Griffin Shelor

6-5-2024

Ecological Forecasting

Forecasting Snow Water Equivalent (SWE) Using 3 Sites in Alaska as a Case Study

Snowpack is a critical resource, particularly in the mountain west of the United States. It is crucial to properly understand the conditions and fluxes in snowpack in the west due to the region’s reliance on this resource for agriculture, sustaining a growth in residential population, and resource development (Elder et al. 1991). Much of the water fueling these industries and localities originally falls as snow in mountainous regions such as the Sierra Nevada and the Rocky Mountains. However, there is currently a dearth of continuous snow sampling sites at high elevation regions (Molotch and Bales 2006). This lack of sampling due to the difficulty of accessing these areas produces high levels of uncertainty in estimating snow properties at high elevations. Snow Water Equivalent (SWE) in particular deserves extra focus, given its importance to agriculture and relevance to the water usage by regular citizens. SWE represents the depth of a snowpack is the entire mass of snow were to melt and is commonly measured in millimeters, though this can be interpreted as liters per square meter (Fierz et al. 2009). Though it falls as snow, its impact as spring and summer flow in rivers, streams, and reservoirs is felt throughout entire regions as it affects water that is available for irrigation (Qin et al. 2020). The snowpack also provides valuable reinforcement to streamflow for fisheries as well as fueling the growth of forage cover for ranchers and wildlife (Siirila-Woodburn et al. 2021). The western U.S. is particularly dependent on this melting process and the continuous, reliable availability of melted snowpack during the growing season (Qin et al. 2020). Snowpack is also considered to be a resource under threat, as climate change and related forces have resulted in decreases in snowpack storage in over 25% of western mountainous watersheds (Hale et al. 2023). However, there is still much uncertainty as to how snowpack will respond to a changing climate (Musselman et al. 2017). As climate change continues to progress, it is vital that snow sampling sites provide informative, actionable insights which reflect the larger region of interest they are in, provide context to how the snow cover is affected by climate change in a given region. As such, forecasting SWE values is a useful way of evaluating our understanding of the dynamics and fluxes of snowpack as a system, identifying important influences on shifts within SWE values, and evaluating our ability as scientists to predict future water supplies in a region dependent on this resource.

After gathering all of my data together, I set about my model-building process by first examining the correlation of all of my predictor variables with my dependent variable, SWE. I use this to identify variables that are likely to be important factors when building a model. Before building models, I split out the range of data to be forecasted and then took the rest of the data and split it into training and testing datasets for pre-forecast evaluation. The first models I built were linear deterministic formula implemented in Random Forest models. Random Forest is a machine learning algorithm which selects random subsets of a training dataset, makes a prediction based on the values of that subset, then builds many trees following this same pathway, with the “final” decision of the forest being the average result among all of the trees. I chose this pathway to start because most implementations of Random Forest models in R or Python include a variable importance function, which can help provide useful context to what the primary influences are on a model. While the variable importance values themselves are somewhat arbitrary, I compared the relative order of variable importance with the relative order of absolute values of correlation. Finding that the previous day’s SWE measurement, previous minimum temperature, previous mean temperature, previous cumulative precipitation, and elevation were all both relatively highly correlated with SWE and considered important in Random Forest models drove a majority of the thinking throughout the rest of my model-building process. I found particular utility in using the previous day’s SWE value as an autoregressive term in both my Random Forest models as well as the Stan models I built afterwards.

A screenshot of a computer

Description automatically generatedAfter identifying prominent variables using multiple Random Forest models, I began constructing models in Stan using the lognormal distribution. I chose the lognormal distribution because statistically, it is impossible for SWE to be less than 0 and the data itself did not appear to have a normal distribution, with most SWE observations between 0 and the low 200s in millimeters. This makes the data well suited in theory to a lognormal or gamma distribution. However, I chose to focus on lognormal-based models due to its easier explainability and lower computational costs since moment matching would be necessary to properly use the gamma distribution.

Figure 1: A table displaying model type, covariates, which distribution was used, and scores for the evaluation metrics used.

Once I settled on a distribution, my Stan models took on a familiar form, with the µ parameter of the lognormal distribution being equivalent to a linear deterministic function and the sigma parameter representing the variance. I created 2 different models using the lognormal, testing the effect of different combinations of variables with each model (Figure 1). After evaluating my first two lognormal models, I switched to using a normal distribution

out of a curiosity as to whether it would better fit the data and provide a forecast with less uncertainty.

To evaluate my models, I plotted them to visually examine how well the model fit the data as well as to identify the range of uncertainty in the forecasts (Figures 2,3,4). I also calculated RMSE and Coverage (percentage of forecasted points within the forecast’s credible intervals). The values in Figure 1 reflect the RMSE and Coverage of each model when applied to the data being forecasted.

A graph of different types of data

Description automatically generated with medium confidence When evaluating my model, I found that including an autoregressive term was particularly impactful in my forecast, particularly for Random Forest models and models using a normal distribution (Figures 2,4). Given the strong correlation between the previous SWE value and the current SWE value (0.99), this is unsurprising. For the Random Forest models, the biggest impact was on the levels of uncertainty, which dramatically increased when the autoregressive terms were removed. RMSE also increased as well, with the RF model being unable to regress to the overall SWE trend after encountering either large, sudden jumps in SWE values (likely due to winter storms) or large temporal gaps in the data.

Figure 2: Forecasts produced by the Random Forest models, with 90% credible intervals

While one of the lognormal Stan models had relatively high coverage, a quick examination of the plots of the forecasts makes it somewhat obvious that this is largely because of extreme uncertainty issues with the upper bound of the credible intervals (Figure 3). Strangely, including an autoregressive term appears to have made the model worse in both coverage and RMSE, which could indicate that this dataset was perhaps not well suited to a lognormal distribution despite the theoretical support for such a distribution (Table 1).A graph of a graph of a graph

Description automatically generated with medium confidence

Figure 3: Forecasts produced by the lognormal distribution models, with 90% credible intervals

After encountering issues with the lognormal distribution, I then implemented two models with the normal distribution, one with an autoregressive term and one without (Figure 4). This is where the greatest forecast success was found, despite the theoretical support of a normal distribution not being well-suited to a variable such as SWE which can only be non-negative real numbers. An autoregressive model using a normal distribution achieved phenomenal metrics for both coverage and RMSE, outperforming all other models. Without the autoregressive term, a model using a normal distribution was able to do a reasonable job approximating the general trend up until the forecast ran into a temporal gap in data. A graph of a graph of a model and a diagram of a model

Description automatically generated with medium confidence

Figure 4: Forecasts produced by the normal distribution models, with 90% credible intervals

Given the dramatically increased uncertainty when the autoregressive term is removed and the process by which I built my non-RF forecasts, using a randomly generated value from the relevant distribution, I attribute most of the uncertainty to process error, with parameter error comprising a smaller proportion of the overall forecast uncertainty. I have also learned that a normal distribution can still create a reasonably decent forecast, and even create an excellent one with an autoregressive term, despite not being the most likely distribution according to Bayesian statistical theory.

In the future, I would hope to apply these models to a wider variety of sites and include more data in the building of the models so that I could determine whether these forecasts are specific to the 3 sites in Alaska which were chosen for this project, or whether the lessons learned here are more widely applicable throughout the mountain west. There are also many opportunities for improvement with the lognormal distribution to reduce the uncertainty levels, either through the evaluation of other covariates not included here such as slope and aspect or through the adjustment of priors and hyperparameters for these models.

References

Elder, K., Dozier, J., & Michaelsen, J. (1991). Snow accumulation and distribution in an Alpine Watershed. *Water Resources Research*, *27*(7), 1541–1552. <https://doi.org/10.1029/91WR00506>

Fierz, C., Armstrong, R.L., Durand, Y., Etchevers, P., Greene, E., McClung, D.M., Nishimura, K., Satyawali, P.K. and Sokratov, S.A. 2009. *The International Classification for Seasonal Snow on the Ground.* IHP-VII Technical Documents in Hydrology N°83, IACS Contribution N°1, UNESCO-IHP, Paris.

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*(4), 723–739. <https://doi.org/10.1002/hyp.6128>

Musselman, K. N., Molotch, N. P., & Margulis, S. A. (2017). Snowmelt response to simulated warming across a large elevation gradient, southern Sierra Nevada, California. *The Cryosphere*, *11*(6), 2847–2866. <https://doi.org/10.5194/tc-11-2847-2017>

Qin, Y., Abatzoglou, J. T., Siebert, S., Huning, L. S., AghaKouchak, A., Mankin, J. S., Hong, C., Tong, D., Davis, S. J., & Mueller, N. D. (2020). Agricultural risks from changing snowmelt. Nature Climate Change, 10(5), 459–465. <https://doi.org/10.1038/s41558-020-0746-8>

Siirila-Woodburn, E. R., Rhoades, A. M., Hatchett, B. J., Huning, L. S., Szinai, J., Tague, C., Nico, P. S., Feldman, D. R., Jones, A. D., Collins, W. D., & Kaatz, L. (2021). A low-to-no snow future and its impacts on water resources in the western United States. *Nature Reviews Earth & Environment*, *2*(11), 800–819. <https://doi.org/10.1038/s43017-021-00219-y>