sampling designs used in the previous section. 47

4.1 Hierarchical k-means cluster analysis of high-resolution snow depth. Binary cluster trees are traced onto boxplots of the terminal nodes. Nodes are labeled by index in a complete binary tree. Histograms corresponding to the boxplots show the distribution with in each of the final nodes along with relative frequency of occurrence of snow depths. 62

4.2 Log-log semivariograms of high-resolution data at each site and its clusters. Numbers correspond to clusters in 4.1. No results shown for FS where spatial structure was considered superflous. The gap in the variograms at short range is due to calculations on a regular grid. 63

4.3 Snow depth clustering and cluster prediction at low resolution. Cluster prediction is performed for ‘‘best’’ predictors on LiDAR-measured and downloaded variable sets. 70

4.4 True R^2 comparison for binary regression tree estimates using random (pastel colors) and cluster predicted (primary colors) sampling schemes with LiDAR-measured and downloaded predictor sets. Cluster prediction is based on qualitative predictor selection. Cluster prediction was judged unnecessary at FS, thus its prediction skill is 0. 73

4.5 R^2 comparison of spatial modeling by cluster prediction (primary colors) and binary regression trees (pastel colors), for both LiDAR-measured (blue) and downloaded (red) predictor sets. Cluster prediction is based on qualitative predictor selection. Cluster prediction was judged unnecessary at FS, its prediction is skill is 0. 75

5.1 Location of Senator Beck Basin in Colorado and relative to the CLPX sites. 85

5.2 Overview of Senator Beck Basin, Ouray County, Colorado. This first order watershed is delineated in red. Meteorological stations and stream gage locations shown on map. Credit: Jeff Deems 86

5.3 Snow water equivalent time series at the upper Senator Beck study plot (SBSP) and the lower Swamp Angel study plot (SASP) in 2005 and 2006. 88

5.4 Senator Beck stream gage daily average stream flows in water years 2005 and 2006. 88

5.5 Daily temperatures with a 7 day running mean at SBSP and SASP for 2005 and 2006. 89

5.6 Senator Beck basin as divided 3 ways. Elevation Bands: elevation terciles (3 regions); Clusters: predictor clustering (5 regions); Elevation + Clusters: The intersection of the previous (11 regions). 94

5.7 The set of 1407 unique depletion curves sampled in the ADC and elevation+ADC parameter spaces. Snow-17 selects A_i to calculate the effective mean water equivalent, A_i is the smaller of a) the largest maximum water equivalent during the accumulation period or b) the free SI parameter which describes the water equivalent above which 100% snow cover always exists. 100

5.8 RMSE densities for the sampled parameter spaces of each model setup. The vertical black line indicates the skill of the null model when used to simulate runoff, this is a single number and not a distribution. 102

5.9 Nash-Sutcliffe model efficiency densities for the sampled parameter spaces of each model setup. The vertical black line indicates the skill of the null model when used to simulate runoff, this is a single number and not a distribution. 102

5.10 Example hydrographs shown for the best, median, and worst RMSE scores for each model setup. Note the different y-axis for the worst hydrographs. 104

5.11 RMSE density of the elevation+ADC model compared to the elevation and cluster models. 105

5.12 RMSE as a function of parameter strength when only individual parameters use in the cluster model. The skill of the point model is shown by the black line in both years. The individual predictors are elevation (ele), accumulated potential accumulated solar (sun), maximum upwind slope (sx), and vegetation height (veg). 107

5.13 RMSE density when individual parameters use in the cluster model. The individual predictors are elevation (ele), potential accumulated solar (sun), maximum upwind slope (sx), and vegetation height (veg). 107

5.14 RMSE as a function of parameter strength when any two parameters are used in the cluster model in 2005. The first variable in the y.x pair is plotted on the y-axis and the second on the x-axis. 108

5.15 RMSE as a function of parameter strength when any two parameters are used in the cluster model in 2006. The first variable in the y.x pair is plotted on the y-axis and the second on the x-axis. 108

5.16 Comparison of elevation band model and cluster model when only the elevation strength parameter is used. 110

5.17 Comparison of original parameterization to several other model situations. **No Routing:** the model is run assuming runoff from each spatially distributed model arrives immediately at the basin outlet. **Lapse Rate:** a lapse rate of $-8^oC/km$ is applied to distribution of model temperature threshold parameters. **Param Switch:** considers the effects of alternate parameter choices on model skill. Calibrated parameters in each year are traded for those in the other year. 112

A.1 Measured snow depth at each ISA weather station between date of manual, in-situ measurements and LiDAR observation. Manual measurements were taken on lowest DOY shown for each site and LiDAR on the highest. 125

Chapter 1

Introduction

Snow is a critical resource and environmental variable. It has been estimated that nearly 40% of the world's population derives water from snowmelt (Beniston, 2003). While certainly important at larger scales, spatial distribution of snow at scales of hundreds of meters can have significant effect on ground water recharge, surface water runoff and timing (Luce et al., 1998), surface energy balance (Segal et al., 1991), ecosystem response (Liston, 1999), and validation of space-borne observations (Derksen et al., 2008).

Because the spatial distribution of snow is difficult to observe at even the smallest hydrologic scales and because downstream communities often rely on vast tracts of unobserved, snow-covered landscapes, research in this thesis focuses on the fundamental problems observing and predicting the spatial distribution of snow accumulation and melt processes. The problem of observation is critical to advancing our understanding of snow. Theories need reliable observations to be tested. Due to difficulties associated with observation, significant questions about inference from sparse observations remain. A variety of these questions are posed here for the first time in the literature. Because an overwhelming majority of snowmelt dominated watersheds will remain unobserved, or effectively unobserved, prediction of spatial distribution of snow accumulation and ablation represents one avenue by which hydrologic forecasts may be improved. No prior research exists on predicting the specific spatial distributions of snow accumulation at the first-order basin scale. A simple hypothesis about the spatial distribution of snow depth provides a foundation for predicting spatial patterns of snow depth accumulation. Predicted patterns and their relationships to controlling variables

provide a spatial foundation for runoff simulation in a new, minimal-complexity extension of a 1-dimensional temperature index snow model. This approach is presented as a physically motivated alternative to the standard areal depletion curves used for extending point snow models in areal runoff simulation.

Chapter 2 presents the high-resolution, LiDAR-measured snow depth data used in chapters 3 and 4. This data set offers a unique view into snow depth distribution, providing snow depth at $1.5m$ resolution at six mountainous, $1.17km^2$ sites. Observations, taken in early April, 2003, represent peak-accumulation snow depth and its distribution, which are important during melt for both hydrologic and energy balance reasons (e.g. Liston, 1999).

The problem of estimating snow depth spatial distribution from sparse observations is considered in Chapter 3. We implement a sampling design which attempts to maximize the distance between otherwise random observations. Subsampling experiments are performed on the LiDAR data to investigate the how each of the following factors affect estimation of the spatial distribution:

(1) How many observations are needed for what purposes? Multiple model skills are considered:

   (a) Snow volume or mean depth estimation: Percent error.

   (b) Standard error: RMSE.

   (c) Spatial distribution: R^2.

(2) What is the effect of model resolution on model skill? Do predictors relate best to snow depth variability at a particular scale?

(3) How does predictor quality affect model skill? LiDAR-measured predictors are compared to readily-available, downloaded predictors at the later resolution.

(4) What is the difference between apparent (cross-validated) model skill and true model skill?

(5) What model performs the best considering each of these factors: binary regression trees, bagging trees, or multiple linear regression?

Chapter 3 is written as rough guide for those who might collect between 25 or more (3200 is the largest observation count considered) point-scale snow depth observations in a several square kilometer area for the purposes of estimating the total snow depth distribution. Results apply to manual probe measurements as well as LiDAR measurements.

In chapter 4, we shift focus to the problem of prediction. The high-resolution LiDAR data is used to develop a theory of spatial distribution of snow as governed by the spatial distribution of predictor variables. This simple hypothesis is verified and then applied to predict the spatial structure of snow depth distribution without observations using readily-available predictor variables downloaded from the internet. The following questions are posed:

(1) Is snow depth "structured"; does decomposition of spatial non-stationarity of mean snow depth relate to predictor variable controls ?

(2) Is decomposition of spatial non-stationarity in a multivariate set of predictor variables related to that of snow depth; are predicted patterns useful for spatial modeling of peak-accumulation snow depth?

Results indicate that the first-order structure of snow depth spatial distribution can be predicted without the need for observations, though predictor selection remains a subjective task. This chapter is pertinent to designing representative observational campaigns and spatial modeling of snow depth or snow water equivalent when less than a few hundred observations are available in a few square kilometer area. Spatial pattern prediction may also find application to statistical downscaling of snowfall or in data assimilation. The prediction methodology is used in chapter 5 to delineate hydrological response units (e.g. Leavesley et al., 2002) for snowmelt modeling.

Chapter 5 considers the problem of areal representation by the 1-dimensional, temperature-index snow model, snow-17, developed by the National Weather Service River Forecast Center for simulating snowmelt runoff. Operationally, areal snow accumulation and melt are described by an areal depletion curve (ADC). An ADC is a subgrid parameterization of a highly non-linear process which is decoupled from the physical considerations which actually govern it. We modify

the prediction methodology in the previous chapter to describe spatial distribution of both snow accumulation and melt processes. Predicted patterns are used as a basis for parameterization of a spatially distributed conceptual model of snowmelt runoff. Spatially distributed free parameters and model forcings are not introduced. Requiring a single parameter for each independent variable used to spatially distribute the model, this approach introduces fewer or equivalent parameters than an ADC.

ADC parameters are highly uncertain and have been shown to vary dramatically from year to year (e.g. Shamir and Georgakakos, 2007). Using an essentially random selection (uniform prior) of parameters for both approaches, we compare the ADC to our proposed parameterization by asking which approach is more likely to reproduce observed streamflow timing for simulations in two water years. The methodology we outline requires some up-front investment in spatial analysis but results in virtually guaranteed and significant improvements in streamflow simulations at lower parmeter cost than the simple ADC parameterization we use. The proposed methodology will be of interest to other researchers studying the problem of snowmelt runoff when limited observations are available and may be of practical interest to managers of small, upstream reservoirs.

Chapter 2

CLPX LiDAR Sites

2.1 CLPX overview

NASA's Cold Lands Processes eXperiment (CLPX) is responsible for all data used in chapters 3 and 4 of this thesis. Light Detection And Ranging (LiDAR) was flown in aircraft at 9 Intensive Study Areas (ISAs) (Miller, 2003). Of these we've selected the 6 which lie in more complex, mountainous terrain. Locations are shown in figures 2.1 and 2.2. Each ISA is roughly 1 square km ($(1080m)^2 = 1.17km^2$ as used in this study) and they are grouped by geographic proximity into 2 groups, F (Fraser) and R (Rabbit Ears) which are occupy the first position in each site's two letter code. The sites are named (and coded) Alpine (FA), Fool Creek (FF), St. Louis Creek (FS), Buffalo Pass (RB), Spring Creek (RS), and Walton Creek (RW). This study also uses measurements from automated weather stations located within each ISA as indicated on figure 2.3. Meteorological measurements (Elder and Goodbody, 2004) were gathered from the end of September, 2002, through September, 2003. In situ manual snow depth measurements were also gathered in March and April but significant ablation and accumulation during the time between these and LiDAR measurements renders meaningful comparison impossible. Change in snow depth at each ISA's weather station between in situ and LiDAR observations are shown in appendix figure A.1.

2.2 High-resolution LiDAR data

LiDAR measurements were flown in early April and late September, 2003. In both months, surface and vegetation top elevations were measured simultaneously by the instrument. (Data

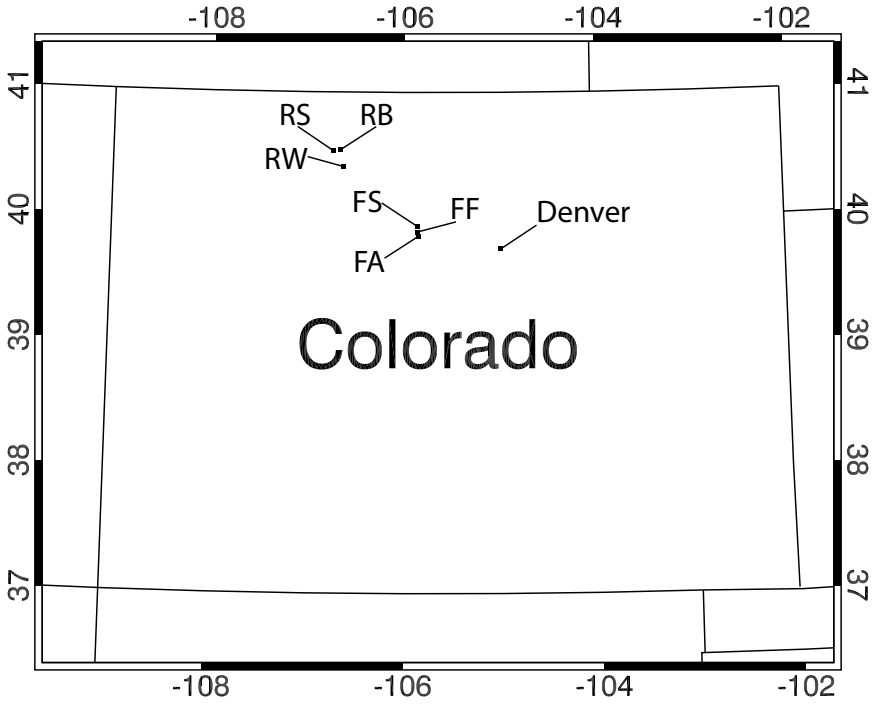


Figure 2.1: Location of sites in Colorado, USA. Latitude and longitude shown on figure border.

are supplied with proprietary filtering and classification applied, raw returns are not available.) Snow depth is calculated as the difference of surface elevations between the dates of observation. Dates were chosen so that this difference would represent peak-accumulation snow depth and its distribution, which are important during melt for both hydrologic and energy balance reasons (e.g. Liston, 1999). At high-elevation sites, like those in this study, snow accumulates from fall through winter and then ablates in spring. While there is certainly ablation during the accumulation phase and vice-versa, the two phases can be distinguished climatologically by the time at which snow depth (or SWE) usually reaches its maximum value. Early April is considered peak accumulation in the Colorado Rockies (Serreze et al., 1999) and the measurements used in this study were gathered on April 8 and 9, 2003. Bare earth (snow-free) observations were made in September when snow cover is at a minimum. Predictor variables of elevation and vegetation height are taken from the September observations.

Because LiDAR observations are not collocated for separate observation dates, data are transformed to a grid prior to calculating snow depth as the April-September surface elevation difference. The mode of spacing between surface elevations measurements was close to 1.5 m (average observation density $376,000/km^2$) and data were interpolated to a grid at this spatial resolution using locally-fit power-law variograms in ordinary Kriging estimation. This approach minimized cross-validated (leave one out) RMSE over all regridding approaches considered in this study. RMSE for all sites ranged from 6cm to 19cm, with an average 11.1cm over all sites which is comparable to the stated LiDAR error, of less than 30cm RMSE absolute error with 5cm vertical tolerance. Vegetation heights from LiDAR, which do not vary as smoothly in space at this resolution, are gridded using a nearest neighbor approach. See appendix A for details on interpolation to the regular grid. Snow depth, elevation, and vegetation height are the 3 primary spatial variables used in this study and this set on the 1.5 m grid is referred to as ‘‘high-resolution’’ observations.

Because there is no way to independently verify error of the LiDAR measurements from each flight, we briefly address this topic from our experience with the data. It generally appears that LiDAR error is random and uncorrelated. Because 2 LiDAR flights are used to obtain snow depth,

this error combines when calculating snow depth, some may cancel and some may amplify. The random error from their total seems to roughly approximate random variability of snow depth at meter scales. Of course, such random variability of snow depth varies by site and even across sites. We address the magnitude of this error more specifically momentarily.

There is evidence of systematic errors in LiDAR measurements which may depend on vegetation height, surface slope, vegetation type, and flight direction (H.P. Marshall, personal communication). Results presented in the following chapters often highlight dependence of snow depth on vegetation height. There is some concern that this dependence in the LiDAR data may in fact be due to systematic error. At the FF site however, very similar dependencies of snow depth on vegetation height at the FF site can be seen in both the results of Erxleben et al. (2002) and in the in situ CLPX snow depth observations (not presented here). We do not find any particularly strong evidence for this kind of systematic error and believe that observed variability is true. Due to simple geometrical considerations, it is quite possible that extra error is introduced in to LiDAR observation by increasing surface slope. Results in this work do not appear to be affected by this error in any way and it would be extremely difficult to asses if any extra variability on steep slopes is real or artificial. Systematic LiDAR error in the presence of certain kinds of brush has been reported in other locations where such brush is wide-spread (H.P. Marshall, personal communication). In this study, only at the RW site does this kind of brush exist and it is at this site we find unexplainable patterns in the snow depth field and unexpected difficulties estimating the observed snow depth distribution. We suspect that systematic LiDAR error in the presence of certain kinds of vegetation may be to blame but cannot be sure at this time.

At the site with the lowest suspected true variability in terrain, vegetation, and snow depth, there is evidence of systematic LiDAR errors dependent on flight direction. Because there is no appreciable true variability at this site they do not affect our study and are actually instructive to understanding the LiDAR error. In figure 2.3 one can see LiDAR scanning errors in the vegetation map at FS and some corresponding structure in the snow depth map. This is the only site where LiDAR errors rival, but do not dominate, variability in the vegetation and snow depth. While such

large, scan direction dependent errors are not obvious at any other sites, we can not be completely certain they dont exist. There may be some hints at the RS site where they do not appear to affect analysis at all. Magnitude of error is much less than overall variability at that site. Though we cannot infer the magnitude of the LiDAR errors at the other sites, we suspect them to generally have less variability than seen at FS because they are not visible. For snow depth, 28cm standard deviation is reported at FS. This implies an upper bound of roughly $20cm(= \sqrt{\frac{.28^2}{2}})$ standard deviation for each LiDAR scan as combined LiDAR error appears to roughly approximate snow depth variability. We believe errors for most sites are typically less than this, but cannot be sure. Modeling results in the following chapter and results not presented in this thesis on the amount of variability explained when parsing spatial non-stationarity of snow depth distributions in chapter 4 indicate that random variability does not dominate these observations any more than would reasonably be expected for snow depth at these sites.

Overall, LiDAR error does not appear to interfere with analysis in the following sections except perhaps at the RW site where it is not certain if poor results are attributable to LiDAR error or some other poorly understood mechanism such as wind. Problems modeling at this site are described in the following chapters, 3 and 4.

2.3 Low-resolution data

Because we wish to predict snow depth spatial structure using readily-available variables, we must also work at their spatial resolution. The same basic variables from the LiDAR, elevation and vegetation, are available for the entire contiguous United States from the US Geological Survey (USGS) at seamless.usgs.gov. We downloaded the National Elevation Dataset (NED) and the 2001 National LandCover Dataset (NLCD2001) for each ISA, both at 1 arc second resolution (approximately 30m) and reprojected them to a 30 m grid nested around the high-resolution grid. Landcover class is shown for each site in figure 2.4. Vegetation height is inferred from NLCD2001 landcover class (see appendix A for details.) Independent variables obtained in this way are referred to as ‘‘downloaded predictors.’’ A second set of predictor variables at this low-resolution are

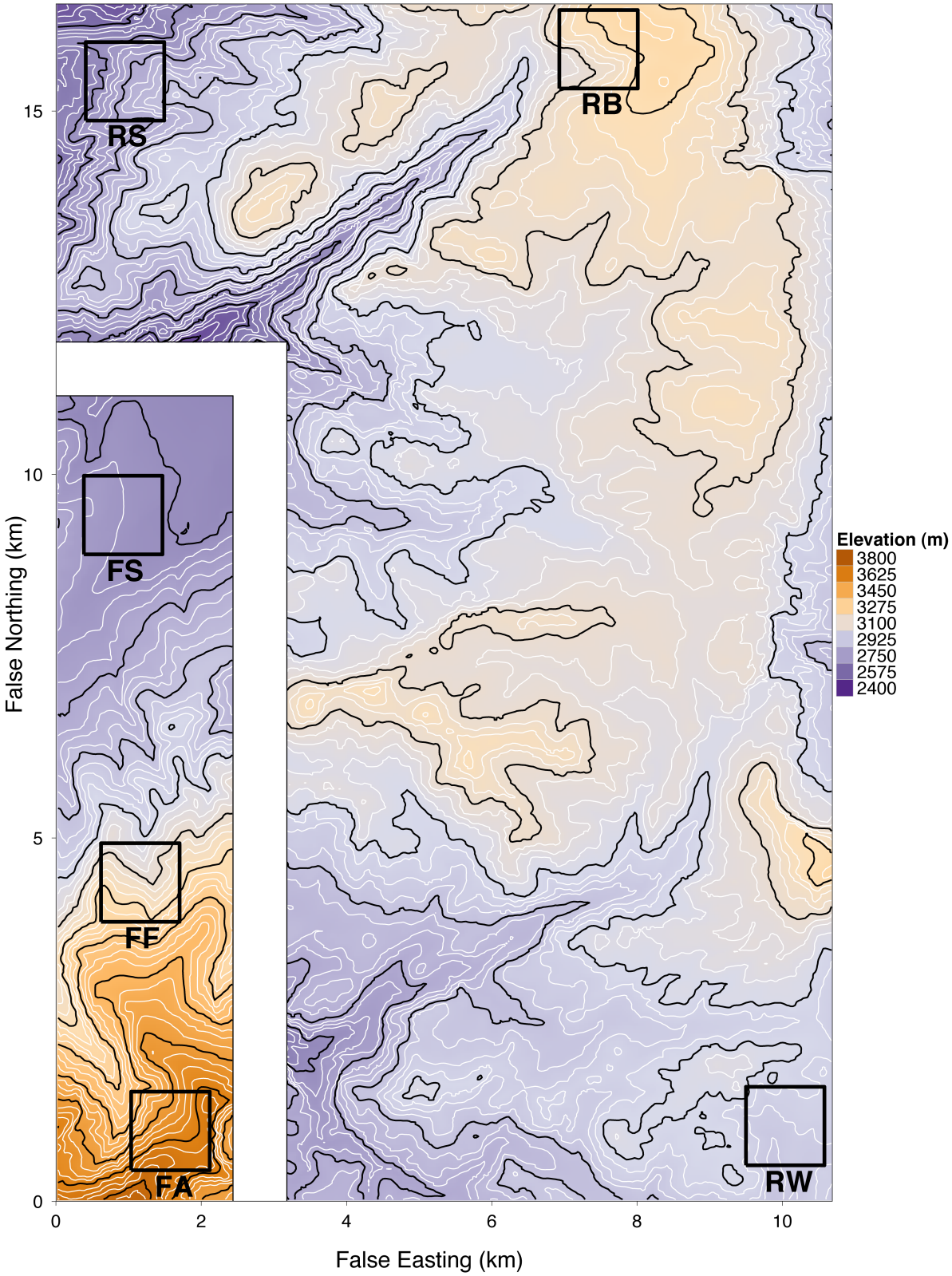


Figure 2.2: Overview of the separate areas containing the F and R sites. Contours 25m, bold contours 100m.

derived by averaging high-resolution predictor variables (elevation and vegetation height) over nested grid cells. Elevation and vegetation height averaged from high-resolution, LiDAR-measured observations result in more accurate prediction than the downloaded predictors and are termed "LiDAR-measured predictors." Averaging of high resolution predictors is also used in chapter 3 to obtain predictor variables at variety of other resolutions. In deriving the low-resolution data set, only elevation and vegetation height are averaged before calculating the derived predictor variables described in the next section. When many resolutions are considered in chatper 3, derived predictor variables at high resolution are averaged as described in that chapter.

To obtain corresponding snow depth at low resolution, we average snow depth over nested grid cells in the same way. The change of resolution is substantial (400:1) and averaging reduces variance (evident in the snow depth histograms in figure 2.6) though it preserves total snow volume.

2.4 Derived predictor variables

For each predictor set (high-resolution, downloaded, and LiDAR-measured), the following variables are derived from the primary predictors (elevation and vegetation height): vegetation density, aspect, slope, northness, integrated potential solar radiation, and maximum upwind slope. The following are also derived from the raster sum of elevation and vegetation: aspect, slope, northness, integrated potential solar radiation, and maximum upwind slope. This results in 13 total predictor variables. Though our study does not focus on which predictor variables are most significant, we found the set of variables derived from elevation+vegetation height to consistently improve estimates.

Vegetation density is the only variable not actually derived from the primary predictors within each predictor set. Vegetation density for LiDAR-measured predictors is calculated using windowed averaging of vegetation height at high resolution which assumes that vegetation density is proportional to average vegetation height . High-resolution density is then averaged to low resolution over nested grid cells. For the downloaded predictor set, NLCD2001 canopy fraction is used as vegetation density.

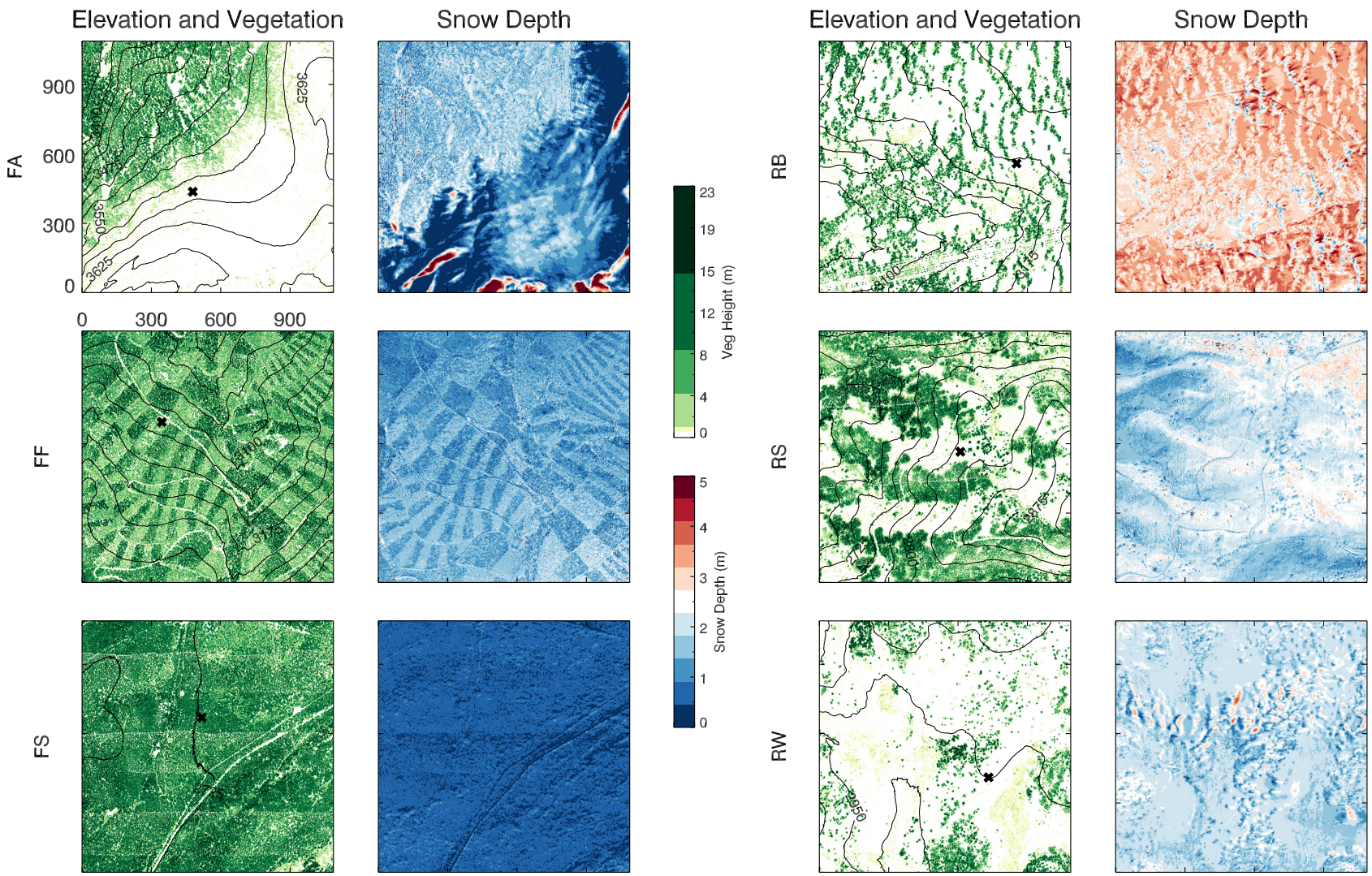


Figure 2.3: High-resolution elevation, vegetation, and snow depth at each site. Site dimensions shown on the first plot, all other sites are the same dimension. Contours shown at 25m. Location of weather stations indicated with a bold x in each. Location of each site can be seen in both figures 2.1 and 2.2, including surrounding terrain and distances between sites in the latter.

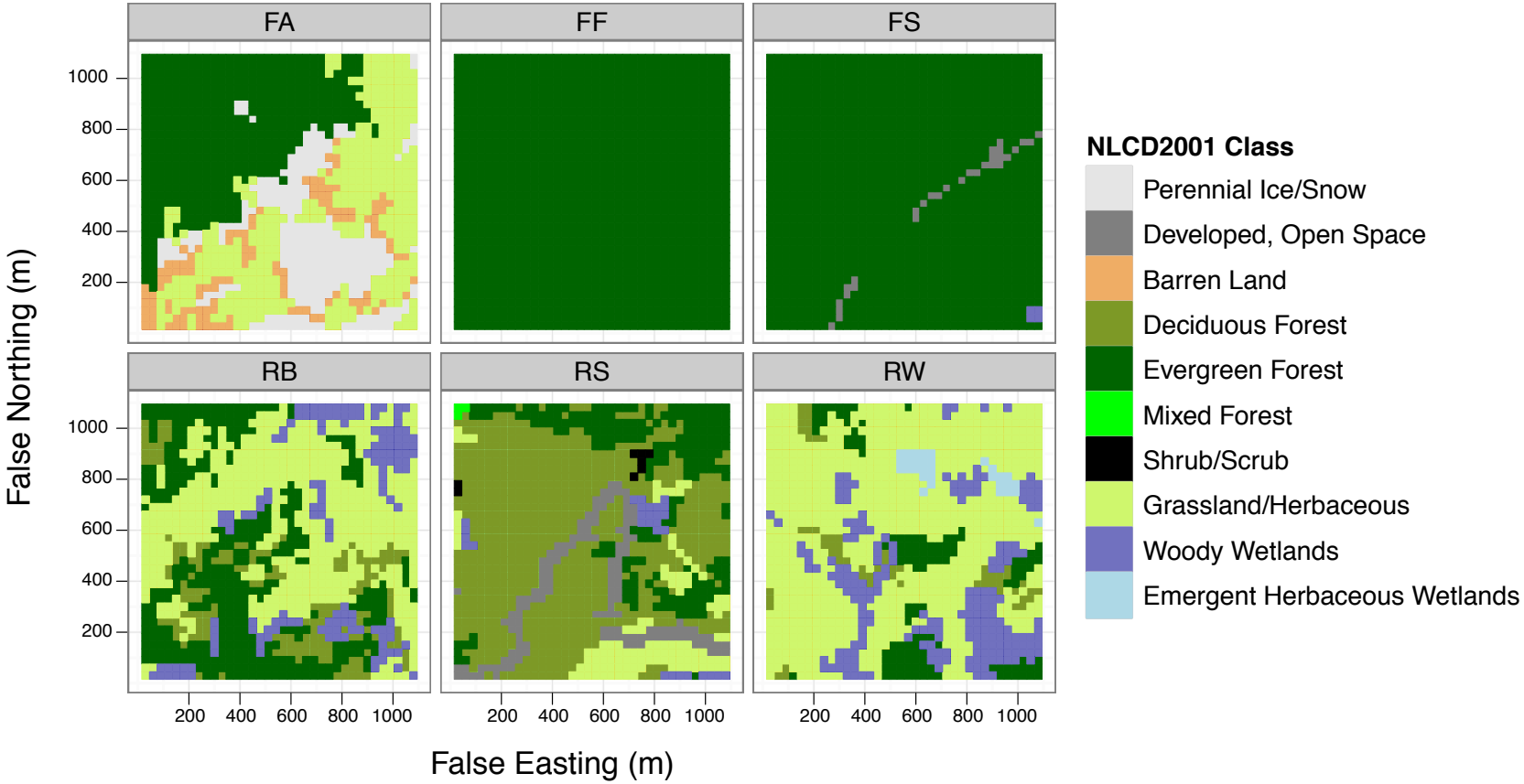


Figure 2.4: NLCD2001 landcover classification at each site. Developed areas at FS and RS are roads.

Aspect, slope, northness, and integrated potential solar radiation were calculated from the primary predictors (and also from their sum) in each predictor set using the GRASS GIS software (GRASS Development Team, 2010). Aspect describes the direction a slope faces. Given slope measured from horizontal and aspect measured counter-clockwise from east, northness is the product of $sin(slope) * sin(aspect)$. Steep and north-facing slopes have the greatest ''northness.''

When using aspect in chapter 4 in the context of cluster prediction, we transform calculated aspect via the sine function because aspect with domain $[0, 360)$ degrees (CCW from east) implies an unrealistic discontinuity in values at due east which is undesirable. This transform maps north to 1 and south to -1 with east and west in between. This is a realistic representation as the sites in this study lie at approximately $40^0 N$ where the north and south aspects are more dichotomous in terms of solar exposure than east and west. Though inappropriate if we want to consider aspect's relationship to wind, we leave summary of wind effects to the maximum upwind slope predictor.

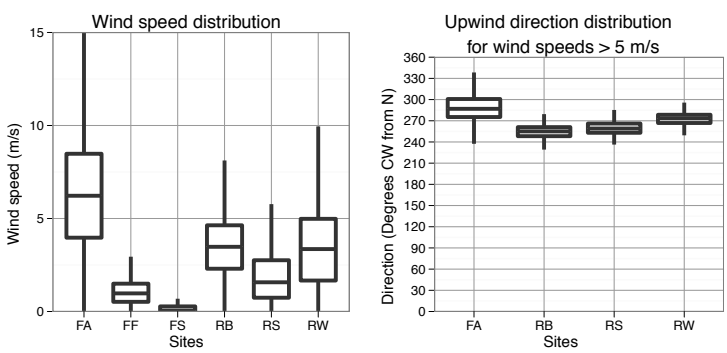


Figure 2.5: Ten meter wind speeds and directions measured at each site's weather station from November 1, 2002, through April 8, 2003. Directions summarized are for only wind speeds of at least $5ms^{-1}$.

Solar radiation is an important variable affecting peak accumulation snow depth at the high-elevation sites in this study. Potential solar radiation describes shortwave radiation effects at a given location resulting from slope, aspect, vegetation, and terrain shading while assuming clear-sky radiances. It was calculated in GRASS by the *r.sun* routine (based on TOPORAD of

Dozier (1980)) for each data set on a daily basis, from Dec 1st through March 31. Values were then summed at each pixel for all days during this approximate accumulation period to give our predictor, integrated potential solar radiation over the accumulation period. The algorithm calculates direct beam, diffuse, and reflected radiation accounting for day length, declination angle, incidence angle, and shadowing affects. Surface albedo used in the calculation was set to .7 for areas with less than a meter of vegetation height and to .2 elsewhere. These approximate values are used to incorporate vegetation effects on distribution of reflected solar radiation.

The maximum upwind slope predictor (Winstral et al., 2002) describes a location's exposure to prevailing wind. If negative, terrain slopes downhill away from the point in the upwind direction and there is greater exposure to wind. If positive, then terrain slopes uphill in the upwind direction, potentially sheltering the point from wind. Maximum upwind slope is calculated along an angle, A, in the upwind direction of the climatological wind vector and over some upwind search distance, $dmax$, from the point of interest, (x_i, y_i), as formulated by Winstral et al. (2002)

$$Sx_{A,dmax}(x_i, y_i) = max\left[\arctan\left(\frac{ELEV(x_v, y_v) - ELEV(x_i, y_i)}{\sqrt{(x_v - x_i)^2 - (y_v - y_i)^2}}\right)\right]$$

where (x_v, y_v) is the set of search points determined by raster resolution and choice of A and $dmax$. Based on wind direction histograms (e.g. figure 2.5) at a location, the angle A is actually supplanted by multiple angles in ± 5 degree increments from the climatological wind vector and the maximum upwind slope along each of these angles is combined by weighted average based on the distribution of wind direction during the accumulation period. The upwind search distance was approximated by half the period of the primary terrain undulations at each site. Software for computation was supplied by Adam Winstral.

2.5 Overview of sites, F and R groups

We provide a quick, general picture of the 6 sites by considering similarities and differences in their *mean* snow depths and factors controlling these. Snow depth histograms in figure 2.6 reveal large differences in mean snow depth between the F and R groups. Snow depth at the R sites

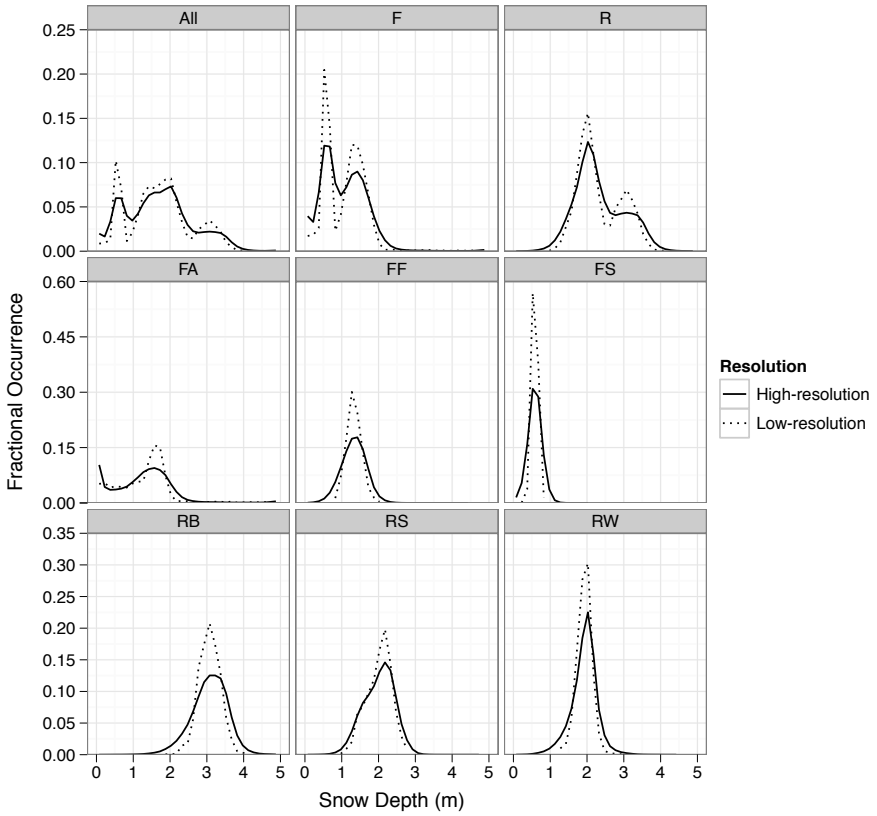


Figure 2.6: Snow depth histograms at all sites, the F sites, the R sites, and all individual sites. Solid line is high-resolution and dotted is low-resolution snow depth. Bin size is 15cm.

Variable	**Site**					
	FA	FF	FS	RB	RS	RW
Mean Snow Depth (m)	1.3	1.3	.6	3.1	2.0	1.9
Std. Dev. Snow Depth (m)	.82	.37	.28	.49	.45	.35
Mean Elevation (m)	3567	3145	2725	3160	2780	2951
Elevation Range (m)	335	287	59	183	247	94
Mean Wind Speed (m/s)	5.5	1.1	.2	3.1	1.9	3.0
Mean Temperature ($^{\circ}C$)	-9.8	-7.6	-5.9	-8.6	-5.6	-7.0
Degree Day Fraction ($^{\circ}C$)	6.5	41.1	69.8	11.8	44.0	30.1
Mean Incoming Shortwave (Wm^{-2})	148	102	74	114	114	145
Mean Potential Accumulated Solar ($Wh/m^2/day$)	2138	1832	2809	3036	2326	2994
Mean Potent. Accum. Solar w/ Veg ($Wh/m^2/day$)	1784	1067	1285	2265	1537	2617

Table 2.1: Observed and derived statistics of snow depth and physiographic variables at each site.

averages roughly a meter more snow than sites in the F group due to differences in local climatology between the two groups. We assume within-group variability can be explained by physiography and that within-group differences in total precipitation from other causes are negligible over the accumulation period.

A standard assumption is that snow depth increases with elevation due to both temperature and precipitation effects. Table 2.1 reveals that mean site elevation does not predict within-group mean snow depth. To properly understand variation of mean snow depth between sites, tree line must be accounted for in the F group and solar exposure must be considered in the R group. Alphabetical order of the F sites (FA, FF, FS) corresponds to decreasing mean site elevation. Below tree line, the F sites have similar vegetation cover and aspect and mean snow depth increases with elevation until tree line, above which wind removes significant amounts of snow from the FA site. This greatly impacts the mean snow depth at FA. In the trees, we see the greatest snow depth of all F sites but above tree line we find the lowest snow depths of all F sites. The result is that FA has mean snow depth equal to that of FF which lies at approximately 400 m lower elevation.

The R sites decrease in elevation from RB to RW to RS. However, RS has higher mean snow depth than RW even though it lies almost 200 m below RW. The RW site faces towards mostly southward over its entire extent while there are significant terrain undulations at RS which result many areas of different aspect including significant areas of north-facing terrain. These factors affect the integrated potential solar radiation incident at the surface. The sites' difference in accumulated potential solar radiation (with and without vegetation shading, table 2.1) illustrates the importance of solar exposure in comparing these locations.

Differences in mean snow depth between sites in this study illustrate the role of climatology (between groups), elevation (tree line), wind (above tree line scour), vegetation (tree line and solar shading), and aspect (solar exposure) in shaping snow cover distribution. Local controls on snow depth upset simplistic assumptions of how snow is distributed at larger scales.

Chapter 3

Estimation from Sparse Observations

3.1 Introduction

Estimation from sparse observations is the classic snow depth problem. Typically, several hundred observations are used to estimate spatial distribution of snow depth on a regular grid of a few square kilometers and estimation skill is obtained via cross-validation (e.g. Molotch and Bales, 2005; Erickson et al., 2005; Winstral et al., 2002; Balk and Elder, 2000). Numerous studies have examined this problem at our scale of interest but these have been limited to relatively small and sparse data sets, especially when compared to the CLPX LiDAR data. Because results have varied widely from study to study, significant questions linger about factors affecting snow depth estimation. In conclusion of their study using manual measurements (taken prior to CLPX) at the sites FF, FS, and RW, Erxleben et al. (2002, p 3627) ''recommended that the differences in the characteristics of the study sites, the characteristics of the sampling strategy, and the characteristics and resolution of the independent variables need to be explored further in order to evaluate the causes for the results presented.'' Here, we take advantage of the wealth of LiDAR observations to explore these and related topics using data subsampling experiments.

This study examines several broad questions in estimation from sparse observations which have not been systematically explored previously in the snow depth literature:

(1) How many observations are needed and for what purpose?

(2) How does the choice of model resolution affect model skill?

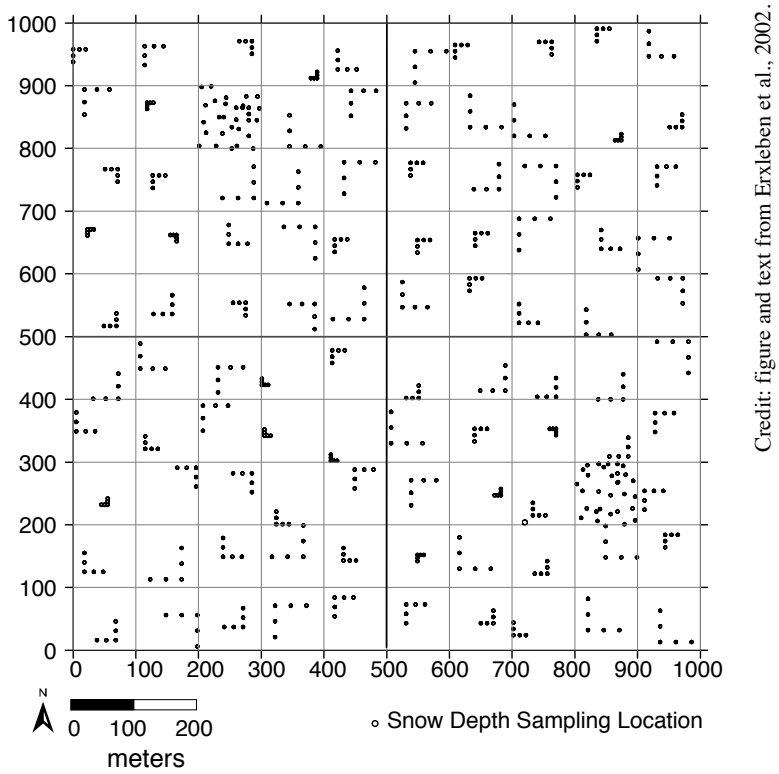
(3) How does predictor variable quality affect model skill?

(4) How do cross-validated and true estimation skills compare?

We also compare the standard snow depth estimation model, binary regression trees, against its more modern variant, bagging trees, and a multiple linear regression model.

In investigating these basic questions, our study aims to provide a reference point for understanding how these factors are likely to affect future sampling campaigns, a rough guide to making practical sampling decisions which satisfy estimation accuracy requirements. Our experiments assume point-scale sampling. Given the expense of LiDAR measurements and their shortcomings in certain environments, manual probing is likely to remain the standard method of snow depth measurement for some time. But because LiDAR observations are similar in support to manual measurements and because they may be randomly distributed as well, this study may be of interest for future airborne LiDAR measurement campaigns. Results presented here may illuminate trade-offs between LiDAR observation density and extent and inform selection of flight altitude to balance these. Though an initial, dense snow-free measurement would be required, subsequent snow-covered measurements could sample less densely without sacrificing estimation accuracy. For example, new observations of snow cover acquired at the CLPX sites could be one or two orders of magnitude less dense.

In this chapter we focus on random sampling. Stratified sampling recognizes that variations in snow depth and SWE are controlled by collocated physical variables such as elevation, vegetation, and aspect, and spatially distributes samples in relation to these. Though studies indicate that stratified sampling is more efficient than random sampling (e.g. Watson et al., 2006), especially at large scales, snow surveys at the kilometer scale typically implement sampling strategies which ignore such considerations. Most commonly, sampling is organized by some regular geometric pattern which disregards controlling variables as illustrated in figure 3.1 which are the same designs used for in situ snow depth sampling during CLPX. Other examples can be found in, Molotch and Bales (2005) and Balk and Elder (2000). Here, observations are selected essentially at random



Sampling strategy for snow depth. Five snow depth sampling sites are located within each 100 m x 100 m cell. In addition, 25 random samples were added to two of the cells. There are a total of 550 snow depth sampling locations.

Figure 3.1: Snow depth sampling design from Erxleben et al. (2002).

though in a way which attempts to maximize coverage of the domain of interest. This approach is different than that shown in figure 3.1 and the two are compared at the end of the study to understand their different implications for cross-validated and true statistics. In the following chapter (4) on snow depth structure, stratified sampling is touched upon.

The number of observations or sample size has a primary effect on estimation accuracy and serves as a comparison against which other effects on model skill can be measured. (For example, by just changing model resolution can we obtain the same model skill as when doubling the amount of observations?) Though we report the number of observations used in each 1.17 sq km site, the term ‘‘observation density’’ is used for convenience while only reporting the number of observations. The term density is not meant to imply that estimation at larger extents requires the same density for equivalent estimation accuracy. Rather, results depend on the scale of the domain and the surface being estimated. This study is limited by the extent of available data but will still provide insight into the estimation problem at similar scales. Results are case studies intended to provide insights into how a variety of factors might be expected to affect estimation from sparse observations.

3.2 Background

Many approaches have been pursued to estimate snow depth and SWE spatial distribution from sparse observations. We focus on snow depth in this study. Though approaches to these variables are similar, it can be difficult to compare results because SWE observations are usually an order of magnitude more sparse than in situ snow depth observations (e.g. Erxleben et al., 2002). The disparity is much greater in this study, where we might have 10 SWE observations at a given ($1km^2$) site, LiDAR provides over 10^5 snow depth observations.

The binary regression tree model (BRT, Breiman et al., 1984), first applied to snow depth interpolation by Elder et al. (1991), has become the most popular model for snow depth estimation at kilometer scales. Applications of BRT to snow depth interpolation and their pertinent details are listed in table 3.1. BRT is often followed by residual kriging (Balk and Elder, 2000). Though this virtually always improves cross-validated skill, improvements have often been small (Erxleben

et al., 2002; Molotch et al., 2005).

Intuitively, BRT is a more appropriate model than a spatial interpolator such as inverse distance weighting or various kinds of kriging, which vary smoothly between observations and disregard the relationship of snow depth to controlling variables which commonly cause it to vary abruptly in space. BRT is described in detail in section B.2.1 of the appendix. Though only Erxleben et al. (2002) has demonstrated that BRTs outperform such interpolators for snow depth (and Balk and Elder (2000) with Balk and Elder (1998) provide some evidence as well), we do not investigate smooth interpolators here. Instead, we forge ahead with tree methods and investigate the bagging tree (Breiman, 1996) innovation of the BRT model, detailed in section B.2.2 of the appendix. Bagging is short for ''boot strap aggregating.'' The model averages multiple fits over much smaller subsamples of the data than it would use in regular cross-validation to obtain more robust estimates. A second variant of BRT, random forests (Breiman, 2001), was also investigated though we only offer occasional comments on its performance relative to that of bagging trees as the two yielded similar results.

Carroll and Cressie (1996) and Fassnacht et al. (2003) employed linear regression in combination with geostatistical techniques to estimate SWE from sparse observations at scales of tens of kilometers, using 1km model resolution. Hosang and Dettwiler (1991) used linear regression with residual kriging to estimate distribution of SWE in the 2.1 km^2 Schwangbach basin, from 33 observations. They used elevation, direct solar, and existence of forest as predictor variables. We are not aware of any study using multiple linear regression to estimate snow depth at the scales of our study. Any such model must include non-linear cross terms accounting for interaction of variables in order to be competitive with BRT and bagging trees which model such relationships via their hierarchical structure. Such linear regression has the benefits/drawbacks of being able to estimate beyond the range of observed data and to model spatial trends, unlike tree methods.

Study	**Site**	**Year**	R^2	**Area** (km^2)	**# of Obs**	**Predictor Source**	**Resolution** (m)	**Notes**
Balk and Elder (2000)	Loch Vale	1997 1998	.54 .6-.75 .65 .7-.85	6.9	197 173	USGS 7.5min	10	BRT. With residual co-kriging. BRT. With residual co-kriging.
Winstral et al. (2002)	Green Lakes Valley	1999	.5 .27-.42	2.25	504	USGS 7.5min+survey	30,10	May results w/ max upwind slope. Without max upwind slope.
Erxleben et al. (2002)	St. Louis Creek Fool Creek Walton Creek	2001	.18 .31 .29	1	550	USGS 7.5min	30	BRT best. res. cokrig=BRT R^2+.01 res. cokrig=BRT R^2+.03
Molotch et al. (2005)	Tokopah	1997	.34 .37	19.1	328	USGS 7.5min, NED, SRTM	30, 30, 27	April, 7.5min DEM, BRT. With residual interpolations.
Molotch and Bales (2005)	Slumgullion Upper San Juan Wolf Creek Sum. Lily Pond	2001 2002 2001 2002 2001 2002 2001 2002	.62 .42 .57 .66 .64 .60 .47 .50	1,4,16	2001: 140,184,184 2002: 140,159,159	USGS 7.5min	30	April results. Obs correspond to area.
This study	Fraser Alpine Fool Creek St. Louis Creek Rabbit Ears Spring Creek Walton Creek	2003	...	1	25, 50 100, 200, 400, 800, 1600, 3200	LiDAR-Measured, NED+NLCD2001	1.5, 2.1, 3, 4.2, 6, 8.4, 11.9, 16.9, 24, 30, 33.8, 47, 67.5, 98.2, 135, 180, 270, 360, 540	$\frac{1}{2}$+ million observations per site used in subsampling experiments. Cross-validated compared to "true" skill statistics.

Table 3.1: Summary of results from previous studies of snow depth estimation using binary regression trees (BRT) at kilometer scales ($< 20km^2$). Predictor sources listed are United States Geological Survey (USGS) 7.5 minute maps, National Elevation Data set (NED), Shuttle Radar Topographic Mission (SRTM), and the 2001 National LandCover Data set (NLCD2001).

Of our four broad questions, only effect of predictor quality has been touched upon by Molotch et al. (2005) who investigated impacts of 3 different digital elevation maps (DEM) on BRT model skill. Their study resulted in selection of a best overall DEM, however choice of DEM resulted in relatively small differences in estimation skill compared to those seen in our study. We use LiDAR-measured predictors (DEM and vegetation) as a high-quality reference set against which we judge the quality of predictors derived from the National Elevation Data (NED) and National Land Cover Data (NLCD2001) sets. Our investigation thus sheds light on the value of greatly improved predictor sets and the associated loss of model skill when readily available predictor sets are used.

A topic we do not address is the formulation of new and better predictor variables for snow depth estimation. This was done by Winstral et al. (2002), who developed a clever independent variable which increased BRT R^2 by .08-.23. We do not formulate new predictor variables but make a comprehensive survey of predictors previously used.

Our study takes novel perspectives on estimation from sparse observations. No study has systematically looked at the effects of observation density or the effects of random observation selection on model skill. No study has considered the impacts of model resolution on model skill and no study has been able to compare cross-validated model skill to skill measured against the ‘‘true’’ snow depth distribution.

3.3 Experimental design and methodology

We consider three measures of estimation skill with the following implications:

(1) Percent error of mean depth or total snow volume (PE), how accurately we might estimate total amount of water in a given area:

$$PE = \frac{1}{n * \overline{d}} \sum_{i=1}^{n} \left[\hat{d}(x_i) - d(x_i) \right] = \frac{1}{V} \sum_{i=1}^{n} a * \left[\hat{d}(x_i) - d(x_i) \right] = \frac{\hat{V} - V}{V}$$

$\hat{d}$ is the cross-validated estimate of the observed depth d, $\overline{d}$ is the observed mean snow depth, V the total snow volume based on observed mean snow depth, $\hat{V}$ the estimated snow volume, and a the area of each grid cell, such that over all grid cells $na/V = 1/\overline{d}$. Estimating the mean depth or

volume is an important step in estimating the total volume of water in a given area. However, the density of snow must also be estimated to make the calculation. We only work with snow depth and thus only look at percent error in this part of the calculation which has a first-order effect on the estimation of water volume.

(2) Root mean squared error (RMSE) gives standard estimation error:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\hat{d}(x_i) - d(x_i)\right]^2}.$$

(3) R-squared, R^2, judges how accurately we estimate the spatial distribution of snow:

$$R^2 = G = 1 - \left(\sum_{i=1}^{n}\left[d(x_i) - \hat{d}(x_i)\right]^2 / \sum_{i=1}^{n}\left[d(x_i) - \overline{d}\right]^2\right)$$

This measure has also been called the G-value (Erxleben et al., 2002) or the Nash-Sutcliffe model efficiency. It makes useful comparison of the model being evaluated against the mean value of data used in fitting the model, $\overline{d}$, the spatial null model. The measure describes the fraction of squared residuals left by the null model which are explained by the model. A value of zero indicates that the model is an equal estimate to the mean value and a value of 1 indicates that the model fits the data perfectly, explaining all residuals left by the null model. Negative values, allowed in this formulation, indicate that the model introduces greater error than the spatial mean and is thus unwarranted.

This study is broken into two sections which consider separately how (1) model resolution and (2) predictor quality affect these skill measures. In both sections we examine:

(1) What factors control the number of observations required to:

(a) optimize each statistic and

(b) minimize uncertainty in each statistic resulting from random selection?

(2) Performance of BRT, bagging, and linear regression models.

(3) True versus apparent, or cross-validated, model skill.

To study the effects of model resolution on estimation, the high-resolution LiDAR data is averaged from its highest resolution, 1.5m, to lower resolutions. Only aspect is transformed to coarser resolution by using the statistical mode because of problems with discontinuity at zero and 360. True snow depth at each resolution is considered to be the corresponding spatial average which preserves true snow volume, though it greatly reduces snow depth variability as resolution increases as will be seen in the results. The algorithm used for changing resolution is sketched in appendix B. In the section on the effects of model resolution, we pose the questions:

For a given set of observations, can one:

(1) optimize the statistic by changing model resolution?

(2) minimize uncertainty in the statistic resulting from random sampling by changing model resolution?

Next, we compare LiDAR-measured predictors to readily-available predictors downloaded from the internet. Estimates based on the National Elevation Data (NED) and National LandCover Data (2001) sets are compared to estimates from LiDAR-measured predictors sets at the 30m resolution of the downloaded sets. We ask: *How much skill is lost by using downloaded predictors versus LiDAR-measured predictors?*

In both sections, the same basic algorithm for estimating from sparse observations is used. As outlined in appendix B, our procedure samples the high-resolution data at a given observation density, collocates observations with predictors at the model resolution (averaging observations within the same model grid cell), fits each model, and then estimates. Estimates of the observations used in fitting are made via 10-fold cross-validation to give the ‘‘apparent’’ skill. Estimates are made of the snow distribution over the entire domain to give the ‘‘true’’ skill. Uncertainty resulting from random observation selection is simulated via Monte-Carlo sampling, this process is repeated using 100 sets of observations at each observation density.

When fitting models and comparing their estimates, each is provided an equal amount of data and fit in a 10-fold manner. Though each model has a particular approach to fitting, all models

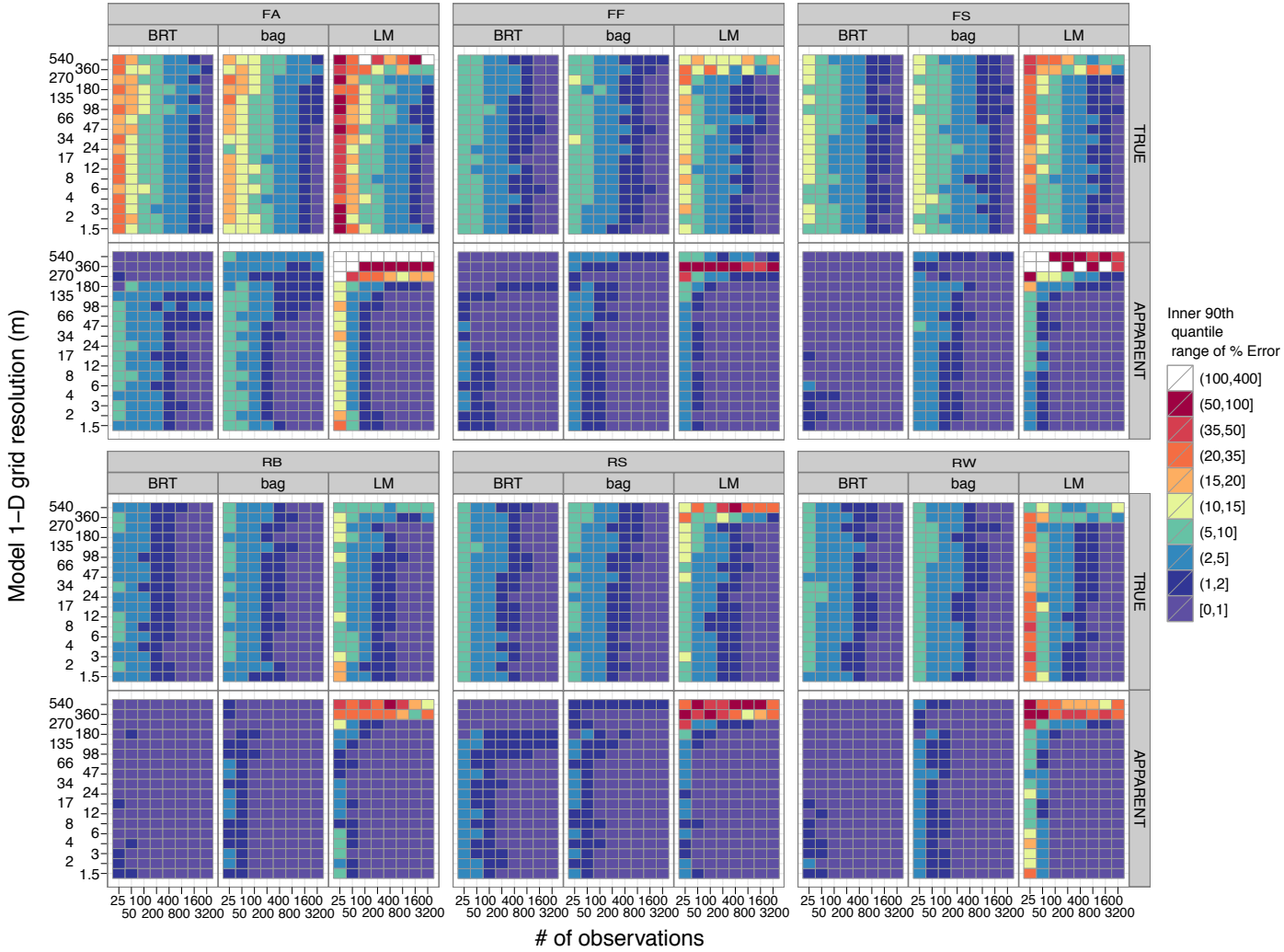


Figure 3.2: Uncertainty (90% confidence) in percent error resulting from random selection of observations as a function of observation count and model resolution. Median percent error is not shown because it is essentially zero at all but the lowest observation densities. Median precent error is shown for 30m resolution in figure 3.5.

are fit by partitioning the observations into 10 parts and predicting each from the others (10-fold). The three models, binary regression trees, bagging trees, and linear regression, and their fitting are discussed in appendix B.

3.4 Skill as a function of observation density and model resolution

3.4.1 Percent error

For a given set of observations, can one minimize percent error by changing resolution?

Models in this study are regression models which formulate a conditional expectation of snow depth for any given set of predictors based on their algorithm and some set of training data. Each model aims for unbiased estimation relative to the training data. Apparent skill is unbiased by design. Minimization of true percent error only requires a large enough data set to attain representative conditional mean values. Neither depends on model resolution because high resolution data are always averaged to formulate the conditional mean values. Figure 3.2 does not show median percent error, only its uncertainty due to random sampling, because all three models used in this study are unbiased at all resolutions for all but the lowest observation densities. There's very little to show in a plot. When comparing predictor sets in the next section, we will see that the number of observations needed for an unbiased estimate depends on the individual site and that the LM model is particularly biased at low observations densities though its apparent PE does not reflect this.

For a given set of observations, can one minimize uncertainty in percent error by changing resolution?

Figure 3.2 displays the range of the inner 90th quantile of percent error for each set of 100 random observations. Because median PE has very low bias, plus-minus half this range is thus the 90% confidence interval, or uncertainty, for the percent error due to random selection of observations which naturally decreases with increasing observation density. Studies with limited observations will not be able to estimate apparent uncertainty in PE except, perhaps, at the lowest observation

densities, while LiDAR surveys (with densities lower than our ‘‘truth’’ data set) may be able to estimate apparent uncertainty at higher observation densities using subsampling experiments.

Uncertainty in volume estimation does not depend on model resolution, it cannot be improved by changing model resolution, it can only be improved by choice of model or by adding observations. Though the LM model appears to depend on resolution, with uncertainty exploding at the largest resolutions (180, 270, 360, and 540m), this is actually a result of poor model fit from using very few data points. These resolutions have 36, 16, 9, and 4 grid cells in our domain, respectively. Fitting with 10-fold cross-validation removes at least 10% of these before model fitting. Poor fits at these resolutions are obvious for LM and closer inspection shows similar, though slight, increases in PE uncertainty for BRT and bagging models as well in this region. Results indicate that roughly at least 64 points (135m resolution) are needed for properly estimating uncertainty in PE instead of looking at error due to improper model fit.

Given each of these regression models fits a conditional mean response and that observations used to fit the models are not affected by choice of resolution, it is no surprise that PE uncertainty is independent of model resolution. Though resolution may affect the relationship between predictor and response, and thus where the model puts more or less snow, it has limited affect on the average response estimated. In particular, tree models are confined to forming a sort of weighted average of the values used in fitting. By contrast, the LM model can extrapolate outside the training set if new predictors fall outside the range of predictors used in training the model, which explains the extreme range or uncertainty of true PE for LM models at lower observation densities (25 and 50 per $1.17km^2$). Only at low observation densities at FA, where there's tremendous variability in snow depth, does true uncertainty of BRT or bagging compare to that of LM. LM should not be used with less than about 100 observations if one cares to minimize uncertainty in estimated PE.

Uncertainty in true percent error at the F sites is typically greater than at the R sites because similar estimation errors account for a smaller percent of the deeper snow depths at the R sites. There is also more variation in uncertainty in PE at F sites resulting from diferences in mean snow depth and large variability at FA. Ignoring the lowest observation densities and coarsest resolutions

where LM is particularly unreliable, the three models give very similar uncertainties in PE at each site. Overall, it is remarkable how few observations are needed for reliable estimation of mean snow depth or total volume at each site. One hundred observations typically result in an estimate within 5% of the true volume.

For all models at low and medium observation densities, apparent uncertainty in percent error is significantly smaller than true uncertainty even though we used 10-fold cross validation. Because apparent uncertainty is better (in the case of PE, lower) than true, we say uncertainty in PE is optimistic. For example, if could estimate your apparent percent error for 200 observations at FF, you would find your estimates within $\pm.5\%$ though the true error is bounded by $\pm 2.5\%$ (with 90% confidence). Though all models give optimistic uncertainty in PE, BRT and LM are more optimistic than bagging because bagging reduces the training data by bootstrapping. Only at FA is bagging more optimistic than BRT (at 200 and 400 observations). We inspect the problem of optimistic uncertainty in PE more closely when comparing predictor sets.

3.4.2 RMSE

For a given set of observations, can one minimize RMSE by changing resolution?

In figure 3.3 colored tiles represent median RMSE (cm) of simulations at each combination of observation density and resolution. Lines represent the 90% confidence range or uncertainty for RMSE about the median value which results from random selection of observations at each combination of observation density and resolution. The uncertainty may not be symmetric about the median. Both true and apparent median RMSE depend on resolution. True RMSE to reduces with resolution by experimental design because of reduced variability in true snow depth with increasing resolution. Apparent RMSE also depends on resolution, by experimental design. Multiple observations within any grid cell are averaged prior to fitting by the model which effectively reduces observed variance and decreases apparent RMSE as the number of observations grows.

All models reduce RMSE with resolution, though for BRT and bagging we see true RMSE start increasing again at approximately 270m resolution at all sites but FS and RW (those with the

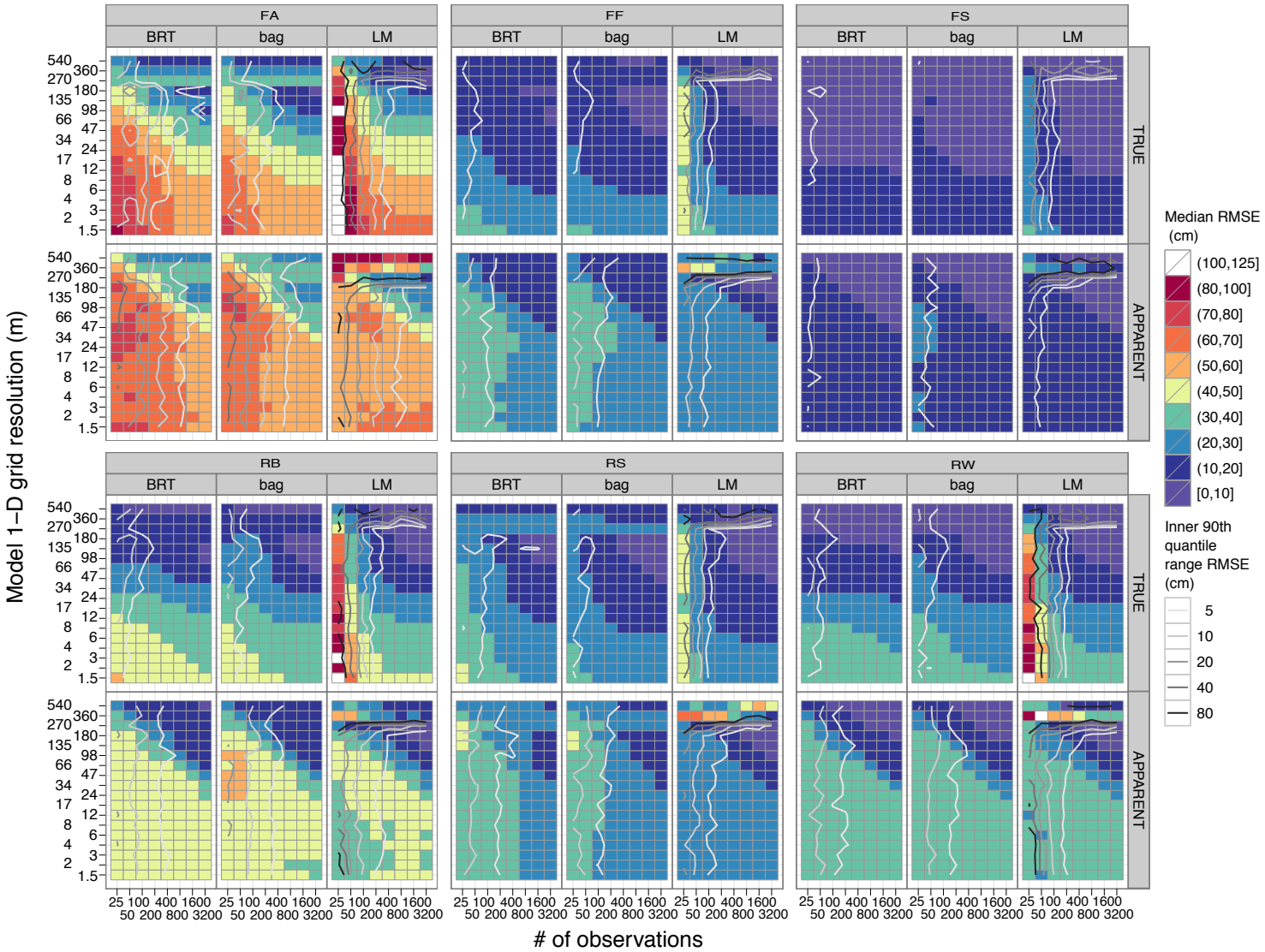


Figure 3.3: Median RMSE and its uncertainty (90% confidence) due to random selection of observations as function of observation count and resolution.

lowest variances). Again, this is caused by models being fit to very few observations. In this region, BRT and bagging RMSE uncertainty (show by lines) drop relatively quickly after a small peak at slightly finer resolutions which indicates that models based on different observations are converging as observations are being averaged to very few training points for the model. True RMSE increases because model complexity diminishes faster than variance because of so few training points. LM model RMSE in this region is extreme, reflecting its behavior resulting from poor fit. (Interestingly, median true RMSE for LM is much more stable in this region though its uncertainty is much greater.) The lesson from all of this is that the minimum number of observations required to fit these models to estimate RMSE occurs near 180m resolution (36 grid cells).

Though RMSE depends on resolution, there is no resolution which particularly reduces RMSE. Improvements in RMSE reflect more on lenient evaluation criteria (less variable snow depth with resolution) rather than on improved model performance. Thus RMSE may be (artificially) minimized for a fixed number of observations by maximizing model resolution. Maximum model resolution may be determined by the minimum number of observations needed to properly fit the model or by other, external considerations, such as dominant scales of variability at a study site at which error is most relevant to the subject of interest.

True RMSE error also depends strongly on the particular distribution being estimated and is not clearly related to variance. For a given model resolution, bagging most often results in the lowest true RMSE at all sites. Also note that, for a given number of observations, true RMSE decreases more quickly than apparent RMSE as resolution is increased. Thus RMSE is pessimistic by experimental design, which is desirable; apparent (median) RMSE provides an approximate upper bound or worst case scenario of true RMSE. If uncertainty in RMSE can be minimized, this bound is more certain. There are several exceptions to the general pessimism of RMSE near the edges of the plots, mostly for LM at low observation densities and for tree models at the finest resolutions at FA, FF, and RB. Pessimism of the statistic is largely due to choice of sampling design, as will be discussed in a later section.

For a given set of observations, can one minimize RMSE ***uncertainty*** *by changing resolution?*

Figure 3.3 shows that both true and apparent uncertainty in RMSE are generally independent of resolution. As observed above, uncertainty increases at resolutions greater than about 100m because very few data points are used in model fitting. For BRT and bagging, RMSE uncertainty peaks around 135 observations and then reduces. The peak is somewhat larger for bagging because it effectively reduces its fitting sample by bootstrapping. For tree models, very few observations cause model divergence over the set of random samples up to some point after which point models converge again. For LM, models simply diverge with fewer fitting potins and uncertainty increases dramatically after 270m resolution. These effects do not result from changing resolution as much as model fitting. From the figures we also note that these effects are greatest at sites with large snow depth variability (FA, RB, RS), as we would expect. (For this same reason, peaks in apparent uncertainty of RMSE move towards smaller resolutions as observation density increases at FA and RB because larger observations sets have greater variability.) Though changing resolution does not appear to affect uncertainty in RMSE, we do get an estimate of how many observations are needed to properly apply BRT and bagging models. At 135m resolution models are fit to no more than 58 =(.9*64) grid points. Larger observation variance will require more fitting points.

Though changing resolution may not improve true uncertainty in RMSE, the bagging model gives the lowest RMSE of any model for virtually all observation densities and resolutions. If your experiment can estimate the uncertainty in RMSE at a given observation density, note that uncertainty in RMSE is pessimistic for BRT and bagging and that the difference between true and apparent uncertainty is highly dependent on site and observation density for LM.

3.4.3 R^2

For a given set of observations, can one maximize R^2 by changing resolution?

In figure 3.4 colored tiles represent median R^2 and lines give its uncertainty arising from random observation selection. At the largest resolutions we again see results degrading due to poor model fitting, as in the previous sections.

For low and medium observation densities, we often see R^2 maximizing at intermediate resolu-

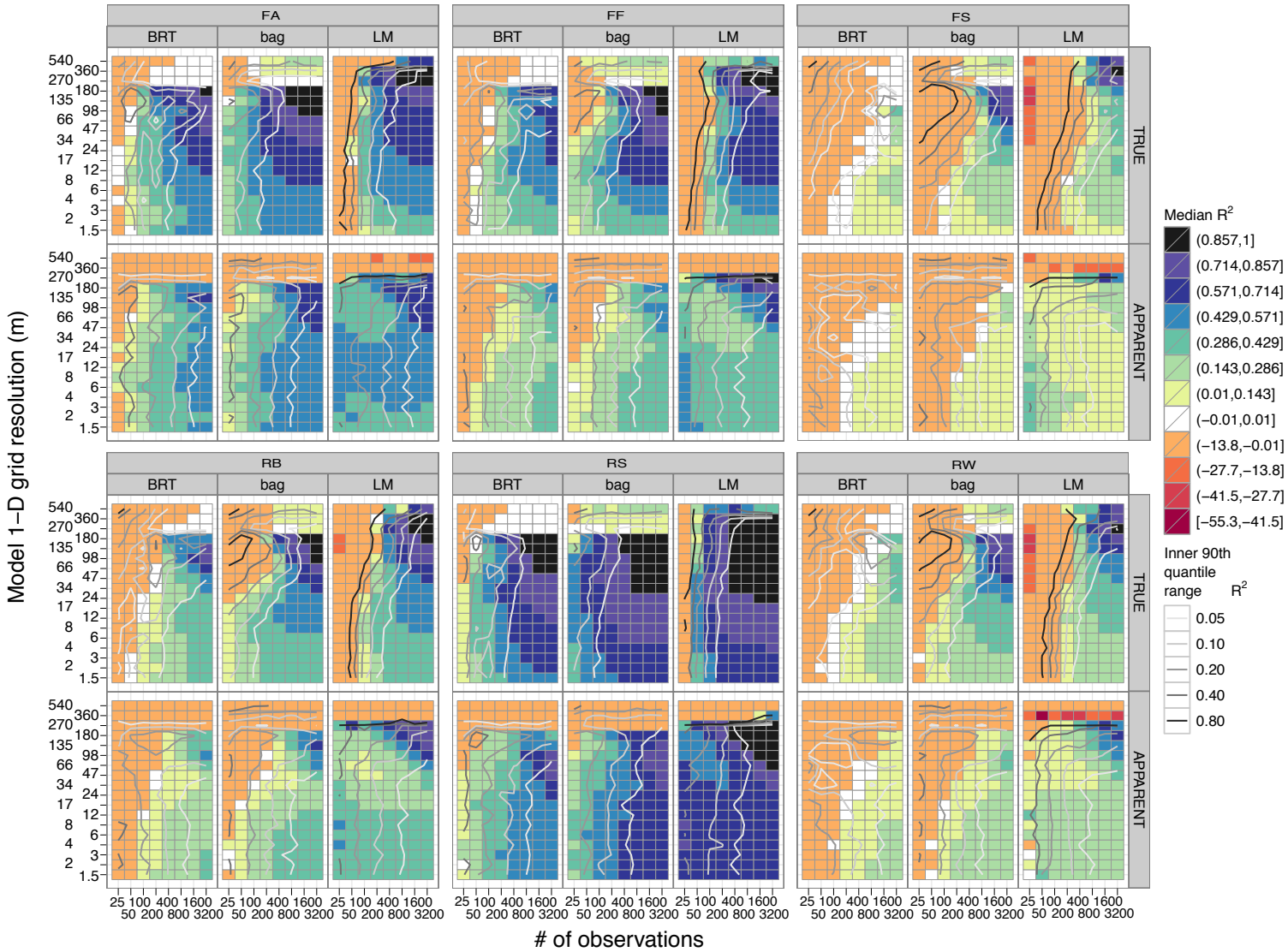


Figure 3.4: R^2 and its uncertainty (90% confidence) due to random selection of observations as a function of observation density and model resolution.

tions whenever it is positive. Because variance decreases with resolution by design, results indicate that the scale of the relationship between predictors and response can have a greater effect on R^2 than reduction in variance. This implies that proper estimation of spatial distribution can depend critically on model resolution. However this relationship also depends on observation density. At most sites, change in R^2 is appears dominated by reduction in variance above approximately 400-800 observations.

Because of the intertwined effects of model resolution, decreased true variance, and observation density on R^2, it is difficult to point to the ''preferred'' model scale at each site. Model choice affects estimated R^2 as well. Still, the general outlines of preferred model resolutions match our understanding of dominant scales of variability at each site. At FF, RB, and RW, where stands of vegetation dominate snow depth distribution, preferred resolution is between 15-20m. At FS, where holes in dense forest determine the dominant variability, R^2 maximizes near 6m resolution. Variability at RS is dominated by aspect, at hill slope scales of around 30-50m, and FA is dominated by even large scale differences between forested and alpine areas on scales of 50m or more.

At low observation densities, choosing too large of a model scale can drastically affect R^2 and send it negative. Using too fine a scale has less dramatic impact on R^2. When important variability exists at fine scales, observation and predictor scales must be matched to avoid introducing noisy relationships into the model. Thus, for a fixed amount of observations, results indicate that R^2 might be optimized by choosing the lower end of the range at which dominant variability is believed to occur. This approach also minimizes uncertainty in R^2, as discussed below.

While apparent R^2 is generally pessimistic, true and apparent surfaces dont correspond well to each other. Their comparison is especially poor for the LM model. If one could experiment with apparent R^2 as a function of resolution for a given set of observations under BRT or bagging, one might roughly estimate the resolution at which R^2 begins to decrease. One would then choose a slightly smaller model resolution and assume true R^2 greater than apparent by pessimism of the measure under this choice of sampling design. Results for alternate sampling designs are discussed later.

For a given set of observations, can one minimize uncertainty in R^2 by changing resolution?

As model scale exceeds the scale of interaction between predictors and snow depth, the two essentially become decoupled and the model is presented with a noisy relationships. Given 100 sets of observations at some observation density, models based on each set will begin to digerve as increased model scale introduces this decoupling noise. Remaining on an uncertainty isopleth provides one way of estimating ''preferred'' model scales. For a given number of observations, increase resolution without increasing uncertainty due to random sampling. Though we only plot 90% uncertainty for R^2 down to .05, this method lands one near the peak R^2 for medium observation densities. In practice, this is similar to the suggestion above for modeling slightly below the smallest scale of important variability.

3.5 Skill of downloaded and LiDAR-measured predictors

In this section we assess loss of estimation skill going from LiDAR-measured to downloaded predictors at the 30m resolution of the downloaded predictors. Though the previous section indicated that R^2 may be compromised by this choice of model resolution for all but the FA and RS sites, there is no way to downscale downloaded predictors.

For each skill metric, we consider true and apparent skill for both predictor sets. In the three figures which follow, the following color scheme is used, LiDAR-measured predictors: true (blue), apparent (cyan); downloaded predictors: true (red), apparent (pink). True colors are ''primary'' and apparent colors are ''pastel.'' In all figures the black horizontal line indicates the optimum of the statistic.

3.5.1 PE

Figure 3.5 illustrates median percent error (bias) in total snow volume and its uncertainty over all simulations. The 90% confidence interval is shown by the extent of the box plot whiskers (same uncertainty as plotted when considering resolution above). For both downloaded and LiDAR-measured predictors, apparent model bias is extremely low (less than 1 %) for all models at all

observation densities because its minimization (10 fold cross-validation not withstanding) is how our regression models are fit. True model bias for BRT and bagging models is also very low (less than 2.5%) at all observation densities. The LM model shows significant bias at 25 observations, over 5% at FA and FS, but is much less biased by 100 observations.

Though bias is generally low, its uncertainty ranges widely. The largest contrasts in the figure are between uncertainty in apparent (pastel colors) and true (bold colors) percent error. Because regression models intentionally minimize their bias to the fitting data set and apparent statistics are (only) using 10-fold cross-validation, it is not surprising that each systematically underestimates percent error at low observation densities. We see that the difference in apparent uncertainty and true uncertainty is much less for bagging relative to the other two models. This results from the bootstrapping nature of the model, which further reduces the number of fitting observations used each time the model is fit. Bagging's apparent uncertainty approximates half that of the true uncertainty. Apparent error from the BRT and LM models vanishes around by 100-200 observations, inaccurately reflecting true percent error. Apparent error from BRT is often negligible at even the lowest observation densities. If one is able to calculate apparent uncertainties, a method like bagging should be employed to estimate more reliable confidence intervals than provided by BRT and LM in this study.

Gross differences in uncertainty exist between the F and R sites due to differences in group mean snow depth. Higher mean snow depth at R sites yields lower uncertainty in percent error (and bias). Uncertainty is plotted in ±12% error which contains inner quartile range (IQR), though not the 90% uncertainty, for nearly all simulations. Where models were observed to be most biased, we see greatest uncertainty at low observation densities. At all R sites, the 90% uncertainty in percent error for all setups is bounded by ±10% at 50 observations and by ±5% at 100 observations for all but one set of simulations. At the F sites, uncertainty depends on snow depth mean and variance. FS (low mean) and FA (high variance) have much greater uncertainties than FF. At FA, uncertainty for nearly all simulations is bounded by ±10% at 200 observations and by ±5% at 800 observations.

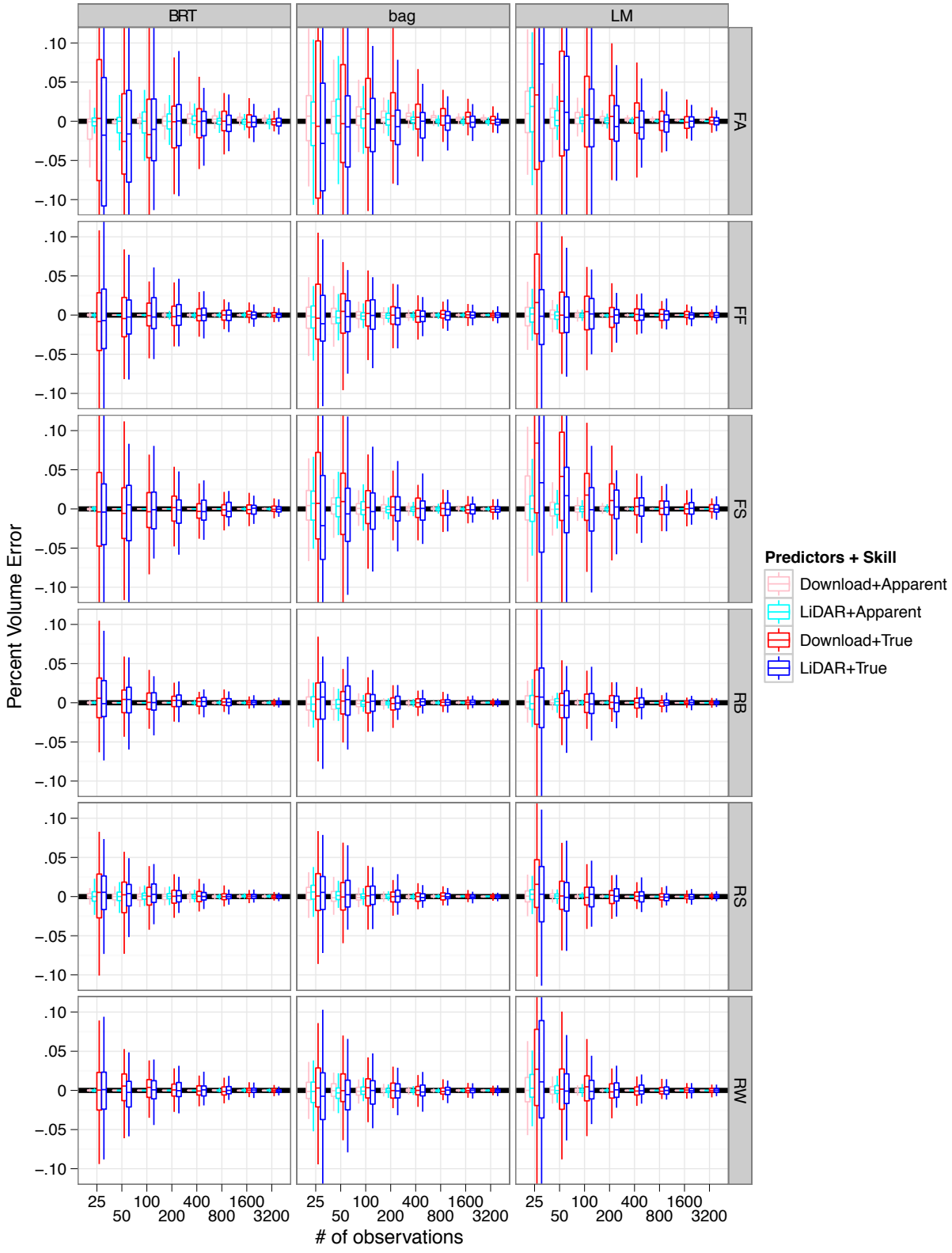


Figure 3.5: Percent error and its uncertainty (90% confidence) as a function of observation density at 30m resolution.

Differences between LiDAR-measured and downloaded predictors are not large. Only at lower observation densities do downloaded predictors give either slightly more biased percent error or a greater uncertainty at some sites. Differences at 100 observations are typically very small. Estimation of percent error and its uncertainty are more sensitive to number of observations and model choice than to predictor quality. Coupled with uncertainty below ±10% by 100 observations for all sites but FA, this is certainly encouraging news for estimating total snow volume over 1km extents.

3.5.2 RMSE

The general pessimism of RMSE is perhaps the most obvious feature in figure 3.6, apparent (pastel) RMSE overestimates true (primary) RMSE most of the time. The FA site and the LM model provide exceptions. At low observations densities, true and apparent RMSE are comparable at FA ostensibly because the overall variability at FA approximates that of the observations used in fitting. LM is actually optimistic at low observation because it produces an unreliable fit to less than about 100 observations as seen in previous sections.

The pessimism of RMSE is highlighted in the figure because differences between true and apparent RMSE dwarf differences between downloaded and LiDAR-measured predictors. Downloaded predictors do not greatly affect RMSE. FA provides the clearest exception to this, true and apparent skill from LiDAR-measured is considerably lower for downloaded predictors. While differences between the actual downloaded and LiDAR-measured predictor sets are not huge themselves, extreme variability at the site means that a few misplaced tree pixels in the downloaded product can result in a 2m estimation error. The failure of downloaded predictors to properly resolve the relatively small terrain pockets in the ridge may induce perhaps even greater error. Smaller differences in RMSE resulting from choice of predictor set can be seen at FF and each of the R sites. At FF, downloaded vegetation misses the LiDAR-measured pattern of logged forest. At the R sites differences in RMSE between downloaded and LiDAR-measured predictors likely result from inaccuracy of vegetation in the downloaded predictor set.

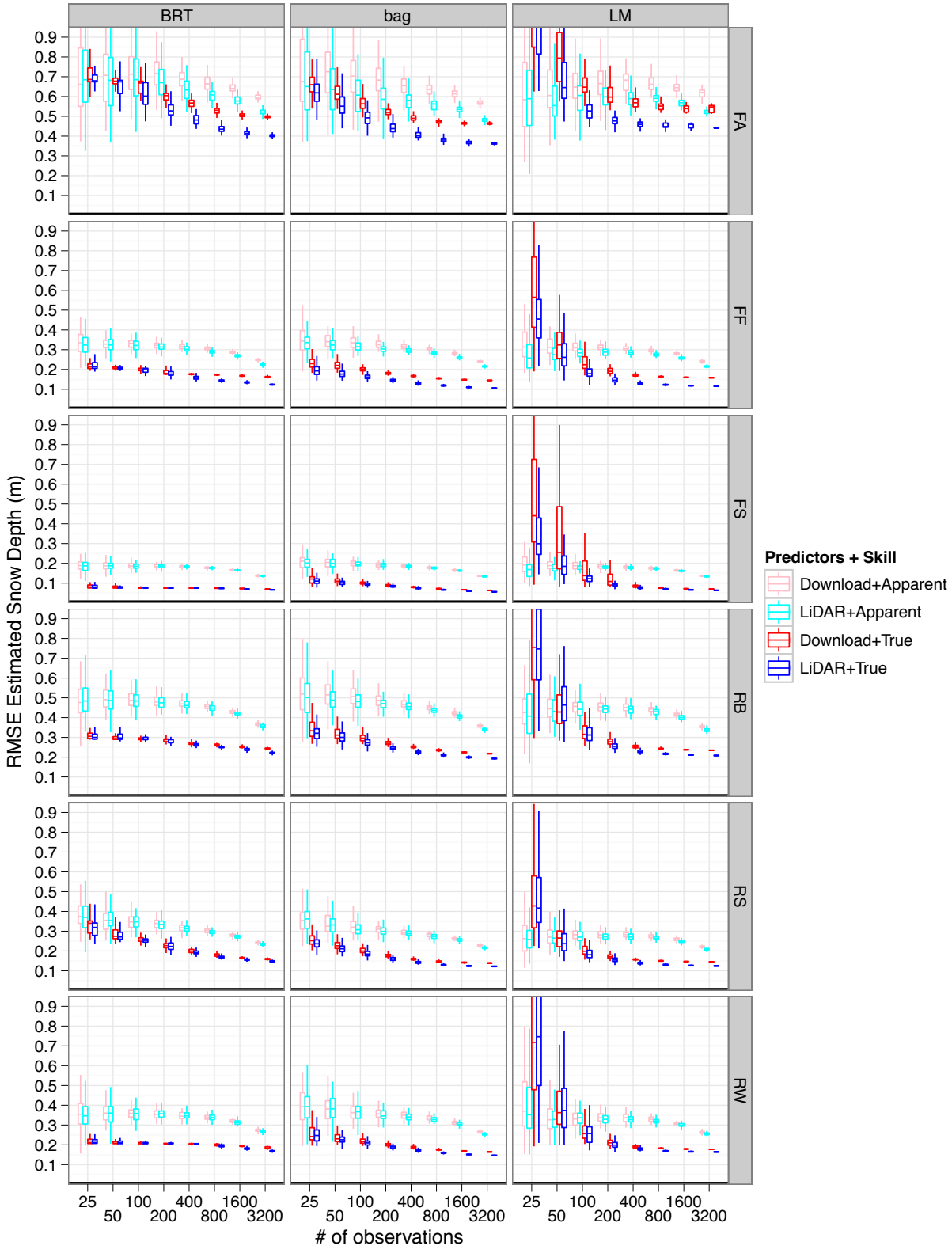


Figure 3.6: RMSE and its uncertainty (90% confidence) as a function of observation density at 30m resolution.

Results indicate that bagging and LM models produce lower RMSE at medium observations densities than BRT. Except at FA, where bagging has large advantage, bagging and LM skill are extremely competitive in the 200-800 observation range. At higher observation densities, bagging starts to edge out LM and at low observation densities LM appears unreliable. Bagging is the overall best model for minimizing RMSE. Though not reported here, the random forests variant on BRT was also investigated. We choose to focus on bagging over random forests because it excelled at intermediate observation densities (200-800). At higher observation densities, random forests typically achieved the best results.

Uncertainty in apparent RMSE arising from random observations can easily be 30cm at the lowest observation densities. By 400 observations we see apparent RMSE typically constrained to a 10cm range. Apparent uncertainty is very similar for downloaded and LiDAR-measured data sets. True uncertainty is much smaller than apparent, again with little differences between predictor sets. Because apparent uncertainty at the lowest observation densities can overlap or cover the range of true uncertainty at several sites in this study, pessimism of the RMSE statistic should not be assumed until medium observation densities are reached.

3.5.3 R^2

Though results vary widely with site, the R^2 statistic in figure 3.7 reveals that correct estimation of snow depth spatial structure depends strongly on the quality of the predictor set used. Apparent R^2 describes how well the model estimate explains residuals left by subtracting the observed mean from the fitting data set. True R^2 is a measure how the model improves upon the flat, uniform distribution of true mean snow depth at each point. When R^2 is positive, the estimate has explained this fraction of residuals left by estimating with the null model. The number of observations required for R^2 to go positive depends on site and model choice, but is generally increasing in this order of sites, RS, FA, FF, RB, RW, FS. This observation density is controlled by 2 factors, the ratio of random to systematic variability in the distribution and the relationship between predictors and response. In the previous investigation of resolution effects on R^2, the

relationship between predictor and response was seen to be maximized at a particular resolution and introduce noise at larger scales. Results depend on choice of resolution.

Median R^2 is generally pessimistic at medium to high observation densities whenever apparent R^2 is positive. Differences in true and apparent R^2 can easily be .2 at medium and high observation densities. This gap, for both predictor sets, partially results from our sampling design which spaces observations widely despite strong spatial autocorrelation in snow depth. This has the effect of deflating apparent R^2 compared to other designs (e.g. figure 3.1) which sample points in close proximity, as explored in the following section. Whenever true R^2 is positive under our sampling design, true R^2 will typically be greater than apparent R^2 and improve faster than apparent as the model improves.

However, care must be taken when interpreting pessimism from median R^2 as large uncertainties are evident, especially at low observation densities. Uncertainty of R^2 is discussed below. Apparent R^2 from the LM model is extremely unreliable due to poor model fits at low observation densities, approximately 200 observations are required for it to become pessimistic and even more for downloaded predictors. For BRT, R^2 is pessimistic whenever at least 50 observations are used or apparent median R^2 is above .05, except at FA where 200 observations are needed. Under bagging, apparent R^2 is pessimistic or equal to true R^2 whenever positive. The bagging model consistently gives the best true R^2 scores at a given observation density even though its apparent skill is not always the greatest.

For identical models and observation densities, at virtually all sites we see decrease in R^2 resulting from use of downloaded predictor sets. At 3200 observations, losses range from about .05 at RS to over .2 at FF for true R^2. Differences for apparent R^2 between predictor sets are typically less, though comparable at FA. One way to view the value of LiDAR-measured predictors is by the observation density required to obtain an R^2 equal to downloaded predictors using 3200 observations. LiDAR-measured predictors typically require only 1/8-1/4 of the observation density to attain the same skill, though this depends on the site being estimated.

Differences in R^2 resulting from choice of predictor set is likely due to poor representation of

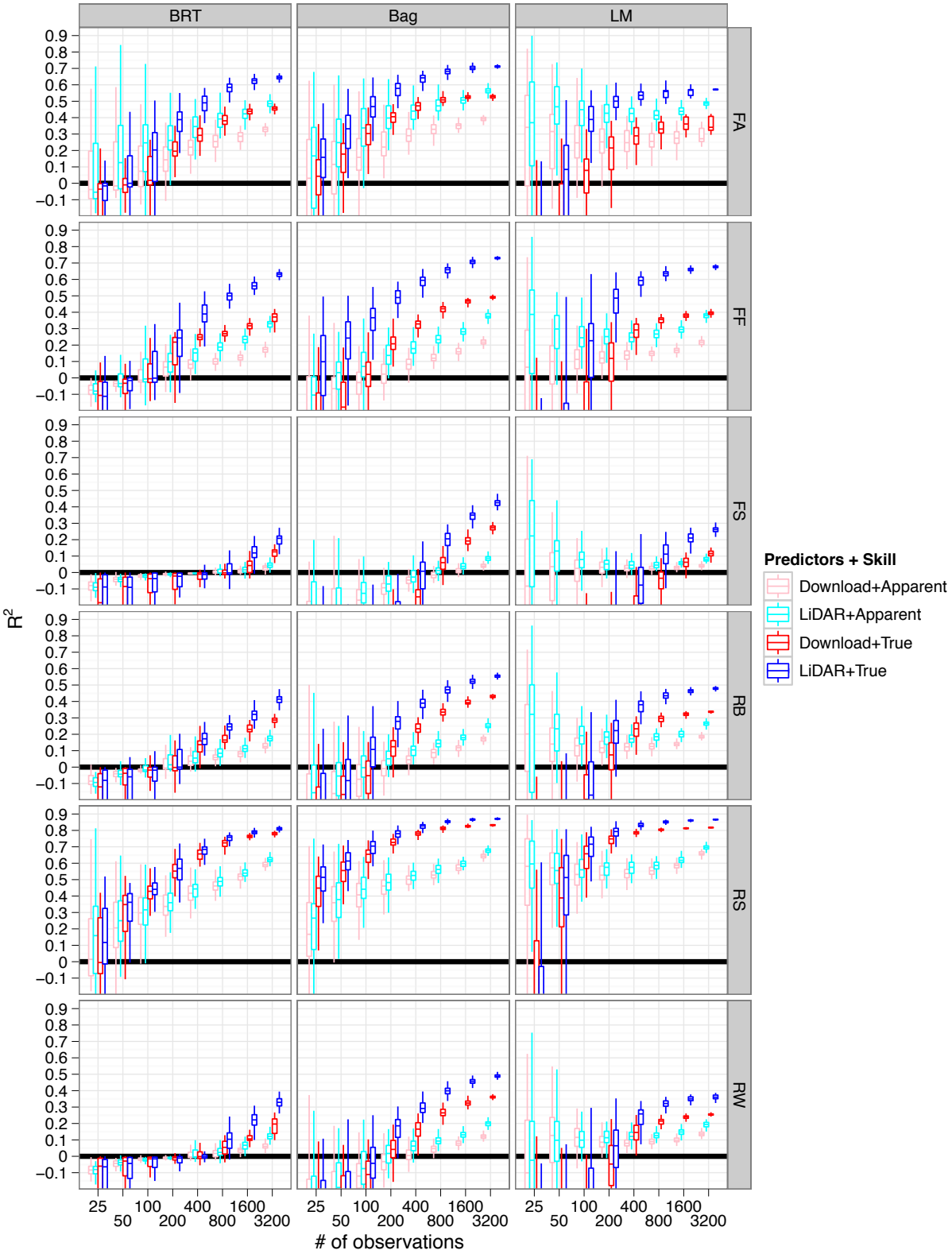


Figure 3.7: R^2 and its uncertainty (90% confidence) as a function of observation density at 30m resolution.

vegetation in downloaded predictors, as seen for RMSE. At the RS site we see virtually no difference in R^2 because snow distribution is dominated by aspect, which downloaded predictors represent relatively accurately. Huge differences at FF result from lack of observed spatial variability in vegetation in the downloaded product. At FA extreme variability in a few difficult to resolve locations is responsible for the disparity in scores between predictor sets rather than a gross mis-representation of physical controls.

The 90% uncertainty in true R^2 due to random sampling is remarkably large. At lower observation densities, it can easily be near .2 when R^2 is positive and often .1 at 800 observations. Thus, true R^2 depends strongly on the particular set of observations used. Even as models tend to converge with more observations, noise limits convergence even when using 3200 observations. Noise may be real, introduced by sampling or by inappropriate model scale. The effects of real noise can be seen at some sites, particularly FA, at low observation densities when models struggle to fit a relationship to large variability in some cases. We expect LiDAR sampling errors and model scale to also contribute to the noise. Though all sampling methods will introduce noise, if another method introduces less noise, our results present an upper bound on the uncertainty in R^2 which can be expected at these sites. Note this large range of uncertainty implies that observation counts often need to be at least quadrupled to ensure a statistically better R^2, regardless of predictor choice. An important question becomes, can we design sampling which ensures the highest R^2 at a given number of observations? We do not tackle this question directly here. Note that apparent uncertainty R^2 is often comparable to true uncertainty. If one is able to estimate the apparent uncertainty, it is likely to be similar or worse than the true uncertainty, particularly at high observation densities.

3.6 Sampling design

Our sampling design has emphasized randomly selected observations which cover the domain. This places more value on independent information than on spatial autocorrelation and results in deflated cross-validated statistics for RMSE and R^2. We have illustrated that pessimistic cross-validated statistics may be preferable, it's often better to know that the true score is better

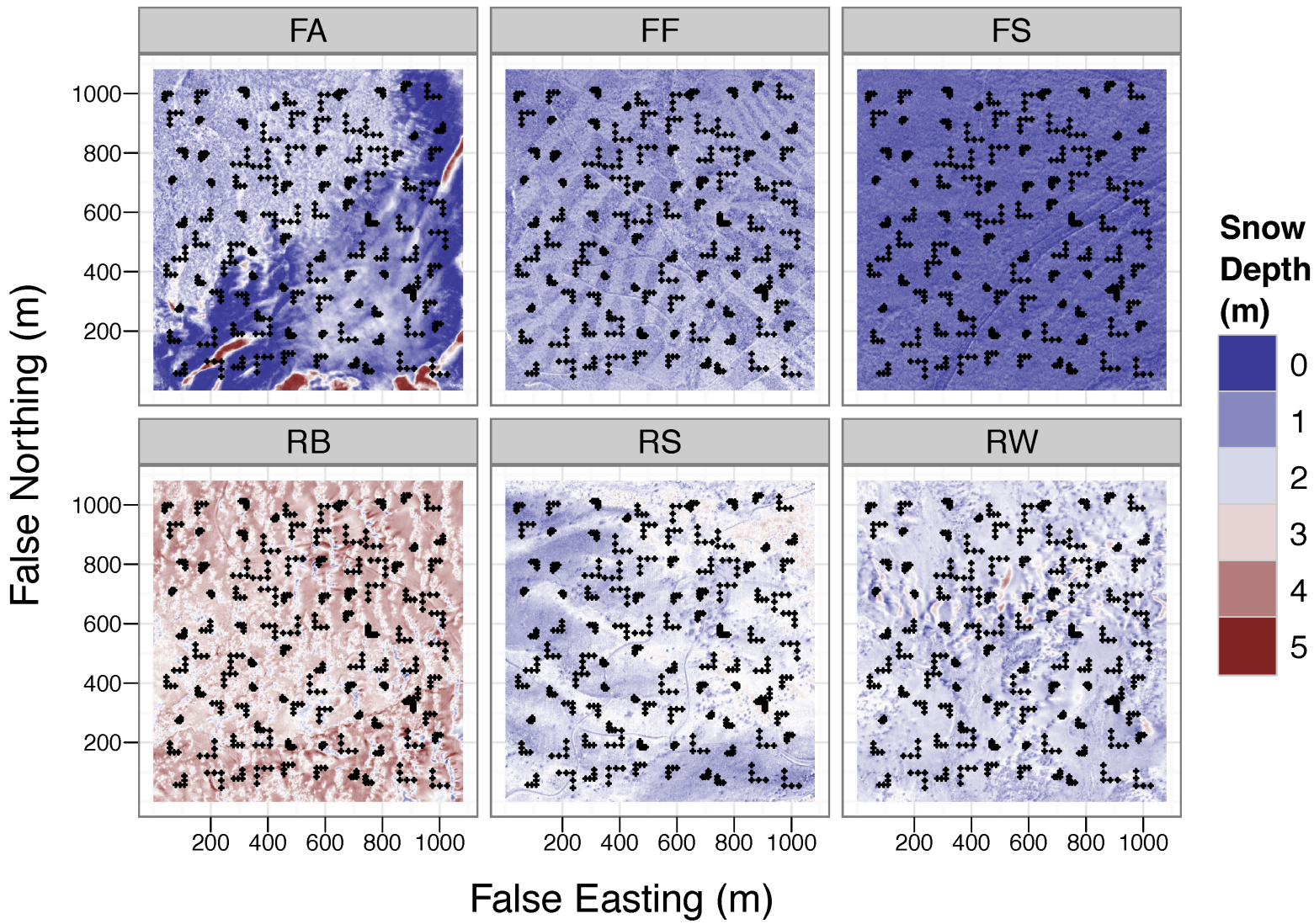


Figure 3.8: Location of ground-based (in situ), manual measurements from CLPX at each site.

than the cross-validated one even if the true score is unknown. To provide more context for our choice of sampling design in this study, we compare with the CLPX sampling design by sampling of the high-resolution LiDAR data using the same patterns reported in the ground-based, manual probing campaigns. Shown in figure 3.8, the CLPX sampling designs carried out were extremely close to and based upon their proposed sampling design of Erxleben et al. (2002) shown in figure 3.1. We compare R^2 from this sampling design to results in the previous section where 30m resolution predictors were used. As mentioned in chapter 2, comparison with the ground-based and LiDAR measurements is difficult due to the ablation and accumulation in the period between their observations (shown in figure A.1). After comparing results using LiDAR measurements for the different designs, cross-validated results for the ground-based measurements are presented as food for thought.

		BRT		bag		LM	
		Apparent	True	Apparent	True	Apparent	True
FA	Download	0.32	0.30	0.45	0.42	0.42	0.26
	LiDAR	0.42	0.50	0.53	0.59	0.51	0.55
FF	Download	0.06	0.24	0.07	0.28	0.11	-0.01
	LiDAR	0.10	0.31	0.16	0.47	0.24	0.42
FS	Download	-0.01	-0.01	-0.13	-0.17	-0.01	-0.02
	LiDAR	-0.01	-0.01	-0.06	-0.11	-0.00	-0.13
RB	Download	0.064	0.15	0.01	0.25	0.13	0.26
	LiDAR	0.10	0.13	0.12	0.30	0.14	0.27
RS	Download	0.47	0.66	0.55	0.77	0.61	0.78
	LiDAR	0.48	0.67	0.56	0.82	0.61	0.85
RW	Download	-0.01	-0.02	-0.03	0.20	0.09	0.08
	LiDAR	-0.01	-0.02	0.09	0.29	0.13	0.20

Table 3.2: Apparent and true R^2 for CLPX LiDAR measurements using identical sampling design to the in situ measurements. Results for downloaded and LiDAR-measured predictor sets.

At each site, 684 in situ observations were made. Sampling these from the 1.5m resolution LiDAR snow depth data set, points in close proximity end up in the same cell. This reduces the number of samples of the LiDAR snow depths to 615, 616, 612, 617, 604, and 612 observations at the FA, FF, FS, RB, RS, and RW sites, respectively. Table 3.2 shows the R^2 scores for this

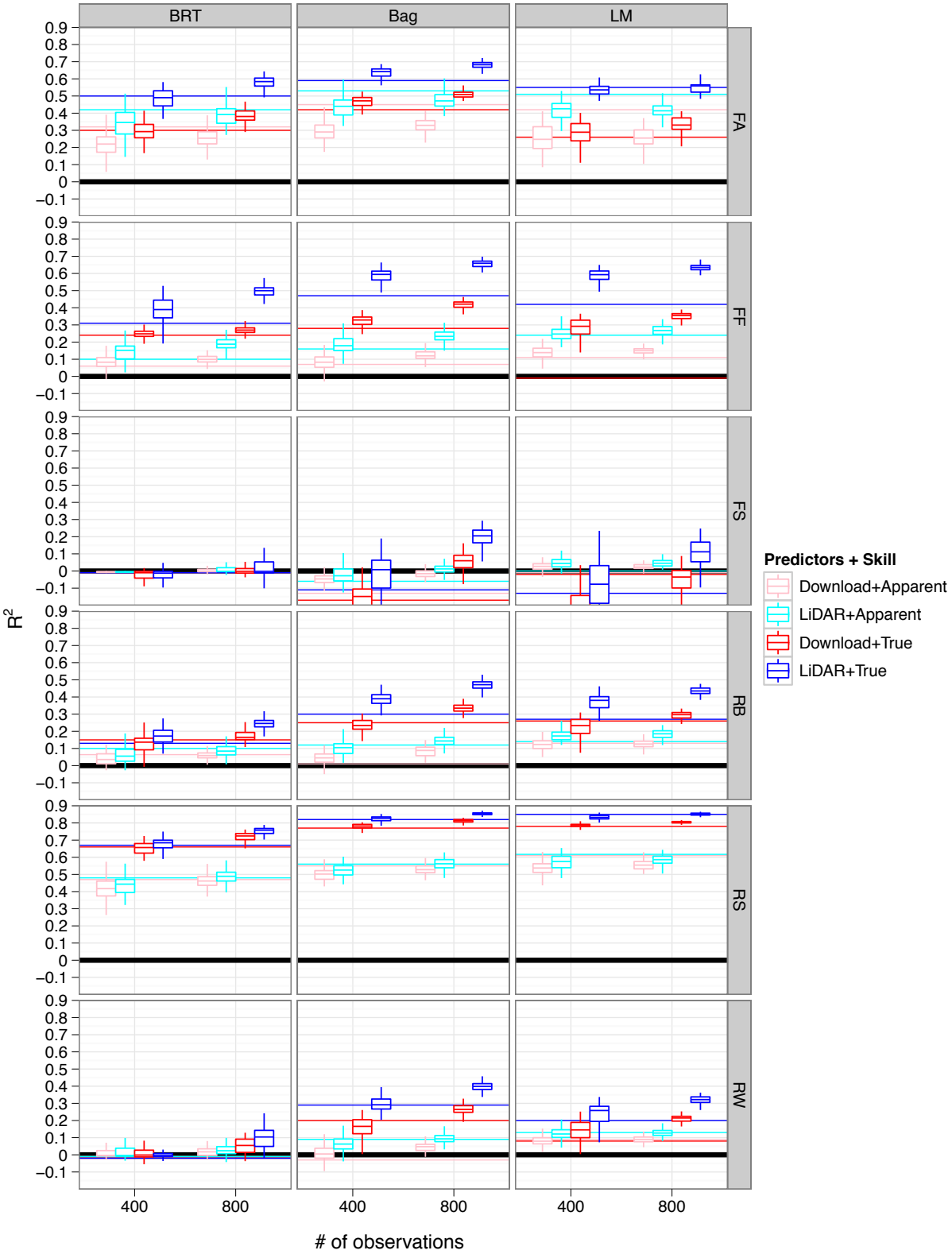


Figure 3.9: Comparison of R^2 from the CLPX sampling design and those from the random sampling designs used in the previous section.

sampling of the LiDAR snow depths and figure 3.9 shows the data in this table as horizontal lines plotted along with results for 400 and 800 observations from the previous section.

Though R^2 can vary considerably with individual observation selection, as shown by uncertainties associated with random selection of observations in the previous section, results generally confirm the differences we expect between the sampling designs. We typically see the true skill for the CLPX design, using both LiDAR and Downloaded predictor sets, much closer to the median R^2 based on 400 observations for our sampling design. This is lower than we would expect for roughly 600 observations and results because the clustered sampling design sacrifices independent information in favor of measuring spatial auto correlation. Figure 3.9 suggests that random sampling is more likely to yield better true R^2 statistics for a fixed number of observations.

Because true R^2 can rarely be estimated as in this study, if it must estimated it has to be inferred from the apparent R^2. By incorporating spatial autocorrelation into measurements, sampling design can attempt to estimate the true R^2. Paradoxically, trying to estimate true R^2 in this way may act sacrifice its score for a fixed number of observations. At many sites, we see that the CLPX sampling design raises the apparent R^2 towards or beyond the cross-validated R^2 from our sampling design. But in doing so, sacrifices true R^2. This is can be seen at FA, RB, RS and RW.

As illustrated by downloaded predictors at FA, there is also an important balance between variance and autocorrelation in measurements which may be tipped (depending on predictor quality as well) so that apparent skill gives an optimistic impression of true R^2. Should the true R^2 need to be estimated, results from the extreme snow depth variance at FA and suggests that trade-offs between independent information and autocorrelation in sampling design should be based on the expected observation variance. If only a lower bound on true R^2 is sufficient, then observation should be gathered to maximize independent information.

3.7 Discussion

To provide further context to this study, we compare our results to those of Erxleben et al. (2002). Their study, quoted earlier as impetus to explore the variety of topics in this paper, used data at the FF, FS, and RW sites from a manual survey in 2001. We make comparison acknowledging the caveat that we are comparing distributions of snow packs in different years. While a variety of studies have found strong similarities in snow distribution in different years at the same location (Kirnbauer and Bloschl, 1994; Sturm and Holmgren, 1995; Deems et al., 2008), this adds a new dimension of uncertainty to the comparison. However, comparison will provide some guidance on how results here may be applied more generally.

Site	Study	Date	Mean	St. Deviation
FF	Erxleben et al. (2002)	April 3-6, 2001	1.09	.211
	This study	April 8, 2003	1.3	.37
FS	Erxleben et al. (2002)	April 1-2, 2001	.58	.115
	This study	April 8,2003	.6	.28
RW	Erxleben et al. (2002)	April 9-11, 2001	1.77	.373
	This study	April 9, 2003	1.9	.35

Table 3.3: Comparison of snow conditions to Erxleben et al. (2002). Depths and standard deviations in meters.

Snow depth mean and standard deviation from the 2001 survey and the CLPX LiDAR data are compared in table 3.3. In table 3.4 we rank $RMSE_{Erx}$ and R^2_{Erx} obtained by Erxleben et al. (2002) using BRT, under 10-fold cross-validation from 550 observations at 30m resolution, to our BRT results using 400 observations from LiDAR-measured predictors at 30m resolution. We choose LiDAR-measured predictors because Erxleben et al. (2002) derive their vegetation map from aerial photography, which resolves location of vegetation more accurately than our downloaded vegetation from NLCD2001. Deficiencies of the downloaded vegetation were cited as the main cause of poor predictor performance at these sites. Their results are ranked with relevant true and apparent results from our study.

RMSE scores are generally comparable, differences appear to depend on differences in standard deviation between the data sets. At FF and FS, where their standard deviation is much lower than ours, $RMSE_{Erx}$ is better (lower) than our apparent RMSE but worse than our true RMSE. At RW, $RMSE_{Erx}$ is within the uncertainty of apparent RMSE as standard deviations are similar.

Site	Comparison						
FF			$RMSE_{True_{max}} = .14$	$<$	$RMSE_{Erx} = .176$	$<$	$RMSE_{App_{min}} = .27$
	$R^2_{True_{min}} = .19$	$<$	$R^2_{App_{max}} = .27$	$<$	$R^2_{Erx} = .298$	$<$	$R^2_{True_{max}} = .53$
FS			$RMSE_{True_{max}} = .075$	$<$	$RMSE_{Erx} = .104$	$<$	$RMSE_{App_{min}} = .165$
	$R^2_{App_{max}} = .005$	$<$	$R^2_{True_{max}} = .04$	$<$	$R^2_{Erx} = .175$		
RW			$RMSE_{App_{min}} = .3$	$<$	$RMSE_{Erx} = .32$	$<$	$RMSE_{App_{max}} = .4$
	$R^2_{True_{max}} = .03$	$<$	$R^2_{App_{max}} = .1$	$<$	$R^2_{Erx} = .261$		

Table 3.4: Comparison of RMSE (m) and R^2 results to Erxleben et al. (2002).

At FF, R^2_{Erx} is slightly better than our best apparent R^2 score. At FS and RW, R^2_{Erx} is significantly better than our apparent R^2 scores. Their R^2 at these sites are even slightly better than our maximum, true R^2 at 800 observations (not shown in the table). Comparison suggests that differences in sampling design and LiDAR measurement error may account for their better results. Of course no conclusions can be made comparing these different years.

Difference in sampling design should increase the R^2 of Erxleben et al. (2002) relative to our apparent scores, as described in the previous section. They sample many points within the correlation length scale at each site while our sampling scheme emphasizes large distances between samples in order to cover the domain. Deems et al. (2006) showed strong correlation at short distances for several of these sites and variograms for all sites in this study (except FS) are presented in the following chapter in figure 4.2, though we do not present scale break analysis here. Differences in sampling schemes may also explain why their observed standard deviation is reduced compared to ours.

As previously discussed, LiDAR observational error is likely inflating the variance in our data. At FS, we believe LiDAR error competes with actual snow depth variability. Low R^2 scores at FS in our study are not surprising. RW is the one site where we suspect systematic error in LiDAR

observations resulting from certain types of vegetation. In both cases, LiDAR error obscures the relationship between snow and predictors and results in particularly poor R^2. At FS and RW, the number of observations required for positive scores is much greater than at all other sites. By contrast, at FF, actual variability is larger than observational error so a clearer relationship between predictors and snow depth is presented to the model. This would explain why results are much more comparable at FF than at FS and RW.

While it is unfortunate the above comparison highlights the main difficulties with the LiDAR data set in this study, comparison is still useful. We believe our results to be much more robust at sites where systematic variability exceeds LiDAR error (random or systematic). These are all sites besides FS and RW. Sites such as RS and FA have strong systematic variability and estimates often are very good despite the presence of LiDAR measurement error.

3.8 Conclusion

Results indicate that proper estimation of spatial snow depth distribution from sparse observations, as judged by R^2, depends critically on choice of model resolution and on the quality of the predictor set used. Estimation of total volume is generally insensitive to these factors while RMSE shows sensitivity to predictor quality but no preference for model resolution. RMSE decreases with resolution as a result of decreasing snow depth variance at larger model scales. These results are not surprising, but our quantification frames effects of model resolution and predictor quality on each skill and its uncertainty in terms of sampling requirements. In the case of R^2, we have shown that compromises in predictor quality or model resolution can easily require a doubling of observation density to maintain the same level of skill.

The most surprising results come from considering uncertainty in skill due to random sampling. Though we consistently see apparent percent error in total snow volume underestimating true percent error by large margins, true percent error is well constrained by a few hundred observations, depending on snow depth variance. Uncertainty in true R^2 is remarkably large at medium observation densities. A range of .15 at 400 observations is a typical, though there is tremendous variability

by site. These uncertainties imply a quadrupling of observation density is required to ensure improvement of true R^2 and highlight the need for further research into sampling design to minimize uncertainty in R^2 due to selection of observations. In all experiments, this study has found the bagging tree model to out perform binary regression trees and linear regression. Bagging not only produces the best scores at each observation density, it also produces the most reliable apparent skill for each measure.

Pessimism of RMSE and R^2 statistics result, in large measure, from our choice of sampling design. We have described how it can be advantageous to have a pessimistic statistic. Under BRT and bagging, apparent RMSE provides a consistent upper bound on true RMSE except at the lowest observation densities and the highest variability site, FA. Large uncertainties dominate R^2 at lower observation densities but, using the bagging model, the statistic is pessimistic at medium observation densities. Comparison of our random sampling design, which emphasizes independent information, to the CLPX sampling design, which incorporates spatial autocorrelation, highlights that true R^2 is sacrificed at medium observation densities whenever more autocorrelation is included in sampling at a fixed number of observations. Thus trying to estimate true R^2 by altering sampling design for a fixed number of observations will produce a worse true R^2 score than if it were left unknown. Comparison of our results with an earlier study at 3 of the same sites revealed potential issues with LiDAR measurements and some idea of how result may be applied beyond the data sets in this study.

Chapter 4

Snow Depth Spatial Structure

4.1 Introduction

Previous studies of the spatial distribution of snow have concentrated on its estimation from sparse observations. Virtually no attention has been given to predicting snow distribution in locations where measurements are unavailable. This emphasis is, of course, attributable to the historical difficulty of effectively measuring snow at even the smallest hydrologic scales. In this chapter we investigate prediction of snow depth spatial distribution based solely on independent predictor variables (such as elevation, vegetation height, slope, etc.) in the situation where snow depth measurements are assumed not to exist. In order to make such predictions, we first build a new view of snow depth spatial distribution based on the high-resolution CLPX snow depth measurements.

Though these measurements offer three orders of magnitude more data at each site in this study than would be available using manual probe measurements (e.g. Erxleben et al., 2002) and a unique window into the spatial distribution of snow depth at incredible spatial density, for practical reasons LiDAR will never offer more than a mere glimpse into earth's snow cover. The relatively small of spatial extent the data (1km) underscores our goal. Thus, LiDAR measurements are viewed as an opportunity to assess the predictability of snow depth distribution in unobserved or minimally observed locations, in the spirit of the Prediction in Ungauged BasinS (PUBS) initiative (Sivapalan, 2003). To this end, we evaluate the skill of ''readily-available'' variables, which are easily obtained via the internet for the entire conterminous United States, for prediction of spatial structure of

peak-accumulation snow depth at six 1 sq km sites in the Colorado Rocky Mountains.

4.2 Background

This chapter examines the intuitive idea that snow depth spatial structure is governed by the spatial structure of relevant independent variables. This notion is examined in the data and then taken to its logical conclusion in the context of prediction, where we see how much is gained from the idea in practice; if spatial structure of predictor variables is a first-order control on snow depth, how well can we predict snow depth spatial structure from these? Note that we are not attempting to predict snow depth, but rather its spatial structure or pattern.

We define the idea of ''spatial structure'' more precisely, starting from the definition of spatial non-stationarity. Spatial non-stationarity of a statistic, mean snow depth, signifies that the statistic depends on the location of the spatial window over which it is calculated. The idea of spatial structure goes slightly further to say this mean does not vary randomly in space but as a response to controlling variables. Structure implies some amount of order and causality. We use the term structure to refer to both the actual structure and its representation by an organized spatial pattern which may represent only the spatial shape and not the values of the structure.

The idea of snow depth spatial structure is present in several themes in the snow depth literature: *a*) Stratified sampling design; *b*) The success of binary regression trees versus other models for estimating snow depth distribution; *c*) Scale breaks in log-log variograms and power spectral densities.

Stratified sampling design begins from the premise that snow depth distribution is controlled by changes in topographic, vegetative, and other variables and combinations thereof. Thus, such changes are important to account for if sampling campaigns wish to accurately estimate areal snow cover. Watson et al. (2006) found optimal snow sampling schemes based on stratifying the landscape in to zones using ranges of time, elevation, radiation and vegetation. Elder et al. (1991) used clustering of elevation, slope, and radiation in conjunction with Bayesian classification to stratify a landscape in to zones used for both determining sampling distribution and for spatial

modeling. Our study follows most closely from the work of Elder et al. (1991) though with the luxury of high-resolution LiDAR observations at multiple locations.

The most prevalent case for snow depth spatial structure is the success of binary regression tree (BRT) models for interpolating or estimating snow depth distribution from sparse measurements (e.g. Elder et al., 1995; Balk and Elder, 2000; Winstral et al., 2002; Molotch et al., 2005). In multiple studies, the BRT model (sometimes in conjunction with residual kriging) has been shown to outperform models which produce spatially smooth (weakly structured) snow depth estimates, such as inverse distance weighting or Kriging (e.g. Erxleben et al., 2002). The BRT model (described in detail in the appendix section 4.3.4) uses similarities of controlling variables an their collocated snow depths to fit a hierarchical model in which snow depth can vary abruptly with changes in controlling variables or their combinations. When applied to a spatial domain, a binary regression tree implies that spatial non-stationarity of snow depth depends on that of controlling variables. The success of this model over other representations of snow distribution (at the small, 1 sq km scales considered in this study) suggests that it most closely matches the physical nature of snow distribution.

Deems et al. (2006) and Trujillo et al. (2007) presented statistical summaries of spatial structure in high-resolution LiDAR snow depths (using most of the same data sets as in our study). Deems et al. (2006) employed log-log semivariograms whereas Trujillo et al. (2007) analyzed the spatial structure in terms of log-log power spectral densities. These two approaches revealed very similar snow depth spatial structures, understood via "scale breaks" in their respective statistical summaries. A scale break is the length at which scaling behavior (variability as a function of distance) of a statistic is seen to change. Results and scale breaks in these studies are compatible with the view that snow depth spatial structure is governed by spatial changes in controlling physical variables. At short scales, where controlling variables tend to be similar or homogeneous, little variability is present. Moving to larger scales, where controlling variables and their combinations become heterogeneous, variability increases quickly to the scale break. After the scale break, variability saturates. Increasing only slightly with distance (up to 1km), processes at scales near

the scale break contribute similarly to the variance as processes at larger scales. Scale breaks of tens of meters correspond intuitively with the scales at which independent variables affect snow depth. For example the scale of stands of trees or of hill slope processes in complex terrain.

Together, the studies of Deems et al. (2006) and Trujillo et al. (2007) indicate that snow depth variability at scales of tens and hundreds of meters accounts for the majority of variability and provide impetus to understand what controls variability at these scales. We note that their analyses are spatially averaged, including that of scale breaks. This feature is perhaps one reason each met limited success in linking snow depth variability to controlling variables. In contrast to using average statistical descriptors, our study aims to understand and predict the explicit spatial structure of snow depth at these scales. Where Trujillo et al. (2009) decomposed spatial structure at one study site using subjectively-determined, rectilinear regions to uncover important differences between them (including vastly different scale breaks for the two regions), our study first decomposes irregular spatial structure via unsupervised learning and then attempts to predict this decomposition based solely on independent variables.

Several studies suggest that spatial structure of snow depth distribution depends primarily on static, controlling variables. Kirnbauer and Bloschl (1994); Sturm and Holmgren (1995) showed that snow distributions follow similar patterns from year to year and are largely independent of snowfall totals. Using variogram analysis in two different years at one site in our study, Deems et al. (2008) concluded that even though the temporal patterns of accumulation were very different in these years, spatial structure of snow depth was"largely determined by physiography and vegetation characteristics and [was] relatively insensitive to annual variations in snowfall."

Understanding and predicting snow depth spatial structure at these scales may have significant implications for modeling the hydrologic cycle and surface energy balance, especially considering the increasingly high-resolution nature of physical models. Where Liston (2004) proposed subgrid snow depth distributions based on regional air temperature, topographic variability, and windspeed, we use local (30m scale) predictor variables and their particular spatial distributions to estimate the spatial structure of snow depth variability below the 1 km scale.

4.3 Methodology

We first use cluster analysis to diagnose spatial structure of observed snow depth and qualitatively examine the hypothesis that snow depth spatial structure is governed by the spatial structure of independent variables. As a logical extension of this hypothesis, multivariate clustering of independent variables is then performed to predict snow depth spatial structure. Predicted spatial structures are then used in spatial modeling and compared against interpolation results from BRT to evaluate their equivalent observational cost in terms of BRT structure and quantify their utility. In this section, we describe the cluster analysis algorithm, it's appropriateness and limitations for revealing spatial structure of snow depth, cluster prediction of snow depth spatial structure, BRTs, and the measure by which we compare predicted and interpolated structures.

4.3.1 Hierarchical k-means clustering

The principle behind cluster analysis is, for k clusters of the data, to achieve a maximum separation in cluster means while minimizing within-cluster deviance. In the multivariate case, a metric (Euclidean distance) in n-dimensional space measures separation and deviance of the k clusters which partition an n-variabile by m-observation matrix. The clustering approach we use is typically referred to as k-means clustering and several algorithms exists (e.g. Everitt et al., 2001). Each selects initial cluster centers at random and then moves points between clusters, retaining only configurations which further optimize the clustering criteria. As such, clustering is not a deterministic process but for data sets with well defined clusters the algorithm usually converges. We performed multiple clustering runs on each data set and found reliable convergence. We also note that though individual variables may be weighted in k-means clustering, all were given equal weight in our analysis.

To solve the problem of how many clusters, k, should be considered, we perform hierarchical clustering in a binary (k=2) tree fashion, clustering is applied recursively in each resulting partition. Clustering continues until the improvement in mean squared error (MSE) resulting from an individual

clustering drops below 5% (10% for cluster prediction) of the MSE of the unpartitioned data set (calculated from residuals from subtracting the mean value of the entire domain). While this approach and even this last criteria are similar to the binary regression tree (BRT) algorithm (Breiman et al., 1984), differences in the overall approach and splitting criteria are fundamentally different and will be noted in the subsequent section on BRTs.

4.3.2 Snow depth structure and clustering

For spatially abrupt non-stationarity of snow depth, cluster analysis is well suited to partitioning the data into regions of different mean value because cluster separation is well defined. If non-stationarity results from spatial trend, cluster separation is not as well defined and cluster selection will hinge on smaller differences as trends become spatially smoother, though cluster analysis may still partition the space. Abrupt spatial non-stationarity of snow depth typically results from obvious, short scale changes in controlling variables. Spatial trends result from trend in a single controlling variable or from a protracted region of change between controlling variables. Though snow depth in this study is structured by each of these spatial mechanisms, cluster analysis is expected to perform well because the scale of investigation limits the importance of large-scale trends relative to that of smaller scale variability. Were data available at larger scales, where spatial trends in snow depth typically dominate the variability, cluster analysis may not perform as in this study.

So far we have ignored the notoriously noisy nature of snow depth while focusing on its structure. Through their difficulty in explaining a relatively large fraction of observed variability, previous studies suggest a large component of random variability in the spatial distribution of snow depth (e.g. variability not resulting from predictor variables) though the real reasons remain unclear (e.g. Erxleben et al., 2002). We refer to unexplained variability as noise and explained variability as ‘‘structure.’’

In terms of cluster analysis, within-cluster variability is noise and its magnitude can often approximate that of between-cluster variability, or structure. Thus, noise may represent a problem

for analysis of snow depth, particularly at high resolution where all original variability is present. Because we want cluster analysis to reveal the structured variability of snow depth and not its noisy nature, we build an nxm matrix for this purpose. Though we are clustering the one variable of snow depth, we build a multivariate matrix where the first variable is snow depth and the other variables are spatially smoothed copies of it. At high resolution, original observations are supplemented with 2 smoothed copies (m=3), with smoothing over 5x5 cells ($7.5m * 7.5m$) and 11x11 cells ($16.5m * 16.5m$). At low resolution, a single smoothed copy is attached to the original data (m=2), smoothing over a window of 3x3 cells ($90m * 90m$). Smoothing results in greater, artificial for our purposes, spatial autocorrelation in these variables, results in greater spatial coherency of cluster results, and serves to filter noise. After clusters are obtained, the original data sets are viewed through these clusters, for example, in figure 4.1.

Cluster analysis of snow depth simply diagnoses its spatial non-stationarity without regard for predictor variables. It is up to the analyst to infer the cause of structure via qualitative relationship with independent variables. For example, two regions with clearly different controlling variables can end up in the same cluster by virtue of having similar mean snow depths. Fortunately, clustering results correspond intuitively and unambiguously to independent variable controls at all but one site in this study.

4.3.3 Predictor non-stationarity and clustering: cluster prediction

If spatial structure of predictor variables and their combinations control snow depth spatial structure, then we should be able to cluster predictor variables to predict snow depth structure. We call this ‘‘cluster prediction.’’ Cluster prediction is well suited to revealing interactions of independent variables but does not involve snow depth values. We predict only the spatial structure or pattern of snow depth, so that it may be applied in the absence of observations. For practical reasons, we predict snow depth non-stationarity only at low resolution, which was determined by that of our readily-available predictor variables. Results using the ‘‘downloaded’’ and ‘‘LiDAR-measured’’ predictor sets will be compared to evaluate the effect of predictor quality on predictive

skill.

As mentioned, predictor clustering is performed on n-variable by m-observation matrices of predictor variables for each study location and cluster prediction terminates when a split explains less than 10% of the pre-partitioning MSE. Because predictor variables are in different units, variables must be normalized in a way which makes them comparable while preserving clustering. The largest range of each variable over all sites is used to normalize the difference of that variable and its minimum at each site. This scales each variable at each site to a sub-interval of the unit interval which begins at 0. At the site with the largest range for a given variable, this variable is scaled to the full unit interval. Cluster prediction suffers the same draw backs as cluster analysis of snow depth, it may not appropriately handle spatial trends in combinations of predictor variables which is not expected to be a problem at the scale of investigation.

At each site we face a poorly defined problem of which predictors should be used in cluster prediction. As no algorithm produced clear results, we compare 3 approaches to this problem. First, a ''basic'' setup uses only integrated potential solar radiation and vegetation density as predictors at all sites. Second, variables identified as important when understanding the non-stationarity of snow depth at each site constitute a ''qualitative'' setup. Last, we use a heuristic approach to discover a ''best'' set of predictor variables, which minimizes modeled snow depth error, though possibly for the wrong reasons. Results focus primarily on the ''qualitative'' setup using the others for comparison. Variables used in each setup are listed in table 4.2.

4.3.4 Binary regression trees and R-squared

We evaluate the skill of cluster prediction in the context of sampling design and spatial modeling experiments (described later) where predicted snow depth structures determine the distribution of samples supplied to a binary regression tree (BRT) model and compete with its spatial distribution structure. BRTs are the standard model used for interpolation of snow depth measurements at the first-order basin scale. In our case, a BRT fits a hierarchical relationship between a spatial response variable (snow depth) and vectors of collocated predictor variables. As

is often the case (e.g. Molotch et al., 2005), values are taken to represent grid areas and not just points. The BRT algorithm and fitting procedures are detailed in section B.2.1 of the appendix.

While hierarchical k-means clustering and BRTs have similarities on the surface, their significant difference is worth emphasizing. A BRT hierarchically stratifies a response variable based on splits or thresholds in predictor variables which minimize the MSE of the response. Cluster prediction works directly with the response while ignoring predictors, splitting the response variable to maximize separation of clusters while minimizing their spread. The splitting or clustering criteria is not only different but is applied in a very different way, without predictors.

However both cluster prediction and BRTs result in a partitioning of their spatial domain into levels which represent snow depth spatial non-stationarity or structure. We introduce values to cluster predicted structures via sampling design, modeling each region by its mean value just as is done with BRTs. To evaluate the skill of these structures and compare cluster prediction with BRT estimates, we focus on the R^2 metric. This is the standard measure of BRTs (e.g. Winstral et al., 2002; Erxleben et al., 2002; Molotch and Bales, 2005), and particularly of their spatial structre. R^2 compares model estimates against a null model, the mean value of all points used in fitting the model. A null model is a level 0 tree for both models. R^2 gives the fraction of unexplained variance from null model which is accounted for by the model being evaluated. The formulation used here, also known as the Nash-Sutcliffe model efficiency and given in section 3.3, goes negative if the model introduces greater residuals than the mean value of observations.

4.4 Site snow depth non-stationarity

At individual sites, interpretation of cluster analysis of snow depth reveals its relationship to independent variables. We primarily focus on high-resolution results in figure 4.1 though we do consider the effect of spatial resolution on the results. In figure 4.3, the first column at each site shows cluster analysis on the low-resolution snow depth. Change of resolution at two sites compromises small-scale structure, though low-resolution analysis clarifies larger scale non-stationarity.

Figure 4.2 reveals how full site variograms are decomposed by cluster analysis. Scale breaks

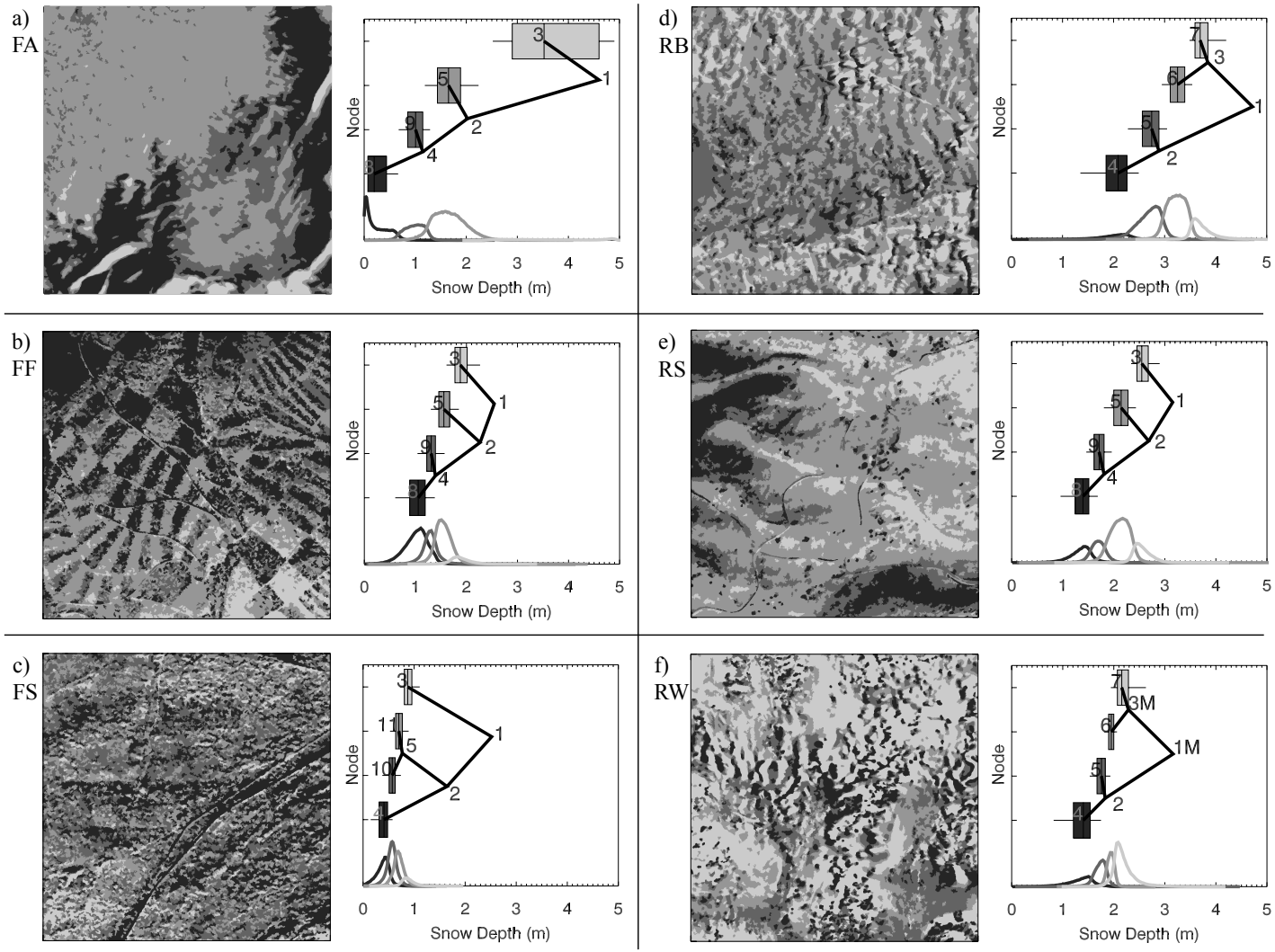


Figure 4.1: Hierarchical k-means cluster analysis of high-resolution snow depth. Binary cluster trees are traced onto boxplots of the terminal nodes. Nodes are labeled by index in a complete binary tree. Histograms corresponding to the boxplots show the distribution with in each of the final nodes along with relative frequency of occurrence of snow depths.

are significantly shorter and semivariance typically constant thereafter for individual clusters. While there are several exceptions in areas of strong variability, results indicate cluster analysis appropriately decomposes snow depth spatial non-stationarity. Table 2.1 provides sites' snow depth and physiographic statistics discussed in this section and table 4.1 provides cluster mean snow depth and values of independent variables.

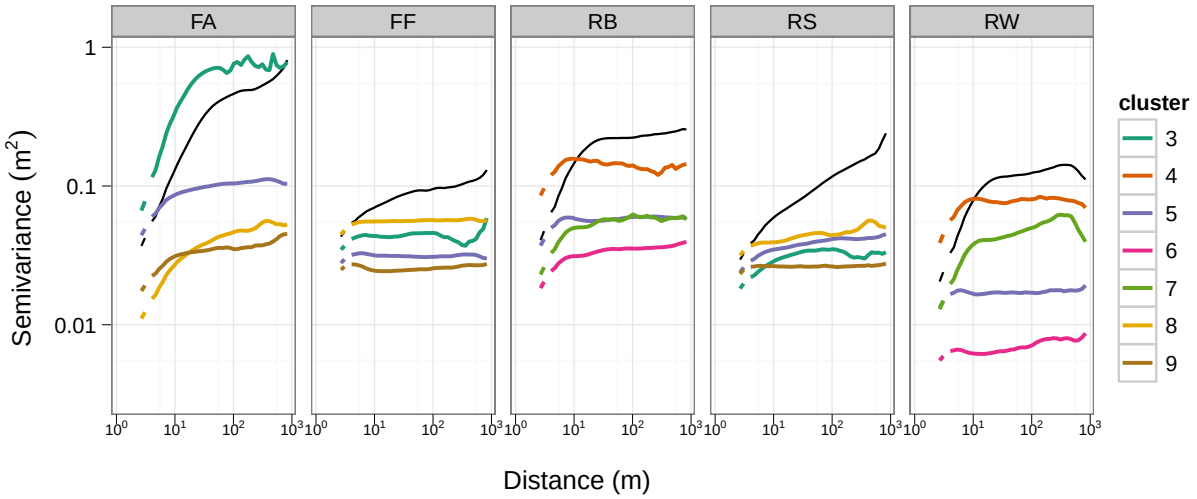


Figure 4.2: Log-log semivariograms of high-resolution data at each site and its clusters. Numbers correspond to clusters in 4.1. No results shown for FS where spatial structure was considered superflous. The gap in the variograms at short range is due to calculations on a regular grid.

4.4.1 FA

The bimodal snow depth histogram (figure 2.6) at FA indicates spatial structure even before looking at the spatial pattern. Spatial structure is essentially unchanged between high and low resolution because there is no significant spatial structure below the 30m scales of averaging.

In figure 4.1a we see cluster analysis partitioning the site into 4 clusters with most of the data falling in clusters 5 and 8. These clusters correspond to the modes of the snow depth histogram and to the primary physical controls on snow depth, wind exposure and sheltering. Cluster 5, with mean snow depth around 1.75m, exists both above and below tree line (at approximately 3575m, as

seen in figure 2.3). Below tree line, trees shelter snow from wind and 5 is the only significant cluster. Some lower depths seen on its western edge may be evidence of spatial trend below tree line (due to elevation trend, i.e. Trujillo et al., 2009) which is not well resolved by clustering (note the large variance of this cluster). Cluster 5 is also found above tree line in a smaller area where snow collects on wind sheltered aspects. These two wind-sheltered areas are in the same cluster because they have similar mean snow depths even though they are controlled by different physical processes.

Cluster 8, with most if its snow depths near 0m, is the largest cluster above tree line and represents where wind-scour dominates snow depth distribution. Wind comes predominantly from the northwest (figure 2.5) and cluster 8 consists primarily of northwest facing slopes (table 4.1) sheltered neither by trees nor by other terrain (as cluster 5 above tree line). From clusters 5 and 8, we see that the primary controls on snow depth at this site are wind, vegetation, aspect, and their interactions.

Clusters 3 and 9 represent smaller areas of secondary importance. Cluster 3 corresponds to deep terrain pockets near the ridge top which fill with snow over the accumulation period both from snowfall and from wind redistribution over the top of the ridge. Cluster 3 has extremely high snow depth and variance but occupies very little area. Because the region maximizes difference in cluster means, compared to any other split, it is the first split made by the clustering algorithm. Cluster 9 lies between clusters 5 and 8 both spatially and in the histograms and boxplots. We assume it represents a transition zone between these areas of wind sheltering and exposure which transition only as abruptly as aspect.

4.4.2 FF

Despite a normal-looking snow depth histogram, figure 4.1b reveals a geometric, man-made vegetation pattern governing spatial non-stationarity of snow depth at FF. The forest at FF was logged in the 1950s (Erxleben et al., 2002), leaving tree stands of different heights. Table 4.1 indicates an inverse relationship between snow depth and vegetation height. Because this site is predominately north-facing and well below the ridge top, vegetation acts to reduce collocated snow

depth via both canopy interception and longwave radiation rather than shielding it from wind and solar radiation.

Closer analysis reveals elevation gradient interacting with vegetation controls. Increasing snow depth with elevation can also be seen in figure 4.1. Sloping downhill to the north, FF has the second largest elevation range of all sites in this study (after FA). On its northern, downhill edge, we see only the lowest snow depth clusters, 8 and 9, representing snow depth structure. On the southern, uphill edge, structure is comprised by higher mean snow depth clusters 3 and 5. Between these elevation extremes and over the majority of the site, snow depth structure is represented by clusters 8 and 5.

Vegetation and elevation non-stationarity and their interaction clearly govern snow depth spatial structure at FF. Solar exposure plays no obvious role because all aspects are northeast or northwest facing on reasonably steep slopes. The scale of snow depth structure is mostly larger than the scales of averaging and snow depth structure is similar at high and low resolution.

4.4.3 FS

Clustering results at FS highlight the need for interpretation. Though the algorithm uncovers 4 clusters (figure 4.1c), no structured variability in snow depth exists and the site should be modeled by a zeroth order tree. Because the site has low overall variance which competes with LiDAR measurement error, there is little separation in clusters which resemble noise in their spatial pattern. Relating individual clusters to physical controls would be meaningless, which matches our understanding of independent variable controls at this nearly flat, densely forested site; there is no appreciable non-stationarity in independent variables to structure snow depth distribution. (Though we do see the lowest snow depths on the plowed road.) At low resolution, cluster separation (not shown) is even less due to the overall reduction in variability though apparent noise is reduced when using the smoothed snow depth.

Because standard deviation of snow depth and vegetation height are very low and approximate the LiDAR instrument error ($< 30cm$), scanning errors dependent on flight direction appear in

striped east-west patterns of snow depth and vegetation plots (figure 2.3). This implies that at least the September (used for vegetation) measurements had measurement bias dependent on flight direction.

4.4.4 RB

High-resolution cluster analysis at RB (figure 4.1d) indicates the importance of ribbon forest in determining snow depth spatial structure. This structure is all but missing in cluster analysis of low-resolution snow depth (figure 4.3) because the scales of variability induced by the ribbon forest are smaller than the grid scale. Though snow depth structure at larger scales resulting from solar exposure and aspect is preserved when degrading resolution, low resolution data misses nearly ubiquitous, small-scale variability at this site.

High-resolution clusters 4 and 5 match the pattern of ribbon forest in the vegetation map for RB in figure 2.3. (Note that the band of lower vegetation height cutting across the bottom of the site in this figure reflects installation of power lines.) As at FF, an inverse relationship is also seen between snow depth and vegetation height and can be seen in table 4.1. This indicates the importance of canopy interception and longwave radiative effects on snow during the accumulation period, even at this site with primarily southwest aspect. Clusters collocated with ribbon forest correspond to the skewed lower limb of the high-resolution snow depth histogram at RB. Change of resolution erases this structure which explains the more symmetric low-resolution histogram and the greatest decrease in standard deviation between high and low resolution for all sites in this study.

Clusters 6 and 7 represent areas with lower vegetation and higher snow depths (table 4.1). Cluster 7 represents the deepest snow in the site. On the windward, west-facing slopes located predominantly in the northern third of the site, cluster 7 represents deep drifts behind stands of trees, resulting from interaction of aspect, vegetation, and wind. In the southern quarter of the site and in a small area in the northwest corner, northerly aspect combines with drifting to result in more continuous areas of deep snow, also cluster 7. Cluster 6 accounts for the remainder of the aspects and cluster 7 is excluded from the steeper southwest-facing slopes, found across the middle

of the site, where lower overall snow depths result from greater solar exposure. Clustering at low resolution reveals these effects of aspect more clearly.

4.4.5 RS

The snow depth histogram at RS suggests spatial non-stationarity. Degrading spatial resolution has little effect on the spread of the histogram and indicates snow depth structure at scales larger than $30m$. The RS site has the most complex and undulating terrain of all the sites, aspect and solar radiation play important roles at this site. There is also strong association between vegetation and aspect, with vegetation primarily on south and southwest facing aspects. Canopy interception on these aspects greatly reduces snow depth relative to the north-facing, unvegetated areas at the site.

The lowest snow depth cluster figure 4.1e, 8, corresponds to southerly vegetated aspects (table 4.1). The second lowest cluster, 9, corresponds to vegetated transition to other aspects from south (mode happens to be west-southwest). The highest snow depth cluster, 3, corresponds to north-facing and unvegetated northwest-facing slopes. Cluster 5 picks up snow of intermediate depth, found mostly on sparsely vegetated, northwest-facing slopes, between the more extreme south and north aspects. Clustering is essentially identical for high and low resolution peak accumulation snow depths. The principle predictors at RS are potential solar radiation, aspect, and vegetation. Non-stationarity in these controlling variables is at hill slope scales at RS, on scales larger than the 30m low-resolution pixels.

4.4.6 RW

Snow depth at RW has the least variance of any site in the R group. With relatively flat terrain and sparse vegetation, the primary controls over snow depth structure at RW appear to be wind interaction with vegetation or small-scale terrain features, though interpretation of results is not clear. As discussed in chapter 2, we suspect systematic LiDAR measurement error at RW due to existence of shrubbery known to produce inaccurate surface elevations.

		FA		
	Node 8	Node 9	Node 5	Node 3
Snow Depth	0.3	1.0	1.7	3.7
Elevation	3625.0	3590.5	3529.0	3625.6
Veg Density	0.1	1.0	3.5	0.1
Aspect	120	130	140	320
Sx	-1.7	-0.5	-2.8	14.6
		FF		
	Node 8	Node 9	Node 5	Node 3
Snow Depth	1.0	1.3	1.6	1.9
Elevation	3121	3136	3165	3210
Veg Density	8.2	6.0	4.3	3.1
		FS		
	Node 4	Node 10	Node 11	Node 3
Snow Depth	0.4	0.6	0.7	0.9
Veg Density	8.7	9.1	7.6	6.6
		RB		
	Node 4	Node 5	Node 6	Node 7
Snow Depth	2.0	2.7	3.2	3.7
Veg Density	3.4	1.5	.5	.6
Potential Solar	577128	560576	520648	440734
Aspect	260	250	230	150
Sx	-0.4	-0.2	0.1	0.9
		RS		
	Node 8	Node 9	Node 5	Node 3
Snow Depth	1.4	1.7	2.1	2.6
Veg Density	6.1	6.3	3.6	1.3
Potential Solar	595465	524375	393467	293722
Aspect	260	190	150	100
		RW		
	Node 4	Node 5	Node 6	Node 7
Snow Depth	1.4	1.7	1.9	2.2
Veg Density	3.4	1.5	0.5	0.6
Potential Solar	526235	529854	529313	513039
Sx	3.3	6.4	9.7	8.5

Table 4.1: Statistics of independent variables within cluster nodes at each site as shown in figure 4.1. All are mean values except for aspect which is the mode over 10 degree bins and measured in degrees CCW from east. Snow depth and vegetation density are measured in meters. Potential accumulated solar radiation is measured in $Wh/m^2/day$ and maximum upwind slope in degrees from horizontal.

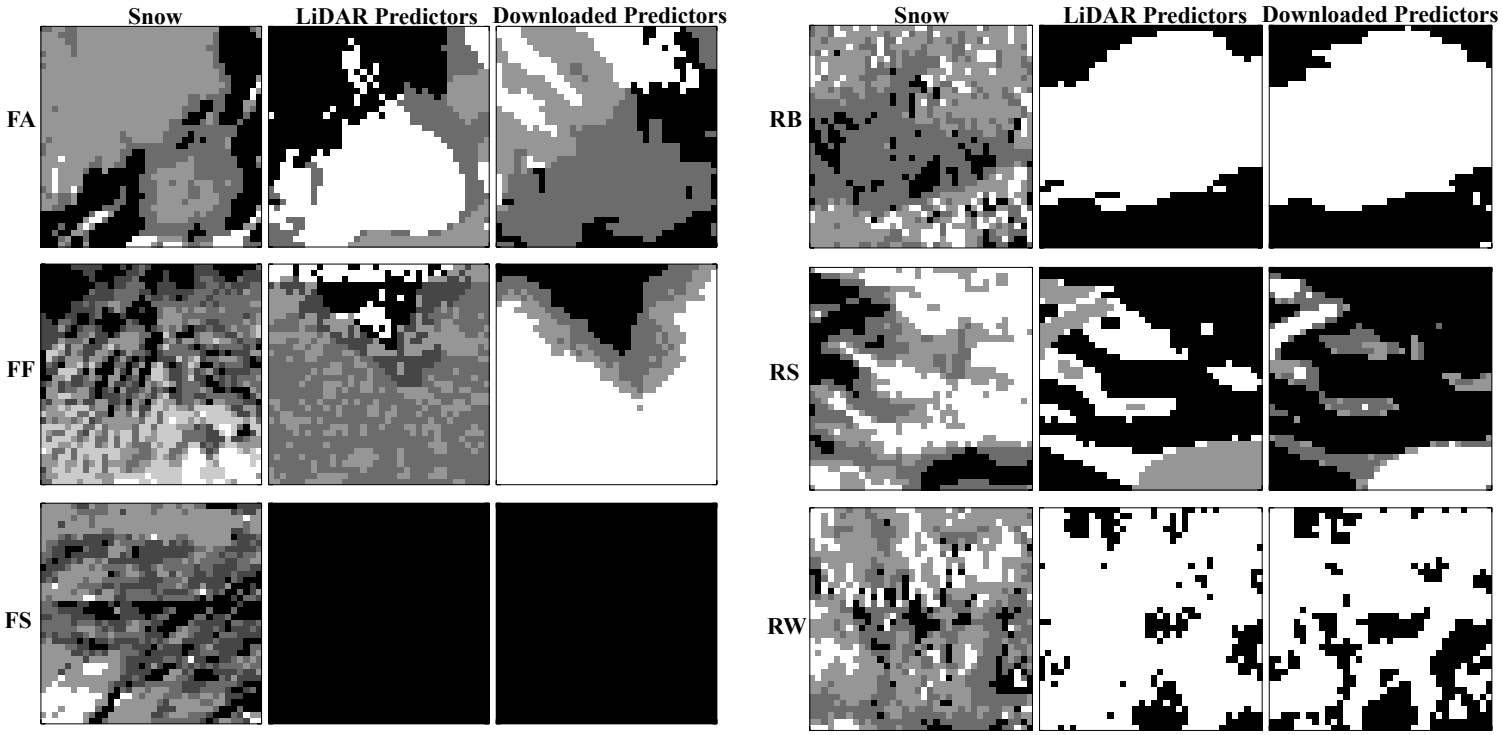
The lowest snow depths (cluster 4) often, but not always, correspond to the location of vegetation. Throughout the central and western portion of the site, the deepest snow depths (cluster 7) are found immediately on the leeward side of the shallowest snow depths (cluster 4) in short-scale undulations with a range of snow depth approximately 65cm. It is difficult to discern if these undulations are caused by small scale terrain or vegetation features or if these result from wind-snow interaction as with sustrugi or if they may result from LiDAR errors.

Other large scale controls on snow depth structure are not apparent at RW. The site is nearly flat with average slope of about 6 degrees and a generally south-facing aspect. Low-resolution cluster analysis (figure 4.3(b)) only hints at a relationship with vegetation and does not reflect the strong, small-scale variability present in the high-resolution clustering. For this reason, as at RB, change of resolution reduces skew in the lower limb of the high-resolution snow depth histogram.

4.5 Predictor non-stationarity and prediction of snow depth non-stationarity

In this section, we cluster independent variables as a means of predicting snow depth spatial structure. Figure 4.3 shows clustering of low-resolution snow depth next to cluster predicted structure. On the left, low-resolution snow depth is clustered. Predictions in the middle column are based on ‘‘LiDAR-measured’’ predictors and those in the right column from ‘‘downloaded’’ predictors. Predictions in the figure are based on the best predictor variables found by heuristically checking all combinations of predictors. Best predictors are visually most similar to clustered snow depth, but we will see that qualitatively selected predictors perform nearly as well by quantitative measure. Variables used in each of these predictors sets are listed in table 4.2. Note that snow depth clustering in figure 4.3 is ordered from black to white by increasing mean snow depth but the color of predicted clusters is arbitrary because their order is not predicted.

While this figure indicates that predicted patterns may be useful for estimating snow depth structure, we quantitatively evaluate their connection in this section. The primary question we



(a) F clusters

(b) R clusters

Figure 4.3: Snow depth clustering and cluster prediction at low resolution. Cluster prediction is performed for ‘‘best’’ predictors on LiDAR-measured and downloaded variable sets.

pose is how does the spatial structure of cluster prediction compete with spatial structure inferred from observations by a binary regression tree model? In cluster prediction, a model's spatial structure is determined a priori of snow depth data and fixed, but snow depth values are not predicted. To bring values into predicted clusters, we take snow depth observations and use the mean value in each cluster to model its snow depth. This is analogous to how a BRT estimates snow depth in its terminal nodes, using the mean value of observations fit to that node. However a BRT arrives at its spatial structure in a different way which evolves with data. Because the spatial structure and skill of a BRT depends on the number of observations whereas the spatial structure of our cluster prediction model is fixed in advance of observations, our question is: how many observations does the a priori structure of a cluster prediction model represent in terms of a binary regression tree? Or: How many observations does a BRT need to attain an equivalent structure to that of a priori cluster prediction; how many observations do we get for ''free'' from cluster prediction in terms of its structure? Comparison will also indicate how much better a BRT can do if we are willing or able to collect more observations.

In our sampling design experiment, observations are drawn from the high-resolution data, mimicking measurements with a snow probe. (Each observation actually represents $(1.5m)^2 = 2.25m^2$, which may still have lower variability than true point measurements. We'll assume variability at these scales is small enough to be negligible for our purposes.) For each model, we collect observations in slightly different ways, with the following justification. For the BRT model, we sample randomly over each site because this approach assumes no prior knowledge of the snow depth distribution. Because the premise of the cluster prediction model is that each region has a stationary mean value and we predict these in advance of data, we divide up the total number of observations at each site by the number of clusters predicted by the model. (For lack of information, we assume equal variance within clusters.)

Because different sampling designs are used for BRT and cluster prediction, we first investigate the degree to which results are affected by sampling design: Is BRT put at a disadvantage by using random sampling? This is answered by comparing BRT results using both sampling designs, which

considers if cluster predicted sampling design might improve BRT estimates when the two are used in conjunction.

Figure 4.4 displays differences in true R^2 (not apparent R^2) resulting from random (pastel colors) and cluster prediction (primary colors) sampling designs. Results for the two designs are similar up to 100-200 observations, depending on site, for both predictor sets. After this, the random sampling method results in increased skill, particularly for LiDAR-measured predictors. Because observations are divided evenly between clusters, small clusters end up being sampled redundantly at high observation counts at the cost of larger clusters being sampled less thoroughly. This compromises independent information presented to the model. Degradation of skill is typically less for downloaded predictors because they already lack the accuracy and detail needed by the model to fully exploit observations.

While we might expect cluster based sampling to improve R^2 at lower observation counts and not higher ones, results also indicate that the BRT model needs some minimum observation count to correctly infer the spatial structure behind in sampling design, regardless of sampling design. Because BRTs are fit using 100 runs of 10-fold cross-validation, sample sizes must be large enough to overcome uncertainty in all these runs. Thus, at no time does random sampling put the BRT at a disadvantage.

The above does not imply that sampling design based on cluster prediction has no utility at all. Random sampling covering the domain is simply competitive. However, such coverage of the spatial domain is often not possible in practice. In such cases, cluster prediction might be used in advance of sampling to delineate regions with different expected controls on snow depth and help direct observations into regions which may otherwise be skipped, so that more representative sampling is performed. We do not consider such application of cluster prediction here.

4.5.1 Application to spatial modeling

Figure 4.5 compares BRT and cluster prediction model skill as a function of the number of observations using both LiDAR-measured and downloaded predictor sets. Cluster prediction skill

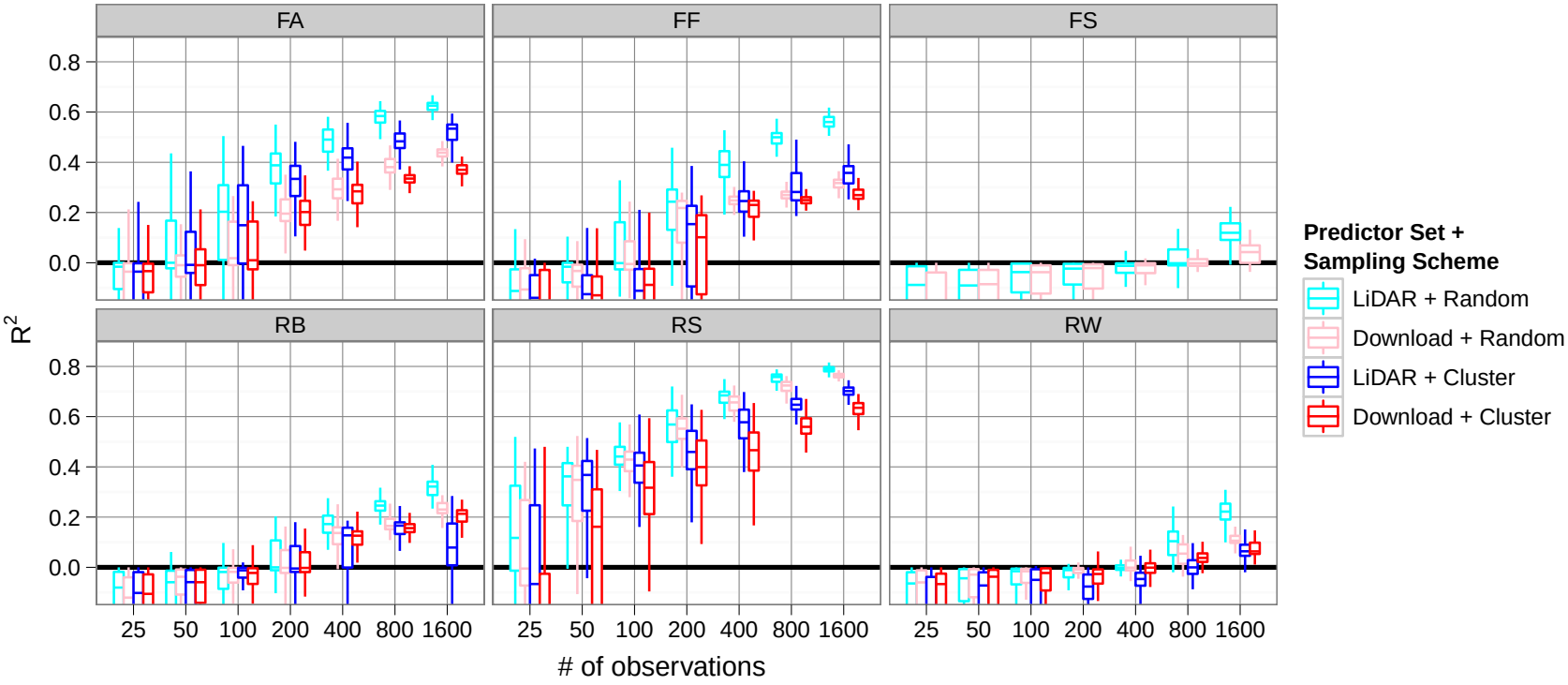


Figure 4.4: True R^2 comparison for binary regression tree estimates using random (pastel colors) and cluster predicted (primary colors) sampling schemes with LiDAR-measured and downloaded predictor sets. Cluster prediction is based on qualitative predictor selection. Cluster prediction was judged unnecessary at FS, thus its prediction skill is 0.

is shown in primary colors and BRT skill in pastel colors.

Revealing the fixed and evolving spatial structures of the respective models, figure 4.5 shows median cluster prediction skill increasing much less than BRT skill with the number of observations. Though its structure is fixed, cluster prediction skill increases somewhat with the number of observations. As cluster means become more accurate with more observations, median R^2 and its uncertainty due to random sampling diminish quickly. The skill of cluster predictions for 1600 observations always lies within the range of cluster predictions for 25 observations, and often within its inner-quartile range. The same cannot be said for the BRT model, highlighting dependence of its spatial structure on the number of observations.

For downloaded predictors, median R^2 of BRT models intersects that of cluster prediction between 100-200 observations at FA, RS, and FF. At RB and RW, more observations are required for BRT to match the R^2 of cluster prediction, 200-400 at RB and 800-1600 at RW (and FS where the null model is the cluster model). Results are similar for the LiDAR-measured predictors, with minor differences at FA, FF, and RW.

At sites with low BRT skill versus number of observations, cluster prediction presents an efficient way of estimating snow depth structure. Poor BRT skill means that more observations are required to match the skill of cluster prediction and implies at least one of poor predictors or lack of snow depth spatial structure. At RB and RW, there's evidence that poor predictors result because subgrid variability translates to a noisy predictor-response relationship at low resolution. FS represents lack of spatial structure and estimation from downloaded predictors requires 1600 observations to raise median R^2 above the null model, which is actually the model chosen by cluster prediction.

In contrast, at sites such as FA, FF, and RS, with clear spatial structure and where BRT performs well, very few observations are required for BRT to match the skill of cluster prediction. Still, cluster prediction model structures always result in positive R^2 worth at least 100-200 observations per square kilometer. This observation count represents a fair amount of effort on the ground. Manual probing of snow depth often comes up short of this observation density and

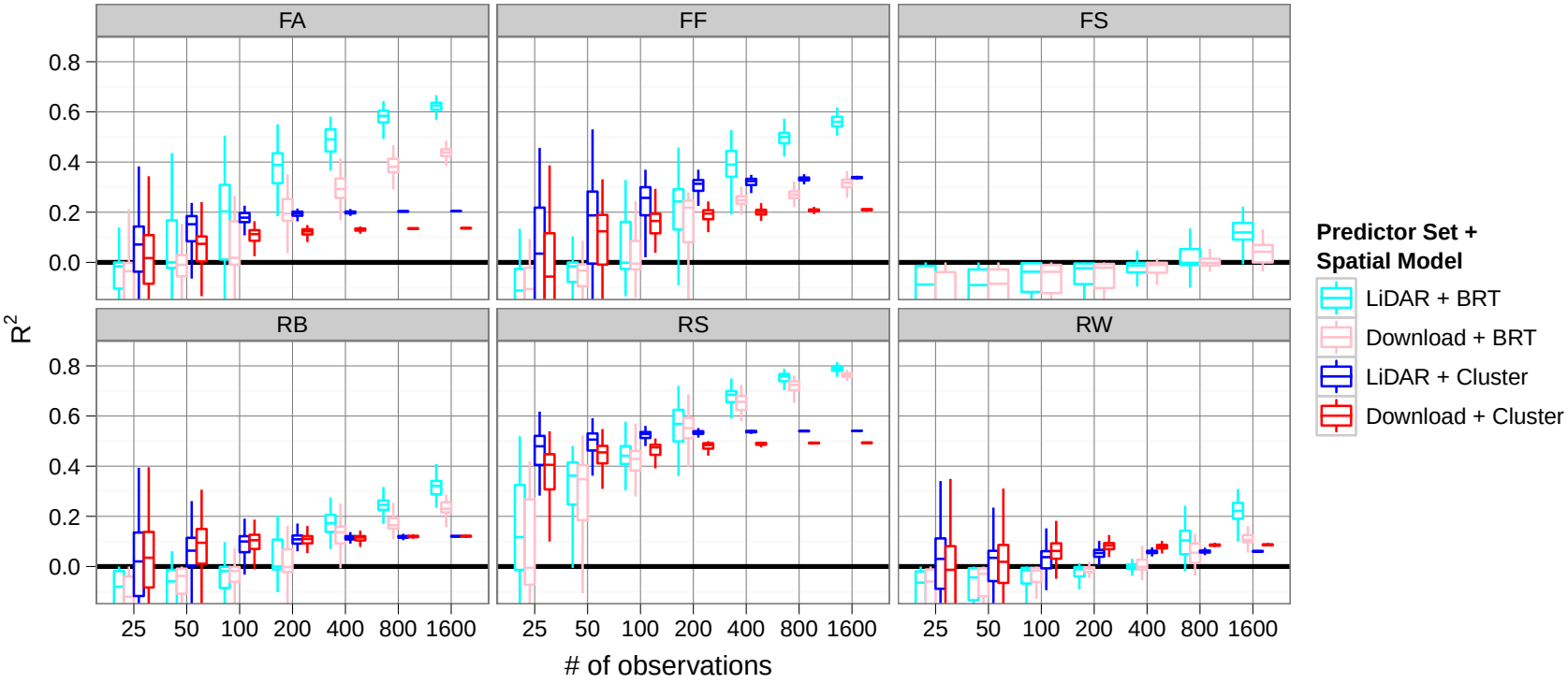


Figure 4.5: R^2 comparison of spatial modeling by cluster prediction (primary colors) and binary regression trees (pastel colors), for both LiDAR-measured (blue) and downloaded (red) predictor sets. Cluster prediction is based on qualitative predictor selection. Cluster prediction was judged unnecessary at FS, its prediction is skill is 0.

cluster prediction may still be of use for spatial modeling in such situations, especially when only downloaded predictors are available - as described momentarily.

Given the low skill of cluster prediction at several sites, it is reasonable to question if cluster prediction skill results merely from regioned averaging in a domain with significant spatial autocorrelation. Figure 4.5 suggests skill because spatial patterns appear appropriate. To quantify how much skill can be expected due to spatial autocorrelation, we divide snow depth at each site into blocks larger than their correlation length scales, shuffle these blocks randomly 1000 times, and calculate skill of the cluster predicted patterns for each shuffle. As spatial autocorrelation is at scales less than 50 m (figure 4.2), we shuffle $2x2$ ($60mx60m$) blocks. Median R^2 from all shuffles is approximately 0 at all sites and 95th percentiles are reported in the ''shuffle .95'' column of table 4.2. In all but 2 cases (each using basic predictors), skill obtained from cluster prediction well exceeds the confidence limits of the R^2 resulting from shuffling, indicating that cluster prediction skill does not result from mere chance.

Differences in skill between LiDAR-measured and downloaded predictors indicate importance of predictor quality and the fact that better predictors are leveraged more by BRT. Except for cluster prediction at RW, LiDAR-measured predictors result in more skill in both prediction and estimation. Differences between LiDAR-measured and downloaded predictor skill are largest for BRT estimates at high observation densities which indicates that more accurate, detailed information is used by BRT when data support it. Downloaded predictors provide only slightly less skill in cluster prediction apparently because they contain most of the first-order structure used by the model. For cluster prediction, differences between LiDAR-measured and downloaded predictors are largest at FF where downloaded data lacks the geometric logging pattern seen in the previous section.

In the common scenario when detailed, LiDAR-measured data is not available for fitting detailed models, cluster prediction represents a more useful tool. At all sites but FA, cluster prediction spatial structure derived from qualitatively selected, downloaded predictors captures at least 50% of BRT skill at 1600 observations. Even provided more data, R^2 scores at 1600

Predictor Set	**Predictors**	R^2		**Shuffle .95**		% BRT R^2	
		LiDAR-measured	downloaded	LiDAR-measured	downloaded	LiDAR-measured	downloaded
	FA						
Basic	veg, sun	8.9	15.4	3.9	5.5	14	35
Qualitative	ele, vegdense, asp, sx	20.5	13.7	5.5	4.6	32	31
Best	ele, sx	22.8	17.2	6.5	5.6	36	39
	FF						
Basic	veg, sun	24.4	0.3	5.6	.23	42	0
Qualitative	vegdense, ele	31.0	21.3	7.4	6.3	53	65
Best	veg, vegdense	35.8	5.7	6.9	2.2	61	17.2
	FS						
Basic	veg, sun	1.6	0.0	1.2	0.0	10	-
Qualitative	vegdense	0.0	0.0	0.0	0.0	-	-
Best	ele, n	11.6	5.6	4.2	2.8	73	-
	RB						
Basic	veg, sun	9.5	0.0	3.0	0.0	28	0
Qualitative	sun, asp, sx	12.2	12.2	4.1	4.2	36	50
Best	ele, sun	25.6	24.4	5.7	5.8	76	97
	RS						
Basic	veg, sun	38.3	22.7	7.8	5.9	48	30
Qualitative	vegdense, sun, asp	53.2	49.5	10.8	10.8	67	70
Best	sx, sun, slp, asp	53.6	52.2	10.2	11.2	67	69
	RW						
Basic	veg, sun	8.6	6.5	3.1	2.7	34	53
Qualitative	vegdense, sun, sx	6.2	8.8	2.4	3.1	25	71
Best	veg, sun, n	15.9	6.5	4.1	2.6	63	53

Table 4.2: R^2 for three methods of predictor selection at each site for LiDAR-measured and downloaded predictor variable sets. Variables are elevation (ele), vegetation height (veg), vegetation density (vegdense), aspect (asp), slope (slp), potential accumulated solar radiation (sun), maximum upwind slope (sx), and northness (n). The "Shuffle .95" column gives a 95% confidence limit on the maximum amount of R^2 resulting from autocorrelation. Percent of BRT skill is calculated at 1600 observations and compared for identical predictor sets.

observations are virtually as well as a BRT will be able to do with downloaded predictors. By contrast, no observations were required for cluster prediction. When LiDAR-measured predictors are available, cluster prediction captures a smaller fraction of BRT skill at 1600 observations because BRT has more accurate information and makes better estimates. Using LiDAR-measured predictors, comparing results for cluster prediction and BRT at 1600 observations is equivalent to comparing tree structures with 2-5 and 10-12 terminal nodes.

We have focused on results for qualitatively selected predictors. Table 4.2 presents results for basic and best predictor sets as well. Generally, the qualitative set improves upon the basic set and the best set improves upon the qualitative. Best predictor sets were found heuristically using LiDAR-measured predictors. Thus, in 2 cases, the qualitative set performs better than ''best'' when using downloaded predictors. This partially indicates that ''best'' predictors may not be so for the right reasons, they are simply combinations which minimize modeled snow depth residuals. Overall, qualitative predictor selection performs very competitively with the best predictor set. Though basic predictor selection does reasonably well, results indicate that substantial skill is gained by qualitative variable selection in most cases.

We note that even though cluster analysis of high-resolution snow depth informed our choice of ''qualitative'' predictor sets used in cluster prediction, the selection of predictors was generally intuitive and might have been determined independently of the analysis by someone familiar with basic knowledge of the sites. The serious research challenge lies in automating selection of these variables and developing a quantitative way of selecting the most appropriate predictors.

4.6 Conclusions

Results validate the hypothesis that spatial structure of independent variables controls that of snow depth. Cluster analysis of snow depth revealed spatial structure matching qualitative understanding of independent controls. Cluster analysis of controlling variables, termed cluster prediction, showed significant skill in predicting snow depth spatial structure, capturing first order-structure without the need for snow depth observations or even LiDAR-measured measured

predictor variables. Binary regression trees required 100-200 observations per square kilometer to match skill of cluster prediction (and considerably more in locations with less snow depth spatial structure). Cluster prediction is particularly useful when only crude predictor variables are available. In this case, at all but one site, skill of cluster prediction matched at least 50% of BRT skill using approximately 1600 observations per square kilometer, a very high sampling density for ground-based campaigns. Though the problem of predictor selection for cluster prediction was not solved, this represents an important direction for future research if results are to be automated and applied over large areas. For small studies, qualitative predictor selection appears to be satisfactory.

This chapter presents a particular view of spatial distribution of snow cover and its variability at scales of 10-100 of meters which lends itself to prediction of spatial patterns of snow depth distribution. While this view is not 100% accurate, we demonstrate its significant utility in prediction of snow depth spatial structure from readily-available, independent variables. Results can be applied without snow depth observations to estimate 1st-order spatial distribution patterns of snow depth, with skill level requiring approximately 100-200 observations in the context of a binary regression tree model, anywhere 30m resolution predictor variables are available. We hope this view and approach to prediction can be applied, refined, and extended in other applications. More case studies in a variety of snow packs and physiographic situations will be needed to establish general utility. We have also noted that results may be strongly dependent the scale of investigation. Moving to larger scales where spatial trends represent a greater share of variability is expected to change results.

We've highlighted application to spatial modeling in order to give quantitative evaluation of our prediction method. Cluster prediction presented here has potential application to statistical downscaling of snow precipitation for hydrological models and to snow-hydrologic modeling. However a crucial step will require prediction of cluster mean snow depth or at least its order.

Chapter 5

Spatial Structure and Hydrologic Modeling

5.1 Introduction

The problem of prediction in ungaged basins (PUBS) (Sivapalan, 2003) is perhaps the greatest challenge to theoretical hydrology. In the PUBS setting, free model parameters pose a significant challenge because they must be determined without calibration. While objective methods exists for parameter transferability and regionalization, parameters are often tuned by hand in practice. In this case, not only does the number of free model parameters need to be minimized because of their often competing effects, it is desirable that free parameters might be intuitively selected to produce accurate results.

In this study, a simple semi-distributed model of snow melt runoff spatial variability is presented. Because the spatial distribution of this model rests only upon free parameters added to a 1-dimensional, point model of snow, its parameterization is comparable to the sub-grid areal depletion curve (ADC) approach typically used to extend point scale modeling areally for runoff estimation. Spatial modeling units in our model are derived from predicted spatial patterns developed in chapter 4 using readily-downloaded predictors and independent of snow observations. Modeling is performed in a $2.9km^2$ basin in southwest Colorado, USA.

Our study begins from the premise that reliable parameters can be obtained for a point model of SWE at some location nearby the study basin where model forcings and response have some physical connection to the modeled basin. Though we use point observations from the bottom of our basin, this study might also be conducted with data from a nearby SNOTEL site. We take

data at such a point as given and focus on how to most reliably and accurately extend calibrated parameters from that point to the basin of interest while adding minimal parameters with clear interpretation.

We demonstrate that a linear relationship between snow model parameters (controlling spatial distribution of melt rates and SWE) and mean predictor variables in key physiographic regions extends the point model in a more intuitive way, while using fewer parameters, than an ADC. Results indicate that the proposed method produces a significantly better result virtually all of the time. This is important when model paramaters must be subjectively determined but is hardly surprising considering the amount of spatial predictor information introduced into the model.

5.2 Background

Entering the era of distributed snow models, Kirnbauer et al. (1994) made two key observations, one for distributed models and the other against, which we attempt to balance in this work. First, the argument against distributed models warns of using massive amounts of parameters in a model. Uniquely determining parameters against the integrated measure of streamflow is unlikely as their number grows and parameter calibration using optimization methods has no physical basis; we almost assuredly will produce the ‘‘right answers for the wrong reasons’’ by tuning a large number of parameters in a distributed model (e.g. Beven, 2006). Even the problem of distributing forcing variables to a spatially distributed model introduces assumptions and parameters, such as lapse rates and elevation gradients, which are then balanced against choice of other model free parameters.

The second observation of Kirnbauer et al. (1994, p 10), in favor of distributed models, is that ‘‘effective parameters (and an equivalent homogeneous system) do not always exist, particularly when the processes are non-linear (as is the case here).’’ In the context of our paper, the non-linear subgrid parameterization relating modeled snow water equivalent (SWE) to snow covered area (SCA), the ADC, is traded for spatially distributed, linear parameterizations. The aim of this trade is to clarify intuitive or physically based selection of free model parameters while simultaneously reducing sensitivity of results to sub-optimal choices.

Helping to balance these two mandates, of minimizing tunable model parameters while moving to a spatially distributed model, parameter complexity is independent of spatial resolution in the model we develop. Because overall model efficiency is also important, we turn to the previous chapter when designing spatial modeling units. Cluster prediction, developed for peak-accumulation snow depth, is extended in this chapter to account for spatial distribution of both accumulation and melt rates at a scale which resolves primary variability in these processes between grid elements while assuming homogenous subgrid variability. Cluster prediction presents a method for decomposing variability and process heterogeneity in a way similar to hydrologic response units, HRUs (Leavesley et al., 2002), but for snow.

The first study to spatially distribute a temperature index model in a way similar to that considered in this paper was by Cazorzi and Fontana (1996). They explicitly avoided ''troublesome'' depletion curves by distributing melt rates over a regular grid on an hourly basis via the following equation

$$M_{h,j} = CMF * EI_{h,j} * T_{h,j}$$

where CMF is a common, hourly melt factor to all grid elements, $T_{h,j}$ hourly temperature at grid cell j, and $EI_{h,j}$ is the energy index applied hourly, though its value is only based on monthly means of daily total solar radiation at each point and is set to a minimum value at night to simulate the diurnal cycle. Note that CMF is the only parameter in this formulation, which replaces the standard melt factor. Their study simulated snow pack observations and the inclusion of radiation presented an advantage over a spatially distributed temperature-index model. Both used spatially distributed forcing variables.

Hock (1999) proposed a method nearly identical to this, using a slightly different formulation

$$M_{h,j} = (MF + aI_{h,j})T_{h,j}$$

where the melt factor, MF, is adjusted by a radiation coefficient times the parameter I which is the hourly accumulated potential solar radiation. A second formulation, altering I to include global measured radiation, was not found to greatly improve simulations. Using this formula,

their calibrated, distributed model at 30m resolution achieved streamflow R^2 efficiency of .8 and reproduced diurnal cycles observed in the hydrograph.

By spatially distributing model melt rate parameterization based on expected distribution of solar radiation, Cazorzi and Fontana (1996) and Hock (1999) avoid the use of depletion curves. But their approach requires calculation of distributed temperature and precipitation forcings. Instead of generating distributed forcings, we choose to distribute model parameters in a way which describes expected spatial variability in forcing variables as well. Instead of stashing assumptions in an ''external'' set of forcing variables, placing them squarely on free parameters associated with spatial distribution of model parameters results in a spatially distributed model extended from a point model via its extra free parameters. As far as we are aware, this is a novel take on distributed modeling. Unlike Cazorzi and Fontana (1996) and Hock (1999), our approach introduces a free *strength* parameter for each predictor used to control the spatial distribution. Though this adds parameter complexity, it is independent of the spatial resolution of the model and allows it to be adjusted for temporal non-stationarity in controlling variables. Parameter complexity of the model in this study is also comparable to that of areal depletion curves typically used to represent spatial distribution of a point model and can be designed to simulate affects of both accumulation and ablation which can both be important in running simulations for full water years.

Areal depletion curves (ADC) relate 1-dimensional model mean snow water equivalent to a snow covered area. As implemented in practice, curves are associated with a SWE threshold above which the basin is 100% snow-covered. Thus, the simplest parameterization of an ADC, using a cubic function for the curve, requires 3 values. Physical basis for deriving ADC has been given by several studies. Liston (1999) presented a theory relating SCA, SWE, and melt rates, requiring knowledge of any two to calculate the third. Depletion curves based on parametric SWE distributions were presented by Luce and Tarboton (2004). Several observational methods have been used for estimating or updating ADC estimates. Methods based on in situ data were offered by Buttle and McDonnell (1987) and Shamir and Georgakakos (2007). Remote sensing approaches have been presented by Andreadis and Lettenmaier (2006) and Kolberg and Gottschalk (2006). The

ADC represents an area of active an on-going research in snow-hydrology. Despite much past and continuing effort, determination of ADC is far from intuitive and requires intensive work. Further, Shamir and Georgakakos (2007) showed that ADC exhibit tremendous inter-annual variability in the same basin.

We compare our approach to extending a 1-dimensional model of snow to the ADC parameterization in a small, high-elevation basin with significant relief and variation in aspect. Results depend on the scale of investigation and nature of the study basin, as Gurtz et al. (2003, p 310) observed "the smaller the catchment and greater the elevation ranges, the more detailed the spatial structure needs to be."

5.3 Data: Senator Beck basin

Senator Beck basin (SBB), shown in figures 5.1 and 5.2, is a $2.9km^2$, first-order basin in southwest Colorado with large annual snow accumulations and large spatial variability of snow depth. Observational data at SBB were supplied by Chris Landry at the center for Snow and Avalanche Studies. These include hourly meteorological observations from two study plots in the basin, Senator Beck (SBSP) and Swam Angel (SASP), stream flows at the basin outlet, and density measurements during the 2005 and 2006 water years. Location of all observations at Senator Beck basin are shown in figure 5.2.

Data at SASP are used for modeling in this study. Temperature data were quality controlled and a very small amount of missing data were filled using a k-nearest neighbor procedure, weighting on day of year and similar temperature at SBSP (missing observations at the two sites did not overlap in time). Precipitation observations were quality controlled and a handful of missing observations set to zero. Snow depth measurements were quality controlled and the few missing data were filled in using a regression spline. Average density, measured routinely at approximately 3 week intervals at both SASP and SBSP, was linearly interpolated in time and multiplied by snow depth to give snow water equivalent time series at SASP. This time series was then smoothed in time to reduce the effect of freshly fallen snow on the calculation of SWE. SWE time series at SASP

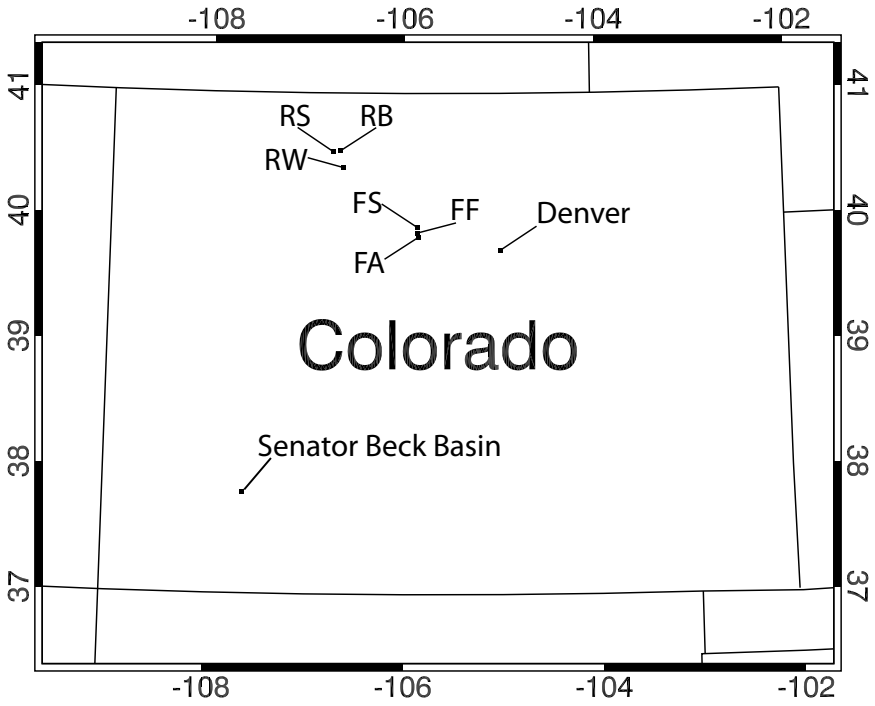


Figure 5.1: Location of Senator Beck Basin in Colorado and relative to the CLPX sites.

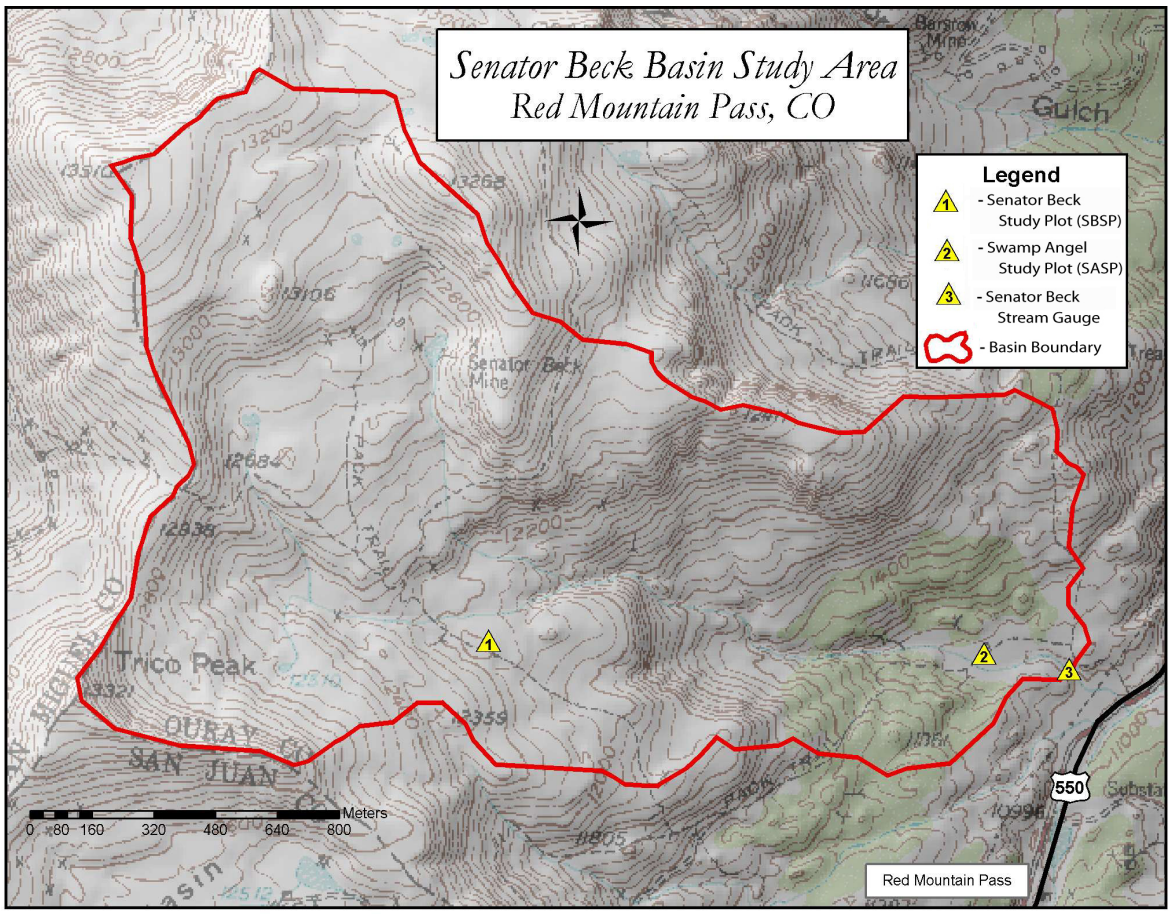


Figure 5.2: Overview of Senator Beck Basin, Ouray County, Colorado. This first order watershed is delineated in red. Meteorological stations and stream gage locations shown on map. Credit: Jeff Deems

and SBSP are shown in figure 5.3. SWE at SBSP are only used to aid interpretation of results.

Hourly observed stream flow was averaged to daily streamflow for 2005 and 2006, shown in figure 5.4. Missing data were left from the record and not used in calculation of model skill. Significant differences can be seen in the hydrographs for the two years. Temperatures in 2006 were generally colder prior to March, then warmer through June. Temperatures during the model period are shown in figure 5.5. June 2006 was approximately 3 deg C warmer than that of 2005, on average. Thus, recorded flows in May 2005 are much lower than the same time in 2006 and remain low until mid June. After this time, flows from 2005 generally exceed those of 2006 because more water was stored up to this point.

The digital elevation map and land cover map used for the basin are the National Elevation Dataset (NED) and the National Land Cover Data (2001). From these, we obtain elevation and vegetation density at 30 meter resolution and infer vegetation height, aspect, slope, northness, potential solar radiation, and maximum upwind slope as described in chapter 2.

5.4 Methodology

5.4.1 Experimental design

Beginning from a calibrated point model at SASP, we investigate its areal extension over the entire basin for full water year simulations in 2005 and 2006. The ADC approach to spatially distributing snowmelt is compared against and our ‘‘parameter-region’’ method in the following model setups:

(1) elevation: three elevation bands (+ 1 parameter),

(2) cluster: three elevation bands with five predicted regions, total 11 regions (+4 parameters),

(3) ADC (+5 parameters),

Several other cases are considered as modifications of these three initial cases.

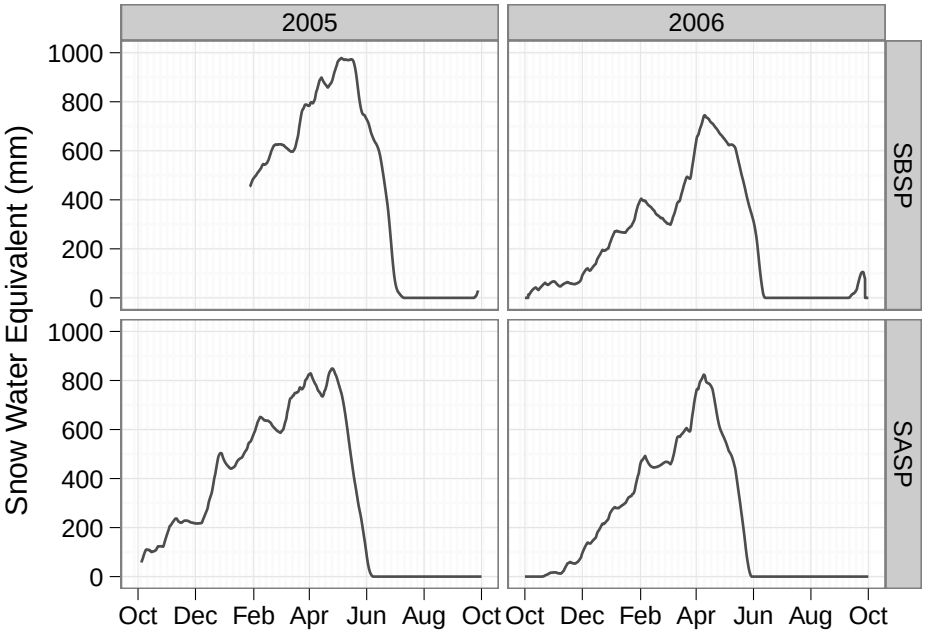


Figure 5.3: Snow water equivalent time series at the upper Senator Beck study plot (SBSP) and the lower Swamp Angel study plot (SASP) in 2005 and 2006.

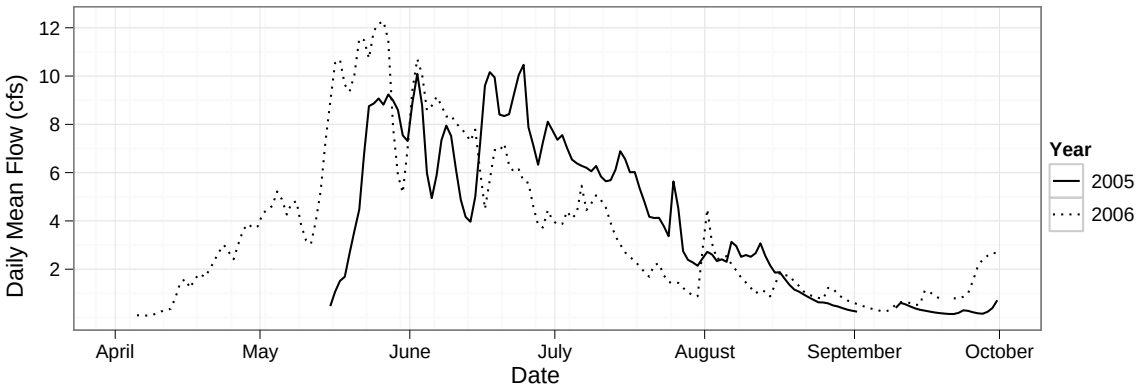


Figure 5.4: Senator Beck stream gage daily average stream flows in water years 2005 and 2006.

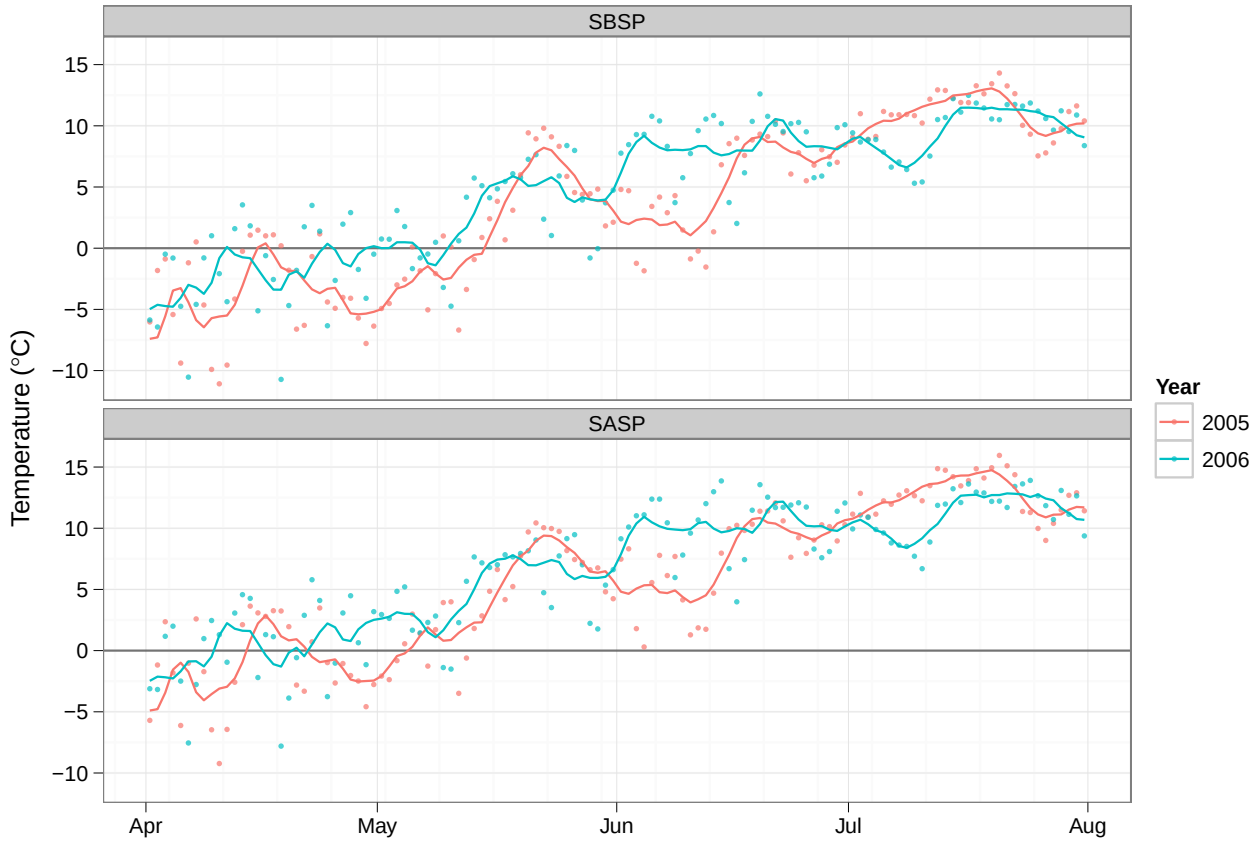


Figure 5.5: Daily temperatures with a 7 day running mean at SBSP and SASP for 2005 and 2006.

Cases are listed in increasing order of free parameter complexity and the calibrated point model is considered our null model. Because we do not calibrate free parameters in this PUBS-style setting, we understand the real-world problem of parameter selection by uniformly sampling a reasonable parameter space of each model and running the model for each combination of selected parameters. After each run, simulated flows are normalized by their total volume and their RMS error is calculated against the normalized observed hydrograph between April 15th and August 31st. (Thus RMSE is in units of % of total volume.) The resulting set of RMSE describes the probability of a certain model skill given a random selection of parameters. Models are judged by their ability to produce more accurate flows more often. In each case, the same hydrologic routing model is used to bring spatially distributed runoff to the basin outlet with no tunable parameters.

Calculation of RMSE based on normalized hydrographs ignores simulated volume and emphasizes resolving how water is distributed in time at the basin outlet. We focus on understanding flow timing because of insufficient knowledge of spatially and temporally varying infiltration rates which have a tremendous effect on simulation of total volume. We do not attempt to understand this problem here bu focus on stream flow timing which is important for management of small, upstream reservoirs.

5.4.2 Snow-17 model and calibration at SASP

Snow-17 (Anderson, 1973) is a snow model developed at the National Weather Service River Forecast Center. It is a temperature index model; it requires only 2 inputs, temperature and precipitation, and calculates snow accumulation and melt as a function of air temperature. Snow-17 is a vast simplification of myriad processes governing snow pack, but can be run with reasonable accuracy in areas where observations or forecasts of temperature and precipitation are available. In remote, mountainous locations it can still be challenging to estimate the two required inputs with much certainty.

For hydrologic purposes, total water equivalent is the primary variable of interest estimated by snow-17. Water is modeled in solid and liquid phases accounting for liquid storage and percolation

through the snow pack to simulate ripening and delays between surface melt and runoff. This is an essential feature for a snow model coupled to a runoff routing model in SBB, where 2m of snow is common. Snow-17 also handles liquid precipitation accounting for fraction of snow covered area.

Snow-17 has 22 free parameters (referenced in italics). Twelve of these, SI and the ADC vector (length 11), describe its areal depletion curve. Running at a point, depletion curve parameters are suppressed by setting SI to 1mm SWE and the remaining 10 parameters are calibrated against SWE for both years at SASP. Calibration is carried out using simulated annealing (Kirkpatrick et al., 1983) for 40 different choices of its own parameters, the maximum temperature and cooling schedule. Calibrated snow-17 model parameters at SASP are listed in table 5.1.

Year	scf	mfmax	mfmin	uadj	nmf	tipm	pxtemp	mbase	plwhc	daygm
2005	1.28	0.89	0.85	0.16	0.25	0.21	0.29	0.46	0.18	0.60
2006	1.43	0.95	0.61	0.06	0.15	0.93	0.06	-0.29	0.08	0.46

Table 5.1: Table of calibrated snow-17 model parameters at SASP in 2005 and 2006. See Anderson (2006) for details on the parameters.

We further describe snow-17 via the 3 parameters used for spatially distributing snow accumulation and melt rates in the parameter-region models and via its 12 parameter ADC implementation. Other details are left to Anderson (2006). Snow accumulation is governed by the SCF parameter (snow correction factor) which is simply a multiplier on the fraction of input precipitation determined to be snowfall. Typically, this parameter is thought of as a gage ‘‘undercatch’’ correction, which typically increases modeled snow fall relative to observed to compensate for this effect.

A sinusoidal, seasonal melt factor, MF, simulates the affect of seasonal variation in solar radiation by adjusting the melt factor for day length and zenith angle,

$$MF = \frac{\Delta t_t}{6} * \{[.5 * \sin(\frac{2\pi N}{366}) + .5] * (MFMAX - MFMIN) + MFMIN\}$$

This equation is governed by the tunable parameters $MFMAX$ and $MFMIN$. N is the number of days since March 21 (in the northern hemisphere) and Δt_t the time step in hours.

We isolate these parameters because, beginning with a calibrated point model of SWE, we implement spatial distribution by adjusting these parameters to account for expected distribution of SWE accumulation (SCF) and melt rates ($MFMIN, MFMAX$), as described in the next section.

For runoff simulation and river forecasting, snow-17 includes an areal depletion curve. Its ADC formulation relates modeled effective snow water equivalent to snow covered area using 11 values of effective snow water equivalent, one for each 10% (in 0-100%) snow covered area. Effective water equivalent is given by the mean areal water equivalent divided by the smaller of a) the largest maximum water equivalent during the accumulation period or b) the free SI parameter which describes the water equivalent above which 100% snow cover always exists.

The ADC formulation in snow-17 represents a large amount of parameter uncertainty which we choose to reduce using a 4 parameter curve where effective SWE values can be specified at 20, 40, 60 and 80% SCA and 0 and 100% SCA are assumed to have equal effective SWE. Linear interpolation from curves in this parameter space provides the 11 values needed by snow-17. Together with the SI parameter, we use a 5 parameter ADC model with snow-17. We note that it is possible to simplify representation of the ADC but difficult to uniformly sample the space of depletion curves with such representations (e.g. a cubic function).

5.4.3 Parameter-region modeling

Motivating this study is a belief about the existence of ‘‘ideal’’ spatial modeling units. Somewhere in the continuum between a lumped model, with only 1 spatial unit, and a fully distributed model, with model units on a regular grid, lies a parsimonious division of the model domain into spatial units which distill spatial variability in runoff processes to a sufficient degree for purposes of hydrologic simulation. The cluster prediction methodology in chapter 4 was developed for decomposing variability in peak-accumulation snow depth. For snow-hydrologic modeling, we need to decompose spatial variability of melt rates in the ablation phase as well. Spatial decomposition will depend on what variables are believed to control the spatial variability of SWE accumulation and melt.

The most important and easily mapped control on runoff is elevation dependence of melt rates. Higher elevations typically result in lower temperatures which result in lower melt rates. Especially when temperatures in the basin hover around the melting point (e.g. SBB in May and June, 2005), elevation plays a critical role in determining the spatial distribution of melt and associated runoff. The second control over melt rates which we consider is the distribution of accumulated potential solar radiation (described in chapter 2). In the context of peak-accumulation distribution, potential solar radiation is already thought of in terms of melt acting over the period of accumulation. This same idea applies to the melt phase. The predominantly east-west orientation of SBB, with steep slopes on both sides, implies large differences in radiation exposure and melt rates for the north and south sides of the basin during the ablation phase.

In SBB, we know from experience that the primary controls on distribution of snow water equivalent are wind and vegetation. Though elevation is typically associated with greater precipitation rates and more snow, this relationship is upset by the dominance of wind redistribution in the vast majority of the basin which lies above tree line. We consider maximum upwind slope (described in chapter 2) the primary predictor of snow accumulation. Though vegetation accounts for only a small fraction of the basin and canopy interception will play a relatively small role in reducing coincident SWE, vegetation height is also considered in our spatial model.

The first panel in figure 5.6 shows SBB divided in to elevation terciles. The second panel shows cluster prediction from chapter 4 applied to the basin using the potential solar accumulation, maximum upwind slope, and vegetation predictors. Cluster mean predictor values, listed in table 5.2, indicate where each predictor variable dominates. The first cluster corresponds to the forested area in the lower basin, with average 7m elevation heights as inferred from NLCD2001. Cluster 2 is a south facing portion of the basin with significant slope which greatly increases accumulated potential solar radiation. Cluster 3 represents areas which are relatively wind sheltered and which receive an average amount of solar radiation. Clusters 4 and 5 have greater wind exposure and cluster 4 is north-facing with the lowest mean potential solar radiation of all clusters.

The basin is divided into elevation terciles separately from cluster prediction because of

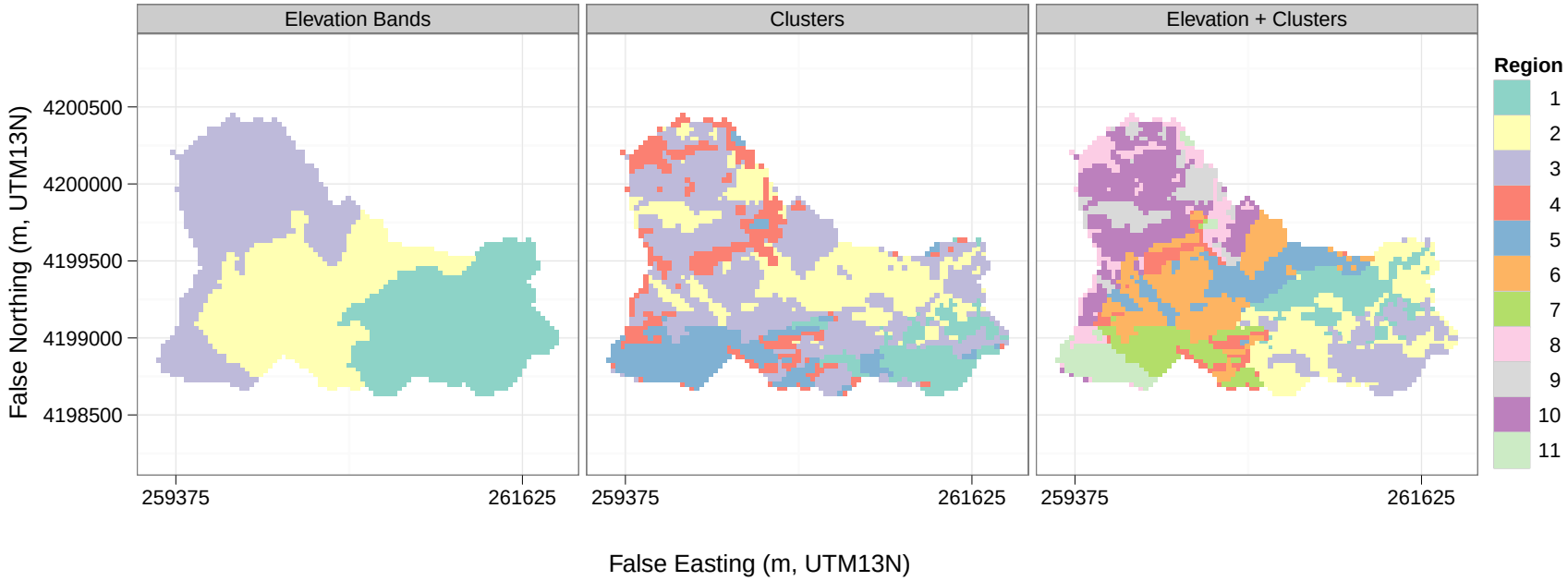


Figure 5.6: Senator Beck basin as divided 3 ways. Elevation Bands: elevation terciles (3 regions); Clusters: predictor clustering (5 regions); Elevation + Clusters: The intersection of the previous (11 regions).

difficulties with such smooth spatial gradients when applying the methodology, as discussed in chapter 4. In the third panel, cluster prediction is combined with elevation bands using spatial intersection, as in a Venn diagram. This results in 11 total clusters or regions (after removal of several very small regions created by the intersection).

In spatially extending our point model, we use mean predictor values in each of the 11 regions to adjust model parameters in a way we believe consistent with how each affects accumulation or melt. This approach may be applied to any partition of the basin and any set of predictor variables believed relevant. Parameter are added for each predictor variable used, independent of the number of regions modeled. We consider the 4 predictor variables for SBB outlined above. Region mean elevation and potential accumulated solar radiation affect melt rates in each region by adjusting MFMIN and MFMAX. Maximum upwind slope and vegetation affect accumulation of SWE by adjusting the SCF parameter in each region.

	SBB				
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Potential Solar	477343	800202	628624	579447	325087
Sx	15.5	26.6	22.1	10.5	8.5
Veg Dense	4.3	0.005	0.02	0.0002	0.01
Aspect	40	300	330	350	50
Elevation	3463	3733.4	3775.5	3884.9	3806.9

Table 5.2: Predictor variable statistics in each SBB cluster region shown in the middle panel of figure 5.6. All are mean values except aspect which is the mode over 10 degree bins. Potential accumulated solar radiation is in units of $Wh/m^2/day$, maximum upwind slope (Sx) is in units of degrees from horizontal, vegetation density and elevation are in units of meters.

Unlike Cazorzi and Fontana (1996) and Hock (1999), our model does not use fixed empirical relationships for the distribution parameters over the model domain, though we certainly could include such specific relationships, as appropriate. Using fixed relationships avoids introduction of new model parameters, on one hand. On the other, because we do not spatially distribute forcing data, new paramaters in our formulation describe uncertainty in the spatial distribution of accumulation and ablation processes resulting from spatially distributed controls on snow pack.

Because there is significant inter-annual variability in these processes and their relative effects, free parameters appear to be worth their added complexity. Our model establishes a linear relationship between model spatial unit mean predictor variables and their corresponding distributed model parameters where each is governed by a *strength*. Though the functional relationship between predictor means and parameters remains fixed, *strength* varies the spread of distributed parameters. Using a linear relationship, the relative distances between parameters remain constant as *strength* varies. The linear relationship assumption, however, may be traded for any more appropriate functional form, though not investigated here.

For convenience in setting up our method, which limits *strength* parameters to the unit interval, we use bounds on the values which the adjusted parameters may take. Such bounds were used in the calibration stage to prevent simulated annealing from wandering in to irrelevant parameter space. Similar or identical bounds can be used in this step. For SCF, we limit it between its lower physical minimum of zero and a maximum of 2.5. This says that any region has at least zero snow and at most 2.5 the observed forcing in the point model. We limit MFMAX and MFMIN between .01 and 2. For simplicity, we adjust MFMAX and MFMIN within these bounds in a way which maintains the calibrated amplitude, MFMAX-MFMIN, in the point model. Our approach could be modified to consider changing the mean and amplitude of the melt factor instead of these individual parameters, but is not considered here. The lower limit on the melt factor is near the physical limit of zero and the upper limit represents a reasonable guess at the maximum melt rate needed in extending the calibrated values. Because the point model parameters are calibrated at the lower elevation of the basin, the range of this parameter needs only to accomodate slightly greater melt rates due to increased potential accumulated solar radiation in other parts of the basin. Using bounds on our parameters does not introduce any more free parameters than determining a reasonable range of values for ADC.

From a vector of mean predictor values in each spatial modeling unit, $\overline{\mathbf{pred}}$, a calibrated model parameter, $param_{calib}$, and bounds on adjustment of this parameter when distributing spatially, $param_{min}$ and $param_{max}$, we calculate the parameters in each model unit for a given choice of

strength in the following steps. First, predictor means are normalized to span the unit interval,

$$\mathbf{z} = \frac{\overline{\mathbf{pred}} - min(\overline{\mathbf{pred}})}{max(\overline{\mathbf{pred}}) - min(\overline{\mathbf{pred}})}.$$

Next, we find the largest range which keeps adjusted parameters within the specified bounds,

$$range = min\left(\left\{param_{max} - param_{min}, \frac{param_{max} - param_{calib}}{max(\mathbf{z} - z_{calib})}, \frac{param_{min} - param_{calib}}{min(\mathbf{z} - z_{calib})}\right\}\right)$$

where z_{calib} is the predictor mean value in the region where calibration was performed. The strength parameter then selects how much of this range is used to get parameter values from predictor means in the linear equation

$$\begin{aligned}\mathbf{params} &= strength * range * (\mathbf{z} - z_{calib}) + param_{calib} \\ &= (strength * range) * \mathbf{z} + (param_{calib} - strength * range * z_{calib}) \\ &= b * \mathbf{z} + a.\end{aligned}$$

The previous equations frame the unknown *strength* parameter as influencing the slope in an assumed linear relationship between the model parameter and the region mean predictor values. Given observations of parameter values and cluster mean predictor variables, this parameter might be inferred by a linear regression and we could test the validity of this relationship. For now, we investigate if it has any practical application. If the linear relationship were suspected to be some other function of the region mean variables, this might be explored. Because elevation and vegetation have negative relationships with the parameters they control, each is multiplied by -1 prior to input. Because the model is calibrated at a very specific elevation, we introduce an extra, dummy region at that elevation to $\overline{\mathbf{pred}}$ and remove it from the resulting **params**. All other region mean predictor are assumed to be applicable at the point of calibration. If the calibrated point were outside the basin, such a dummy point would be needed for all variables.

When more than one predictor adjusts a single model parameter, we apply the above for each predictor separately and then average the resulting parameters. The individual *strength* parameters for each predictor determine the competition between the predictors and resulting parameter.

5.4.4 Runoff model

Now we describe how resulting runoff finds its way to the basin outlet. In each modeled region, snow-17 gives total water discharge, including that from snow and rainfall on bare ground. At each hourly time step, this discharge is applied at every 30m cell in a modeled region. All water exiting snow-17 (for all models) then finds its way to the basin outlet via the same routing model developed on the 30m grid.

Though a physical model of distributed runoff-routing is attractive, run times (approximately 20 minutes) required to accurately simulate hydrographs in such a small basin are prohibitive for producing a large number of simulations. Instead, we employ such a model to derive water time delay to the basin outlet as a function of rainfall intensity for all 30m cells in the basin using a Lagrangian approach based on the instantaneous unit hydrograph.

The CASC2D model (Julien et al., 1995) was used to get travel times from each individual raster cell to the basin outlet (instantaneous unit hydrographs) for a variety of steady-state rainfall intensities. For each cell, a linear regression provided excellent fit to $traveltime = a + b/intensity$. Because this relationship only considers individual cells in insolation, neglecting the non-linear affects of spatial aggregation of water due to routing, each cell's input intensities were multiplied by its number of contributing cells to get an effective intensity and effective travel time. From here a Lagrangian view is taken, residence times for each cell are found by subtracting the effective travel time of its downhill neighbor from its own. Travel times are then recalculated by summing residence times along each water parcel's path to the basin outlet. This is done for a variety of intensities and the same inverse relation is fit again via linear regression. Using this relationship, water at any input intensity finds its way to the outlet in some amount of time. This relationship still neglects affects of antecedent precipitation on the travel time from each cell. A variety of approaches were attempted but marginal increases in model skill were not justified by the added parameter complexity.

The above model was derived assuming an overland Manning coefficient of .5 and an overland

no-flow threshold of 1mm. These values were found to give reasonable results for SBB rainfall events in August 2005 and 2006. No channel routing is performed within CASC2D, though it is approximated by effective intensity in the derivation of travel time. The hydrologic component has no tunable parameters. Water losses due to infiltration, evaporation, and transpiration are neglected in this study as we compare flows normalized by their total volume, which puts emphasis on timing instead of volume. Our complete ignorance of these processes necessitates the assumption that none has a large enough effect on flow timing to warrant inclusion in conceptual modeling. Because modeled flows are approximately double that of observed flows, these processes clearly are important. Effects of infiltration and evaporation/sublimation are expected to be important to modeling the correct volume.

5.4.5 Parameter sampling

For the elevation model, with a single parameter, we evenly sample 1600 points of elevation strength in the unit interval. The cluster model, with 4 parameters corresponding to elevation, potential solar, maximum upwind slope, and vegetation strengths, 8 values are uniformly spaced in the unit interval of each parameter for a total of 4096 samples. The ADC model is constructed by letting effective SWE at 20, 40, 60, and 80% increase by any of 0, 5, 10, 20, 40, or 60mm though only attaining a maximum of 99. These same numbers were used to decrease SWE from 100 to result in the symmetric set of 1407 unique depletion curves depicted in figure 5.7. These curves are evaluated at 8 choices of the SI parameter, 100, 200, 300, 400, 500, 600, 700, 800 mm for a total of 11,256 points sampled in the domain of the ADC parameterization.

These subsampling numbers are used in further modifications of the 3 setups listed (elevation, cluster, and ADC0. When the elevation model is combined with the ADC model, we use the same set of SI with a less dense set of ADC curves in conjunction with only 5 samples of elevation strength, for a total of 20,120 points. The ADC is assumed to be the same in each elevation band to reign in parameter complexity.

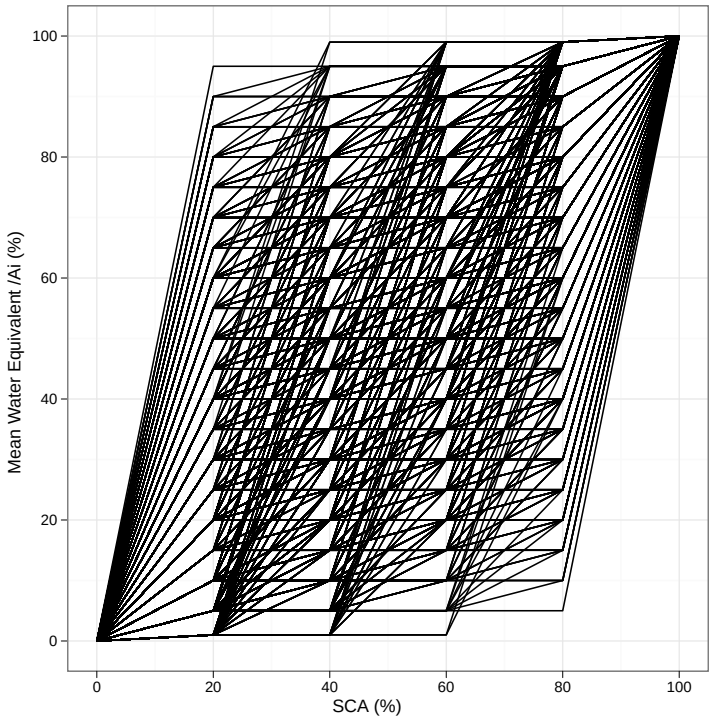


Figure 5.7: The set of 1407 unique depletion curves sampled in the ADC and elevation+ADC parameter spaces. Snow-17 selects A_i to calculate the effective mean water equivalent, A_i is the smaller of a) the largest maximum water equivalent during the accumulation period or b) the free *SI* parameter which describes the water equivalent above which 100% snow cover always exists.

5.5 Results

Figure 5.8 shows density of RMSE scores for the ADC, elevation, and cluster model setups where each has been normalized to by its maximum density. The vertical black line in each year indicates the RMSE when the null model is used to simulate streamflow. Figure 5.9 shows Nash-Sutcliffe model efficiency densities instead of RMSE densities. The two figures tell essentially the same story, there's no appreciable difference between the two skill measures. It is immediately clear from the areas occupied by each density curve that the ADC model has very limited effect on improvement of RMSE scores. In 2005, the ADC model is most often worse then the null model. When it does improve upon the null model, the gains are slight and infrequent. Though the ADC model performs better than the null model a majority of the time in 2006, significant improvement is improbable if parameters are selected randomly, density appears to taper off towards lower RMSE. Closer inspection of the largest improvements by the ADC also reveals that the correspond to unrealistically low values of the SI parameter.

By contrast, elevation and cluster models affect strong improvements in RMSE which are far more likely when parameter selection is uncertain. In 2005, the elevation model almost never results in a worse score than the null model and its second mode is near its best RMSE scores. In 2006, the mode of the distribution is very near its best RMSE. We see that using elevation has a particularly dramatic effect on RMSE in 2005, where it performs better than the cluster model by itself. While this partially results from denser sampling of the strength parameter in the elevation model, the lower temperatures in May and June 2005 resulted in a very important elevation temperature gradient where it appears the upper portion of the basin remained solid and stored water which was released in late June and into July. Basin temperatures at the upper (SBSP) and lower (SASP) sites are shown for 2005 and 2006 in figure 5.5. Because the ADC does not resolve spatial distribution of melt rates, it struggles to simulate this effect, as we will see when discussing the associated hydrographs below.

The cluster model shows very different behavior in the 2 years. In 2005, a significant portion

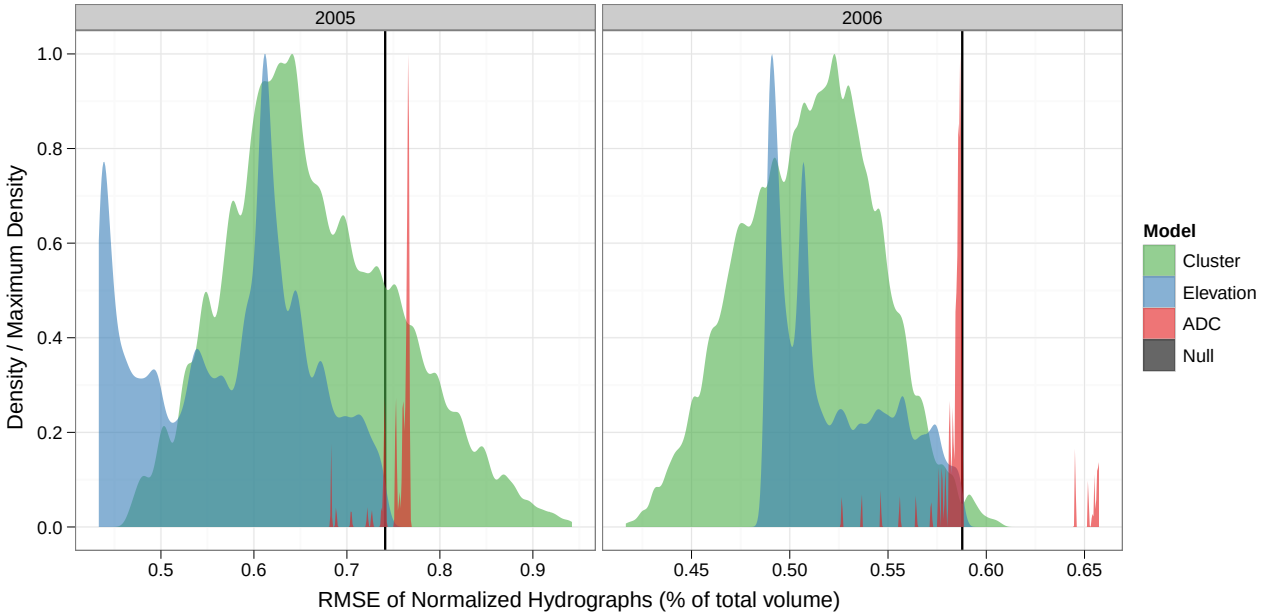


Figure 5.8: RMSE densities for the sampled parameter spaces of each model setup. The vertical black line indicates the skill of the null model when used to simulate runoff, this is a single number and not a distribution.

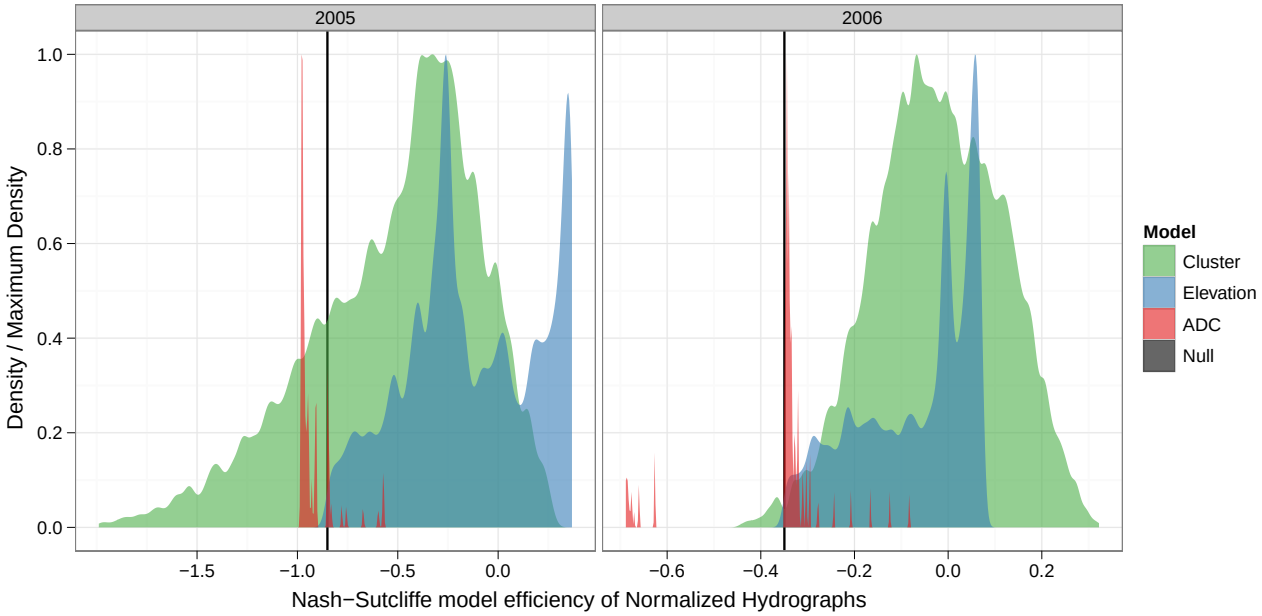


Figure 5.9: Nash-Sutcliffe model efficiency densities for the sampled parameter spaces of each model setup. The vertical black line indicates the skill of the null model when used to simulate runoff, this is a single number and not a distribution.

of its RMSE distribution is worse than the null model while in 2006 only a small portion is. This indicates that the assumed relationship between predictors and parameters may have been inappropriate in 2005, or simply that elevation affects dominated strongly over all other combinations of parameters and that its exclusion in part of the cluster parameter space was a poor choice. In 2006, the relationship of the cluster model to the runoff distribution appears to be appropriate most of the time, improving RMSE by a wide margin over the elevation model. In both years, the cluster model is likely to improve upon the null RMSE as well as that of the ADC when given uncertainty in these free parameters.

Figure 5.10 displays each model's hydrograph for its best, median, and worst RMSE scores. These indicate the overall accuracy and range of accuracy given by each model. In 2005, improvements of the cluster and elevation models come from reduction of the initial peak in late May and from more sustained flows in July. These features are related to the low temperatures in late May and June 2005, when melt rates appear to have been strongly dependent on elevation and snow was stored in the upper basin and then released in late June and July. The best ADC model neither reduces the initial peak nor sustains flows in July. In 2006 we see a similar problem with the ADC model, too much runoff is generated early and too little flow results even in June. This repeated scenario, which the more physically based parameter-region models address, results from the lumped model having no physically-based way of distributing melt rates. Under the ADC, when snow is melting in the point model, it is melting everywhere at the same rate.

5.5.1 Spatially distributed melt rates and the ADC

To remedy the ADC model's lack of spatial variability in melt rates, we combine the elevation model with the ADC model, though using the same curve in all regions to keep the number of free parameters to 6. Results are compared against the elevation and cluster models in figure 5.11. While its best RMSE scores have dramatically improved with inclusion of spatially distributed melt rates, we see that most of its density lies at higher RMSE than for the other models. Thus, if parameter selection is uncertain we are not likely to obtain a good value. While estimates are improved over the

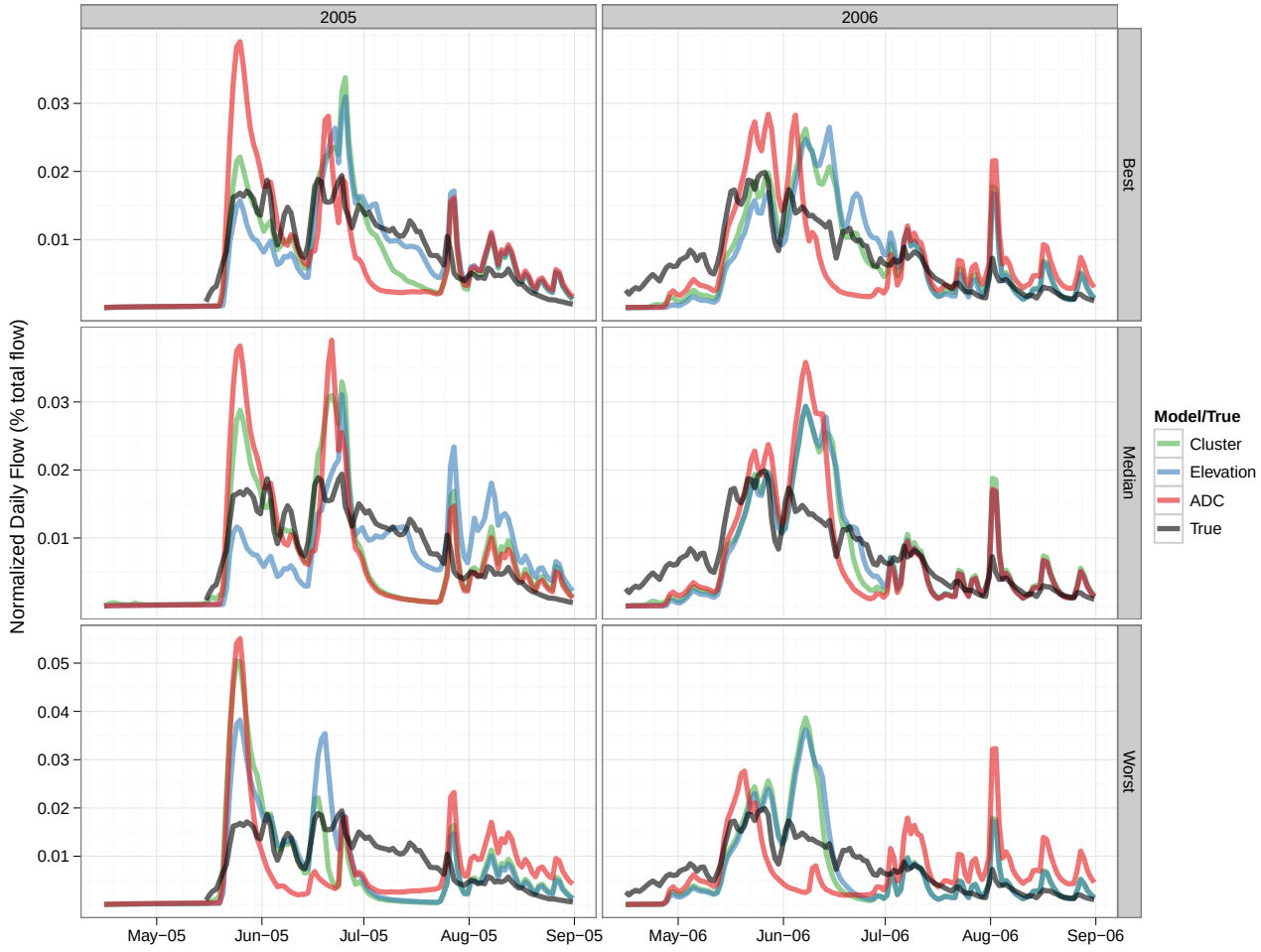


Figure 5.10: Example hydrographs shown for the best, median, and worst RMSE scores for each model setup. Note the different y-axis for the worst hydrographs.

ADC-only model for 2005, estimates in 2006 become much worse, with a much lower probability of improving upon the null model. The ADC formulation appears to be an expensive parameterization which is likely to degrade model skill relative to the null model even when spatially distributed melt is used in conjunction.

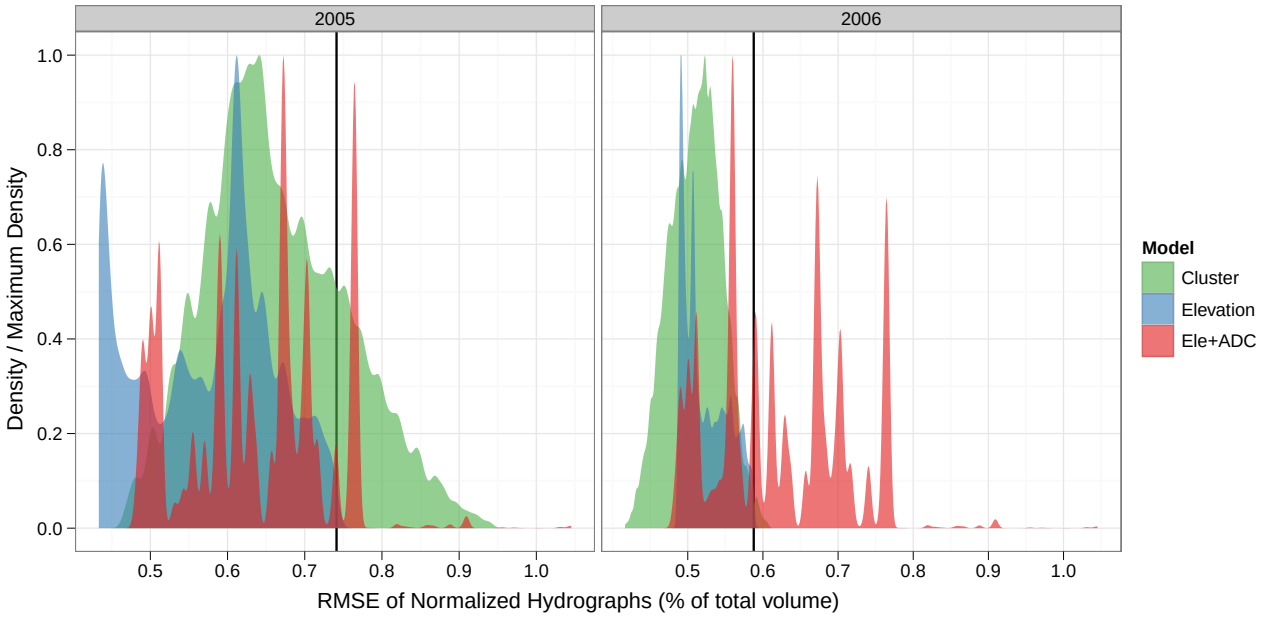


Figure 5.11: RMSE density of the elevation+ADC model compared to the elevation and cluster models.

5.5.2 Cluster model parameters and streamflow simulation

Figure 5.12 shows the dependence of cluster model skill on the strength parameters of individual predictor variables while the others are held to zero. Figure 5.13 shows the density of RMSE corresponding to each of these curves. In 2005, the figures confirm the dominance of the elevation strength parameter on model skill. When used alone, vegetation is the only other predictor which improves simulation RMSE. Both accumulated potential solar radiation and maximum upwind slope parameterizations degrade null model skill. In 2006, all individual parameters eventually

improve model skill. The parameter for maximum upwind slope degrades model skill at first but eventually results in the greatest reduction of simulation RMSE. The elevation and solar parameters also provide consistent reduction in simulation error while the vegetation parameter provides improvement only up to some value.

Figures 5.14 and 5.15 display the dependence of cluster model RMSE on any two of the four parameters while the other 2 are held to zero, for 2005 and 2006. In 2005, we see elevation dominating improvements in RMSE overall. However, improvements in RMSE using accumulated potential solar radiation in conjunction with elevation are unexpected based on results for individual parameters where the potential solar parameter produced no improvements in simulations. Moving from left to right from x values of zero, we see potential solar improving RMSE above about .3 but only when elevation strength is greater than about .3 as well. The interaction of the elevation and vegetation parameters is similarly non-linear, vegetation has most effect at values of about .2, but also above .8 when elevation strength is less than approximately .5. Using maximum upwind slope (sx) with elevation degrades simulation RMSE and whenever elevation is neglected RMSE is poor overall. In 2006, the RMSE surfaces indicate that no single variable dominates its improvement, as concluded when looking at individual parameters. Interactions are similar to what we might expect from the individual parameters though non-linear interaction of variables is apparent from the plots as well.

Previously (e.g. figure 5.5), we provided evidence for why the elevation parameter dominates improvement of simulated streamflow RMSE in 2005. Painter et al. (2007) examined the effects of dust on springtime snow melt at SASP and SBSP in 2005 and 2006 using energy-balance point models of snow pack. In 2006, they report a much stronger (2x) radiative forcing term from 8 dust events as compared to the typical number of 4 in 2005. They also remark on smaller peak snow water equivalent in 2006 relative to 2005, particularly at SBSP, as can bee seen in figure 5.3. These observations of physical conditions in the basin correspond to the important model parameters in 2006. In figure 5.12, incorporation of solar radiation distribution in 2006 improves model simulations where it appears inappropriate, decreasing model skill, in 2005. Though it is difficult to conclude

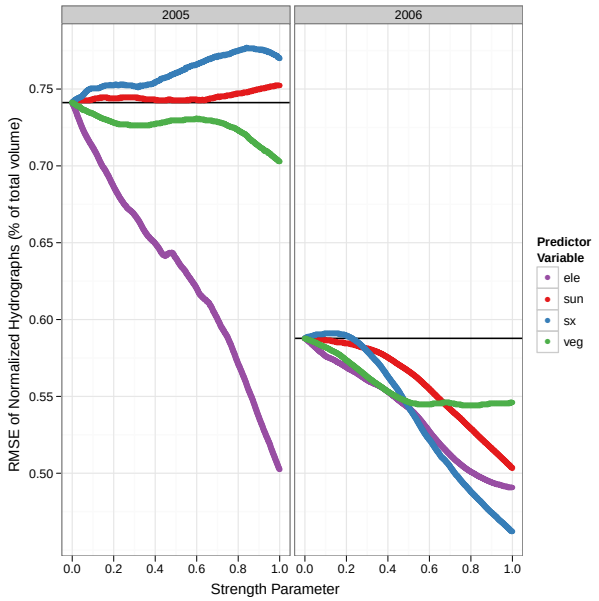


Figure 5.12: RMSE as a function of parameter strength when only individual parameters use in the cluster model. The skill of the point model is shown by the black line in both years. The individual predictors are elevation (ele), accumulated potential accumulated solar (sun), maximum upwind slope (sx), and vegetation height (veg).

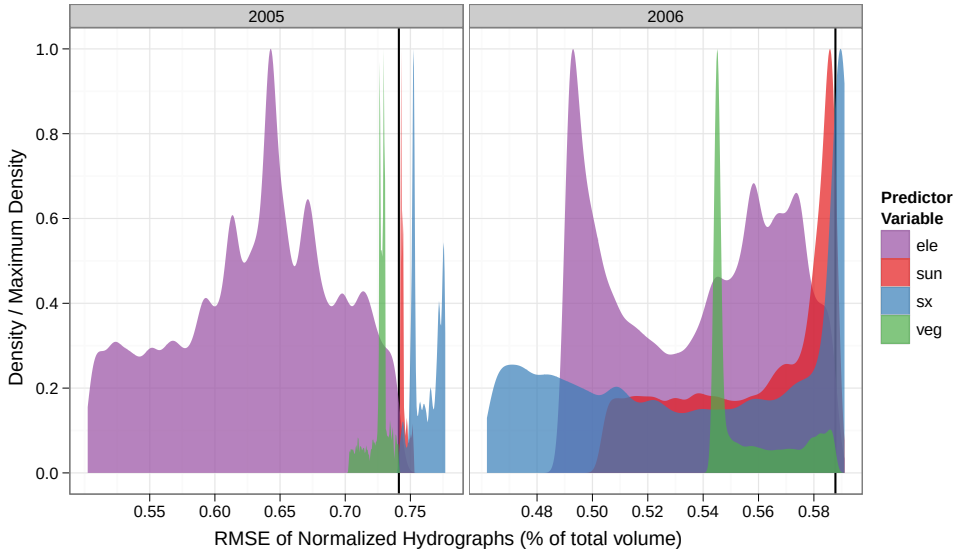


Figure 5.13: RMSE density when individual parameters use in the cluster model. The individual predictors are elevation (ele), potential accumulated solar (sun), maximum upwind slope (sx), and vegetation height (veg).

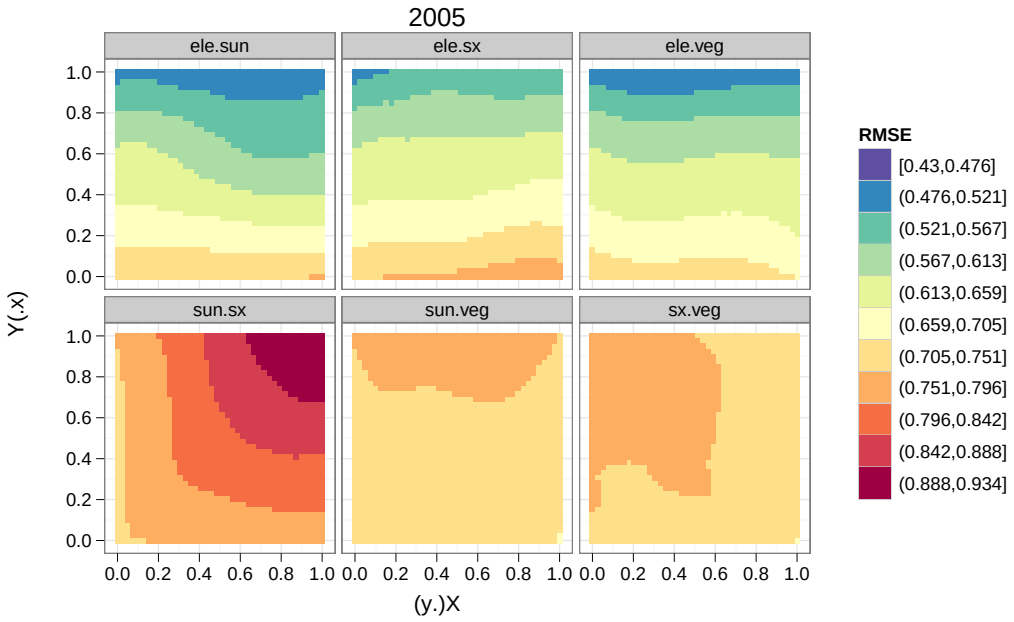


Figure 5.14: RMSE as a function of parameter strength when any two parameters are used in the cluster model in 2005. The first variable in the y.x pair is plotted on the y-axis and the second on the x-axis.

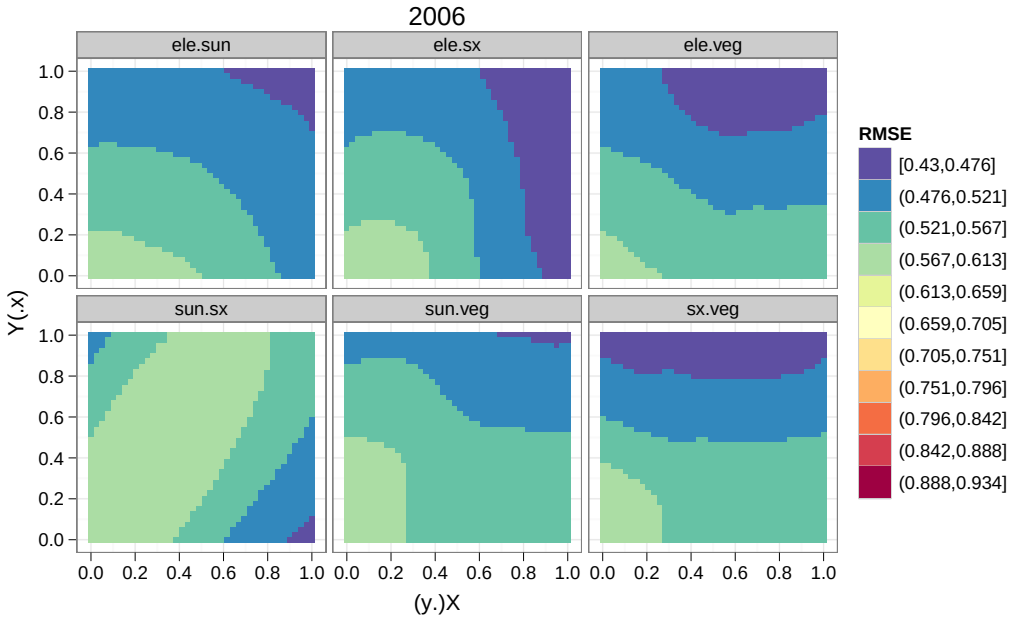


Figure 5.15: RMSE as a function of parameter strength when any two parameters are used in the cluster model in 2006. The first variable in the y.x pair is plotted on the y-axis and the second on the x-axis.

about the upper basin as a whole from the point measurement at SBSP, SWE in the upper basin appears to have been significantly lower than in the upper basin in 2006. Losses in this part of the basin typically result from wind redistribution. Because the upper portion of the basin is more exposed to wind, the maximum upwind slope parameterization decreases snow accumulation in the upper basin relative to the lower basin as its strength parameter increases. Maximum upwind slope is the most important parameter in 2006 according to both figures 5.12 and 5.15 whereas its inclusion in 2005 degraded simulation skill.

Though use of maximum upwind slope and potential solar radiation predictor variables result in poorer simulations in 2005, we cannot conclude if the parameterizations are truly inappropriate in that year. More generally, the dominance of the elevation predictor in 2005 indicates that the primary model structural deficiencies in extending the point model over the basin in 2005 are most effectively corrected by the elevation parameterization. Perhaps naively, limits on parameters have been imposed by physical considerations on each predictor variable. Figure 5.12 indicates that simulations are most improved when certain strength parameters are set to their maximum values which correspond to the upper physical limits for each predictor. Because of the physically extreme nature of the resulting distributed parameters, results suggest that the physically-based parameterizations are correcting for other model deficiencies besides those intended in their formulation. This is not surprising to anyone familiar with model parameters and the equifinality thesis (Beven, 2006). On the other hand, the parameters which result in greatest improvement of simulated streamflow appear related to observed physical conditions in each year, as outlined above, which is an extremely encouraging result.

In 2005, if we let the elevation parameter act over a greater range, we may eventually see other parameters help improve the model. Though we don't explore alternate parameter ranges, figure 5.16 reveals how results depend on spatial structure as well as parameter ranges. In 2005, results for the elevation band model improve upon the cluster model using only the elevation parameter because the extreme values of the spatial parameter distribution apply to a much larger portion (1/3) of the basin than when the same parameterization is used with the spatial pattern

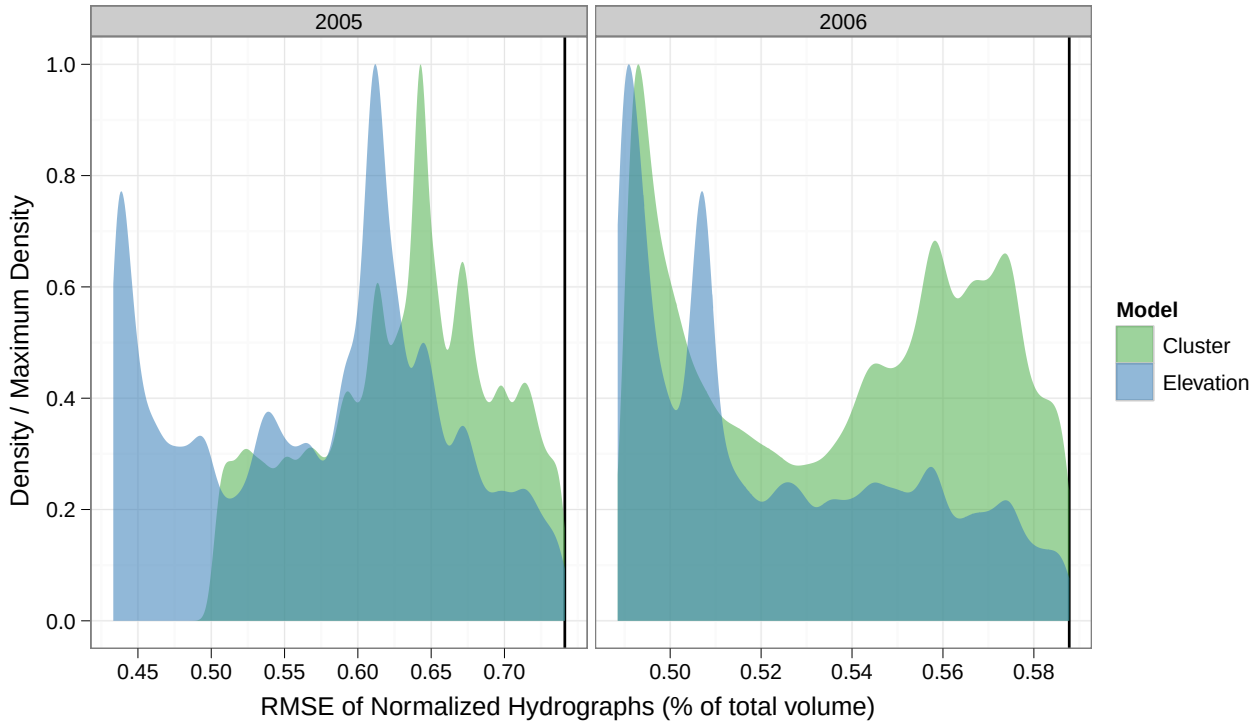


Figure 5.16: Comparison of elevation band model and cluster model when only the elevation strength parameter is used.

in the cluster model. In 2006, the two different spatial structures yield very similar improvements in simulated streamflow. While it would be interesting to try to understand the most appropriate spatial structure for each predictor and its corresponding parameter distribution, the problem of compensation for other model deficiencies makes this a difficult problem absent spatial data for verification.

5.5.3 Three model deficiencies

In this section we explore three changes to the models and their effect on simulation skill. First, models are run without runoff routing. We assume that runoff from each spatial region arrives instantly at the basin outlet. Second, we introduce a parameterization of model temperature thresholds as a function of elevation and an assumed lapse rate. Third, we switch calibrated model parameters between the two years to understand sensitivity of the parameterized spatial distribution to degraded point model parameters. In the first and last case, we intentionally degrade the models as constructed. In the second case, we try one approach to correcting a perceived model deficiency. In all cases we learn something useful about our model parameterization.

Snow-17 includes two temperature thresholds, $PXTEMP$ and $MBASE$, which determine the phase of precipitation and the melting point. The extreme dependence of simulation skill on elevation in 2005 suggests that an alternative approach to spatially distributing melt rates might improve model skill. Temperature dependence on elevation is modeled using a lapse rate. The dry adiabatic lapse rate is has a value of $-9.8^oC/km$ and a standard environmental lapse rate is typically assigned $-6.49^oC/km$. We assume a lapse rate of $-8^oC/km$ and distribute the lapse rate using the mean elevation of regions in the cluster model and a positive relationship with elevation for adjusting thresholds. This parameterization introduces no free parameters since physical reasoning is used to select the strength parameter, as in the formulations of Cazorzi and Fontana (1996) or Hock (1999).

Figure 5.17 shows these three individual adjustments to the model underneath the original results (figure 5.8) in the top panel. The RMSE of the null model from the original setup is indicated

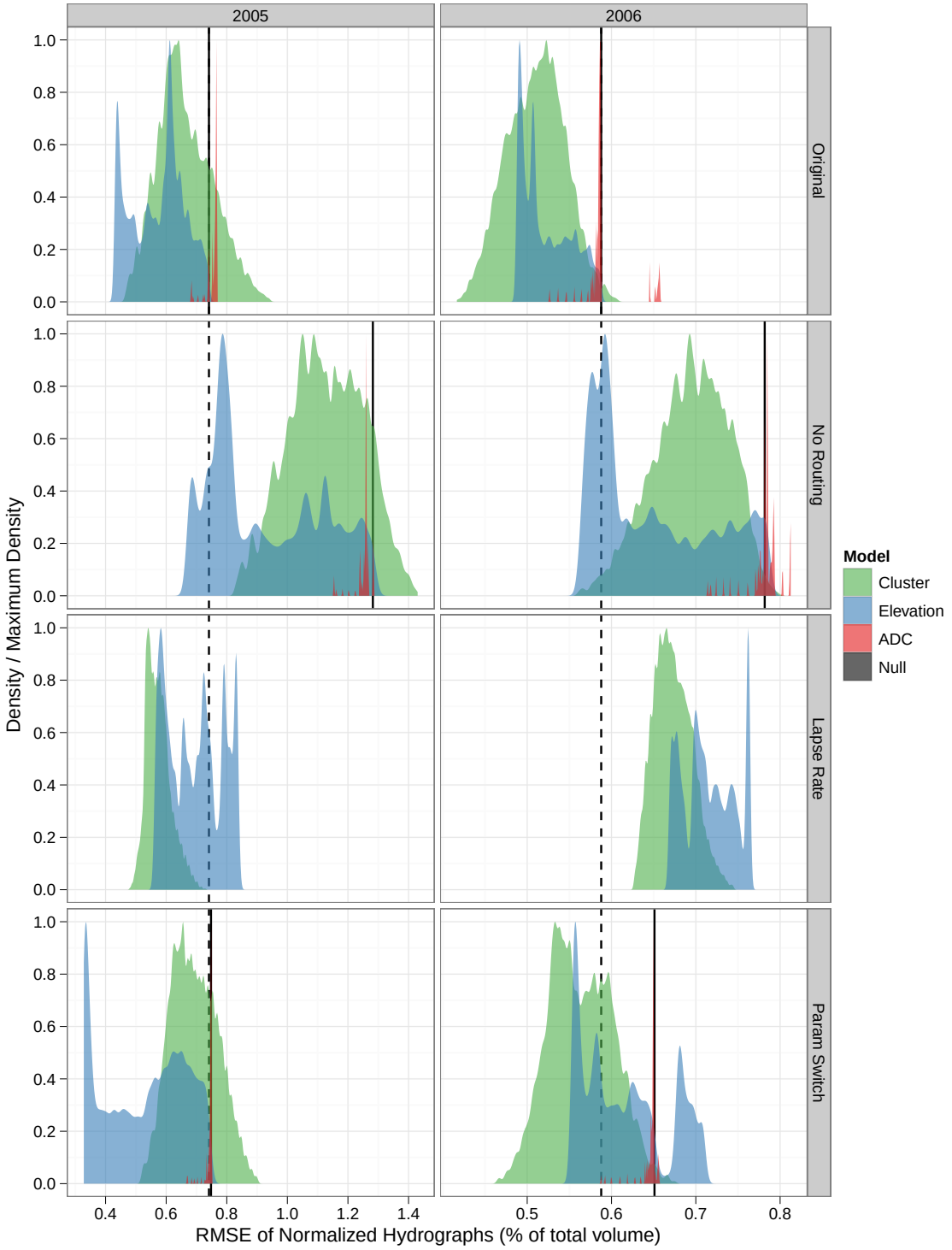


Figure 5.17: Comparison of original parameterization to several other model situations. **No Routing:** the model is run assuming runoff from each spatially distributed model arrives immediately at the basin outlet. **Lapse Rate:** a lapse rate of $-8^oC/km$ is applied to distribution of model temperature threshold parameters. **Param Switch:** considers the effects of alternate parameter choices on model skill. Calibrated parameters in each year are traded for those in the other year.

by vertical dashed lines in the following figures. Removing routing from the model, the null model RMSE drops significantly in both years. Removing routing also puts a much stronger emphasis on the elevation model in both years. In 2005, the elevation model shows a significant improvement over the cluster model and in 2006 the cluster model is no longer able to improve upon the elevation model as seen when routing was included. This indicates that the elevation parameterization (in both the cluster and elevation band models) may compensate for deficiencies in the routing model, even when it is included. The ADC model continues to hug the null model simulation RMSE closely.

There is no null model when the lapse rate parameterization is introduced because there is no introduced parameter to vary it (from zero). The parameterization is also not combined with the ADC. The figure indicates that including the lapse rate increases simulation RMSE for the elevation band model, especially in 2006, and turns the distribution towards a more uniform one. In 2005, the lapse rate parameterization in the cluster model sharpens the distribution of simulated RMSE. With the lapse rate included, the cluster model now performs better than the elevation model though both of their minimum RMSEs are increased relative to results without the lapse rate parameterization. In 2006, we see a similar affect on the simulated RMSE but skill is always and significantly degraded compared to the non-lapse rate model. Given the apparently important elevation distribution of melt rates in 2005, it is somewhat surprising that this formulation does not yield improvements in simulation RMSE. However, the physical basis of choosing a lapse rate is questionable because of temperature inversions in the basin and the importance of solar radiation, especially in 2006. Results indicate that a lapse rate parameterization is not a physically appropriate approach to include without free parameters.

In the bottom panels of figure 5.17, results are shown for switching point model parameters between years. Null model skill under the alternate parameter set is shown in the solid black lines. In 2005, null model skill falls only slightly. (Note that point models were calibrated to observed SWE and not to streamflow.) Structural shortcomings of the areal model dominate simulation RMSE instead of point model parameters. Interestingly, the alternate set of point model parameters allows the elevation band model to improve simulation skill more than the original, calibrated model

parameters. This highlights that the skill of the spatial parameterization depends strongly on the calibrated point model parameters and that point calibrated parameters may not represent the optimal choice. This does not happen for the cluster model because its ability to compensate for model structural deficiencies is limited by its more precise spatial structure. In 2006, switching parameters has some detrimental affect on the elevation model, a substantial amount density lies at greater RMSE than the null model, while the cluster model is less affected. In both years, most of the ADC model RMSE density again lies very close to the null model.

5.6 Conclusions

At this scale of study, our parameter-region method presents an attractive alternative to the areal depletion curve approach typically used to model spatial variability in snow melt runoff. We have demonstrated that our method is more likely to improve simulated stream flow timing in SBB, and also to result in much larger improvements in skill, in two different years when selection of free model parameters is uncertain. This is an important result in a PUBS scenario as parameter selection is uncertain and calibration impossible. The number of parameters is comparable between the methods, depending on the number of predictor variables used and the depletion curve model used. We argue that our method is more physically intuitive than the ADC approach. For our basin, we showed that a model using only elevation could significantly improve simulated streamflow and that inclusion of 3 extra variables could potentially improve estimates even more. While individual predictor variables relate well to physical conditions in the study basin in the two years investigated, further work in other years is needed to understand each more specifically. The impacts of deficiencies due to alternate model formulations and point model parameter selection were also investigated. Results indicate that just as with any other model, the new parameterization can compensate for other model deficiencies.

Study of additional years in conjunction with detailed observational data at a variety of scales to relate basin conditions to model parameterizations may permit automated selection of predictor variables and their strength parameters. Model parameters might then be conditioned in real time

based on climate variables affecting snow distribution and melt rates. Because a wide range of parameters improve upon the null model, application of ensembles of parameters might also be used for probabilistic streamflow forecasting. If important systematic deficiencies in runoff simulation can be overcome by similar parameterizations, ensemble methods will be most useful.

The most important direction this research can be taken is out, over greater areal extents. Results here are clearly dependent on the size and relief characteristics of SBB. Will the ADC formulation prove more competitive at larger scales where hydrographs are much smoother? Does the proposed methodology scale up? We have not considered the problem of how forcing uncertainty will affect simulated streamflow. Can we predict streamflow reliably using this method over an entire mountain range where model parameters are based on a scattering of SNOTEL sites?

Chapter 6

Summary, Conclusions, & Directions

High resolution LiDAR data, presented in chapter 2, offer an important view of the spatial distribution of snow depth. The data were used in chapters 3 and 4 to illuminate two very different problems. In chapter 3, data were leveraged to quantify the effects of multiple factors in the problem of estimating snow depth distribution from sparse observations. In chapter 4, the data permitted a simple theory of snow depth spatial distribution to be tested and verified. The logical conclusion of this theory is that first-order spatial structure of snow depth distribution can be inferred from predictor variables without need for snow depth observations. In chapter 5, spatial structure of both snow depth and melt rates were predicted from this theory and applied to improve snow-hydrologic simulation.

Results highlight the value of these novel snow depth observations to both the observation and prediction of spatial snow depth distribution. Though conclusions are limited by the scale of the available observations and by some uncertainties with LiDAR measurement, results appear to have important consequences and warrant new data sets wherein their scale dependence may be examined. Chapter 3 outlined the accuracy and uncertainty associated with a given number of point observations in each square kilometer in this study, which could be used to design an efficient sampling campaign for examining scaling issues. Results should represent an upper limit on the necessary observation density at larger scales. Scale dependence of clustering, of both snow depth and predictor variables, also needs to be examined within larger scale data sets. How would one use cluster prediction on a $100km^2$ or a $1000km^2$ basin? Will it need to be divided into sub-domains

and applied at scales used in this study or will it scale up in a convenient way?

CLPX LiDAR data represent peak-accumulation snow depth, which is an important hydrologic quantity. But results in chapter 5 affirm the importance of melt rate spatial distribution. While we have used cluster prediction to map spatial variability of melt rates based on results for the accumulation phase, LiDAR has potential to be applied during the melt phase to obtain a clearer picture of the spatial distribution of melt, especially in deep snow packs with large variability in melt rates as at Senator Beck basin. Finally, to ensure reliable results, sufficient ground-truth of LiDAR should be performed for a variety of vegetation types and surface slopes. While point comparison will remain difficult, statistics in a variety of areas should be compared to ensure systematic biases are not present in the LiDAR data. LiDAR data has significant potential to advance our understanding of spatial processes of snow accumulation and melt if approached iteratively and with care towards its shortcomings.

Results in chapter 3 put numbers on how volume estimation (percent error), standard error (RMSE), and inference of spatial distribution (R^2) are influenced by a variety of factors when regression models are used with sparse observations. Factors examined included number of observations, uncertainty due to random selection of observations, predictor resolution, and model resolution. Cross-validated skill measures based on the observations were also compared against skill measures against the full distribution. Of the three models considered, bagging trees proved to be the best for several reasons. For those designing snow depth sampling campaigns using either manual probes or LiDAR technology, this study illustrates potential caveats associated with inference from sparse observations and frames potential loss of model skill in terms of observational cost. The sampling design used in this study is a random selection which aims to cover the domain of interest. Though this design does not account for the spatial autocorrelation of snow depth, focus on independent information provides robust results while ensuring that true statistics are better than cross-validated statistics (except for volume estimation). Comparison with the CLPX sampling design highlighted that introduction of autocorrelation into sampling design degrades true R^2 and the paradox that trying to measure to estimate true R^2 necessarily decreases for a fixed

number of observations.

Chapter 4 examines intuitive hypothesis that the spatial structure of predictor variables governs the spatial structure of snow depth distribution. Prior to the high-resolution LiDAR view of snow depth, verifying this simple hypothesis would have been problematic. Cluster analysis of the LiDAR data revealed spatial non-stationarity of snow depth which qualitatively matched intuitive understanding of independent variables controls at each site. Taking the hypothesis further, predictor variables were clustered to predict snow depth spatial non-stationarity. In a sampling design experiment, a binary regression tree required 100-200 observations to infer an equivalent spatial structure given by cluster prediction without observations. At most sites, cluster prediction patterns explain more than half the R^2 attained by binary regression trees when using 1600 observations.

Patterns given by cluster prediction represent first-order snow depth spatial non-stationarity but with no associated values. They yield only a set of unordered spatial regions. As such, cluster prediction can be used for sampling design and spatial modeling from sparse observations when too few data are available to improve upon the predicted structure using a regression approach. Practical sampling design is often limited by accessibility and critical areas go unobserved, often unintentionally. Cluster prediction maps important regions with the assumption that each has a stationary mean snow depth and can help direct observations to avoid dropping key areas and indicate the easiest way to potentially sample difficult regions. For most purposes, snow water equivalent (SWE) is a more important quantity than snow depth. The relevance of cluster prediction to anticipating SWE spatial structure should be investigated, especially considering the labor intensity of gathering SWE observations.

The problem of ordering the regions by mean snow depth, as well as estimating the differences between regions, is extremely important. Solving this problem has application to data assimilation of sparse snow depth observations via quantification of the areal representativeness error and to statistical downscaling of snow precipitation. Downscaling may be approached using ensembles of differences between regions, given simply that their order is correct.

The main difficulty with cluster prediction is the problem of predictor selection. Attempts to identify a predictor set without the use of observations were not successful. On the other hand, predictor variables selected using knowledge of a given site appear to perform nearly as well as the set of predictors identified heuristically to provide the highest R^2. Effectively, this limits cluster prediction to small studies where an area may be analyzed in advance. Because it's highly desirable that prediction can be applied in an automated fashion to large regions, a significant advancement of cluster prediction will be in understanding how to automate predictor selection. Again, data covering larger scales and even more diverse environments is needed to study this problem correctly and provides yet another objective for future sampling campaigns. Similarly, the two parameters (upwind search distance and azimuthal range of averaging) used in calculation of maximum upwind slope deserve to be studied so that they may be automated. The importance of this predictor is affirmed in this study and its parameters are estimated using knowledge of the site.

In chapter 5 we have proposed an alternative approach to extending a point model of SWE over a basin for simulating streamflow timing. At our scale of investigation (3 sq km), this approach compares extremely favorably with the standard areal depletion curve (ADC) approach. Particularly when parameter selection is uncertain, our new parameterization virtually guarantees significantly better results than the ADC. Our approach uses cluster prediction from the previous chapter (4) to map regions where independent variables cause non-stationarity in snow accumulation and melt rates. Point model parameters are distributed over the regions via a simple linear relationship between individual model parameters and the regional means of its controlling independent variable. For each predictor variable used, a single free parameter is introduced describing the strength or slope of the linear relationship. Our parameterization is extremely flexible. Independent variables governing snow accumulation and melt rates may be included or excluded when distributing the point model spatially by simply adjusting individual strength parameters. The new semi-distributed model is much more flexible and physically intuitive than an ADC approach.

Results for two years indicate that the importance of individual independent variables used in the model is related to physical conditions in the basin in each year. In 2005, elevation dominates

the spatial distribution of melt rates for clear reasons. In 2006, inclusion of elevation, potential solar radiation, and maximum upwind slope are most important. In this year, parameters are again related to elevation effects as well as dust-on-snow events and lower snow accumulations in the upper basin. Though parameters are physically grounded, we still see them compensating for other model deficiencies. Several model deficiencies were explored to illuminate the models overall behavior. Compensating effects of parameters and dependence of simulation skill on model spatial structure make investigation of ''ideal'' spatial units impossible without detailed spatial data on accumulation and melt rates.

The new model has been successful at the very small scale of our study and presents an exciting opportunity to condition spatial response of basin snowmelt on knowledge of only very general of basin conditions such as temperature, wind, cloudiness, dust events, etc, which may be gathered from remote sensing and weather forecasts or reforecasts (hindcasts). The model parameterization is also attractive to probabilistic forecasting, particularly if deficiencies in runoff simulation can be addressed by similar parameterizations in a systematic way. However, this approach must be studied at larger scales, where streamflow is less peaky; will the ADC formulation prove more competitive at larger scales where hydrographs are much smoother? Or does the proposed methodology scale up with similar improvements in simulation skill at larger scales?

Bibliography

Anderson, E. A., 1973: National weather service river forecast system - snow accumulation and ablation model. *NOAA Technical Memorandum NWS HYDRO-17*, 217.

Anderson, E. A., 2006: Snow accumulation and ablation model snow-17. *www.nws.noaa.gov/oh/*.

Andreadis, K. and D. Lettenmaier, 2006: Assimilating remotely sensed snow observations into a macroscale hydrology model. *Advances in Water Resources*, **29**.

Balk, B. and K. Elder, 1998: Using geostatistical methods to estimate snow water equivalence distribution in a mountain watershed. *Report 66, Western Snow Conference*, 100--111.

Balk, B. and K. Elder, 2000: Combining binary decision tree and geostatistical methods to estimate snow distribution in a mountain watershed. *Water Resour. Res.*, **36 (1)**, 13--26.

Beniston, M., 2003: Climatic change in mountain regions: a review of possible impacts. *Climatic Change*, **59**, 5--31.

Beven, K., 2006: A manifesto for the equifinality thesis. *Journal of Hydrology*, **320**, 18--36.

Breiman, L., 1996: Bagging predictors. *Machine learning*, **24 (2)**, 123--140.

Breiman, L., 2001: Random forests. *Machine Learning*.

Breiman, L., J. Friedman, R. Olshen, and C. Stone, 1984: Classification and regression trees. *Wadsworth Books*, 358.

Buttle, J. and J. McDonnell, 1987: Modelling the areal depletion of snowcover in a forested catchment. *Journal of Hydrology*, **90 (1-2)**, 43--60.

Carroll, S. and N. Cressie, 1996: A comparison of geostatistical methodologies used to estimate snow water equivalent. *Journal of the American Water Resources Association*, **32 (2)**, 267--278.

Cazorzi, F. and G. D. Fontana, 1996: Snowmelt modelling by combining air temperature and a distributed radiation index. *Journal of Hydrology*, **181 (1-4)**, 169--187.

Cressie, N., 1993: *Statistics for spatial data*. Wiley, New York.

Deems, J., S. Fassnacht, and K. Elder, 2006: Fractal distribution of snow depth from lidar data. *J Hydrometeorology*, **7 (2)**, 285--297.

Deems, J. S., S. R. Fassnacht, and K. J. Elder, 2008: Interannual consistency in fractal snow depth patterns at two colorado mountain sites. *J Hydrometeorology*, **9 (5)**, 977.

Derksen, C., A. Walker, and P. Toose, 2008: Estimating snow water equivalent in northern regions from satellite passive microwave data. *Cold Region Atmospheric and Hydrologic Studies. The Mackenzie GEWEX Experience*, M. Woo, W. Rouse, R. Stewart, and J. Stone, Eds., Springer, Berlin Heidelberg, chap. 12, 195--212.

Dozier, J., 1980: A clear-sky spectral solar radiation model. *Water Resour. Res.*, **16 (4)**, 709--718.

Elder, K., J. Dozier, and J. Michaelsen, 1991: Snow accumulation and distribution in an alpine watershed. *Water Resour. Res*, **27 (7)**, 1541--1552.

Elder, K. and A. Goodbody, 2004: Clpx-ground: Isa main meteorological data. *[Available online at http:// nsidc.org/data/nsidc-0172.html.]*.

Elder, K., J. Michaelsen, and J. Dozier, 1995: Small basin modeling of snow water equivalence using binary regression tree methods. *Biogeochemistry of Seasonally Snow-Covered Catchments (Proceedings of a Boulder Symposium)*.

Erickson, T., M. Williams, and A. Winstral, 2005: Persistence of topographic controls on the spatial distribution of snow in rugged mountain terrain, colorado, united states. *Water Resour. Res*, **41**.

Erxleben, J., K. Elder, and R. Davis, 2002: Comparison of spatial interpolation methods for estimating snow distribution in the colorado rocky mountains. *Hydrol. Process.*, **16**, 3627--3649.

Everitt, B., S. Landau, and M. Leese, 2001: *Cluster analysis*. Arnold, 237 pp.

Fassnacht, S., K. Dressler, and R. Bales, 2003: Snow water equivalent interpolation for the colorado river basin from snow telemetry (snotel) data. *Water Resour. Res.*, **39 (8)**, 1208.

Grantz, K., B. Rajagopalan, M. Clark, and E. Zagona, 2005: A technique for incorporating large-scale climate information in basin-scale ensemble streamflow forecasts. *Water Resour. Res.*, **41 (10)**, W10 410.

GRASS Development Team, 2010: *Geographic Resources Analysis Support System (GRASS GIS) Software*. USA, Open Source Geospatial Foundation, URL `http://grass.osgeo.org`.

Gurtz, J., M. Zappa, K. Jasper, H. Lang, M. Verbunt, A. Badoux, and T. Vitvar, 2003: A comparative study in modelling runoff and its components in two mountainous catchments. *Hydrological Processes*, **17 (2)**, 297--311.

Hock, R., 1999: A distributed temperature-index ice-and snowmelt model including potential direct solar radiation. *Journal of Glaciology*, **45 (149)**, 101--111.

Hosang, J. and K. Dettwiler, 1991: Evaluation of a water equivalent of snow cover map in a small catchment area using a geostatistical approach. *Hydrol. Process.*, **5 (3)**, 283--290.

Julien, P. Y., B. Saghagian, and F. L. Ogden, 1995: Raster-based hydrologic modeling of spatially-varied surface runoff. *Water Resources Bulletin (Journal of the American Water Resources Association)*, **31 (3)**, 523--536.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi, 1983: Optimization by simulated annealing. *Science*, **220 (4598)**, 671--680.

Kirnbauer, R. and G. Bloschl, 1994: How similar are snow cover patterns from year to year? *Dtsch. Gewaesser. Mitteil.*, **7**, 113--121.

Kirnbauer, R., G. Blöschl, and D. Gutknecht, 1994: Entering the era of distributed snow models. *Nordic Hydrology.*

Kolberg, S. and L. Gottschalk, 2006: Updating of snow depletion curve with remote sensing data. *Hydrological Processes*, **20 (11)**, 2363--2380.

Leavesley, G., S. Markstrom, P. Restrepo, and R. Viger, 2002: A modular approach to addressing model design, scale, and parameter estimation issues in distributed hydrological modelling. *Hydrol. Process.*, **16 (2)**, 173--187.

Liston, G., 1999: Interrelationships among snow distribution, snowmelt, and snow cover depletion: Implications for atmospheric, hydrologic, and ecologic modeling. *Journal of Applied Meteorology*, **38**, 1474--1487.

Liston, G., 2004: Representing subgrid snow cover heterogeneities in regional and global models. *Journal of Climate*, **17**, 1381--1397.

Luce, C. and D. Tarboton, 2004: The application of depletion curves for parameterization of subgrid variability of snow. *Hydrol. Process.*, **18**, 1409--1422.

Luce, C., D. Tarboton, and R. Cooley, 1998: The influence of the spatial distribution of snow on basin-averaged snowmelt. *Hydrol. Process.*, **12 (10-11)**, 1671--1683.

Lumley, T. and A. Miller, 2009: *leaps: regression subset selection.* R package version 2.9.

Miller, S., 2003: Clpx-airborne: Infrared orthophotography and lidar topographic mapping. *National Snow and Ice Data Center, Boulder, CO. Digital Media.*

Molotch, N. and R. Bales, 2005: Scaling snow observations from the point to the grid element: Implications for observation network design. *Water Resour. Res.*, **41**.

Molotch, N., M. Colee, R. Bales, and J. Dozier, 2005: Estimating the spatial distribution of snow water equivalent in an alpine basin using binary regression tree models: the impact of digital elevation data and independent variable selection. *Hydrol. Process.*, **19**, 1459--1479.

Painter, T., A. Barrett, C. Landry, J. Neff, M. Cassidy, C. Lawrence, K. McBride, and G. Farmer, 2007: Impact of disturbed desert soils on duration of mountain snow cover. *Geophys. Res. Lett*, **34**.

Peters, A. and T. Hothorn, 2009: *ipred: Improved Predictors.* R package version 0.8-8.

R Development Core Team, 2009: *R: A Language and Environment for Statistical Computing.* Vienna, Austria, R Foundation for Statistical Computing, ISBN 3-900051-07-0.

Segal, M., J. Garratt, R. Pielke, and Z. Ye, 1991: Scaling and numerical model evaluation of snow-cover effects on the generation and modification of daytime mesoscale circulations. *Journal of the Atmospheric Sciences*, **48 (8)**, 1024--1042.

Serreze, M., M. Clark, R. Armstrong, D. McGinnis, and R. Pulwarty, 1999: Characteristics of the western united states snowpack from snowpack telemetry(snotel) data. *Water Resour. Res.*, **35 (7)**, 2145--2160.

Shamir, E. and K. Georgakakos, 2007: Estimating snow depletion curves for american river basins using distributed snow modeling. *Journal of Hydrology*, **334 (1-2)**, 162--173.

Sivapalan, M., 2003: Prediction in ungauged basins: a grand challenge for theoretical hydrology. *Hydrol. Process.*, **17 (15)**, 3163--3170.

Sturm, M. and J. Holmgren, 1995: A seasonal snow cover classification system for local and global applications. *Journal of Climate*, **8**, 1261--1283.

Therneau, T. M. and B. A. R. port by Brian Ripley., 2009: *rpart: Recursive Partitioning.* R package version 3.1-45.

Trujillo, E., J. Ramirez, and K. Elder, 2009: Scaling properties and spatial organization of snow depth fields in sub-alpine forest and alpine tundra. *Hydrol. Process.*, **23**, 1575--1590.

Trujillo, E., J. A. Ramírez, and K. J. Elder, 2007: Topographic, meteorologic, and canopy controls on the scaling characteristics of the spatial distribution of snow depth fields. *Water Resour. Res.*, **43 (7)**, 1--17.

Watson, F., T. Anderson, and W. Newman, 2006: Optimal sampling schemes for estimating mean snow water equivalents in stratified heterogeneous landscapes. *Journal of Hydrology*, **328**, 432--452.

Winstral, A., K. Elder, and R. Davis, 2002: Spatial snow modeling of wind-redistributed snow using terrain-based parameters. *J Hydrometeorology*, **3 (5)**, 523--538.

Winstral, A. and D. Marks, 2002: Simulating wind fields and snow redistribution using terrain-based. *Hydrol. Process.*

Appendix A

CLPX LiDAR: Details

A.1 LiDAR and in situ measurements

Figure A.1 shows significant accumulation and ablation between dates of in-situ and LiDAR observations at each ISA.

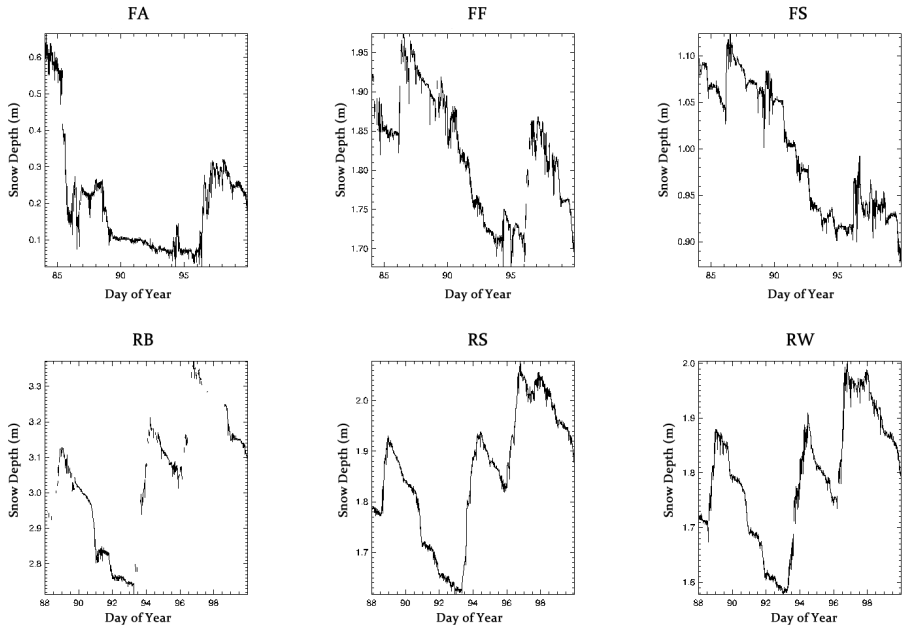


Figure A.1: Measured snow depth at each ISA weather station between date of manual, in-situ measurements and LiDAR observation. Manual measurements were taken on lowest DOY shown for each site and LiDAR on the highest.

A.2 Power-law variogram

Variograms used in ordinary Kriging interpolation of LiDAR observations to a regular grid are obtained by fitting a power-law relationship between semivariance (y) and distance (x) using the 10 nearest observations within a 10m radius of each grid point:

$$y = a * x^b <=> log(y) = log(a) + b * log(x)$$

. For a description of ordinary Kriging, see Cressie (1993).

A.3 Nearest neighbor regridding

Vegetation top elevation does not necessarily vary smoothly in space at scales of meters. Using Kriging interpolation for this variable has potential to increase the total area of vegetation by smoothing between tall and zero vegetation heights when the transition is in fact abrupt. Nearest neighbor interpolation was used for regridding vegetation height, the nearest observation within a $2.1m$ radius of each grid point as the value of that grid point.

A.4 Derivation of vegetation height from NLCD2001 class

Class descriptions accompanying the NLCD2001 data set (http://www.epa.gov/mrlc/definitions.html#2001) often give ranges for vegetation height within each class. In regression, it is important that the heights or corresponding class numbers be correctly ordered even if the values are not accurate. Vegetation height as a function of NLCD vegetation class is formulated:

$$h(c) = \begin{cases} 0m & : c \in [12, 31] \\ .5m & : c \in [21, 71, 95] \\ 2m & : c \in [52, 90] \\ 7m & : c \in [41, 42, 43] \end{cases}$$

Appendix B

Estimation from Sparse Observations: Details

B.1 Algorithm outlines

When investigating effect of change of resolution and predictor set quality on estimation skill, the same basic algorithm is used and is sketched in pseudo code here for clarity. This ‘‘inner’’ set of loops is common to all runs, it performs 100 runs to describe uncertainty resulting from random selection of observations over a set of observation densities. At 30m resolution, it is run separately for the different predictor sets. When changing resolution, it is run at the given resolution.

```
## This loop accounts estimates uncertainty due to random sampling
For iSimulation in {1, 2, ..., 99, 100}:
    Select 1600 observations from the high resolution data set.

    ## This loop degrades the maximum observation count in a comparable way.
    For nObservations in {1600, 800, 400, 200, 100, 50, 25}:

        Select nObservations from the previously selected set of nObservations.
        Locate these observations in the coarse resolution grid.
        If more than one observation exists in a coarse resolution cell, average.

        BRT:
            ## This loop acounts for uncertainty in fitting due to 10-fold partitioning of observations.
            For iRun in {1, 2, ..., 99, 100}:
              Fit 10-fold cross-validated models, track residual mean deviance.
            EndFor  ## iRun
            Prune using residual mean deviance and the 1-SE rule.

        bag:
            Fit and average 50, 10-fold bagging estimates.

        LM:
            Primary variable selection, pick 5 best predictors.
            Calculate cross-terms, select the best 15.
            Use AIC to select the final predictor set.
            10-fold cross-validation.

        Estimate the observations from 10-fold fitting.
        Calculate "Apparent" (cross-validated) skills: PE, RMSE, R2.
```

```
        Estimate the entire coarse resolution grid.
        Calculate "True" skills: PE, RMSE, R2.

    EndFor ## nObservations

EndFor ## iSimulation
```

To change resolution, the following outer loop sets the desired resolution and true snow depth.

```
## This loop changes resolution of predictor variables and of true snow.
For iResolution in {1080m / [720, 512, 362, 256, 181, 128, 91, 64, 45, 32, 23, 16, 11, 8, 6, 4, 3, 2]}:

    For iPredictor in {elevation, vegetation height, vegetation density,
                       aspect, aspect ele+veg, slope, slope ele+veg,
                       northness, northness ele+veg,
                       maximum upwind slope, maximum upwind slope ele+veg,
                       potential solar radiation, potential solar radiation ele+veg}:

        If Not(aspect or apsect ele+veg): Average predictor to coarse resolution from 1.5 m res;
            Else: Use mode over 10 deg bins to transform aspects to coarse resolution.

    EndFor ## iPredictor

    Inner Loop on iSimulations, ..., etc here.

EndFor ## iResolution
```

B.2 Models and fitting

In this section we describe each of the regression models used in this study and their individual fitting procedures. Each model is considered in conjunction with its standard method of fit. Binary regression tree (BRT), bagging tree (bag), and multiple linear regression (LM) models are supplied equal amounts of training used and this data is then estimated by the models under 10-fold cross-validation; fitting data are partitioned (as equally as possible) into 10 parts and each is withheld in turn from the estimation procedure which uses the remaining 90% of the data. In this way, 10 fits are required for estimating the entire training data set. Tough each methodology may compute a different number of runs or use the available data in different ways, none is permitted any advantage in the amount of data used to fit or estimate.

B.2.1 Binary regression trees

Starting from the supplied training data, the BRT algorithm hierarchically and recursively partitions the response in two groups via associated thresholds in the predictor variable set. This results in a hierarchical binary tree structure which partitions response in to the terminal nodes of the tree. At each node in the tree, the response variable is split via its association with *the* threshold in the predictor variables which most reduces response mean squared error when resulting groups of observations are modeled via their average value. After the algorithm terminates (under certain criteria) and the tree is pruned (to account for over fitting of the model), we are left with a set of terminal nodes or leaves of the BRT. In the spatial context, the set of leaves is a partition of the space where each leaf has a set of hierarchically similar predictor values and is represented by the mean value of all observations which were assigned to it when the tree was ''grown'' or fit. The range of a binary regression tree is a discrete set with as many values as terminal nodes of the tree, each leaf comprising apart of the spatial pattern or structure, though not necessarily contiguous. To estimate the response at a new point, its vector of predictor variables is sorted down through the bifurcating tree structure according to the fit thresholds which eventually selects a terminal node and the mean value of fitted points is the assigned response.

In a toy example, at level 0 (node $1 = 2^0$, no splits) MSE of the response is calculated on residuals from the mean value over the entire space. At level 1, with one split and two regions, MSE is calculated based on residuals in each terminal node. By definition of how the split was selected, this is the lowest achievable MSE of the response when it is partitioned on a threshold of any predictor variable.

Binary regression trees (BRT) are fit using the *rpart* package (Therneau and port by Brian Ripley., 2009) in R (R Development Core Team, 2009). ''Full'' regression trees are grown and then pruned using 100 runs 10-fold cross-validation (e.g Winstral and Marks, 2002). For each run, the standard error of fit is kept along with the residual deviance at every level of partitioning. The 1-se rule of Breiman et al. (1984, p. 78) is applied to prune the trees to the first node which is at least as accurate the node of the minimal residual devaince plus the standard error of fit. This is the common method for determining the appropriate fit of regression tree models. Nodes which

increase the mean residual deviance are considered over fit to the data and pruned away. Once the tree size is selected via this procedure, the training data are estimated with this size of a tree under 10-fold cross-validation.

While pruning, actual removal of leaves, happens after the 100 runs, there are also two *a priori* parameters which affect tree growth and effectively serve to prune trees. First, BRTs are grown under the ANOVA method to a minimum complexity parameter (cp) value of .005. This parameter has less effect on tree growth than the choice of minimum leaf size, which we have selected to be at least 1/40 of the number of total points in the data set. For both tree methods care was taken to set these growth parameters to the same values.

B.2.2 Bagging trees

Bootstrap **agg**regat**ing**, or bagging, trees (Breiman, 1996) recognize that error in BRTs depends on the training set used. Using smaller training sets and averaging the results, bagging trees aim to reduce spurious estimation errors by emphasizing agreement between a wider range of outcomes. Bagging trees can be thought of as an ensemble modeling approach where the same model is provided different initial conditions and the ensemble average provides the final estimate. Bagging bootstraps samples from the input data set, selecting approximately 63% (by bootstrap sampling statistics) of the data for fitting the model in each run. The selected data are considered ''in-bag'' and the ''out-of-bag'' data are estimated in a cross-validated way from these. In this study, for consistency in the amount of information supplied to models during the fitting process, bagging trees are grown using the same 10-fold cross validation procedure as BRTs. This reduces the training set in each run to approximately 63% (do to bootstrap statistics) of 90%, or 57%, of the full data set. Because runs are averaged, individual trees are not pruned, variance over model estimates is collapsed by averaging and we only need as many model runs as necessary to achieve estimate convergence. In this study, this number was typically 25, and 50 were runs performed at each site to ensure convergence. Parameters governing tree growth were kept to the same values as in BRT. Calculations were performed using the *ipred* package (Peters and Hothorn, 2009) in R.

B.2.3 LM - linear regression model

Multiple linear regression, or linear modeling (LM), and the theory of least squares are well-known. To obtain a model competitive with BRT and bagging, we must formulate a linear regression model which includes terms describing interaction the fundamental predictor variables used in this study. To formulate such cross-terms we first reduce the set of 13 potential predictor variables. The *regsubsets* package in R (Lumley and Miller, 2009) is used to exhaustively search for the 5 best models using any combination of 5 predictor variables. From these, the 5 variables which occur most often are selected for construction of all cross terms through order 3, which yields a total of $\binom{5}{3} + \binom{5}{2} + \binom{5}{1} = 25$ total terms. At this point we again reduce this set by using *regsubsets* to select the 5 best models using 10 predictors. This time we select the 15 most frequently occurring variables among these. Finally, we employ a stepwise search to find a predictor set which minimizes the Akaike information criteria (AIC). Note that in the previous steps, because the model size was fixed when selecting variables, selection was independent of AIC. Once we have a set of variables suspected to be most useful, then we balance model complexity against accuracy using AIC for the final variable selection.