Comparing Snow Gauge Readings to Snow Density

Navin Souda
University of California, San Diego
nsouda@ucsd.edu

Evan Velasco
University of California, San Diego
emvelasc@ucsd.edu

Shu-Ho Yang
University of California, San Diego
shy096@ucsd.edu

Yu-Ming Lin
University of California, San Diego
yul857@ucsd.edu

Andrew Chavez
University of California, San Diego
adchavez@ucsd.edu

1 Introduction

The Forest Services of the United States Department of Agriculture is responsible for monitoring northern California's water supply in the Sierra Nevada Mountains. While snow density is often measured manually, in high altitudes and especially in the winter, such data collection can be very difficult, if not impossible. A solution to this is to use the radioactive snow gauge, which records snow density automatically[2]. Another major benefit in using a gamma transmission gauge is that the snow is not interfered with or disturbed when readings are being taken. This then means scientists can take multiple readings of the same area without worry of alteration or limiting the number of measurements able to be taken.

Since the measurements can be taken multiple times, the volume of the snow remains constant, allowing scientists to study the snowpack settlement over the entire course of a winter. During this season, due to the significantly low temperatures, when rain falls, the snow absorbs it until a certain threshold is met. Once enough rain has fallen to reach the maximum threshold allowed by the snow, flooding can then occur. It is believed that denser snow is able to absorb less water than in compression to less dense snow. It can therefore be advantageous to study the snowpack and its density with the goal of in turn being able to monitor the water supply to northern California.

The gamma transmission snow gauge is unique in that it does not directly measure the density of the snow instead it calculates the density by analyzing the measurement of gamma ray emissions from the snow. There are a few downsides to using gamma transmission snow gauges - they are expensive, and care must always be taken when handling radioactive equipment. Furthermore, due to the gamma ray emission, wear and tear on the gauge can cause the readings to diminish over time. To prevent this from misrepresenting data, a calibration must be taken at the beginning of each winter. This calibration is done by placing polyethylene blocks of specific densities between the ends of the gauges to mimic the snow, and recording the gain in the gauge for a particular density.

This paper will attempt to devise a procedure in calibrating the gamma transmission snow gauge with the efforts of ensuring accurate measurements and calculations.

2 Data

The data is collected by the USDA Forest Service. The snow gauge is in the Central Sierra Nevada mountain range near Soda Springs. The run places polyethylene blocks of known densities between two poles and uses them to take readings. In the middle 10 of 30 measurements are taken for each block. An amplified version of the gamma photon count are the measurements recorded by the gauge, referred to as the "gain". We find that as the density gets lower, the gain becomes higher, as gamma particles have more space to get through. The data available for investigation consist of 10 measurements for each of 9 densities in grams per cubic centimeter (g/cm3) of polyethylene.

While our data are taken from calibration, when hydrologists actually look to record snow density, we find some issues arise. Specifically, there are three challenges with the data. First, as mentioned earlier, most of the snow gauges need to be in the mountains or the arctic areas, where the weather is typically severe and hence makes people hard to operate and measure. Also, the networks are sparse, and measurement sites are unevenly distributed. Second, there may be differences in data quality and compatibility

across national boundaries. Measurements of solid precipitation create a potential bias due to the difference in how instruments are processed. Finally, it can be hard to validate precipitation data to match with the measurements taken.

3 Background

The snow gauge is a device that spontaneously and continuously records the water equivalent of snow on a given surface throughout the time. Gauges can be found in areas all over the world where it snows, including the Americas, Japan, Russia, and elsewhere. Snow gauges can be used to record various aspects of a snowfall, such as snow-pack settling and avalanche occurrences. The gauge we are interested in is in California at the center of a forest opening. The laboratory sits at an altitude of 2099 meters, and is located in an areas that is affected by all high-altitude storms in the Sierra Nevada.

Before looking at how to calibrate the gauge, we must first understand how it functions. A cesium-137 radioactive source emits gamma photons, or gamma rays, at 662 kilo-electron-volts (keV). A crystal detector sits 70 centimeters from the radioactive source and counts the photons that it receives. It then sends these to an amplifier, which after amplifying the signals, sends its information to a lab, where the signals are stabilized and corrected as necessary. The final measurement which is recorded is what we call the gain. It is proportional to the emission rate. The densities of the polyethylene blocks range from 0.001 to 0.686 grams per cubic centimeter (g/cm3). Practically, it ranges between 0.1 and 0.6 grams per cubic centimeter (g/cm3).

While the true relationship between the density of the polyethylene blocks and the photon count can be quite complex, due to the number of photons emitted and the directions they can travel, we can create a fairly simplified model for the calibration that we wish to perform. That is, we assume that a gamma photon has probability p to be deflected or absorbed by a polyethylene molecule. Then if there are m molecules in a straight line path from the radioactive source to the detector, there is a p^m chance that a single photon makes it from the source to the detector. We can express this as $e^{mlog(p)} = e^{bx}$, where x is the density of the block and b is some undetermined constant.

4 Investigation

4.1 Fitting

Before fitting our line, we want to make sure that it makes sense to even try to fit a regression line in this situation. Luckily, since our regression only contains one covariate, we can simply graph the data and see if it seems to follow a distinct pattern. Looking at the graph below, we can see it clearly has some form of curve.

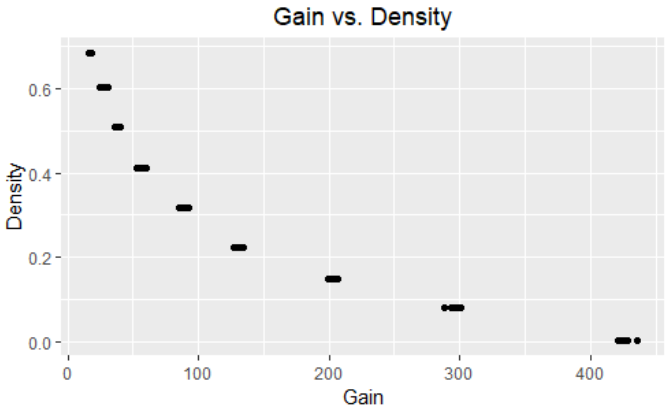

Figure 1: Plotting the gain versus the density

To remove this curve, we will take the logarithm of the gain and plot that versus the density.

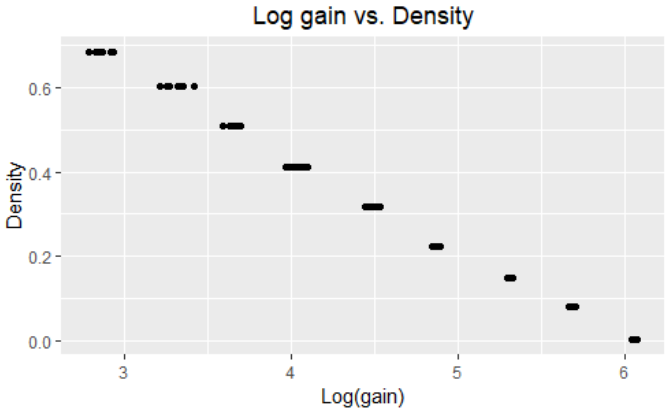

Figure 2: Plotting the log of gain versus density

We can see that this creates what seems to be a fairly straight line, at least from a purely visual standpoint. We calculate the correlation coefficient and get a value of -.903, which affirms what we saw - there is a very strong linear relationship between log gain and density. This gives us good reason to believe that our regression will be able to predict these values well. Thus, we fit our line and get the result below. It has an intercept of 1.298, which has no practical meaning, and a slope of -0.2162, meaning that when the log of the gain increases by 1, the density is on average .21 g/cm^3 lower.

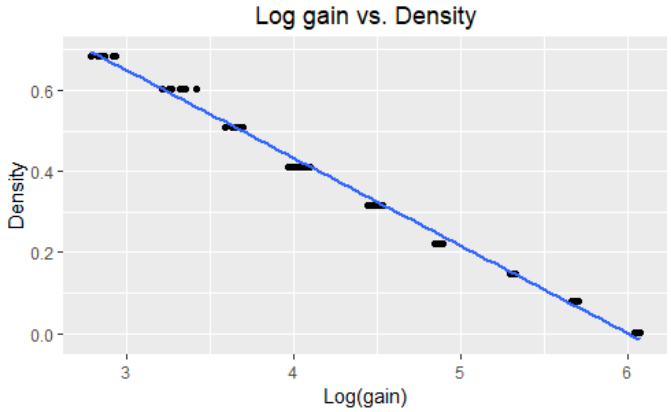

Figure 3: The least squares line of the fit between log gain and density

Now, to judge the accuracy of our model we can consider the residuals. Looking at the plot below, we see there appears to be some pretty clear pattern in the residuals - they certainly are not randomly scattered. They start out high (positive), get smaller in the middle of our point cloud, and then become positive again at the end of our point cloud. This may indicate that our errors are actually dependent on our gain value. We also see the residuals start to cluster together as gain increases, meaning the variance of the residuals may depend on the gain as well. This implies that the errors are not homoscedastic.

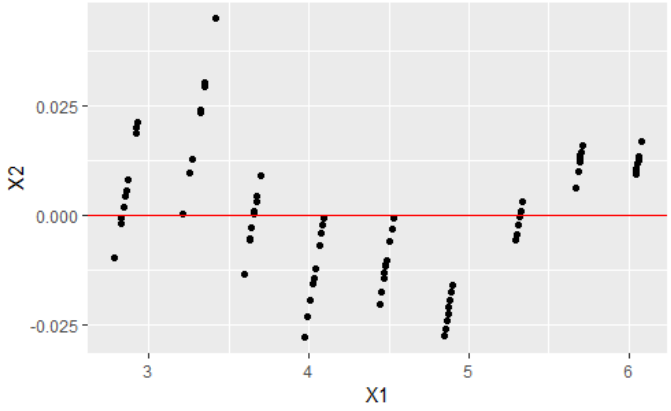

Figure 4: Plotting gain versus residuals shows a pattern and curvature in the residuals - their distribution is not random.

Then, looking at the quantile-quantile plot and the histogram below, we can see that our residuals do not appear to be normal. The quantile-quantile plot shows a fairly heavy right tail, as well as some curvature in the center. Furthermore, the histogram shows what may be a multi-modal distribution. Certainly the distribution does not appear Gaussian. Thus, our initial assumptions to perform least squares seem to be violated. This means we should be reluctant to perform inference with our results - our confidence intervals and p-values will likely be inaccurate.

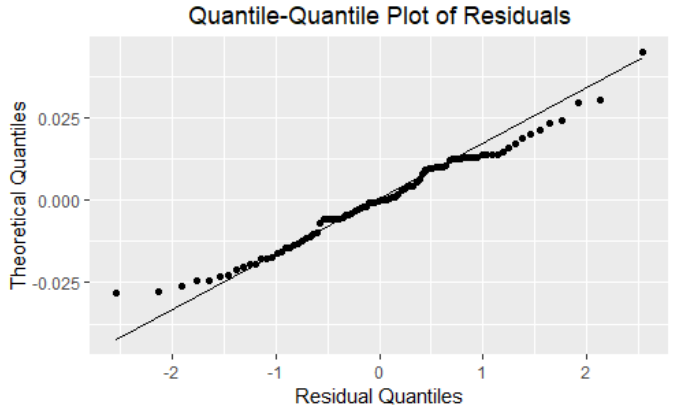

Figure 5: Quantile-Quantile plot comparing quantiles of residuals with quantiles of normal distribution

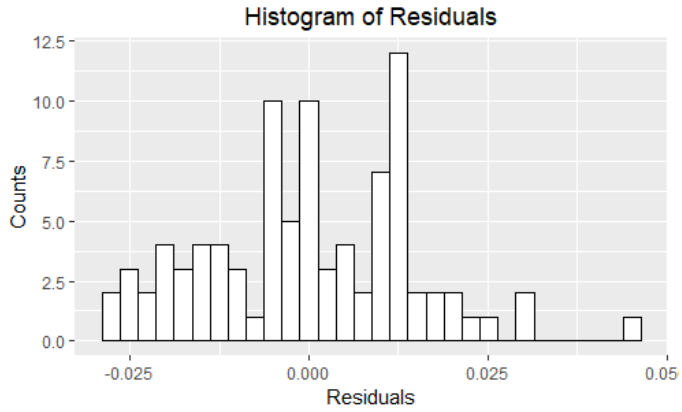

Figure 6: Histogram of residuals

Now that we have fitted the data, we must ask ourselves how our model would react to measurement errors in the explanatory variable. Standard regression models assume that the regressors have been precisely measured and have no measurement or observation error. If gain was measured incorrectly, there would be an impact on our model. These errors would lead to inconsistent estimates, meaning that the parameter estimates do not tend to the true values. This is true even in large samples. What this means for our linear regression is that the effect of measurement error would be an underestimate of the coefficient. This is known as the attenuation bias. For these reasons, it is very important to obtain accurate data when conducting a simple linear regression.[1]

4.2 Prediction

Given a gain reading of 38.6, we predict the density of the snow pack on average to be .5082, which follows our expectation, as 38.6 is the average gain reading for snow with density .508. However, we find that the expected density for a gain reading of 426.7 is -0.011. This obviously isn't possible, as densities cannot be negative, meaning our model is quite accurate here. Since 426.7 is the average gain for snow with density 0.001, it is somewhat understandable that this hap-

pened, if a little disappointing.
We have made a band for our line using the upper and lower bounds of the 95% confidence interval of the intercept. This is a possible method for making interval estimates of our predictions. Note that this does not take into account differences in slope. It is difficult to take into account differences in both slope and intercept without a method for creating confidence bands, such as Scheffe's method. Indeed, Scheffe's method would hold at a 95% confidence level over the whole band, whereas any method like ours which is so simplistic would not.

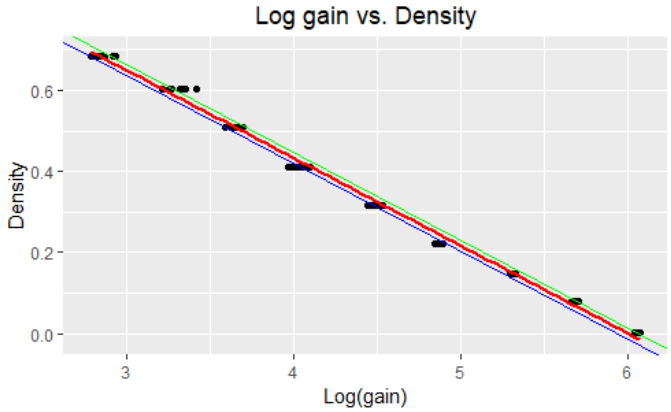

Figure 7: Bands around fitted line

4.3 Cross-Validation

We will now cross validate our results to ensure their accuracy. In order to do this, we will start by omitting the set of measurements corresponding to the blocks of density 0.508. We apply our logarithmic transformation to the variable gain and obtain a new set of data. We use this new data to construct a new linear model. A visualization of this model is shown below.

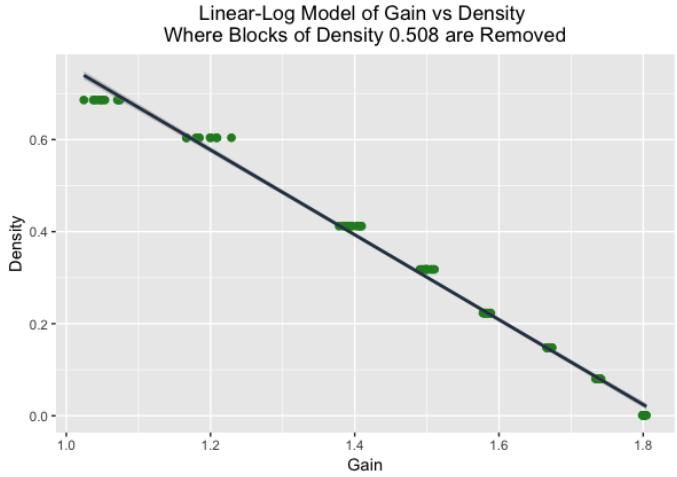

Figure 8: Linear-log model where blocks with density 0.508 have been removed.

Our linear model gives us an intercept of 1.6843 and a coefficient for the gain variable of -0.9222. We will use this model to predict an interval estimate for the density of a block with gain of 38.6. Using a 95% confidence level, we find that our interval is [0.4771654, 0.539442]. We can see that the actual density of 0.508 does indeed fall into this region. This provides evidence to support the accuracy of our model for a gain measure of 36.8.
Next we will test the accuracy of our model in the far right tail by removing blocks with density 0.001 from out data. We apply the same logarithmic transformation to our gain variable and obtain another new set of data. A visualization of this model is shown below.

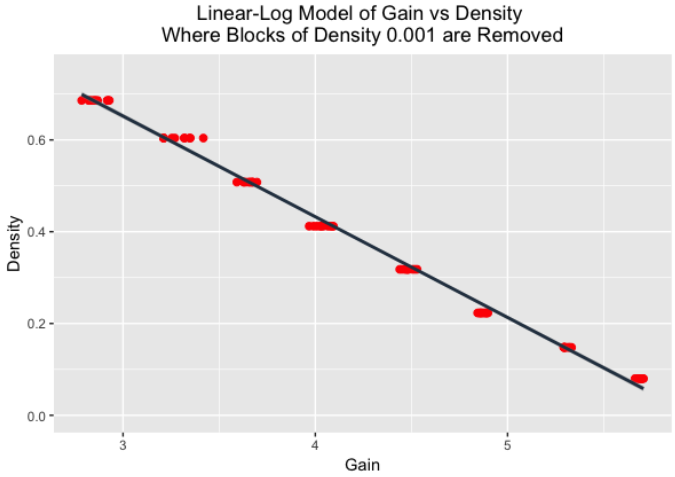

Figure 9: Linear-log model where blocks with density 0.001 have been removed.

This linear model gives us an intercept of 1.3101 and a coefficient for the gain variable of -0.2194. Once again we will use this model to predict an interval estimate for the density of a block with gain of 426.7. Using a 95% confidence level we find that our interval is [-0.04844409, 0.01132649]. We can see that the actual density of 0.001 does fall into this region. This provides evidence to support the accuracy of our model for a gain measure of 426.7.

4.4 Additional Hypothesis

As our additional question, we will fit a line using least absolute deviations(LAD). Our least squares line was already quite close to the data, so we don't expect something to much better, but since the standard assumptions were violated, it's likely that our least absolute deviations line will be slightly more accurate. We see it fitted below. It has a intercept of 1.2954 and a slope of -.2155, which, as expected, is quite close to our original fit.

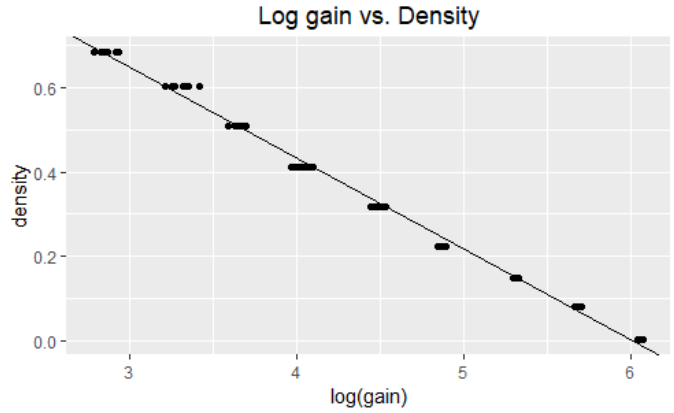

Figure 10: Linear model fitted by least absolute regression

Similarly, the residuals show the same patterns as the least squares line.

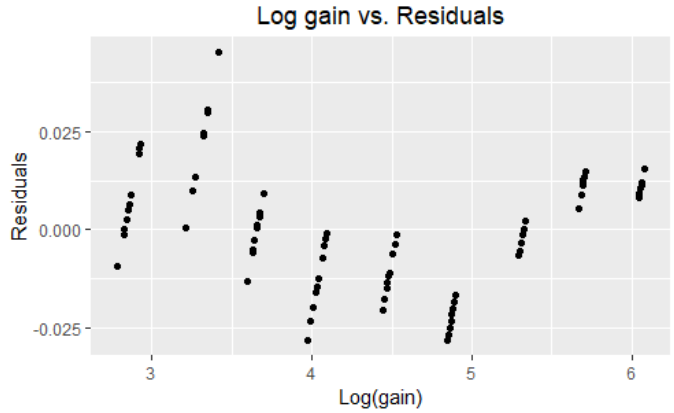

Figure 11: Residuals oof linear model fitted by least absolute regression

5 Theory

5.1 Fitting

We fit the line by the method of least squares. This model assumes that $E[Y|X = x] = \beta_0 + \beta_1 x + \epsilon$. We then find the β_0 and β_1 that minimizes the sum of squared errors, that is $\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2$, which is the sum of the squares of the differences between the predicted response at x_i and the observed response at x_i. When ϵ is normally distributed with mean 0 and some known variance σ^2, the least squares estimator has multiple useful properties. Firstly, they are normally distributed, meaning hypothesis testing can be performed on them. Furthermore, the least squares estimators are the MLE for this particular problem, and finally, by the Gauss-Markov Theorem, the estimators are the best linear unbiased estimator, meaning they have the lowest variance. In fact, this last property holds as long as the errors have zero mean and common variance. The normality provides an even stronger property, that the least squares estimators have minimum variance among all unbiased estimators, not simply linear ones. We use least squares for a few reasons: first, it is the most commonly used estimator in practice. Secondly, it has a closed form solution that is well understood, making it easily computable. Finally, it makes sense intuitively that in many contexts, an error that is twice as large is more than twice as bad - that is, the penalty for mistakes should increase faster than just linearly proportional to the size of the mistake.
We want to verify that out use of least squares was appropriate. To do this, we consider three things: whether or not the data has an actual linear relationship, the normality of the residuals, and finally the homoscedasticity of the residuals. These we checked in the first part of the investigation. Keep in mind that the normality is really only key when attempting to use our fit for inference, such as creating confidence intervals for our coefficients, or making confidence or prediction intervals for values of x.

5.2 Prediction

We make pointwise predictions based on the coefficients and intercepts we calculated using the method of least squares. We simply plug in a value of gain (or rather, it's logarithm), and use that to get a predicted value for density. We were asked to create bands for our data, but as proved in the investigation, since our standard assumptions were violated, there is almost no chance ,even through a statistically proven method, that our intervals would be accurate. Thus, we chose to create a very simple band by simple taking the upper and lower bounds of the 95% CI for our intercept, and bound the line by that. Although it does not take into account possible changes in slope, we feel that due to the sheer inaccuracy of our model, the wrong CIs in slope would only make the interval worse. Thus we felt the interval of the intercept was enough. In reality, one would probably rather choose to use a method such as Scheffe's method, assuming standard assumptions hold.

5.3 Cross-Validation

In order to validate our data, we used a cross-validation technique of removing blocks of certain densities and checking our model's prediction accuracy. For both of the data sets where certain blocks were removed, we used what is called a prediction interval with a significance level of 95%. This significance level looks similar to the significance level we use in confidence intervals, but it's interpretation is slightly different. A 95% prediction interval means that the average predictive success of the entire process is 95%. The mathematical process for determin-

ing this prediction interval for a point is shown below.

$$\hat{y}_h \pm t_{a/2,n-2}\sqrt{MSE(1+\frac{1}{n}+\frac{x_k-\bar{x}}{\sum(x_i-\bar{x}^2)})}$$

Where $\hat{y}_h$ is the fitted response variable at predictor value $\bar{x}_k$ and the critical t-value is $t_{a/2,n-2}$ with $n-2$ being the degrees of freedom. Finally, $\sqrt{MSE(1+\frac{1}{n}+\frac{x_k-\bar{x}}{\sum(x_i-\bar{x}^2)})}$ is the standard error of our prediction. MSE stands for Mean Standard Error, and measures the average squared distance between the estimated values and what is being estimated. The MSE is a measure of the estimators quality. It is always non-negative, and values closer to zero are better. The formula for calculating MSE is shown below.

$$\frac{1}{n}\sum_{i}^{n}(y_i-\hat{y})^2$$

For both data sets where blocks of 0.508 density and 0.001 density were removed, our 95% prediction interval contained the true density values for their respective gains. This supports the validity of our model in determining the density given gain.

5.4 Additional

We fit a line to our data by least absolute deviation. This minimizes the absolute sum of the residuals, that is $|e_1| + |e_2| + ... + |e_n|$. In fact, it assumes that $\text{Med}[Y|X = x] = \beta_0 + \beta_1 x$. While under standard assumptions, least squares should be preferred due to the many properties discussed before, when these assumptions are violated, LAD may be preferred. LAD is more robust to outliers, and has lower variance, making it better in the general case.

References

[1] Errors-in-variables models, Feb 2019. URL `https://en.wikipedia.org/wiki/Errors-in-variables_models`.

[2] A. Lundberg and S. Halldin. Snow measurement techniques for land-surface-atmosphere exchange studies in boreal landscapes. *Theoretical and Applied Climatology*, 70(1):215–230, Sep 2001. ISSN 1434-4483. doi: 10.1007/s007040170016. URL `https://doi.org/10.1007/s007040170016`.