Inductive Machine Learning for Improved Estimation of Catchment-Scale Snow Water Equivalent

David Buckingham[a,*], Christian Skalka[a], Josh Bongard[a]

[a]*Department of Computer Science, University of Vermont, Burlington, VT 05405, USA*

Abstract

Infrastructure for the automatic collection of single-point measurements of snow water equivalent (SWE) is well-established. However, because SWE varies significantly over space, the estimation of SWE at the catchment scale based on a single-point measurement is error-prone. We propose low-cost, lightweight methods for near-real-time estimation of mean catchment-wide SWE using existing infrastructure, wireless sensor networks, and machine learning algorithms. Because snowpack distribution is known to be highly nonlinear, we focus on genetic programming (GP), a nonlinear, white-box, inductive machine learning algorithm.

Because we did not have access to near-real-time catchment-scale SWE data, we used available data as ground truth for machine learning in a set of experiments that are successive approximations of our goal of catchment-wide SWE estimation.

First, we used a history of maritime snowpack data collected by manual snow courses as our ground truth estimate of mean catchment SWE. Second,

*Corresponding author

Email addresses: `dbucking@uvm.edu` (David Buckingham), `skalka@cs.uvm.edu` (Christian Skalka), `josh.bongard@uvm.edu` (Josh Bongard)

we used distributed snow depth (HS) data collected automatically by wireless sensor networks. Thus HS served as an alternative to SWE. Because HS variability is significantly greater than density variability, the primary requirement for estimating SWE over an area is an understanding of HS. We compared the performance of GP against linear regression (LR), binary regression trees (BT), and a widely used basic method (BM) that naively assumes non-variable snowpack. In the first experiment set, GP and LR models predicted SWE with lower error than BM. In the second experiment set, GP had lower error than LR, but outperformed BT only when we applied a technique for determining training and testing datasets that specifically mitigated the possibility of over-fitting.
Keywords: snow water equivalent, machine learning, wireless sensor network, snowpack modeling, genetic programming

1. Introduction

There has been extensive research on techniques for measuring and modeling snowpack because it affects many hydrological, atmospheric, and biological processes (Tappeiner et al., 2001). The accurate estimation of snowpack at the catchment scale is useful in many applications, including agricultural planning, metropolitan use, flood risk evaluation, planning of hydropower production potential, weather forecasting, and climate monitoring (Marofi et al., 2011; Schmucki et al., 2014). More than 1/6 of people globally depend on snowpack for water supplies (Bales et al., 2006), and in the western United States the majority of surface water resources is derived from snowpack (Serreze et al., 1999). However, snowpack has declined across much of the US over the last half-century (Pierce et al., 2008). The current severe drought in California, with record low snowpack measurements, threatens water supplies throughout the state (Boxalla, 2014) and highlights the importance of snowpack research. Snowpack both influences climate and responds directly to climate change (Engeset et al., 2004). While climate change warrants increased snowpack monitoring, existing techniques perform poorly under extreme climatic conditions (Molotch et al., 2005; Balk and Elder, 2000), and it has been argued that the stationarity of hydrological processes can no longer be assumed (Milly et al.). Furthermore, high costs of data gathering constrain the temporal and spatial granularity of estimation methods. New techniques are needed.

We propose new low-cost techniques for modeling snowpack using machine learning algorithms, especially genetic programming. These algorithms use data gathered from existing sensor infrastructure, and possibly short-term deployments of wireless sensor networks. The manipulation of large data sets in order to gain insight into snow accumulation, melt, and runoff has been highlighted as a necessary next step in mountain hydrology (Dozier, 2011). The long-term, overarching goal of our research project is to achieve better near-real-time (NRT), estimation of SWE at the catchment scale. By NRT, we mean automated reporting at fine-grained timescales, for example hourly. By better, we mean more accurate estimation without significantly increased infrastructure cost. Our strategy is to generate snow telemetry datasets using short-term, low-cost field campaigns that can be used by machine learning algorithms to generate snowpack models. Following field campaigns and the termination of associated measurement techniques, these models can be used for NRT SWE estimations with no new instrumentation overhead.

The key idea behind our approach is that machine learning models are able to induce mathematical relationships between input variables and some sort of "ground truth", given adequate training datasets. The machine learning method we emphasize is genetic programming (GP), which generates equations relating a dependent variable to some set of independent variables. Machine learning draws connections between input parameters and an output value, if such exist, on the basis of the ground truth data it is provided.

In our case, we argue that if we obtain multiple years of "true" average SWE for a catchment, machine learning will be able to induce a meaningful mathematical relation between telemetry, such as proximal snow pillow reading(s), and true average SWE. Then, in years when true average SWE is not available, inputs such as snow pillow readings can be translated into average SWE estimates for the catchment. This approach assumes interannual continuity in snow distributions over a catchment, which has been demonstrated by previous research (Scipión et al., 2013; Tappeiner et al., 2001; Schirmer et al., 2011).

Thus, the ideal we aim for is a generally applicable technique for inducing models that take as input parameters existing infrastructure NRT telemetry, such as snow pillow readings, meteorological data, and date/time information, and output accurate estimates of mean catchment SWE. This would allow more accurate SWE estimation to be provided without additional cost beyond that of the initial field campaign for obtaining a ground truth dataset (Figure 1).

Several theoretical and practical challenges exist on the way to achieving this goal. The purpose of this paper is to address them and make progress in three particular ways.

First, we explore the issue of what sort of machine learning approaches are best in this context. In general, we argue that techniques that are able to learn nonlinear relationships are needed due to the known non-linear nature of snow distribution in alpine environments (Tappeiner et al., 2001; Marofi et al., 2011). We also argue that so-called white-box tools are best, since these can provide physical insights for scientists (Schmidt et al., 2011). Furthermore, we emphasize resiliency against over-fitting, which is especially important given that the datasets available for machine learning may be relatively small.

Second, we investigate what sort of input parameters should be used by SWE estimation models, especially in light of practical concerns, i.e. available telemetry and datasets. In fact, we have learned that availability of data is a key issue in this effort, and defines what is possible. We acknowledge the importance of terrain effects in determining snowpack distribution, influencing both accumulation and ablation patterns (Winstral et al., 2013; Fassnacht et al., 2003; Marks et al., 1999). However, because we were unable to precisely geolocate the key snow sensors that we used with respect to topographic maps, we did not include topographic data as explicit inputs to our models. We emphasize the flexibility of inductive machine learning, which can accommodate arbitrary new input modalities. Only those that are predictive of the dependent variable of interest will be significantly incorporated into the generated models. In this paper we focus on several potential snow telemetry and meteorological inputs in order to demonstrate the applicability of our techniques to catchment-scale SWE estimation, while considering the potential for future work to explore other inputs such as topographic data.

Third, we grapple with the issue of ground-truth for catchment-scale SWE and usable datasets. Constraints on our goal were imposed by the availability of snowpack data for the training and evaluation of machine learning models. We are not aware of catchment-wide SWE datasets with sufficiently fine time granularity to support our ideal scenario. Although datasets such as those provided by the Cold Land Processes Field Experiment (National Snow & Ice Data Center) and numerous others provide catchment-scale snowpack measurements, their time granularity is on the order of several months at least. Airborne techniques in general are cost-prohibitive for real-time reporting (Bühler et al., 2011). Although satellites are used to measure snow-covered area and albedo (Dozier and Painter, 2004), satellite retrievals of SWE are not feasible. Manual snow courses provide better temporal resolution than airborne methods (e.g. biweekly) but at low spatial resolution: snow courses measure SWE at a single location. We emphasize the Snowcloud wireless sensor network, which measures HS (an effective predictor of SWE) in NRT (e.g.. hourly) at multiple locations distributed over an area of interest. However, this technology is new, and available data collected by Snowcloud deployments is limited.

2. Background and contributions

Here we briefly define and summarize the machine learning methods used in this work. These techniques are described in more detail, with special emphasis on GP, in Section 4. The basic method (BM) assumes the spatial homogeneity of SWE. It naively estimates mean catchment-wide SWE to be the same as the single-point SWE measurement taken at a snow pillow.

Linear regression (LR) fits a least-squares linear model to training data (Hastie et al., 2009). The prediction is a weighted linear combination of the input variables. Binary regression trees (BT) are nonlinear models which are generated using training data (Hastie et al., 2009). A BT model partitions a set of predictions according to the input variables such that a given set of input values results in a specific prediction. Genetic Programming (GP) is a symbolic regression algorithm that uses training data to iteratively improve a population of nonlinear models through a combination of stochastic variation and performance-based selection (Koza, 1992).

Our goal is to develop models that predict mean catchment SWE in NRT. Therefore in our ideal situation we would use a large set of accurate measurements of mean catchment SWE as ground truth data to train and evaluate models. However, the only SWE measurements available at this spatial scale are generated by airborne techniques with time resolutions that are insufficient for machine learning (e.g. twice per year). Because machine learning needs a large number of samples for model training and because we want to predict SWE in near-real-time, we require much more frequent measurements. We therefore developed a series of experiments using *available* snowpack data in lieu of NRT catchment-scale SWE measurements to explore successive approximations of our ideal scenario. Approximations of average catchment SWE, obtained via snow courses and distributed ground-based sensor readings, serve as ground truth for machine learning in our experiments. Implicit in our work is the importance of new methods for obtaining NRT catchment-scale SWE ground-truthing via low-cost distributed sensor networks.

First, we used snow course measurements, which involve the manual collection of SWE and/or HS at a single location, as a proxy for catchment-wide SWE. Although snow courses do not directly measure snowpack distribution at the catchment scale, they are likely to provide estimates that are *closer* to mean catchment SWE than do snow pillows. Snow courses take multiple measurements over approximately 200 meters, so they involve a much larger sample size than the single-point measurements of snow pillows. Furthermore, pillow under-measurement or over-measurement errors may occur when the base of the snow cover is at melting temperature (Johnson and Marks, 2004). Thus, we used snow course data as a first approximation of mean catchment SWE to provide ground-truth data for machine learning. We generated models that use readily available information such as meteorological telemetry and snow pillow measurements as input variables. These models may allow for shorter or less frequent snow courses or for their discontinuation and, because it uses previously collected data, incurs no data gathering costs. This technique is explored in Experiment Set I.

Second, we used HS data collected by the Snowcloud (Skalka and Frolik, 2014) wireless sensor network (WSN) at sites in Norway and California, each for only one snow season, as a proxy for catchment-wide SWE data. Snowcloud is a WSN-based data gathering system for snow hydrology, notable for its low-cost and ease of deployment, developed and operated by the University of Vermont. A network of light-weight sensor towers (nodes) is deployed over an area of interest for a short term field campaign to collect spatially distributed measurements of relevant meteorological processes (Figure 4). In addition to HS, Snowcloud measures air temperature, soil temperature, and solar radiation. Mesh wireless communication allows data from the entire network to be collected wirelessly by communication with a single node.

We used measurements collected from Snowcloud over the course of a single snow season to generate ground-truth estimates for model-training. Note that it may be desirable to collect data over multiple seasons as models trained on multi-year data may be more robust against internal-annual variations in snowpack distribution. Once a model has been obtained, the WSN may be recovered for re-deployment at another site. Unlike pillows and snow courses, Snowcloud collects NRT data from multiple locations, potentially capturing more of the variability of snowpack distribution than is possible with single-location measurements. Thus, we use Snowcloud data as a second approximation of catchment mean SWE to provide ground-truth data for machine learning. This technique is explored in Experiment Set II.

2.1. Suitability of machine learning

Snow pillows are large, expensive, permanent installations that measure SWE at a single location (Figure 2). The infrastructure for the automatic collection of *single-point* SWE is well established. For example, there are 830 Snowpack Telemetry (SNOTEL) sites in the United States (Snow Surveyor, 2014). However, the extrapolation from single-point measurements to surrounding areas is error prone. The spatial distribution of alpine snow cover is highly variable (Balk and Elder, 2000; Elder et al., 1991; Jost et al., 2007), due to a variety of environmental forcing effects, such as topography (Anderton et al., 2004), canopy cover (Moeser, 2010), and wind and solar exposure (Moeser, 2010; Moeser et al., 2011).

Meromy et al. (2013) studied 15 snow stations across the western United States and found that snow station biases were frequently greater than 10% of the surrounding mean observed snow depth. The flat-field areas where snow pillows are commonly located are usually not typical of more complex nearby terrain, causing the vast majority of such stations to overestimate snow depth in their vicinity (Grünewald et al., 2013). Snow cover persistence at SNOTEL sites is generally greater than the mean persistence of the watershed because SNOTEL stations do not exist in terrain classes located in upper elevations (Molotch and Bales, 2006). Molotch and Bales (2005) studied the areas surrounding six SNOTEL stations in the Rio Grande headwaters. They found that only a small fraction of grid elements were representative of mean grid SWE during accumulation, and that no elements were representative of mean grid SWE during both accumulation and ablation. Rittger (2012) found that errors based on statistical relationships between point measurements of snow and streamflow in the Sierra Nevada can reach 25% to 70% in one out of five years.

The relative importance of separate processes which govern snow distribution varies over the course of a snow season. Elder et al. (1991) summarize the various processes and explain how their influence changes over time. During the winter, accumulation and redistribution processes dominate. Precipitation is determined by regional climate and latitude as well as by local orographic effects, and redistribution by wind, avalanches, and sloughs are the primary causes of spatial heterogeneity. In the spring, however, snow distribution is controlled mainly by ablation. Of the many energy sources, solar and long-wave radiation dominate. This decreases water in a basin through sublimation and when runoff leaves the basin. It also redistributes SWE, affecting spatial variability. These dynamics highlight the need for NRT modeling of snowpack, as the forcing effects that establish snow distribution vary drastically over the course of a snow season.

However, the significant *consistency* of snowpack *between* years encourages investment into the development of reusable models. Strong inter-annual consistency in the spatial distribution of snow (Scipión et al., 2013), in *SCA* (Tappeiner et al., 2001), and in the snow depth patterns of maximum accumulation (Schirmer et al., 2011), have been observed in the Swiss and Italian Alps. In the western United States, consistent wind directions produce stable snow accumulation patterns from year-to-year (Winstral and Marks, 2014). These findings suggest a strong link between accumulation patterns and geophysical terrain and indicate that site-specific snow distribution models may be able to accurately characterize snowpack distribution over multiple years.

It may also be desirable to produce non-cite-specific models. Trained at catchments where ground truth data is available, and making use of predictor variables that vary between catchments, such as topography, such models could then be applied to catchments where no ground truth data exists. The precise coordinates of the snow pillows we used in California are not publicly available, preventing us from geolocating them with respect to topographic data. We therefore focus on site-specific models and use model inputs that vary over time at a given catchment.

2.2. *Why GP?*

It has been demonstrated that the relationships between snow distribution and the topographic and meteorological forcing effects include nonlinearities (Tappeiner et al., 2001). The spatial distribution of SWE is nonlinear because it is influenced simultaneously by numerous processes including accumulation, ablation, and snow drifting (Marofi et al., 2011). GP can produce both linear and nonlinear models. If the data used to train GP contain only linear relationships, the resulting models will be linear, and the performance of GP will be similar to that of LR.

White-box models, such as those produced by GP, can be interpreted by human analysis, potentially yielding new information about the modeled data (Schmidt et al., 2011). Some nonlinear regressors, such as artificial neural networks, produce models that are difficult or impossible to interpret. GP trees, however, can be expressed as mathematical equations (Figure 3). It is possible that by examining these equations domain experts could gain novel insight into the processes governing snow distribution.

Unlike regression techniques that constrain the form of the regressor, GP can combine operators, variables, and constants into arbitrary arrangements. GP does not require any assumptions about the form that a model should take: form is left open to inductive search. By generating models that use predictor variables in unexpected ways, GP may help discover previously unknown relationships underlying snowpack distribution.

Finally, as will discuss further, GP may be augmented with multi-objective optimization, which constrains GP to produce parsimonious models. This mitigates against over-fitting, a significant concern in the case that relatively small datasets are used for machine learning.

While many regression techniques possess one or more of these desirable qualities, GP possesses all of them, making it an ideal candidate for snowpack modeling.

2.3. The primacy of snow depth

While SWE is a product of HS and density (ρ), there is significant evidence that HS is the essential determining metric for SWE estimation. Models have been developed to derive ρ estimates from HS measurements (Logan, 1973; Sturm et al., 2010), and measurements of HS are highly predictive of SWE (Adams, 1976). Analysis of the spatial variability of HS and ρ has revealed that the variability of HS is significantly greater than that of ρ (López-Moreno et al., 2012). Variation of SWE is therefore overwhelmingly a product of HS variation (Moeser et al., 2011; Molotch et al., 2005; Sturm et al., 2010; Elder et al., 1991, 1998). The effect of ρ variation on SWE is small by comparison, and estimates of areal SWE derived from one or several SWE measurements can be greatly improved by incorporating a larger number of HS measurements (Elder et al., 1998; Moeser et al., 2011), which are much less labor intensive than manual SWE measurements (Sturm et al., 2010). Snowcloud, which provides ground-truth data Experiment Set II, measures HS. Therefore, as has been done elsewhere (Winstral et al., 2002), we use HS as a "surrogate for SWE".

2.4. Related work

Moeser et al. (2011) explored three models for estimating SWE in the area around a meteorological station using ground based measurements. The first model used meteorological data such as air temperature and solar radiation, tree canopy cover measurements, and *HS* measurements collected by the Snowcloud WSN, as well as a single-point *SWE* measurement. The second model used multiple *HS* measurements and single-point *SWE* measurements, but no meteorological or tree canopy data. The third model used meteorological and tree canopy data, along with multiple *HS* measurements, but no single-point *SWE* measurement. The meteorological and tree-canopy inputs used in these models were obtained through a two-phase statistical analysis using correspondence analysis and LR. It was found that increasing the number of *HS* measurements can improve areal *SWE* measurements because *HS* varies more than snow density. While this work used linear modeling; our work expands upon it by developing nonlinear models.

Grünewald et al. (2013) used LR to model *HS* distribution on the catchment-scale at seven sites using topographic parameters. They found that elevation, slope, and northing are good predictors of snow distribution. Models calibrated to local conditions performed much better than a global model that combined data from all the sites. They suggest that local statistical models of snowpack distribution based on topographic parameters cannot be transferred to different regions. However, models developed one year *are* good predictors at the same site for other years. Instead of LR, our work emphasizes nonlinear regression.

Marofi et al. (2011) compared three methods for modeling *SWE*: multivariate nonlinear regression (MNLR), artificial neural networks (ANN), and a neural network-genetic algorithm (NNGA), where genetic algorithms were used to parameterize ANNs and the learning process. ANN performed better than MNLR, suggesting that computational intelligence approaches may outperform MNLR for modeling SWE. NNGA performed better than ANN, suggesting that evolution-inspired genetic algorithms can be used to develop effective models of SWE. Tabari et al. (2010) estimated HS and SWE using multiple methods and also found that NNGA provided the best results. Unlike neural networks, GP produces white box models.

Tappeiner et al. (2001) compared the performance of LR-based and ANN-based snowpack models, which used topographic and meteorological data to estimate SWE. The authors compared the results of LR with ANN to estimate the degree of necessary nonlinearity in SWE modeling. The ANN performed significantly better than LR, demonstrating nonlinearity in the relationships between topographic and meteorological variables and SWE.

Several studies have used binary regression trees, which are nonlinear, white-box models, to model snowpack. Winstral et al. (2002) derived terrain-based parameters from digital elevation models (DEM) which were used as input variables to binary regression trees. They found that binary tree models based on terrain-based parameters as well as elevation, solar radiation, and slope performed better than models based only on elevation, solar radiation, and slope. Elder et al. (1998) modeled the distribution of SWE by merging remotely sensed snow-covered area data with binary tree models applied to field measurements of HS and SWE. Balk and Elder (2000) combined binary regression trees, which related HS to solar radiation, elevation, slope and vegetation cover, with kriging of manual snow survey measurements and snow-covered area determined by aerial photographs, to estimate SWE. They found that this technique was an improvement over previous methods.

While the tree-based models alone explained 54-56% of HS variance, the combined depth estimates explained 60-85%. Anderton et al. (2004) used binary regression trees to relate HS and disappearance date to terrain indices. They found that the topographic effects on snow redistribution by wind primarily determined SWE distribution at the start of the melt season which, more than melt rates, determined the patterns of snow disappearance. Molotch et al. (2005) compared binary regression tree models using various sources of DEMs. They found that differences in DEMs make significant differences in modeled snowpack distribution.

We observe that the binary regression trees used in this previous work are classifiers which, given a set of input values, select from a finite set of possible values. GP, on the other hand, is a regressor, and uses input values to produce an output value taken from the real numbers. In Experiment Set II we compare the performance of BT to GP. Unlike this previous work which used binary regression trees to produce spatially distributed models of snowpack, our models predict a single value: mean HS measured by a wireless sensor network.

Marks et al. (1999) also developed spatially distributed models. They used topographic data to determine estimates of radiation, temperature, humidity, wind, and precipitation for use in a coupled energy and mass-balance model called ISNOBAL. Simulations conducted at several basins all closely matched independently measured SWE.

Recent research has made significant advances in simulating the effects of wind on snow distribution. Winstral et al. (2009) developed a simplified wind model that uses upwind topography to accurately predict wind speeds.

Winstral et al. (2013) developed a snow distribution algorithm that uses terrain structure, vegetation, wind, and precipitation data to simulate wind-affected snow accumulation. It accurately predicted disparate snow distribution caused by inhomogeneous precipitation and redistribution by wind. Winstral and Marks (2014) analyzed the effects of wind on snow distribution. They found that high wind speeds increased snow depth variability and that forested sites decreased variability by moderating wind effects. Furthermore, consistent wind directions produced accumulation patterns that were stable between years.

Sturm et al. (2010) used HS, day of the year, and climate classes to estimate snowpack density. Estimated snowpack density was used to convert HS measurements into SWE estimates. The use of climate classes, such as Alpine, Maritime, and Tundra, improved density estimates, and 90% of computed SWE values fell within 8 cm of measured values.

SNOWPACK is a numerical model that simulates snowpack layering characteristics such as density, temperature, and crystal type (Bartelt and Lehning, 2002). Schmucki et al. (2014) analyzed the performance of SNOWPACK when predicting HS and SWE given input data commonly available from weather stations. They found that SNOWPACK successfully modeled HS with a mean error of less than 8 cm and SWE with a mean error of less than 55 mm, but that precipitation measurements must be either corrected or calibrated for correct modeling.

Chang and Li (2000) used multivariate regression to model snow distribution using independent variables derived from a DEM. These variables included easting, southing, elevation, slope, and aspect, as well as more complex derived measures such as “shadow”, which considers the angle of solar illumination, and various metrics of ground curvature. This multivariate regression of derived topographic features performed better at estimating SWE distribution than traditional interpolation methods.

Guan et al. (2010) found that atmospheric rivers (ARs), are associated with intense storms that contribute a large percentage of snow during most years. Because AR storms are relatively warm (close to 0.6,$^{\circ}$ C), the participation of AR participation into snowfall versus rainfall is sensitive to minor variation in surface air temperature.

Rittger et al. (2011) combined satellite-based measurements of snow-covered area with energy balance calculations to retroactively calculate distributed SWE at the date of maximum accumulation, using the the “reconstruction” technique originally developed by Martinec and Rango (1981). This calculation was then used to evaluate the accuracy of two real-time models. They found that at elevations below 1500 m, the real-time models overestimated SWE because of early season melt, and at elevations above 3000 m, the real-time models underestimated SWE because they do not sample these higher elevations. It is possible that this technique could be used to evaluate the effectiveness of the inductive learning methods that we describe in this work.

3. Training data and model inputs

Inductive machine learning requires substantial datasets for developing and evaluating models, and we acquired extensive hydrological and meteorological data for use in our experiments. Lacking access to accurate measurements of mean catchment SWE with NRT granularity, we focused on two types of available datasets that are approximations of mean catchment SWE. First, we consider a record of SNOTEL snow courses from the Sierra Nevada. We observe that SNOTEL snow courses are intended to provide an estimation of SWE at a particular elevation (United States Department of Agriculture, 2014), though in fact they are linear transects of SWE samples. Second, we consider a record of Snowcloud sensor network readings from Norway and California. Snowcloud sensor networks provide distributed coverage of snow depth readings for the deployment area, as well as fine time granularity, and can support better estimations of mean catchment SWE than periodic snow courses.

3.1. Experiment Set I data

Experiment Set I uses data collected from several sites across California. There were three main types of data: SWE from manual snow courses, SWE measurements from snow pillows, and air temperature data.

The California Data Exchange Center (CDEC) provided an extensive database of snow data. SWE measurements were available from 63,287 snow courses conducted at 404 sites across California between 1930 and 2012. The snow courses that we used, which are described in Table 1, were performed monthly, were about 200 meters long, and consisted of 10 measurements, the mean of which was recorded. These mean snow course measurements serve as ground-truth estimates of mean catchment-wide SWE in Experiment Set I. CDEC also maintains single-point SWE measurement data from snow pillows at sites throughout California. Of the 404 snow course sites, 59 are co-located with snow pillows.

The National Climate Data Center (NCDC) maintains meteorological data, such as air temperature, wind speed, and solar radiation measurements, collected at thousands of weather stations across the United States. Four NCDC stations are located within 20 miles of CDEC snow courses. We arbitrarily chose a 20 mile cutoff because we suspected that meteorological activity within 20 miles of a snow course might be predictive of measurements at the snow course. If this data is not predictive, the models generated by machine learning will not make significant use of it.

Significant gaps exist in the NCDC database, and of the various sensor modalities, air temperature data is the most complete. Using more meteorological inputs and necessarily fewer data samples, we had previously been unable to generate effective models of SWE. For Experiment Set I, therefore, air temperature is the only meteorological input, making possible the composition of the large data sets necessary for effective machine learning and demonstrating the use of readily available meteorological data to augment the prediction of SWE. Air temperature is known to be a highly effective predictor of melt rate because it is correlated with longwave atmospheric radiation, the most important heat source for snowmelt (Ohmura, 2001). Air temperature is made accessible to the models by three variables: *minTemp7*, *maxTemp7*, and *meanTemp7*, which aggregate daily values over the seven days inclusively preceding the day for which SWE is estimated.

We used the temporal and spatial intersection of available data from these three sources (CDEC snow courses, CDEC snow pillows, NCDC air temperature data) to construct eight datasets, based on eight snow course sites. These snow courses were selected because they are coincident with either snow pillow data, NCDC air temperature data, or both, over a range of time that includes a large number of samples points (greater than 100 except for one site). Some days are skipped because one or more data source is unavailable. All sites include snow course data, which serves as a ground truth estimate of mean catchment SWE. Three include snow pillow data but no meteorological data, three include meteorological data but no pillow data, and two include both snow pillow data and meteorological data. The constructed datasets are summarized in Table 2.

3.2. Experiment Set II data

Experiment Set II used HS data collected from multiple sources in Norway and in California. Four Snowcloud sensor nodes have been deployed in Sulitjelma, Norway since January, 2013. Data collected between January and April, 2013 were used in this experiment. During that time, each node sampled HS every six hours. We averaged HS measurements from the four nodes and then over each day to produce 93 estimates of mean catchment HS. For the few days when HS measurements from one or more sensor nodes was missing, the mean of the available measurements was used. These values served as ground-truth HS for experiments at Sulitjelma.

Approximately 16 km away from the Sulitjelma Snowcloud deployment site is Storstilla nedanför Balvatn in Nordland County, station number 164.12.0 (Balvatn). The Balvatn station records both HS and SWE. Daily HS measurements collected at Balvatn compose the HS input variable to models developed for Sulitjelma in Experiment Set II.

Six Snowcloud wireless sensor network sensor nodes were deployed within the Sagehen Creek Field Station, near Truckee, California, from January to

May, 2010. Each node reported daily HS measurements, which we averaged to generated 99 estimates of mean catchment SWE. For the few days when HS measurements from one or more sensor nodes was missing, the mean of the available measurements was used. These values served as ground-truth HS for experiments at Sagehen. Note that the same WSN data was used by Moeser (2010).

In order to assess the significance of the *source* of single-point HS input variables, we developed models for estimating mean HS at the Sagehen Snowcloud deployment using inputs from two different CDEC sites, *Independence Camp* ($\mathcal{IDC}$) and *Huysink* ($\mathcal{HYS}$). Note that in Experiment Set I, snow courses at CDEC sites provide SWE ground truth (dependent) data, while in the California experiments in Experiment Set II single-point HS measurements at CDEC sites provide input (independent) data. $\mathcal{IDC}$ is approximately 5.5 km away from the Snowcloud deployment and, like Sagehen, is on the Eastern side of the Sierra crest. $\mathcal{HYS}$ is approximately 30 km away, on the Western side of the crest.

3.3. Time of year

Because the dynamics underlying snowpack distribution vary over the course of a snow season, for example between periods dominated by deposition and periods dominated by ablation, we introduce *time of year* (TOY) as an independent variable for both experiment sets. This allows models to distinguish parts of the snow season. Time of year is an integer value expressing the number of days since the beginning of the snow season.

3.4. Preparation of datasets

We define a dataset, D, for each experiment (each row of Table 8 and each location in each row of Table 7). Elements of a dataset D take the form of a 3-tuple:

$$< T, \theta, \vec{p} >$$

where T, time, specifies a calendar date, θ is ground truth, an estimate of the true value of the independent variable, and $\vec{p}$ is a vector of predictor variables. T is unique in D so that no two data samples in D have the same T:

$$\forall < T_1, \theta_1, \vec{p_1} >, < T_1, \theta_2, \vec{p_2} > \in D \qquad \theta_1 = \theta_2 \qquad \text{and} \qquad \vec{p_1} = \vec{p_2} \tag{1}$$

In Experiment Set I, θ is an approximation of mean catchment SWE derived by manual snow course. In Experiment Set II, θ is an approximation of mean catchment *HS* derived from Snowcloud WSN measurements.

Depending on the experiment, $\vec{p}$ includes some combination of *HS* measured at a snow pillow, SWE measured at a snow pillow, TOY (an integer representation of T), and air temperature, (which is composed of three variables: *minTemp7*, *maxTemp7*, and *meanTemp7*). The *Model inputs* columns of Table 7 and Table 8 specify the contents of $\vec{p}$ for each experiment.

In order that a model developed from D may be evaluated on new, unseen data, D is divided into training, ϱ, and testing, τ, subsets. The training set is twice as large as the testing set:

$$D = \varrho \cup \tau \qquad \text{and} \qquad \varrho \cap \tau = \emptyset \qquad \text{and} \qquad |\varrho| = 2|\tau| \tag{2}$$

However, GP and BT require that ϱ be further divided into grow, g, and selection, s, subsets:

$$\varrho = g \cup s \qquad \text{and} \qquad g \cap s = \emptyset \qquad \text{and} \qquad |g| = |s| \tag{3}$$

In all experiments, D is first divided into g, s, and τ:

$$D = g \cup s \cup \tau \qquad \text{and} \qquad g \cap s \cap \tau = \emptyset \qquad \text{and} \qquad |g| = |s| = |\tau| \tag{4}$$

For BM and LR, g and s are simply combined into ϱ and used as training data. As discussed in more detail in Section 4, in the case of GP and BT g is used to generate a set of models and s is used to determine which one should be kept and evaluated on τ. In any case, ϱ is used to obtain a single model, which is then exposed to τ to evaluate its ability to predict unseen data.

We explored several methods for dividing D into g, s, and τ. In Experiment Set I and in the first part of Experiment Set II (Experiment Set II: *Random Division*), the chronologically ordered D is randomly shuffled and then divided into thirds, as illustrated by Figure 7a. This method has the effect that a large portion of the training data is likely to be temporally proximal to testing data.

As discussed further in Section 5, we found in Experiment Set II that the temporal proximity between ϱ and τ caused machine learning to map

TOY values to estimates of HS. The models memorized the data rather than capturing the relationships among the data. We therefore conducted Experiment Set II: *4 Bins*. Instead of shuffling D, we maintained its ordering and divide it into four chronologically contiguous bins. Each bin is then subdivided into three chronologically contiguous subsets which are assigned to g, s, and τ. This method is illustrated by Figure 7b. We also conducted Experiment Set II: *3 Bins* and Experiment Set II: *2 Bins*, as illustrated in Figures 7c and 7d. As we move from Experiment Set II: *Random Division* to Experiment Set II: *2 Bins*, the division of D transitions from finer to coarser temporal granularity. As this granularity becomes coarser, it becomes more difficult for machine learning to use TOY to simply memorize data. However, it also becomes more difficult for models to capture the variation of the dynamics of snowpack distribution over the course of a snow season. In the extreme hypothetical example of 1 bin, models would be trained on measurements taken during the first two thirds of the snow season and then evaluated on measurements taken during the final third. It would be impossible to model relationships that are unique to the end of the snow season.

In order to introduce stochasticity into the division D and thus allow the repetition of experiments to produce a distributed sample of results, a randomly generated offset shifts the starting point of the division. Figure 7e illustrates the effect of this offset in the case of three bins.

4. Calculation

In this section we first describe how we compared the performance of different snowpack modeling techniques. We then describe the various modeling techniques that we used, with special emphasis on GP.

4.1. Comparing estimation methods

In order to compare the performance of two machine learning techniques, M and M', on a dataset D, D is divided into complementary subsets ϱ and τ. Methods M and M' are applied to ϱ to produce estimators $\hat{\theta}$ and $\hat{\theta}'$. This process may be deterministic or nondeterministic. In Experiment Set I and Experiment Set II: *Random Division*, nondeterminism is introduced by the random division of D. GP introduces further nondeterminism by the stochasticity of the GP algorithm. The BT algorithm is deterministic when a single input variable is used, but nondeterministic when applied to multiple input variables. Estimators $\hat{\theta}$ and $\hat{\theta}'$ are applied to τ to determine the mean absolute errors of the estimators $\mathrm{MAE}(\hat{\theta})$ and $\mathrm{MAE}(\hat{\theta}')$, as we will discuss in section 4.2.

This process of randomly dividing D and applying M and M' to obtain $\mathrm{MAE}(\hat{\theta})$ and $\mathrm{MAE}(\hat{\theta}')$ is repeated 30 times, resulting in vectors of estimator errors $\vec{e}_M$ and $\vec{e}_{M'}$ each with cardinality 30. We consider $\vec{e}_M$ and $\vec{e}_{M'}$ to be statistical samples of errors drawn from the population of errors that method M and M' could produce given D. We chose to collect 30 samples because a sample size of at least 30 allows the Central Limit Theorem to be safely applied without assuming a normal population distribution, permitting the application of the one-sample t-test to calculate confidence intervals and the paired two-sample t test to test hypotheses.

The means of $\vec{e}_M$ and $\vec{e}_{M'}$ are unbiased estimates of the true population means μ_M and $\mu_{M'}$. To find out if M' outperforms M on dataset D we the pose hypotheses:

$$H_0 : \mu_{M'} = \mu_M \qquad \textit{(Null hypothesis)}$$
$$H_a : \mu_{M'} < \mu_M \qquad \textit{(alternative hypothesis)}$$

and apply the Student's t-test for paired samples to $\vec{e}_M$ and $\vec{e}_{M'}$. If the Null hypothesis is rejected, we say that method M' produces lower error (performs better) on dataset D than does M. We report the p-value, the probability that the we have performed a Type I error by rejecting a true Null hypothesis.

4.2. Evaluating estimator error

Recall that an element d of dataset D takes the form $< T, \theta, \vec{p} >$ and that D has been divided into ϱ and τ. An estimation method M is applied to $\varrho \subset D$ to generate an estimator $\hat{\theta}$, which is a function from predictor variables $\vec{p}$ to dependent variable y, an estimate of θ.

$$\hat{\theta} : \vec{p} \rightarrow y \qquad y \approx \theta$$

The error of $\hat{\theta}$ on an input vector is the difference between the estimate it produces and ground truth.

$$\mathrm{E}_{\hat{\theta}}(\vec{p}) = \hat{\theta}(\vec{p}) - \theta \tag{5}$$

The error is calculated on each sample in τ to determine the mean absolute error of the estimator:

$$\mathrm{MAE}(\hat{\theta}) = \frac{\sum_{i=1}^{k} |\mathrm{E}_{\hat{\theta}}(\vec{p}_i)|}{k} \tag{6}$$

Where

$$\tau = (d_1, ..., d_k) \qquad \text{and} \qquad \vec{p}_i \in d_i \in \tau \subset D$$

4.3. Basic method

The basic method (BM) assumes that SWE as measured at a snow pillow is representative of catchment-wide SWE. It naively estimates ground truth (snow course-derived) SWE to be the same as the independent variable (snow pillow-derived) SWE measurement. Error in the predictive power of BM expresses the difference between snow pillow measurements and snow course SWE measurements. If x represent SWE measured at the snow pillow, then

$$x \in \vec{p} \qquad \text{and} \qquad \hat{\theta}(\vec{p}) = x \tag{7}$$

Unlike the more sophisticated machine learning techniques, BM does not make use of training data to generate a model.

4.4. Linear regression

Linear regression (LR) fits a least-squares linear model to training data which is then evaluated on test data (Hastie et al., 2009). LR expresses the linear relationships between independent and dependent variables. We used the *gsl_multifit_linear* function from the GNU Scientific Library (GSL, 2014) to perform LR. We include LR in order to gain insight into the data we are using. LR will perform less well than nonlinear techniques only if the modeled data contain nonlinear relationships.

4.5. Genetic programming

GP is an evolutionary algorithm, inspired by biological evolution, that iteratively evolves populations of parse trees to perform symbolic regression (Koza, 1992). In this work, the trees are snowpack models, estimator functions, that use available independent variables to estimate mean SWE (Experiment Set I) or HS (Experiment Set II) at the catchment scale. Tree terminals are input variables and constants, while internal nodes are arithmetic operators. The operators we used are listed in Table 5.

We used the lil-gp Genetic Programming System (lil-gp Genetic Programming System, 2013), an open source implementation of GP, in order that we might make any needed modifications. We modified lil-gp to implement multi-objective Pareto optimization.

GP begins by generating a starting population of randomly constructed trees. Each tree in the population is evaluated on training data to determine its fitness, defined as the inverse of mean error. Trees are selected according to their size and fitness to produce the population for the next generation. Genetic operators make stochastic modifications to the new trees, randomly perturbing their fitness values. The genetic operators we used were *mutation* and *crossover*. Mutation, which is applied to 40% of new trees, selects a subtree at random and replaces it with new, randomly generated subtree. In crossover, which is applied instead of mutation 60% of the time, two parent trees exchange subtrees, resulting in two novel offspring. Crossover allows recombination of subtrees from existing models while mutation introduces new subtrees to the population, maintaining genetic diversity. Because it is likely that subtrees taken from existing, partially evolved models will be more useful than new, randomly generated subtrees, crossover is applied more frequently than mutation. This process is repeated for many generations, over time generating populations of increasing fitness.

The average wall-clock time for one experiment using the Vermont Advanced Computing Core (VACC) supercomputer was 333 seconds for Experiment Set I (3000 generations) and 1,207 seconds for Experiment Set II (10,000 generations). The total wall-clock time for all of Experiment Set I was approximately 89 hours. The total wall-clock time for all of Experiment Set II was approximately 321 hours.

One challenge facing GP, like all techniques for deriving a model from training data, is over-fitting. An over-fit model performs well on training data but does not generalize well and fails on unseen data. It memorizes values instead of capturing the mathematical relationships among the data.

The size of a GP model (number of nodes in a tree) constrains its complexity and fitness. Trees that are too small are too simple to accurately model the data and are under-fit. They perform poorly on both training and testing data. Trees that become too large perform extremely well on training data but, due to over-fitting, perform poorly on unseen data. Somewhere between these extremes lies the best, non-over-fit model.

In order to explore the gradient from small, under-fit models to large, over-fit models, we added multi-objective Pareto optimization to lil-gp. Pareto optimization applies evolutionary pressure toward multiple simultaneous goals, in this case low error and small model size, by producing a population (front) of non-dominated models. A tree is dominated by another tree if it is inferior by all objectives, i.e. it is both larger and has lower fitness. A Pareto front (non-dominated front) consists of a set of trees such that no tree is dominated by any other tree on the front. The non-dominated trees are selected at each GP generation so that each population is a non-dominated front, including the final population. The result of GP is therefore a set of trees of various sizes. We set an absolute upper bound at size 30 because we had observed that models with size larger than 30 were consistently over-fit. Arranged from smallest to largest, the error of these trees on the training data decreases monotonically. Error on unseen data, however, will decrease only to a point, and will then increase beyond some tree size as the models become over-fitted.

At this point is the tree size that will maximize performance on ϱ without over-fitting. Models no bigger than this can express features common to both training and testing data but cannot express features that are unique to the training data. However, this size threshold is not known while generating models because test data is not available. It must remain *unseen* for model testing.

One possible technique for selecting a model exploits a common feature of Pareto fronts. Pareto fronts often exhibit a characteristic *knee* point where a small improvement in one objective would lead to a large deterioration in another objective (figure 8). There are several different technical definitions that can be used to automate knee identification (Deb and Gupta, 2011). In many multi-objective optimization applications the knee represents a good compromise among objectives (Das, 1999; Deb and Gupta, 2011). However, our goal is to identify the model that can be expected to perform best on unseen data. We therefore developed a novel *selection set* method for selecting a model from the Pareto front.

In the *selection set* method, the training data is further divided into two subsets of equal size, a growth set, g, and a selection set, s (Equation 3). GP is applied to g to obtain a Pareto front. Each model on the front is then evaluated on s. GP returns the model that performs best (lowest error) on s. We used the *election set* method in all experiments.

4.6. Binary regression trees

We include BT in Experiment Set II in order to compare GP to another nonlinear, less computationally demanding, modeling technique. Erxleben et al. (2002) compared the performances of four spatial interpolation methods to estimate *SWE* and found that a method combining binary regression trees with geostatistical methods was more accurate than other methods. We used the DecisionTreeRegressor class of the Scikit-learn machine learning module for Python (Pedregosa et al., 2011). This software implements the Classification and Regression Trees (CART) algorithm, which is similar to C4.5 (Hastie et al., 2009). BT is parameterized by the maximum tree depth; we used default options for other parameters. As with GP, the data for BT was divided into g, s, and τ. For each experiment, a set of trees was trained on g such that the nth tree had a maximum depth of n. The maximum value of n was determined by incrementing n until further increase did not result in larger trees. The maximum value of n varied between 7 and 13.

Like the Pareto front produced by GP with multi-objective optimization, this methods results in a gradient of models ranging from very small models with high error on g to very large models with low error on g. Each is evaluated on s and the one with the lowest error is returned by BT to be evaluated on τ in order to determine model error. Thus, we apply the same *selection set* method to BT as to GP in order to discourage over-fitting and to provide similar exposure to the data so that the performance of the techniques may be compared. Note, however, that in the case of GP, multi-objective optimization applies pressure toward model parsimony continuously over the course of the evolution of a population of models. In the case of BT, the selection set method is applied once to a set of models after they have been generated.

5. Experiments: descriptions and results

In this section we describe the experiments conducted in Experiment Sets I and II and report the results.

5.1. Experiment Set I

In Experiment Set I measurements from snow courses provided ground-truth *SWE* data. We developed models to predict snow course *SWE* at eight different sites in California where snow courses had been conducted (Table 1). Three sites ($\mathcal{CAP}$, $\mathcal{GRZ}$, $\mathcal{KTL}$) were located at snow pillows but are not near any NCDC weather stations. Three sites ($\mathcal{NTH}$, $\mathcal{SPD}$, $\mathcal{MSH}$) were near NCDC stations but are not at snow pillows. Two of the snow course sites ($\mathcal{HYS}$ and $\mathcal{HIG}$) were located at snow pillows and are also near NCDC stations.

First, we conducted experiments at sites with snow pillows but without weather stations ($\mathcal{CAP}$, $\mathcal{GRZ}$, $\mathcal{KTL}$). These experiments explored how well linear and nonlinear models predict snow course-derived ground truth *SWE* using only snow pillow measurements. Inputs to the models were pillow SWE and TOY. At each site we developed models with three combinations of input variables: TOY alone, pillow SWE alone, and TOY combined with pillow SWE. In each case, we compared the performance of GP, LR, and BM.

Second, we conducted experiments at sites near weather stations but without snow pillows ($\mathcal{KTL}$, $\mathcal{MSH}$, $\mathcal{NTH}$). These experiments explored how well linear and nonlinear models predict snow course-derived ground truth SWE using air temperature data without access to snow pillow SWE measurements. Inputs to the models were *air temperature* and TOY. At each site we develop models with three combinations of input variables: temperature alone, TOY alone, and temperature combined with TOY. In each case, we compare the performance of GP to LR. BM was not evaluated because it requires the pillow SWE variable.

Third, we conducted experiments at sites that are near weather stations and have snow pillows ($\mathcal{HIG}$, HYS). These experiments explored how well linear and nonlinear models predict snow course-derived ground truth SWE using both pillow SWE measurements and air temperature data. Inputs to the models were SWE, *air temperature*, and TOY. At each site we develop models with seven unique combinations of input variables: temperature alone, TOY alone, pillow SWE alone, temperature and TOY together, temperature and pillow SWE together, TOY and pillow SWE together, and, finally, temperature, TOY, and pillow SWE together.

Table 7 summarizes Experiment Set I. Each experiment was repeated 30 times to generate error samples for each method. Figures 9-12 plot the mean values of the samples. Error bars indicate 95% confidence intervals, i.e.

sample mean $\pm$(SEM $\times$ 1.96). GP and LR had similar error, but both had lower error than BM with p-value less than 0.001 in all cases.

The mean ground truth SWE value in inches at each site was: $\mathcal{CAP}$: 45.08, $\mathcal{GRZ}$: 49.47, $\mathcal{KTL}$: 27.08, $\mathcal{MSH}$: 68.78, $\mathcal{NTH}$: 13.29, $\mathcal{SPD}$: 27.47, $\mathcal{HIG}$: 23.39, $\mathcal{HYS}$: 41.95.

5.2. Experiment Set II

In Experiment Set II models predicted HS instead of SWE. While research on the influence of meteorological factors on snowpack distribution is extensive (Logan, 1973; Elder et al., 1991; Schmucki et al., 2014; Hock and Noetzli, 1997), the inclusion of meteorological inputs does not always improve snowpack model performance (Moeser, 2010), and the inclusion of air temperature data did not improve model performance in Experiment Set I. Therefore, in Experiment Set II we focus on TOY and single-point HS measurements as predictors of mean catchment HS. Instead of manual snow course data as in Experiment Set I, ground-truth data are derived from HS measurements collected by the Snowcloud WSN. We compared the performance of three machine learning techniques: LR, BT, and GP.

We developed estimators to predict HS at two sites: Sulitjelma, Norway and the Sagehen Experimental Forest, California. At Sulitjelma, model inputs were combinations of HS at Balvatn and TOY. At Sagehen, model inputs were combinations of HS at $\mathcal{HYS}$, HS at $\mathcal{IDC}$, and TOY. Table 8 summarizes Experiment Set II. We repeated each experiment four times (*Random Division*, *4 Bins*, *3 Bins*, *2 Bins*) and each of these 30 times to generate error samples.

Each experiment was repeated 30 times to generate error samples for each method.

Figures 13-16 plot the mean values of the samples, i.e. the error of the modeling techniques on testing data. Error bars indicate 95% confidence intervals, i.e. sample mean $\pm(\mathrm{SEM} \times 1.96)$. Stars indicate p-values for the Student's paired t-test with the hypothesis the GP does not have lower error than BT, i.e. the probability that GP does not outperform BT. One star, *, indicates that p is less than 0.05, ** indicates that p is less than 0.01, and *** indicates that p is less than 0.001. Similarly, plus signs indicate p-values for the hypothesis that GP does not have lower error than LR, i.e. the probability that GP does not outperform LR. One plus sign, +, indicates that p is less than 0.05, and ++ indicates that p is less than 0.01. The mean ground truth *HS* value at Sulitjelma was 1.1900 m. The mean ground truth *HS* value at Sagehen was 0.728 m.

Figures 17-20 plot the mean sizes of the models whose performance is reported in figures 13-16. In the case of GP and BT, these are the models selected using the *selection set* method. For GP, model size is the number of nodes in the GP tree. For BT, model size is the number of nodes in the binary tree. For LR, model size is the number of operators and values, specifically 5 in the case of a single independent variable and 9 in the case of two independent variables. Stars indicate p-values for the Student's paired t-test with the hypothesis the GP models are not smaller than BT models. One star, *, indicates that p is less than 0.05, ** indicates that p is less than 0.01, and *** indicates that p is less than 0.001.

6. Discussion

In this section we discuss the results of our experiments, offer some hypotheses to explain our findings, and suggest ways to explore and test these hypotheses. We are especially interested in assessing the performance of GP and drawing conclusions that can inform future research.

6.1. Experiment Set I

In Experiment Set I GP performed at least as well as other methods in all experiments. This result was expected because GP is capable of generating the same models as LR and BM. We did not perform hypothesis tests comparing GP with LR because visual inspection of error means and 95% confidence intervals (figures 9-12) suggests that the methods performed similarly. At the sites where a snow pillow was present ($\mathcal{CAP}$, $\mathcal{GRZ}$, $\mathcal{KTL}$, $\mathcal{HIG}$, $\mathcal{HYS}$), the performance of BM was evaluated. At all of these sites, in all of the experiments where pillow SWE was an input variable (b, c, f), both LR and GP performed significantly better (p-value less than 0.001) than BM.

These results suggest that machine learning techniques can be used to develop models that predict mean catchment SWE more accurately than BM. However, GP does not do better than LR in any of these experiments. It is possible that ground truth data generated from snow courses, which measure SWE only at a single location, failed to capture nonlinearities present in the actual snowpack distribution. In general, models performed better when snow pillow data was included then when only TOY and air temperature were used. Neither the inclusion of air temperature data nor of TOY significantly affected model performance.

We did not evaluate BT in Experiment Set I. Because LR performed as well as GP in Experiment Set I, we suspected strict linearity among the explanatory relationships in the data and did not further pursue nonlinear modeling. As Experiment Set II used spatially distributed measurements to generate ground-truth data, it offered a more promising venue for the comparison of nonlinear modeling techniques.

6.2. Experiment Set II

First we conducted Experiment Set II: *Random Division.* GP outperformed LR in every experiment except in Norway when the only model input was *HS* at Balvatn. In every experiment in California where TOY was an input, BT has much lower error than either GP or LR. In all experiments where TOY was an input that the resulting BT models were very large. GP also had lower error and larger model sizes when TOY was used then when TOY was not used. We had originally introduced the TOY variable to allow models to distinguish different parts of the season. However, we hypothesized the BT, and to a lesser extent GP, were abusing the TOY variable to memorize snow data by mapping TOY data to ground truth HS. Even though training and testing data were technically distinct, many of the samples in the testing data were temporally proximal to samples in the training data. The testing data was not truly unseen with respect to the TOY variable. Even though models generalized well to the testing data, they were over-fitting to the TOY variable and would likely not generalize to truly unseen data, e.g. from another snow season.

To test this hypothesis and address the possible problem of over-fitting to the TOY variable, we repeated Experiment Set II three more times. In

Experiment Set II: *4 Bins*, *3 Bins*, and *2 Bins*, we successively decreased the temporal overlap between training and testing data and increase the coarseness of the temporal granularity of the division into training and testing data. Proceeding through this sequence, it became more difficult for machine learning to memorize *HS* data by over-fitting to the TOY variable. At the same time, BT error increased and the performance of GP with respect to BT improved. These results suggest that GP is more resilient against over-fitting than BT, possible as a result of multi-objective optimization. Furthermore, when the ability of machine learning to exploit the TOY variable by memorizing *HS* the data was minimized, GP significantly outperformed both LR and BT.

6.3. Interpreting GP trees

Several example GP trees are shown in figure 3. These were manually selected from the final populations of GP runs conducted for Experiment Set II. The leftmost tree represents a simple linear model. The middle tree is a nonlinear model. The rightmost tree is a more complex nonlinear model.

6.4. Input variable usage counts

Tables 9 and 10 show how frequently each input variable appears in the models generated by GP and BT in Experiment Set II. Only experiments where both *HS* and TOY were input variables are show. In general, the counts are higher for BT than for GP, reflecting the larger size of the BT models. Furthermore, model sizes decrease as the temporal granularity of the division into training and testing data becomes coarser. In Norway (Experiment c), the ratio of TOY to *HS* in GP models is high when this temporal granularity is fine, but decreases as it becomes coarser. This may indicate that GP uses TOY less when datasets are constructed so as to prevent models from abusing the TOY variable. However, this pattern is not repeated in the California experiments or for BT in either location.

6.5. Future work

We believe that the preliminary results discussed in this work are promising and warrant further research into of the applicability of GP to snowpack modeling.

This work should be expanded into a multi-year study. Although Experiment I used snow course data collected over several years, Snowcloud data used in Experiment II was limited to single snow season. A multi-year study would allow models trained on Snowcloud data during one or several years to be evaluated on unseen data from another year. Models trained on multi-year data may be more robust to application in future years than are models trained on single-year data, especially with respect to TOY. Even without collecting more data, Experiment Set I could be modified so that models are trained on data from earlier years and tested on unseen data from later years.

Beyond those discussed here, there are many machine learning techniques that could be applied to the problem of catchment-scale SWE estimation. GP possesses a unique combination of desirable qualities, but its performance should be compared against other methods such as ANNs, nonlinear multiple regression, and FFX (McConaghy, 2011), a non-evolutionary symbolic regression technology.

The only meteorological input to our models was air temperature. Future work should incorporate more predictors of SWE and HS. Meteorological data involving wind, solar radiation, humidity, etc. are available for many locations and have been shown to influence snow distribution (Logan, 1973; Elder et al., 1991; Schmucki et al., 2014; Hock and Noetzli, 1997).

Topographic features significantly shape snow distribution, and models of this relationship have been developed and used extensively (Winstral et al., 2013; Marofi et al., 2011; Chang and Li, 2000; Tabari et al., 2010; Anderton et al., 2004; Grünewald et al., 2013; Molotch et al., 2005; Elder et al., 1998). One challenge would be to make topographic data available to GP in an effective form. Some models (Winstral et al., 2002) derive real values from topographical features that are predictive of snow distributions. These values could be input variables for GP. It is possible that machine learning could use topographic and other data to produce non-cite-specific models. Such models would be trained on data from one or more catchments and then applied to other catchments.

Schwaerzel and Bylander (2006) developed high-order statistical functions for GP to model financial data. These allowed GP models to dynamically select and aggregate a slice of time series data. Future work should apply these techniques to allow GP to determine how to select and aggregate meteorological and topographic data. We made air temperature available to GP by means of functions that aggregate daily measurements over an arbitrary seven day window. Instead, GP could inductively discover how models should dynamically select and aggregate a section of time series data according to changing circumstances. Previous efforts to model snowpack using topographic data have derived explicit model inputs from DEMs. However, the possibility of GP playing an active role in determining which topographical features to use should be explored. It is possible that GP would discover new methods for extracting from digital elevation models information that is predictive of snowpack distribution.

7. Conclusion

In this paper we have described novel, low-cost methods for catchment-scale SWE estimation using machine learning algorithms. The commonly used method of estimating catchment-scale SWE from a single point measurement is error-prone because of the spatial heterogeneity of snowpack distribution. We envision an approach wherein short-term field campaigns collect ground-truth data for generating snowpack models which can subsequently augment existing NRT snow telemetry. Toward this end, we explored a suite of machine learning techniques to extrapolate estimates of mean catchment SWE from single point SWE measurements and other available data and pursued three key research directions. First, we addressed the question of which machine learning approaches are best for this problem. Second, we discussed and pursued the use of a range of possible input parameters. Finally, we grappled with the issue of ground-truthing given limited datasets.

We compared the performance of a basic method (BM) which assumes no spatial variability of SWE, linear regression (LR), genetic programming (GP), and binary regression trees (BT). We emphasize GP because it produces nonlinear, white-box models without requiring assumptions about model form. GP can be augmented with multi-objective optimization to constrain model complexity and mitigate over-fitting. We found that machine learning techniques generally outperformed BM, demonstrating the spatial variability of SWE. Nonlinear techniques outperformed linear models in Experiment Set II, but not in Experiment Set I, suggesting that there are nonlinear relationships among the modeled data used in Experiment Set II. Snowpack distribution at the catchment scale has been shown to be highly nonlinear. It is possible that the spatially distributed sampling technique (Snowcloud wireless sensor network) used for ground-truthing in Experiment Set II captured some of the nonlinearity of snowpack distribution, while the single-location sampling (manual snow courses) used for Experiment Set I did not.

When we naively divided our data at random to generate training and testing data, BT had much lower error than GP in experiments where time of year (TOY) was an input variable. In these cases, BT models were much larger than PG models and we suspected that they were memorizing data by mapping TOY to snow depth. When we instead divided the data into more temporally contiguous training and testing data in order to prevent this behavior, BT model size decreased and GP outperformed BT.

We emphasize that GP can flexibly incorporate new predictors of catchment-scale SWE into the models generated, augmenting its capacity to extrapolate estimates of mean catchment-wide SWE from a single point measurement. Genetic programming will make use of input data that helps explain the dependent variable while ignoring data that doesn't. Our choice of independent variables was a result of intuitive guesses combined with constraints on available data. Topographic information was ruled out because we were unable to determine the precise locations of snow pillows. Multiple forms of meteorological data were available, but air temperature was the most complete, allowing us to compose datasets large enough for effective machine learning. However, the inclusion of air temperature did not have a significant impact on model performance in our first experiment set, and so we did not use any meteorological data in our second experiment set.

Because it has been shown that the forcing effects underlying snowpack distribution change over the course of a snow season, we introduced time of year (TOY) as an independent variable so that models can distinguish seasonal differences. However, we found that nonlinear models used TOY to memorize the data by mapping TOY to ground truth measurements instead of expressing the underlying relationships of snowpack distribution. The ideal solution to this problem would be a multi-year study using spatially distributed data collected by Snowcloud. However, given the limitation of a one year dataset, we modified how data was divided to constrain the temporal proximity of training and testing data.

We conducted two sets of experiments, using available data, as successive approximations of our goal of near-real-time catchment-scale SWE estimation. When ground truth was obtained from distributed sampling techniques and when we were careful to mitigate overfitting to the TOY variable, GP outperformed other techniques.

Acknowledgments

We would like to acknowledge several individuals for contributions to the content of this paper. Dr. Ian Brown, Stockholm University. Dr. Jeff Frolik, University of Vermont. Dr. Jeff Dozier, University of California. Jeff Brown, University of California. David Moeser, WSL-Institut für Schnee- und

Lawinenforschung SLF in Davos, Switzerland. Dr. Keith Klepeis, University of Vermont. Rune Engeset, Norwegian Water Resources and Energy Directorate. Heidi Bache Stranden, Norwegian Water and Energy Directorate.

We acknowledge the Vermont Advanced Computing Core which is supported by NASA (NNX 06AC88G), at the University of Vermont for providing High Performance Computing resources that have contributed to the research results reported within this paper.

We acknowledge the support from DARPA through grants W911NF-11-1-0076 and FA8650-11-1-7155.

We acknowledge the support from NSF through grant #PECASE-0953837.

We acknowledge the support of the Air Force Office of Scientific Research through a YIP grant.

This work was supported by NASA under Cooperative Agreement #NNX10AK67H-S02.

References

Adams, W.P., 1976. Areal differentiation of snow cover in east central Ontario. Water Resour. Res. 12, 1226–1234. doi:10.1029/WR012i006p01226.

Anderton, S., White, S., Alvera, B., 2004. Evaluation of spatial variability in snow water equivalent for a high mountain catchment. Hydrol. Process. 18, 435–453. doi:10.1002/hyp.1319.

Bales, R.C., Molotch, N.P., Painter, T.H., Dettinger, M.D., Rice, R., Dozier, J., 2006. Mountain hydrology of the western united states. Water Resources Research 42.

Balk, B., Elder, K., 2000. Combining binary decision tree and geostatistical methods to estimate snow distribution in a mountain watershed. Water Resour. Res. 36, 13–26. doi:10.1029/1999WR900251.

Bartelt, P., Lehning, M., 2002. A physical snowpack model for the Swiss avalanche warning: Part i: numerical model. Cold Reg. Sci. Technol. 35, 123–145. doi:10.1016/S0165-232X(02)00074-5.

Boxalla, B., 2014. California snowpack hits record low. http://articles.latimes.com/2014/jan/30/local/la-me-brown-water-20140131.

Bühler, Y., Christen, M., Kowalski, J., Bartelt, P., 2011. Sensitivity of snow avalanche simulations to digital elevation model quality and resolution. Ann. Glaciol. 52, 72–80. doi:10.3189/172756411797252121.

Chang, K.T., Li, Z., 2000. Modelling snow accumulation with geographic information system. Int. J. Geogr. Inf. Sci. 14, 693–707. doi:10.1080/136588100424981.

Das, I., 1999. On characterizing the "knee" of the Pareto curve based on normal-boundary intersection. Struct. Optim. 18, 107–115. doi:10.1007/BF01195985.

Deb, K., Gupta, S., 2011. Understanding knee points in bicriteria problems and their implications as preferred solution principles. Eng. Optim. 43, 1175–1204. doi:10.1080/0305215X.2010.548863.

Dozier, J., 2011. Mountain hydrology, snow color, and the fourth paradigm. Eos, Transactions American Geophysical Union 92, 373–374.

Dozier, J., Painter, T.H., 2004. Multispectral and hyperspectral remote sensing of alpine snow properties. Annu. Rev. Earth Planet. Sci. 32, 465–494.

Elder, K., Dozier, J., Michaelsen, J., 1991. Snow accumulation and distribution in an alpine watershed. Water Resour. Res. 27, 1541–1552. doi:10.1029/91WR00506.

Elder, K., Rosenthal, W., Davis, R.E., 1998. Estimating the spatial distribution of snow water equivalence in a montane watershed. Hydrol. Process. 12, 1793–1808.

Engeset, R., Tveito, O.E., Alfnes, E., Mengistu, Z., Udnæs, H.C., Isaksen, K., Førland, E.J., 2004. Snow map system for Norway, in: Proc. Nordic Hydrol. Conf., p. 12.

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

Fassnacht, S., Dressler, K., Bales, R., 2003. Snow water equivalent interpolation for the colorado river basin from snow telemetry (snotel) data. Water Resources Research 39.

lil-gp Genetic Programming System, 2013. http://garage.cse.msu.edu/software/lil-gp/.

Grünewald, T., Stotter, J., Pomeroy, J., Dadic, R., Baños, I.M., Marturià, J., Spross, M., Hopkinson, C., Burlando, P., Lehning, M., 2013. Statistical modelling of the snow depth distribution in open alpine terrain. Hydrol. Earth Syst. Sc. 17. doi:10.5194/hess-17-3005-2013.

GSL, 2014. GNU Scientific Library. http://www.gnu.org/software/gsl/.

Guan, B., Molotch, N.P., Waliser, D.E., Fetzer, E.J., Neiman, P.J., 2010. Extreme snowfall events linked to atmospheric rivers and surface air temperature via satellite measurements. Geophysical Research Letters 37.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., Tibshirani, R., 2009. The elements of statistical learning. volume 2. Springer.

Hock, R., Noetzli, C., 1997. Areal melt and discharge modelling of Storglaciären, Sweden. Ann. Glaciol. 24, 211–217.

Johnson, J.B., Marks, D., 2004. The detection and correction of snow water equivalent pressure sensor errors. Hydrological processes 18, 3513–3525.

Jost, G., Weiler, M., Gluns, D.R., Alila, Y., 2007. The influence of forest and topography on snow accumulation and melt at the watershed-scale. J. Hydrol. 347, 101–115. doi:10.1016/j.jhydrol.2007.09.006.

Kellum, K., 2014. Big EID snowpack. http://www.mtdemocrat.com/media_gallery/big-eid-snowpack/.

Koza, J.R., 1992. Genetic Programming. Massachusetts Institue of Technology, Cambridge, MA.

Logan, L., 1973. Basin-wide water equivalent estimation from snowpack depth measurements. Role Snow Ice Hydrol., IAHS AIHS Publ. 107, 864–884.

López-Moreno, J., Fassnacht, S., Heath, J., Musselman, K., Revuelto, J., Latron, J., Morán-Tejeda, E., Jonas, T., 2012. Small scale spatial variability of snow density and depth over complex alpine terrain: Implications for estimating snow water equivalent. Adv. Water Resour. doi:10.1016/j.advwatres.2012.08.010.

Marks, D., Domingo, J., Susong, D., Link, T., Garen, D., 1999. A spatially distributed energy balance snowmelt model for application in mountain basins. Hydrological Processes 13, 1935–1959.

Marofi, S., Tabari, H., Abyaneh, H.Z., 2011. Predicting spatial distribution of snow water equivalent using multivariate non-linear regression and computational intelligence methods. Water Resour. Manag. 25, 1417–1435. doi:10.1007/s11269-010-9751-4.

Martinec, J., Rango, A., 1981. Areal distribution of snow water equivalent evaluated by snow cover monitoring. Water Resources Research 17, 1480–1488.

McConaghy, T., 2011. FFX: Fast, scalable, deterministic symbolic regression technology, in: Genetic Programming Theory and Practice IX. Springer, pp. 235–260. doi:10.1007/978-1-4614-1770-5_13.

Meromy, L., Molotch, N.P., Link, T.E., Fassnacht, S.R., Rice, R., 2013. Subgrid variability of snow water equivalent at operational snow stations in the western usa. Hydrological Processes 27, 2383–2400.

Milly, P., Betancourt, J., Falkenmark, M., Hirsch, R., Kundzewicz, Z., Lettenmaier, D., Stouffer, R., . Stationarity is dead: whither water management? Science 319, 573–574. doi:10.1126/science.1151915.

Moeser, C.D., 2010. Development, Analysis and Use of a Distributed Wireless Sensor Network for Quantifying Spatial Trends of Snow Depth and Snow Water Equivalence Around Meteorological Stations With and Without Snow Sensing Equipment. Master's thesis. University of Nevada - Reno.

Moeser, C.D., Walker, M., Skalka, C., Frolik, J., 2011. Application of a wireless sensor network for distributed snow water equivalence estimation, in: Proc. West. Snow Conf., Stateline, NV, USA.

Molotch, N., Colee, M., Bales, R., Dozier, J., 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. Hydrological Processes 19, 1459–1479. doi:10.1002/hyp.5586.

Molotch, N.P., Bales, R.C., 2005. Scaling snow observations from the point to the grid element: Implications for observation network design. Water Resources Research 41.

Molotch, N.P., Bales, R.C., 2006. Snotel representativeness in the rio grande headwaters on the basis of physiographics and remotely sensed snow cover persistence. Hydrological Processes 20, 723–739.

National Snow & Ice Data Center, .

Ohmura, A., 2001. Physical basis for the temperature-based melt-index method. Journal of Applied Meteorology 40, 753–761.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Pierce, D.W., Barnett, T.P., Hidalgo, H.G., Das, T., Bonfils, C., Santer, B.D., Bala, G., Dettinger, M.D., Cayan, D.R., Mirin, A., et al., 2008. Attribution of declining western US snowpack to human effects. J. Clim. 21. doi:10.1175/2008JCLI2405.1.

Rittger, K., Kahl, A., Dozier, J., 2011. Topographic distribution of snow water equivalent in the sierra nevada, in: Proc. West. Snow Conf., Western Snow Conference.

Rittger, K.E., 2012. Spatial estimates of snow water equivalent in the Sierra Nevada. Ph.D. thesis. UNIVERSITY OF CALIFORNIA Santa Barbara.

Schirmer, M., Wirz, V., Clifton, A., Lehning, M., 2011. Persistence in intra-annual snow depth distribution: 1. measurements and topographic control 47, W09516. doi:10.1029/2010WR009426.

Schmidt, M.D., Vallabhajosyula, R.R., Jenkins, J.W., Hood, J.E., Soni, A.S., Wikswo, J.P., Lipson, H., 2011. Automated refinement and inference of analytical models for metabolic networks. Phys. Biol. 8, 055011.

Schmucki, E., Marty, C., Fierz, C., Lehning, M., 2014. Evaluation of modelled snow depth and snow water equivalent at three contrasting sites in Switzerland using SNOWPACK simulations driven by different meteorological data input. Cold Reg. Sci. Technol. 99, 27–37. doi:10.1016/j.coldregions.2013.12.004.

Schwaerzel, R., Bylander, T., 2006. Predicting financial time series by genetic programming with trigonometric functions and high-order statistics, GECCO. doi:10.1145/1143997.1144167.

Scipión, D., Mott, R., Lehning, M., Schneebeli, M., Berne, A., 2013. Seasonal small-scale spatial variability in alpine snowfall and snow accumulation. Water Resour. Res. 49, 1446–1457. doi:10.1002/wrcr.20135.

Serreze, M.C., Clark, M.P., Armstrong, R.L., McGinnis, D.A., Pulwarty, R.S., 1999. Characteristics of the western united states snowpack from snowpack telemetry (snotel) data. Water Resources Research 35, 2145–2160.

Skalka, C., Frolik, J., 2014. Snowcloud: A complete data gathering system for snow hydrology research, in: Real-World Wireless Sensor Networks. Springer, pp. 3–14. doi:10.1007/978-3-319-03071-5_1.

Snow Surveyor, 2014. http://www.water.ca.gov/floodmgmt/hafoo/hb/sss/surveyor.cfm.

Sturm, M., Taras, B., Liston, G.E., Derksen, C., Jonas, T., Lea, J., 2010. Estimating snow water equivalent using snow depth data and climate classes. J. Hydrometeorol. 11. doi:10.1175/2010JHM1202.1.

Tabari, H., Marofi, S., Abyaneh, H.Z., Sharifi, M.R., 2010. Comparison of artificial neural network and combined models in estimating spatial distribution of snow depth and snow water equivalent in Samsami basin of Iran. Neural Comput. Appl. 19, 625–635. doi:10.1007/s00521-009-0320-9.

Tappeiner, U., Tappeiner, G., Aschenwald, J., Tasser, E., Ostendorf, B., 2001. GIS-based modelling of spatial pattern of snow cover duration in an alpine area. Ecol. Model. 138, 265–275. doi:10.1016/S0304-3800(00)00407-5.

United States Department of Agriculture, 2014. Snow surveys and water supply forecasting. http://www.nrcs.usda.gov/wps/portal/nrcs/detail/or/snow/?cid=nrcs142p2_046152.

USDA, 2014. Photo gallery of snotel site components. http://www.nrcs.usda.gov/wps/portal/nrcs/detail/or/snow/?cid=nrcs142p2_046152.

Winstral, A., Elder, K., Davis, R.E., 2002. Spatial snow modeling of wind-redistributed snow using terrain-based parameters. J. Hydrometeorol. 3, 524–538. doi:10.1175/1525-7541(2002)003<0524:SSMOWR>2.0.CO;2.

Winstral, A., Marks, D., 2014. Long-term snow distribution observations in a mountain catchment: Assessing variability, time stability, and the representativeness of an index site. Water Resources Research 50, 293–305. doi:10.1002/2012WR013038.

Winstral, A., Marks, D., Gurney, R., 2009. An efficient method for distributing wind speeds over heterogeneous terrain. Hydrological processes 23, 2526–2535. doi:10.1002/hyp.7141.

Winstral, A., Marks, D., Gurney, R., 2013. Simulating wind-affected snow accumulations at catchment to basin scales. Advances in Water Resources 55, 64–79. doi:10.1016/j.advwatres.2012.08.011.

Table 1: CDEC snow course site Descriptions

ID	EL(m)	Name	Asp.	Exposure
$\mathcal{CAP}$	2438	Caples Lake	SW	open meadow, low brush
$\mathcal{GRZ}$	2103	Grizzly Ridge	N	meadow in scattered timber
$\mathcal{KTL}$	2225	Kettle Rock	S	sloping, open meadow
$\mathcal{MSH}$	2408	Mount Shasta	SE	grassy and rocky meadow
$\mathcal{NTH}$	2835	North Lake	SE	grassy meadow
$\mathcal{SPD}$	1585	Lake Spaulding	level	grassy meadow
$\mathcal{HIG}$	1838	Highland Lakes	NW	medium sized meadow in dense timber
$\mathcal{HYS}$	2012	Huysink	W	open meadow on one leg, opening in timber on second leg

Table 2: Experiment Set I data summary by CDEC site.

ID	Pillow	NCDC base	Dist (Mi)	Samples	Years
$\mathcal{CAP}$	YES	N/A	N/A	177	1970-2011
$\mathcal{GRZ}$	YES	N/A	N/A	207	1970-2011
$\mathcal{KTL}$	YES	N/A	N/A	159	1979-2011
$\mathcal{MSH}$	NO	Mount Shasta	5.98	137	1973-2011
$\mathcal{NTH}$	NO	Bishop Airport	18.27	147	1973-2011
$\mathcal{SPD}$	NO	Blue Canyon Nyack	4.56	174	1977-2011
$\mathcal{HIG}$	YES	Mount Shasta	18.31	75	1980-2012
$\mathcal{HYS}$	YES	Blue Canyon Nyack	9.79	111	1984-2011

Table 3: Snowcloud deployment at Sulitjelma, Norway.

Tower	**Latitude**	**Longitude**
1	67.0981	16.0488
2	67.0983	16.0497
3	67.0983	16.0482
4	67.0987	16.0487

Table 4: Snowcloud deployment at the Sagehen Field Station, CA.

Tower	**Latitude**	**Longitude**
1	39.431612	-120.239759
2	39.431556	-120.239369
3	39.431402	-120.239761
4	39.431735	-120.238826
5	39.431734	-120.238644
6	39.432041	-120.238724

Table 5: GP Parameters.

parameter	**value**
population size	1000 (Experiment Set I), 2000 (Set II)
number of generations	3000 (Experiment Set I), 10,000 (Set II)
max tree size	30
mutation operators	crossover (60%), mutation (40%)
binary operators	addition, subtraction, mult., division, power
unary operators	log, exponential, sine, cosine,
terminals	independent variables, constants values

Table 6: Experiment Set I available model inputs by CDEC site.

ID	**Temp.**	$\boldsymbol{TOY}$	**Pillow**	**Temp.** $\boldsymbol{TOY}$	**Temp. Pillow**	$\boldsymbol{TOY}$ **Pillow**	**Temp.** $\boldsymbol{TOY}$ **Pillow**
$\mathcal{CAP}$		x	x			x	
$\mathcal{GRZ}$		x	x			x	
$\mathcal{KTL}$		x	x			x	
$\mathcal{MSH}$	x	x		x			
$\mathcal{NTH}$	x	x		x			
$\mathcal{SPD}$	x	x		x			
$\mathcal{HIG}$	x	x	x	x	x	x	x
$\mathcal{HYS}$	x	x	x	x	x	x	x

Table 7: Experiment Set I summary.

Experiment	**Model inputs**	**Locations**
a	air temp.	$\mathcal{MSH}$, $\mathcal{NTH}$, $\mathcal{SPD}$, $\mathcal{HIG}$, $\mathcal{HYS}$
b	TOY	all
c	pillow	$\mathcal{CAP}$, $\mathcal{GRZ}$, $\mathcal{KTL}$, $\mathcal{HIG}$, $\mathcal{HYS}$
d	air temp., TOY	$\mathcal{MSH}$, $\mathcal{NTH}$, $\mathcal{SPD}$, $\mathcal{HIG}$, $\mathcal{HYS}$
e	air temp., pillow	$\mathcal{HIG}$, $\mathcal{HYS}$
f	TOY, pillow	$\mathcal{CAP}$, $\mathcal{GRZ}$, $\mathcal{KTL}$, $\mathcal{HIG}$, $\mathcal{HYS}$
g	air temp., TOY, pillow	$\mathcal{HIG}$, $\mathcal{HYS}$

Table 8: Experiment Set II summary.

Experiment	**Location**	**Model inputs**
a	Sulitjelma, Norway	*TOY*
b	Sulitjelma, Norway	*HS* at Balvatn
c	Sulitjelma, Norway	*HS* at Balvatn, *TOY*
d	Sagehen, California	*TOY*
e	Sagehen, California	*HS* at $\mathcal{HYS}$
f	Sagehen, California	*HS* at $\mathcal{IDC}$
g	Sagehen, California	*HS* at $\mathcal{HYS}$, *TOY*
h	Sagehen, California	*HS* at $\mathcal{IDC}$, *TOY*

Table 9: Number of time *HS* and *TOY* appear in GP models in Experiment Set II

Experiment	mixed data		4 bins		3 bins		2 bins	
	HS	*TOY*	*HS*	*TOY*	*HS*	*TOY*	*HS*	*TOY*
c	54	61	38	23	43	23	36	10
g	52	80	29	53	20	65	16	43
h	50	69	18	63	19	58	19	33
total	156	210	85	139	82	146	71	86

Table 10: Number of time *HS* and *TOY* appear in BT models in Experiment Set II

Experiment	mixed data		4 bins		3 bins		2 bins	
	HS	*TOY*	*HS*	*TOY*	*HS*	*TOY*	*HS*	*TOY*
c	213	285	161	230	185	239	77	138
g	274	532	128	304	106	242	105	233
h	235	561	114	314	92	239	96	289
total	722	1378	403	848	383	720	278	660

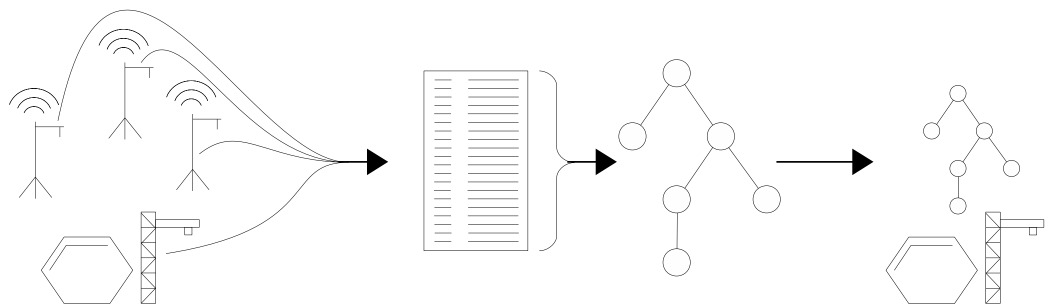


Figure 1: Using machine learning to model snowpack. First, the Snowcloud wireless sensor network is deployed in an area near a snow pillow to collect distributed ground truth data. Next, data generated by Snowcloud, by the pillow, and potentially other sources, is used by machine learning to generate a model of snowpack distribution. Finally, after Snowcloud has been removed, the model is used to estimate snow levels in the area where Snowcloud had been deployed.

Figure 2: SNOTEL site with snow pillow (USDA, 2014).

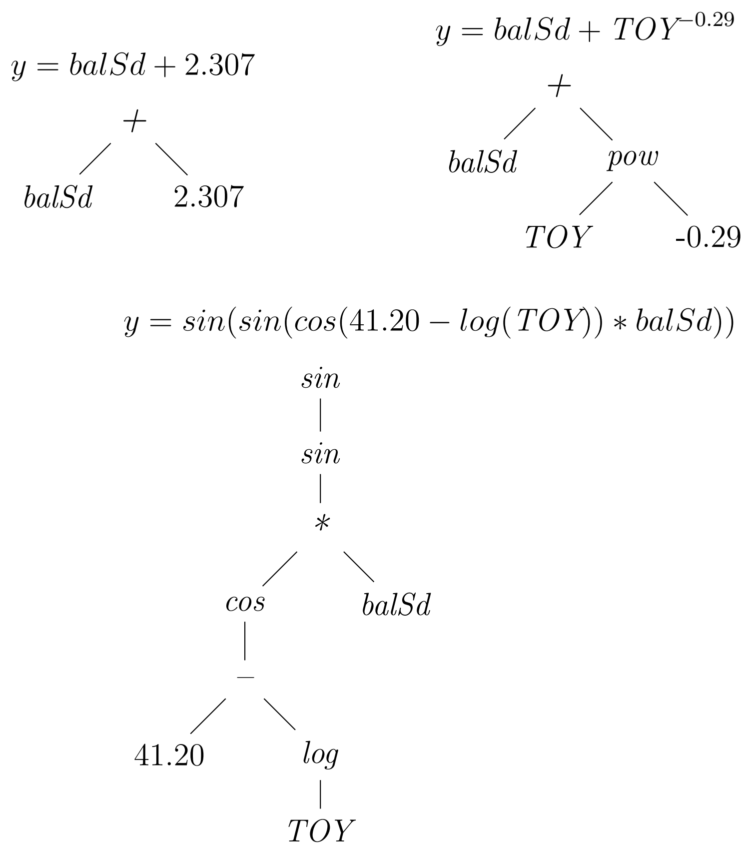
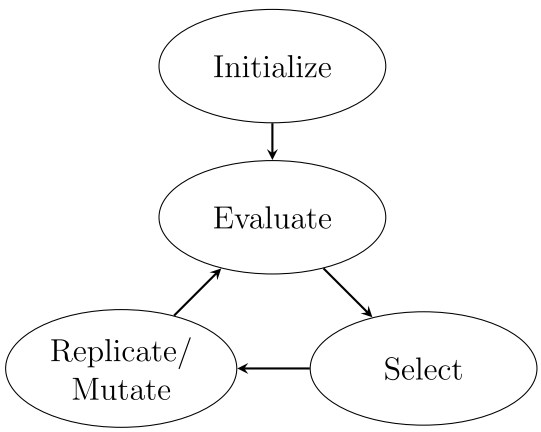


Figure 3: Example GP trees. These trees are models of mean snow depth and can be read as parse trees.

Figure 4: Snowcloud WSN sensor tower. A complete sensor stand with solar-recharged battery power, wireless mesh communication, and multiple sensor modalities. October 2011, Mammoth Mountain, CA.

Figure 5: Manual snow survey. Gene Gutenberger drops a sampling tube into the snow along California's Highway 88 at Carson Pass. Kelly Cross records measurements (Kellum, 2014).



Generation 0

$y = (log(x) + 8.293)^{-2}$
$y = sin(x) + 0.388$
$y = (-x - 0.319)^{x}$
$y = 1.303 * x^{(x^{1.07})}$

Generation 1

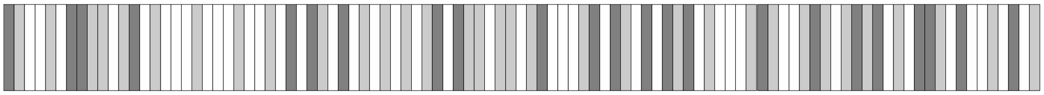
$y = sin(x) + 0.388$
$y = sin(x - 0.026) + 0.388$
$y = 1.303 * x^{(x^{1.07})}$
$y = 0.912 * x^{(x^{1.07})}$

$\vdots$

Generation $\boldsymbol{n}$

$y = cos(x * 1.309) - (x^{0.501})$
$y = ((x - 0.026) * 1.204) + 0.388$
$y = (0.912 * x^{(x^{1.81})}) - 0.441$
$y = (7.337 * (x^{1.81})) - 8.139$

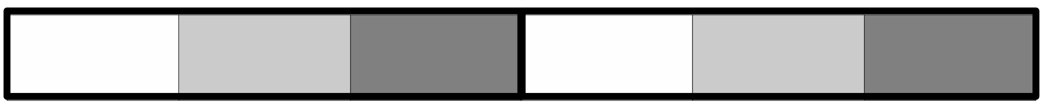
Figure 6: Genetic programming algorithm. The figure on the left demonstrates the iterative process through which GP modifies a population of solutions over time. On the right, a population of four models evolves as each iteration of the GP cycle produces a new generation.



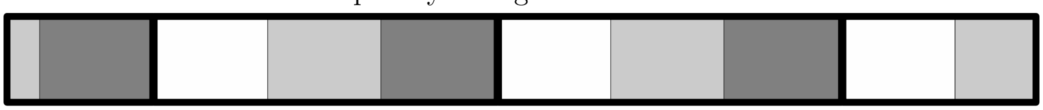
(a) Random division: dataset is randomly divided into three subsets of equal size.

(b) Four bins: dataset is divided into four temporally contiguous bins, which are each divided into three temporally contiguous subsets.

(c) Three bins: dataset is divided into three temporally contiguous bins, which are each divided into three temporally contiguous subsets.



(d) Two bins: dataset is divided into two temporally contiguous bins, which are each divided into three temporally contiguous subsets.



(e) Three bin case illustrating random offset.

Figure 7: Techniques for dividing a chronologically ordered dataset into g, s, and τ (white, light grey, and dark grey respectively).

1209

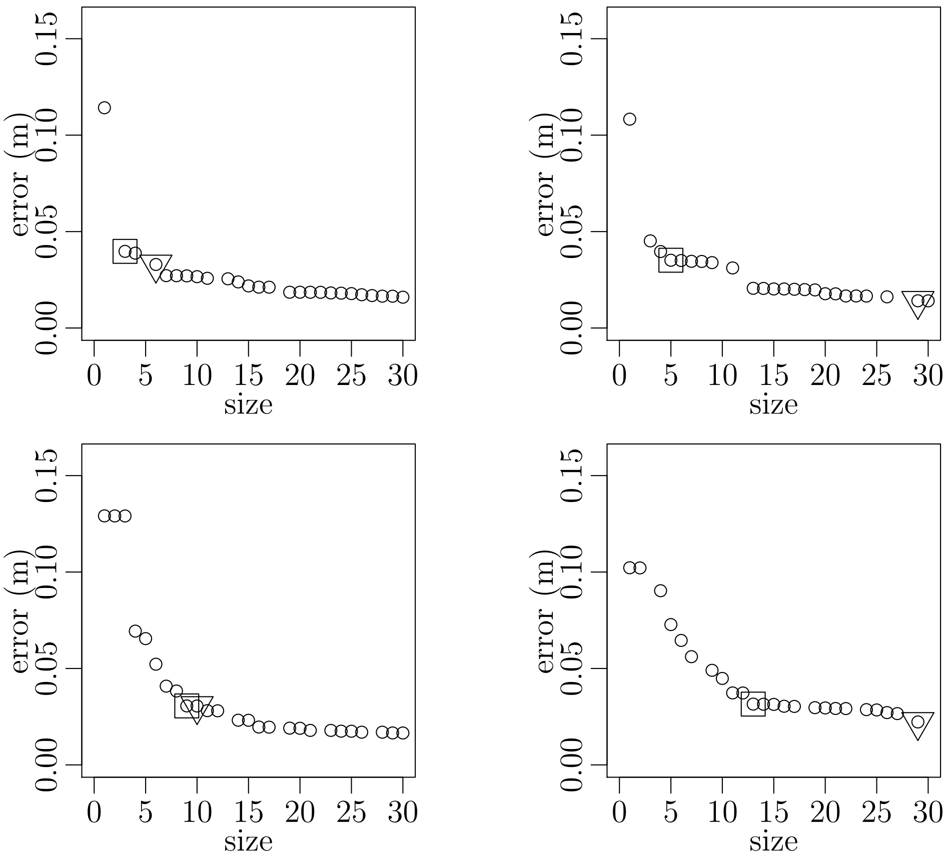
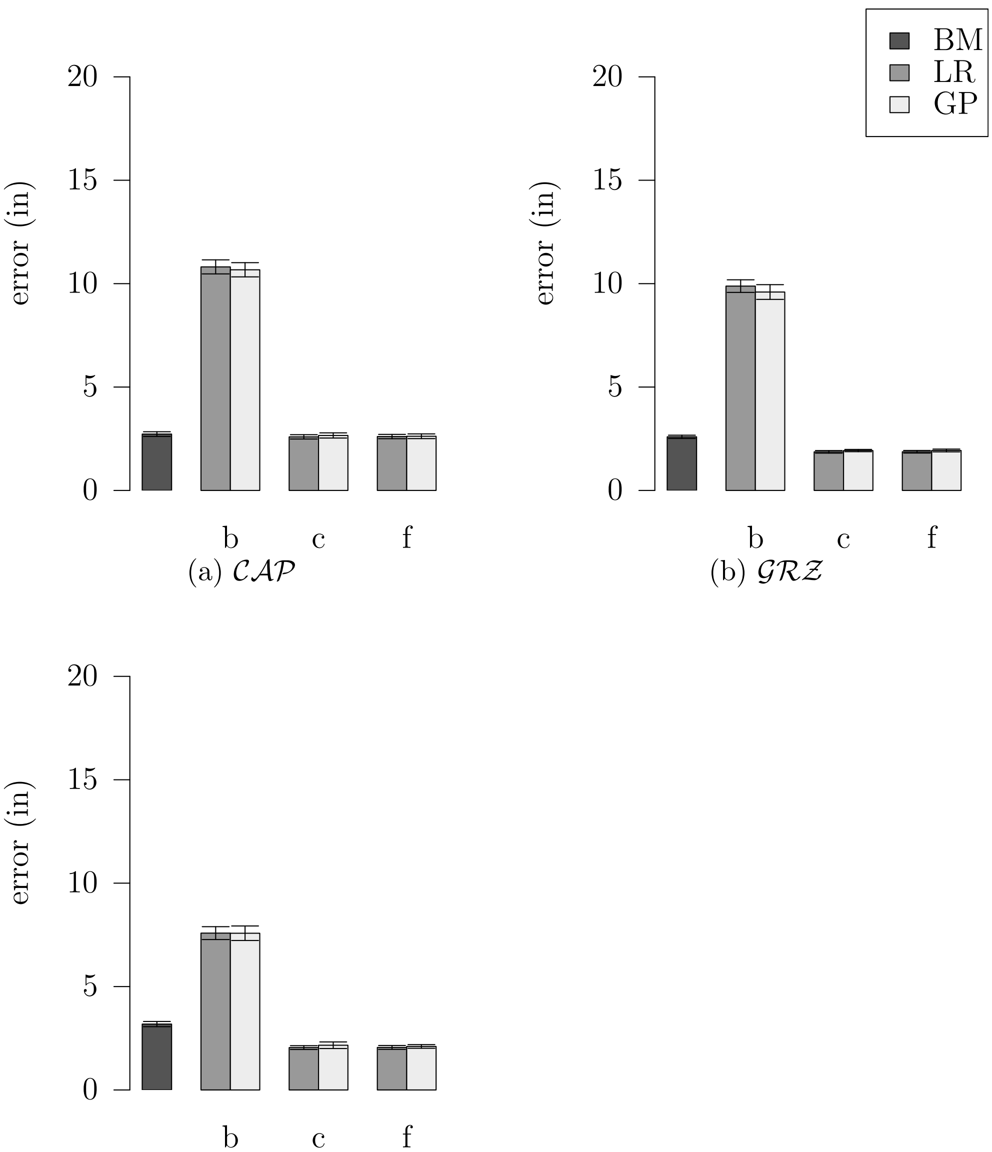
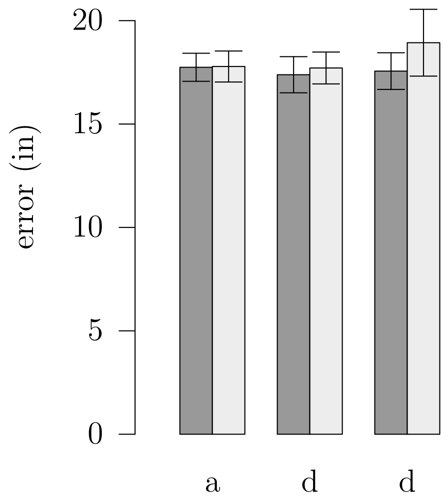


Figure 8: Example multi-objective optimization Pareto fronts. Squares mark the *knee* model. Triangles mark the model returned by the *selection set* method. These plots illustrate that Pareto fronts contain a range of solutions, from small models with high error to large models with low error. It also shows that the model which represents an optimal compromise between size and performance on training data (the *knee* model) may not be the one that performs best on unseen data (the *selection set* model). This sample of four fronts demonstrates the variety of non-dominated populations that multi-objective optimization can generate.

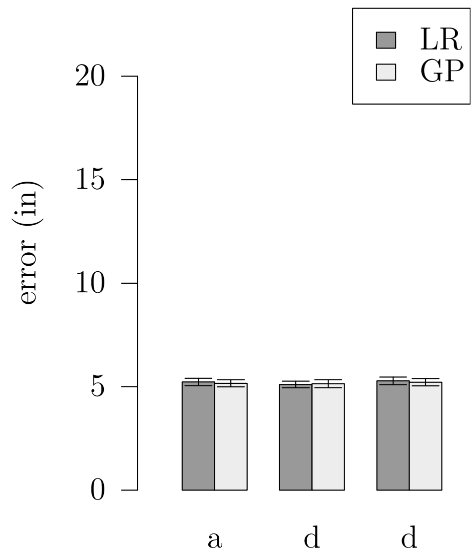


(a) $\mathcal{CAP}$ (b) $\mathcal{GRZ}$ (c) $\mathcal{KTL}$

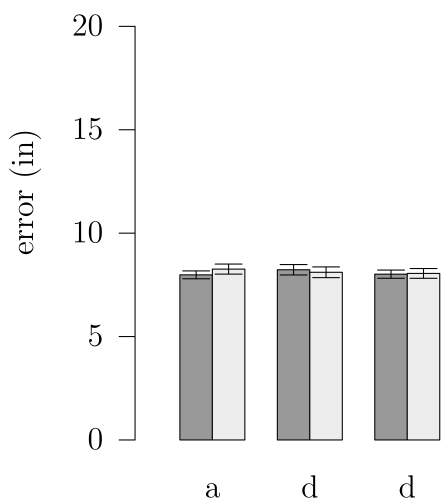
Figure 9: Experiment Set I results: $\mathcal{CAP}$, $\mathcal{GRZ}$, and $\mathcal{KTL}$.



(a) $\mathcal{MSH}$



(b) $\mathcal{NTH}$



(c) $\mathcal{SPD}$

Figure 10: Experiment Set I results: $\mathcal{MSH}$, $\mathcal{NTH}$, and $\mathcal{SPD}$.

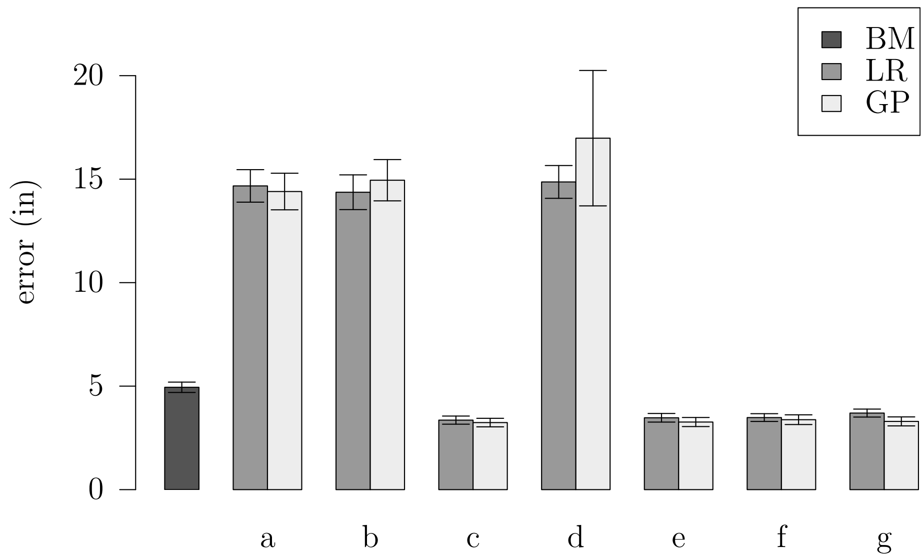


Figure 11: Experiment Set I results: $\mathcal{HIG}$.

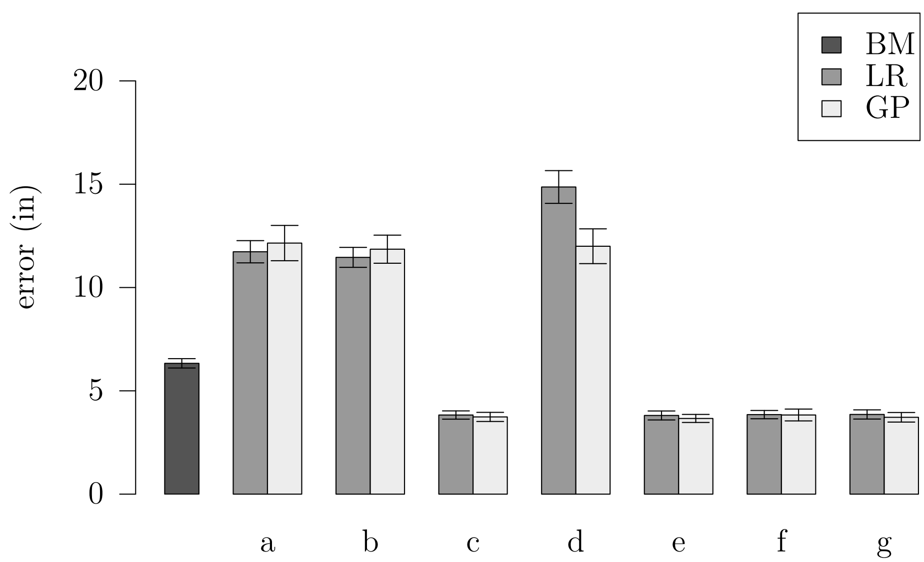


Figure 12: Experiment Set I results: $\mathcal{HYS}$.

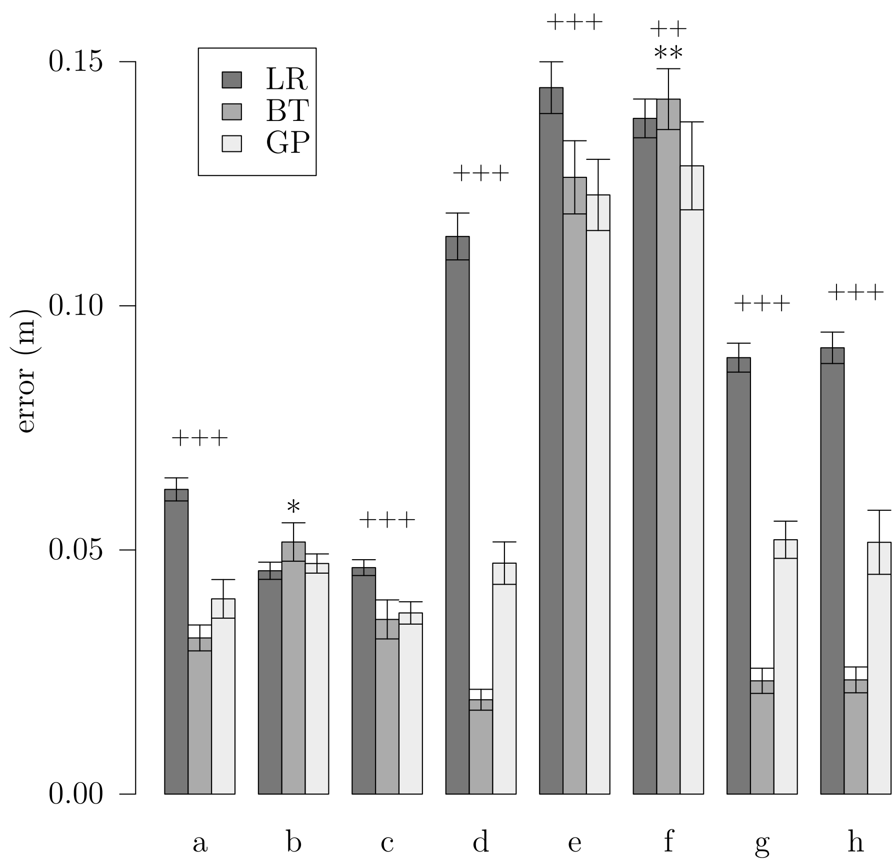


Figure 13: Experiment Set II (random division) model error.

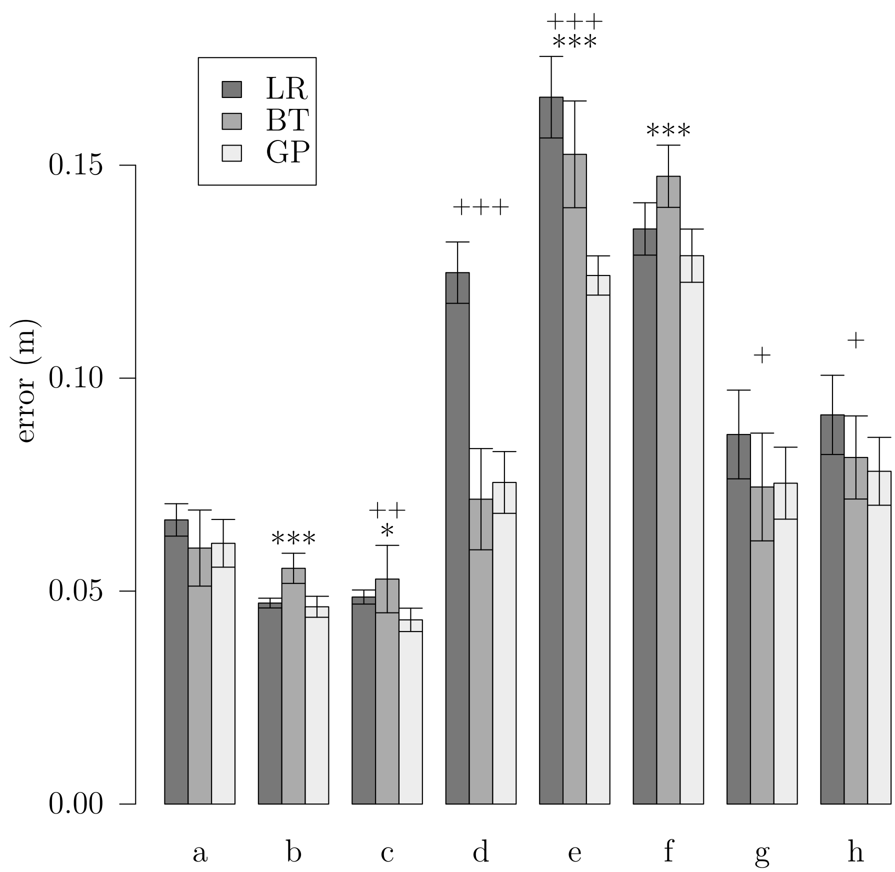


Figure 14: Experiment Set II (four bins) model error.

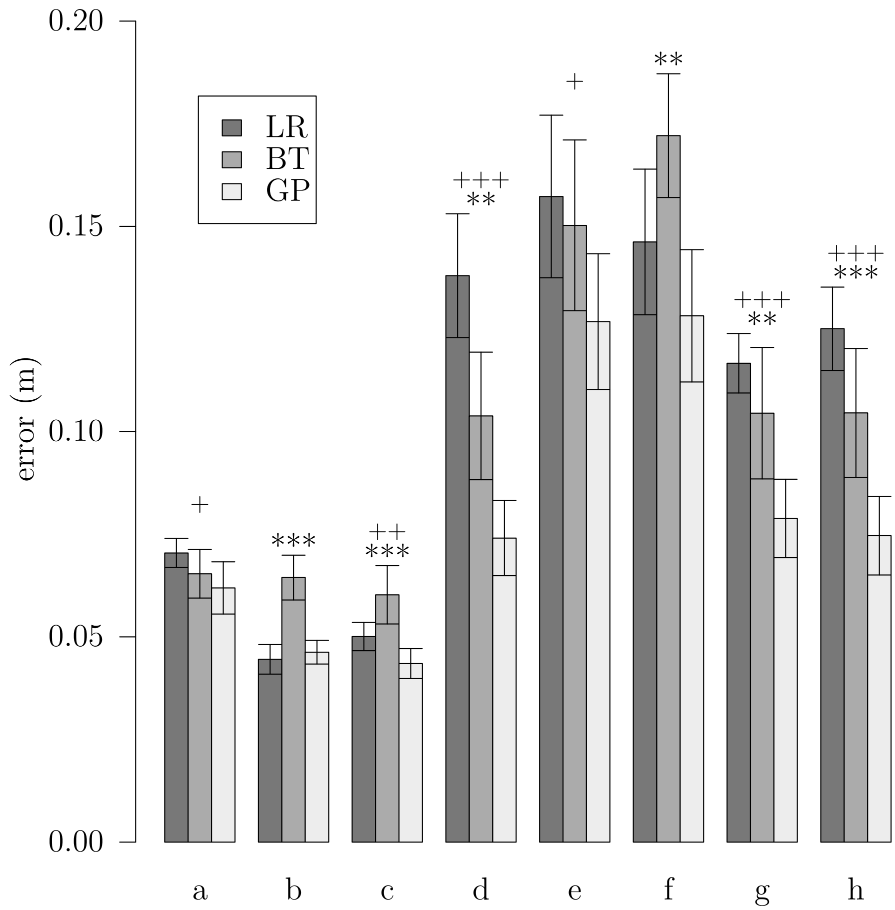


Figure 15: Experiment Set II (three bins) model error.

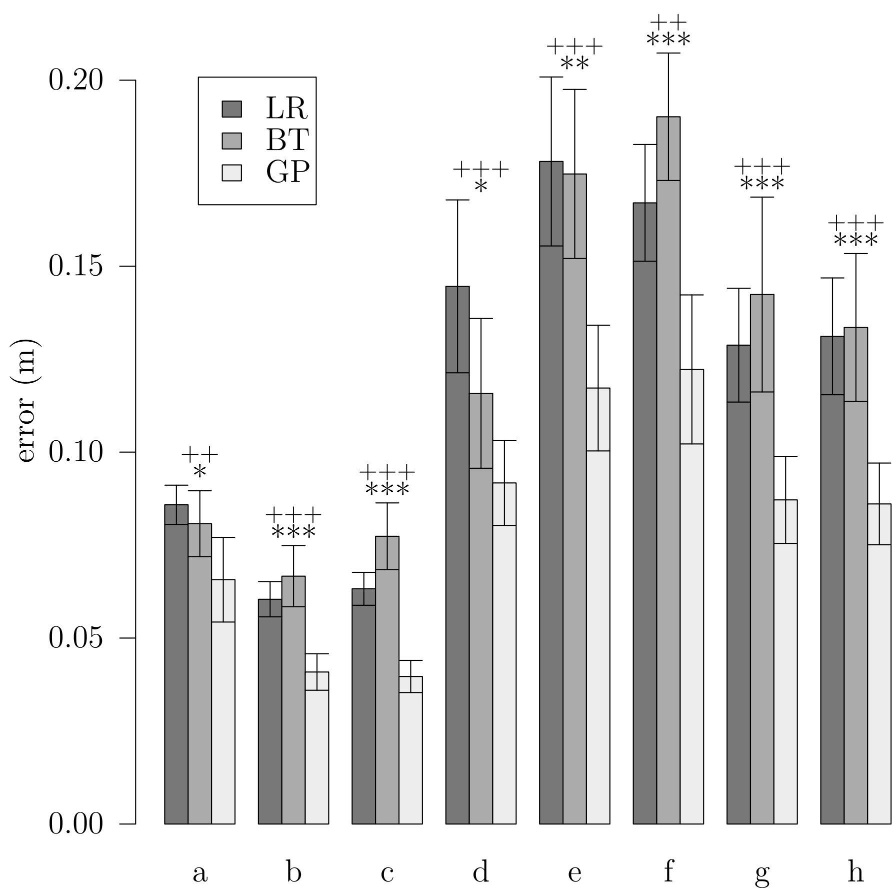


Figure 16: Experiment Set II (two bins) model error.

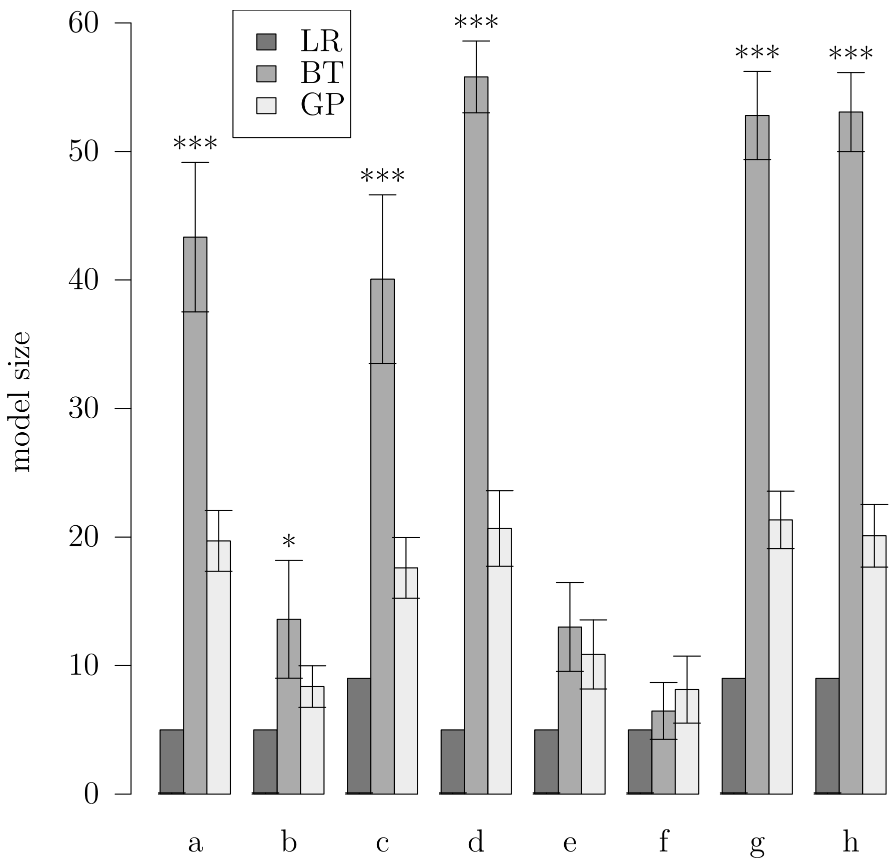


Figure 17: Experiment Set II (random division) model size.

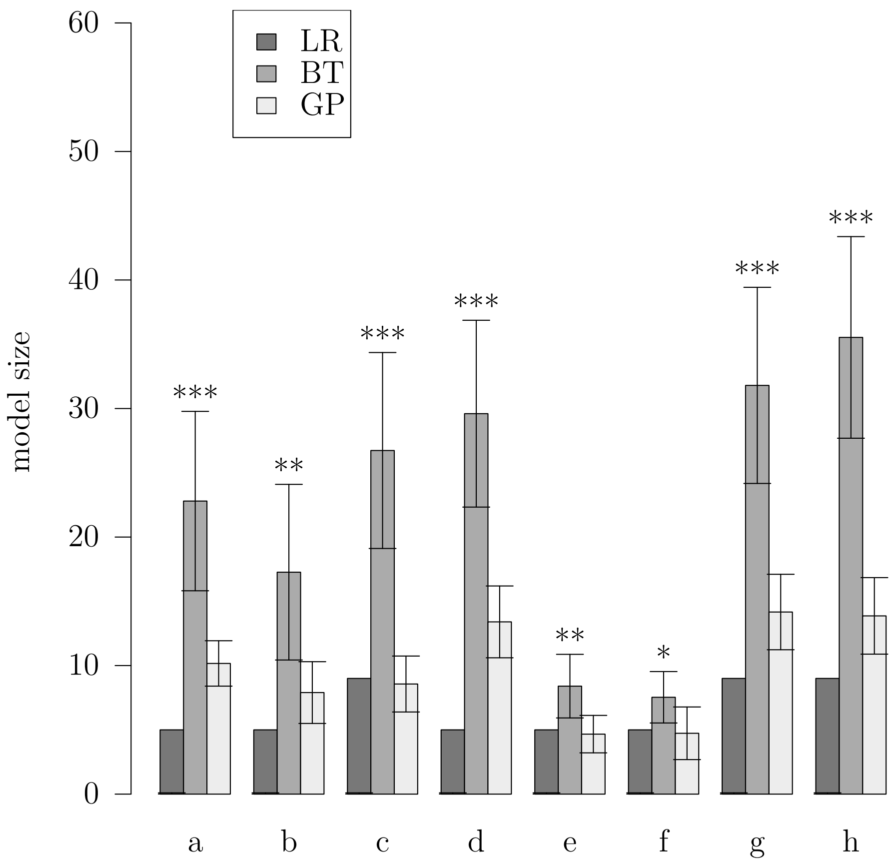


Figure 18: Experiment Set II (four bins) model size.

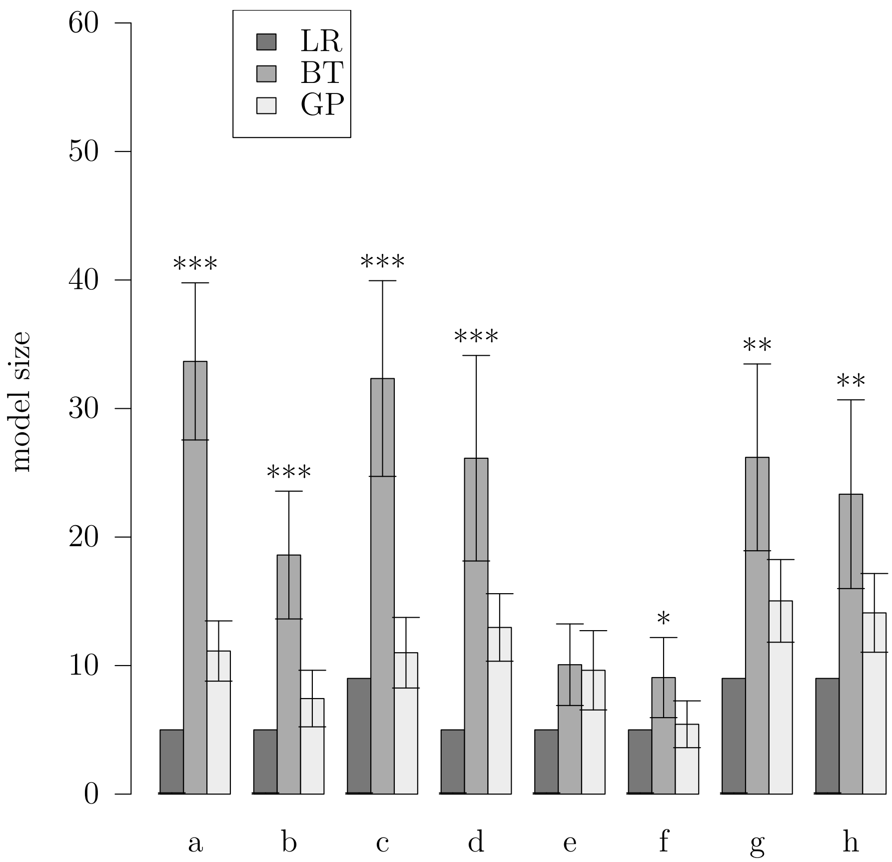


Figure 19: Experiment Set II (three bins) model size.

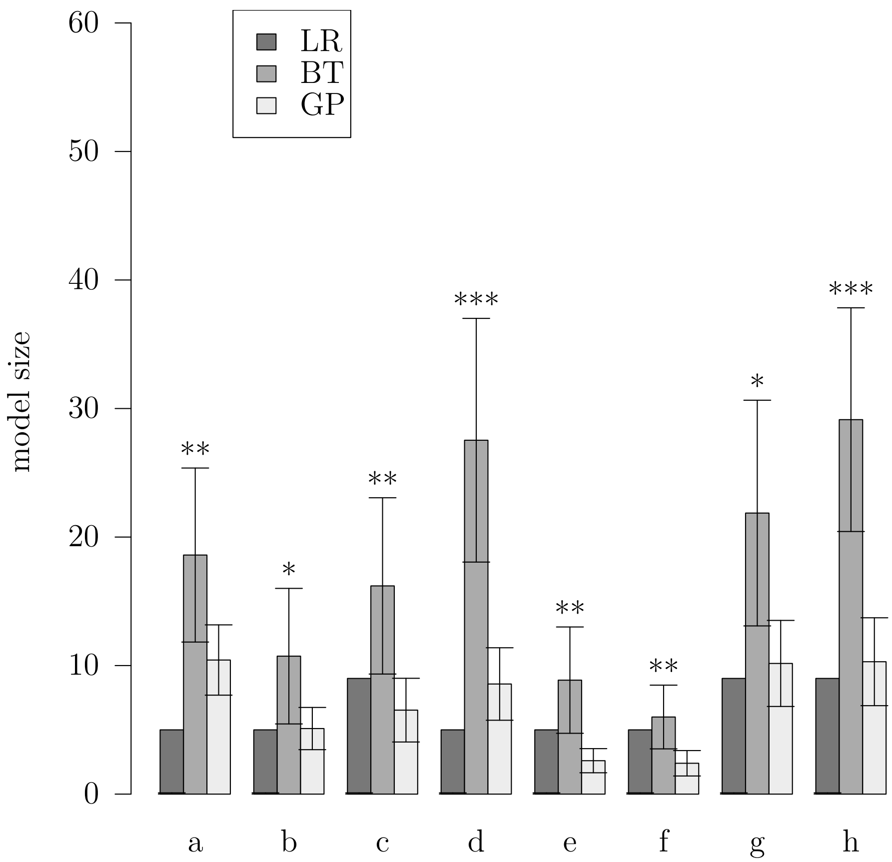


Figure 20: Experiment Set II (two bins) model size.