FinalProject

Chris Corona

2022-12-12

Introduction

The Gallatin River is an important resource for the community of Bozeman, MT. It provides water for irrigation systems of many rural farming families and businesses. It is a renowned fly fishing destination that brings lots of tourist dollars to the area. Low water and droughts will negatively impact these two major economic sectors. River flows are also an indicator for the filling of local reservoirs of drinking water. It is known that the drinking water supply in Bozeman is precariously balanced with its growing population. With more growth and more drought, the water supply in this region may be insufficient. On the flip side, we have seen first hand the damages incurred from recent historic flooding this past spring. The community of Red Lodge suffered massive destruction and the road through the north entrance of Yellowstone National Park was totally destroyed from river flooding. Fortunately the Gallatin River only caused minimal damage, but we might not be so lucky next time. Measuring the river flow is an important metric to be able to predict drought and flooding which have major impacts on everyone in Bozeman.

The data set in this project is a compilation of publicly available data from USGS (https://waterdata.usgs.gov/monitoring-location/06043500/) and SNOTEL websites (https://wcc.sc.egov.usda.gov/reportGenerator/). I downloaded data for the Gallatin River waterflow (CFS) measured at the mouth of the Gallatin Canyon. The weather data includes snow depth (inches), precipitation accumulation (inches), and temperature (F) from three SNOTEL weather stations within the Gallatin River watershed (Carrot Basin, Shower Falls, and Lone Mountain). Figure 1 shows a map locating all the measurement sites. I collected 10 years of hourly data for all measurements. This data was then aggregated into monthly averages so the final data set is 10 variables (3 weather variables x 3 weather stations, and 1 response) with 120 observations (12 months x 10 years).

My research question asks: can we predict the monthly average flow of the Gallatin River at the mouth of the Gallatin Canyon using three weather stations at various locations upstream measuring snow depth, precipitation accumulation, and temperature?

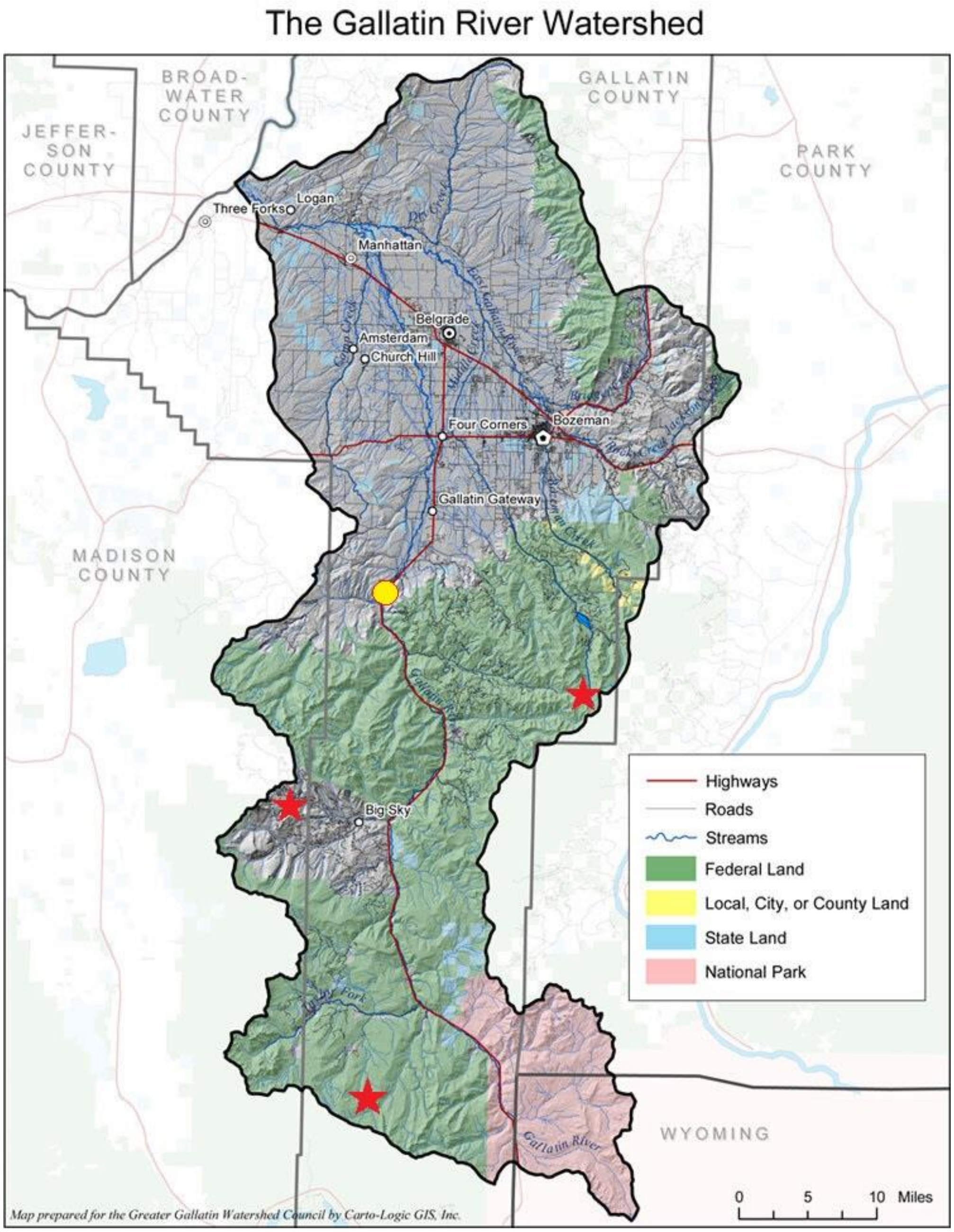


Figure 1: Map of Gallatin River watershed with the three SNOTEL weather stations marked with red stars and the water flow measurement site marked as a yellow circle.

Exploratory Analysis

Before aggregating to monthly averages, the raw data set is hourly measurements. An initial look at the raw data set shows that there is significant amounts of missing data for all measurements. Starting with the response variable `discharge`, we see that the missing data occurs in large blocks of a few days or weeks at a time. Figure 2 shows the missing blocks for one particular year. I imputed these missing values by using a Holt-Winters model trained on the data leading up to the missing block. In some cases where there was a small window of data leading up to the missing block, I resorted to using previously imputed values as part of the training. Figure 3 shows the results of one particular year of imputations. Figure 4 shows the details of the Holt-Winters imputation models. Visually, the imputations look remarkably similar to the existing data. Since I will be aggregating to monthly averages, any slight perturbations to this hourly data set will be smoothed out. Similar patterns of missingness occur in four of the years of `discharge` measurements for a total of 1787 missing observations at the hourly measurement level.

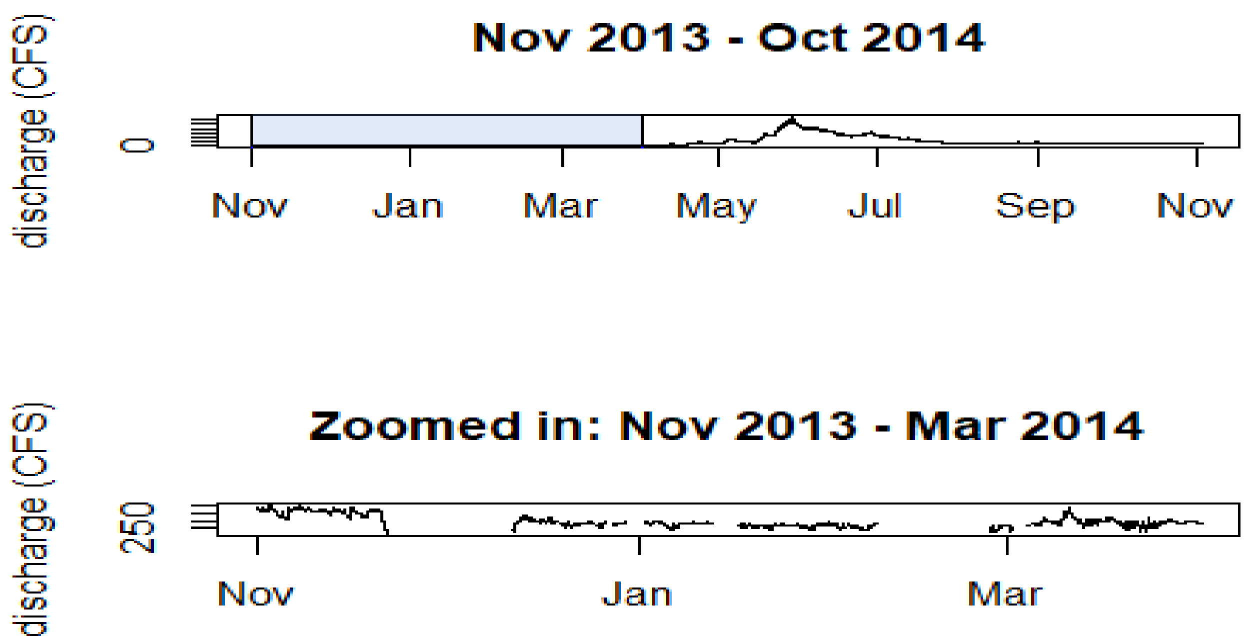


Figure 2: Large blocks of missing data from Nov 2013 - Oct 2014. The top plot shows the full year, and the bottom plot is the light-blue shaded region from above zoomed in to see detail.

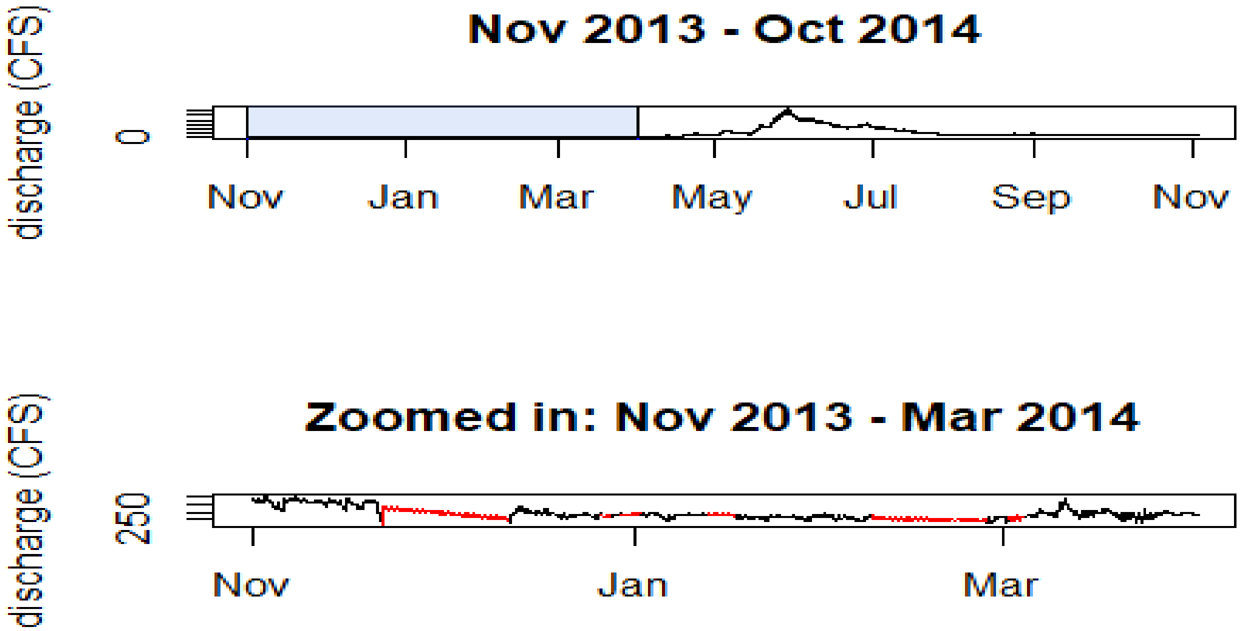


Figure 3: Plot of the existing data in black and the imputations in red for the missing data from Nov 2013 - Oct 2014. The top plot shows the full year, and the bottom plot is the light-blue shaded region from above zoomed in to see detail.

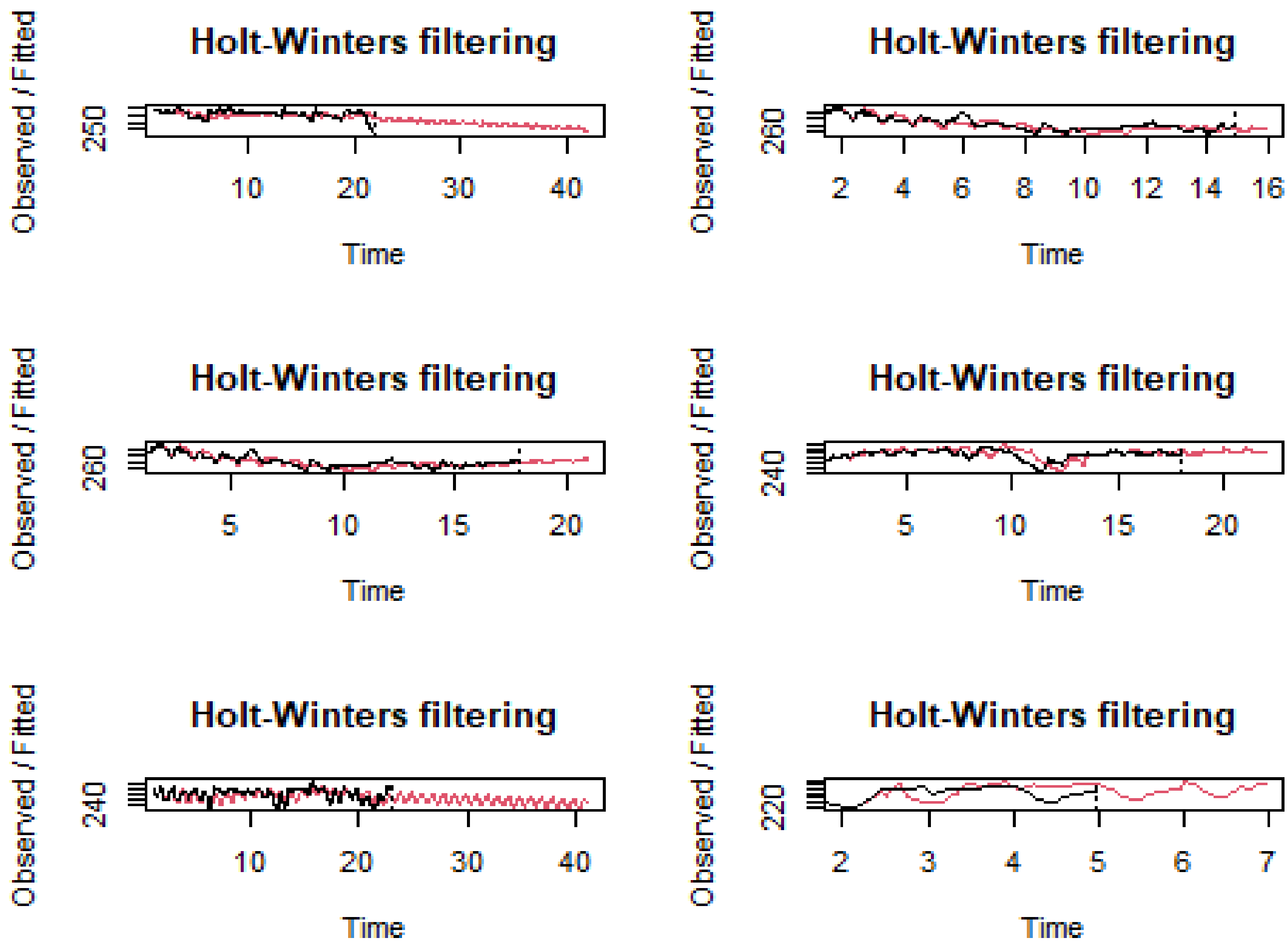


Figure 4: Details of the Holt-Winters imputation models. The solid black line is the training data, the red line is the estimated moving average, and the dotted vertical line shows the boundary where the predictions begin. These models in order from the top of the left-hand column moving down, then to the top of the right-hand column moving down, match the sequence of missing blocks above from left to right.

Next looking at the predictor variables, we also see lots of missing data. In addition, we see unreasonably high measurements probably due to malfunctions in the sensor. I started by first removing these unreasonably high measurements and treating them as missing. As opposed to before, in this case the missing data occurs as just a handful of measurements at a time with high frequency across the entire data set - from one to two missing up to twenty consecutive missing measurements. Instead of fitting models to impute these frequent but few missing values, I simply filled in the missing values with a straight line between the next adjacent existing values. Figure 5 shows the raw data and again with imputations as red points. Again, a visual inspection of the imputations shows a good fit to the existing data. In total across the nine predictor variables there were 26,861 hourly measurements missing.

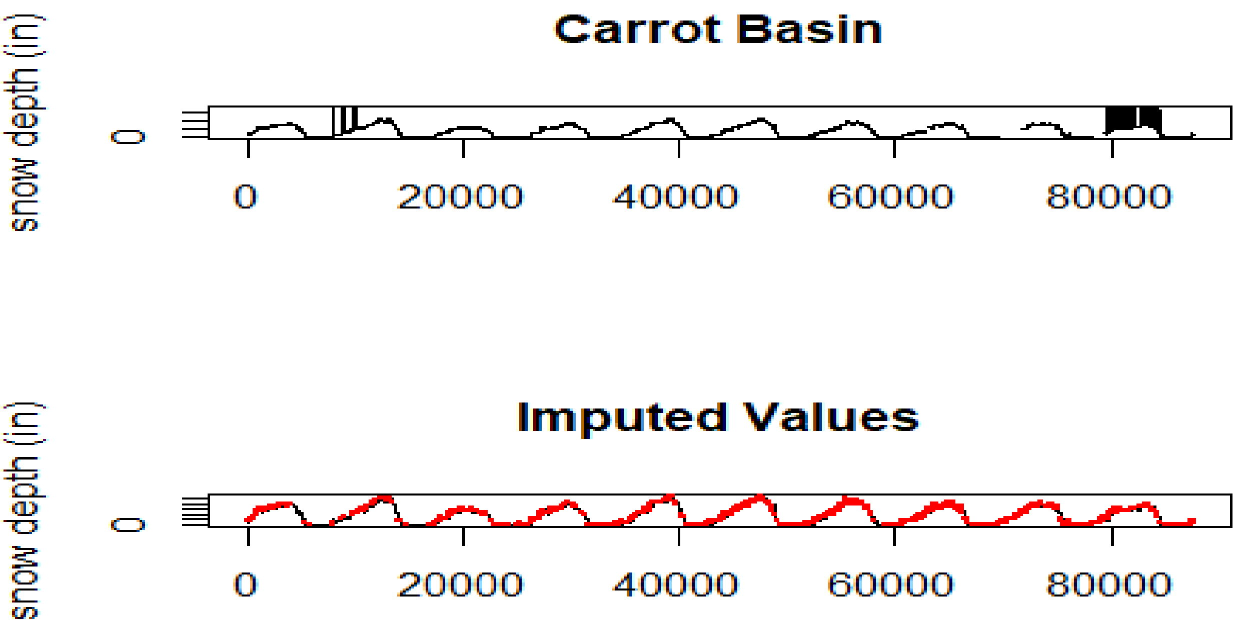
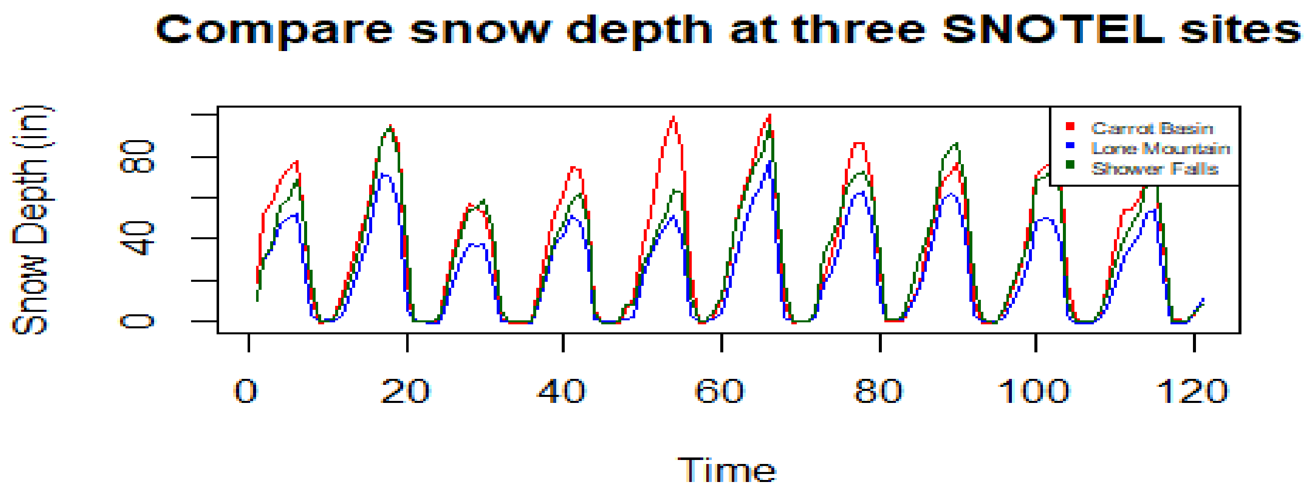
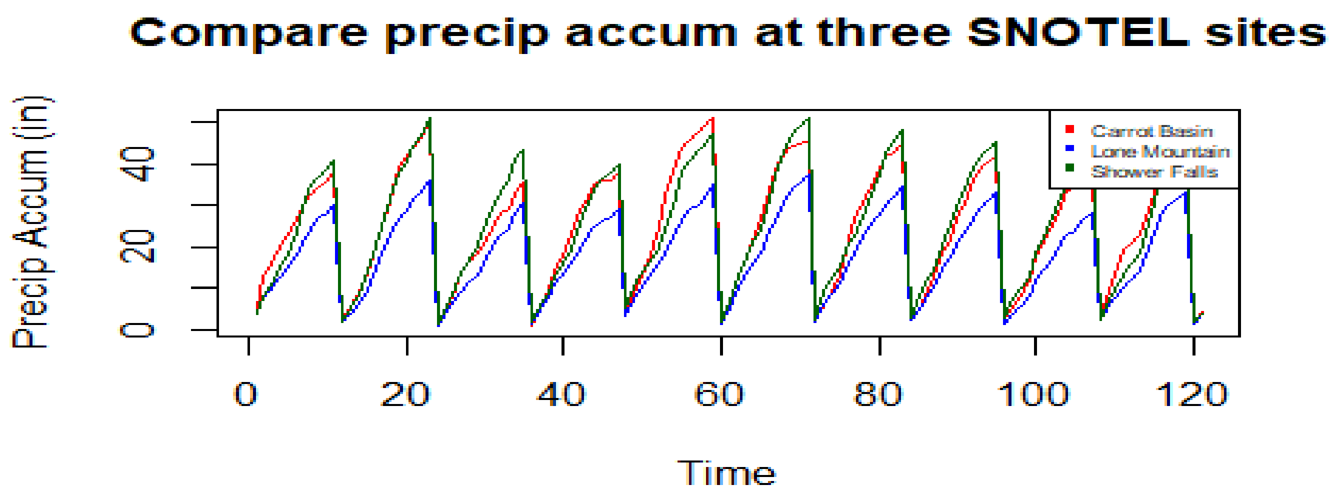


Figure 5: A plot of the snow depth from Carrot Basin with raw data in black and imputations in red.

Once the hourly data is cleaned and has had missing values imputed, I thickened to monthly averages. Now let's look at the set of nine predictors to see if we can reduce this down. Figure 6 shows the three weather measurements from each of three SNOTEL sites. It turns out that each of the three weather stations give very similar weather measurements. Having all three weather stations is redundant, so I will choose just one: Carrot Basin.


Compare snow depth at three SNOTEL sites
Snow Depth (in)
0
40
80
0
20
40
60
80
100
120
Time
Carrot Basin
Lone Mountain
Shower Falls


Compare precip accum at three SNOTEL sites
Precip Accum (in)
0
20
40
0
20
40
60
80
100
120
Time
Carrot Basin
Lone Mountain
Shower Falls

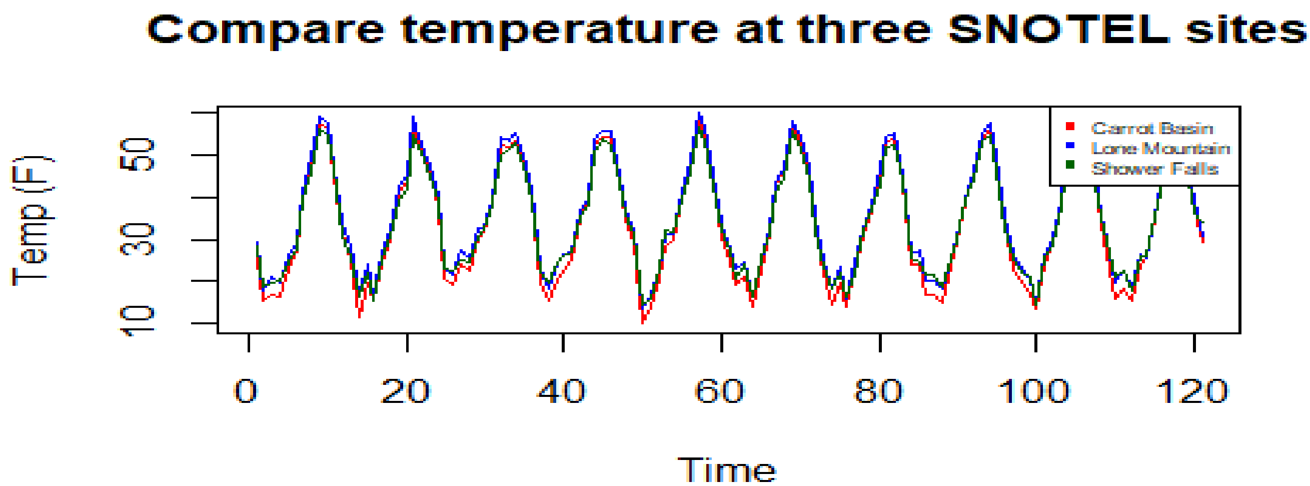


Figure 6: Weather measurements from each of the three SNOTEL sites. Notice each weather station produces very redundant measurements.

Next, we will plot the entire time series as a visual check for any obvious correlation between the covariates and the response. Figure 7 shows that the predictor `precip accum` is extremely correlated with `discharge.snow depth` looks like it has moderate correlation with `discharge`. And `temp` is also highly correlated with `discharge` but with some lag. All of these covariates should have good predictive power. As a final exploratory check, I will look at the autocorrelation of the response. Figure 8 shows the ACF and PACF plots of `discharge`. In the ACF plot we can see that there is definitely autocorrelation. The PACF indicates that maybe an AR(2) model is appropriate with some seasonal components.

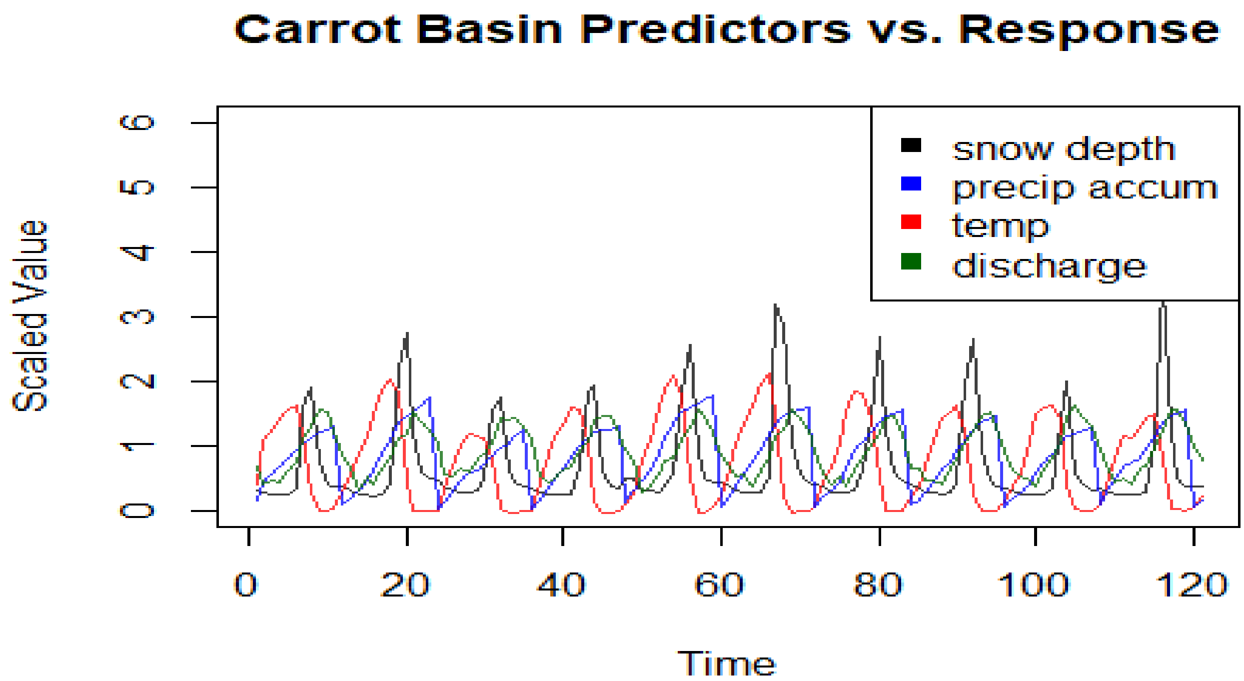


Figure 7: Plot of the time series predictors vs. response from Carrot Basin.

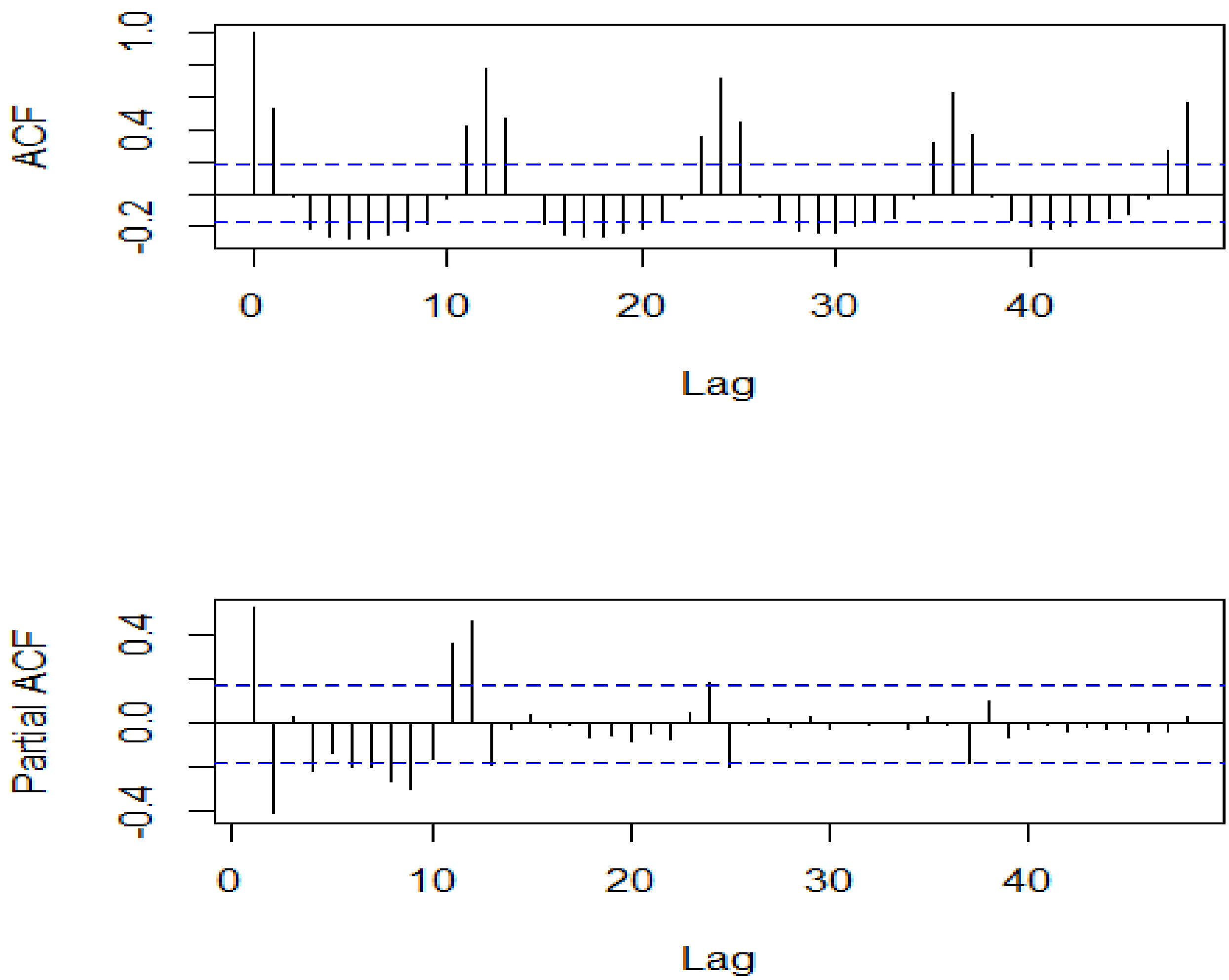


Figure 8: ACF and PACF plots for the response variable `discharge`.

Methods

With ten years of data, I have split the train and test sets in a sequential manner to maintain the time series information.

- Set 1: train = 2012-2017, test = 2018
- Set 2: train = 2012-2018, test = 2019
- Set 3: train = 2012-2019, test = 2020
- Set 4: train = 2012-2020, test = 2021
- Set 5: train = 2012-2021, test = 2022

So the training set will be anywhere from five years (60 data points) to nine years (108 data points). And all the test sets will be 12 month ahead forecasts (one full year). This produces a 5-fold cross-validation that should characterize the various models plenty well. The RMSE will be used as the loss function, so we can interpret the error in terms of the original units (CFS).

I will be considering seasonal ARIMA models for this data set since there is an obvious seasonal component, an AR component, and possible MA and differencing components. These models will include various combinations of the covariates as well. The seasonal ARIMA(p,d,q)(P,D,Q)$_s$ model can be succinctly expressed using the backward shift operator.

$$\Theta_P(\mathrm{B}^s)\theta_p(\mathrm{B})(1-\mathrm{B}^s)^D(1-\mathrm{B}^s)^d x_t = \Phi_Q(\mathrm{B}^s)\phi_q(\mathrm{B})w_t$$

where Θ_P, θ_p, Φ_Q, and ϕ_q are polynomials of order P, p, Q, and q respectively.

I will also be considering a Holt-Winters model as a different approach. This model does not take any covariates. It will simply be a moving average of the response itself plus trend and seasonal components. I expect this model to perform the worst, since it does not include the additional information from the covariates.

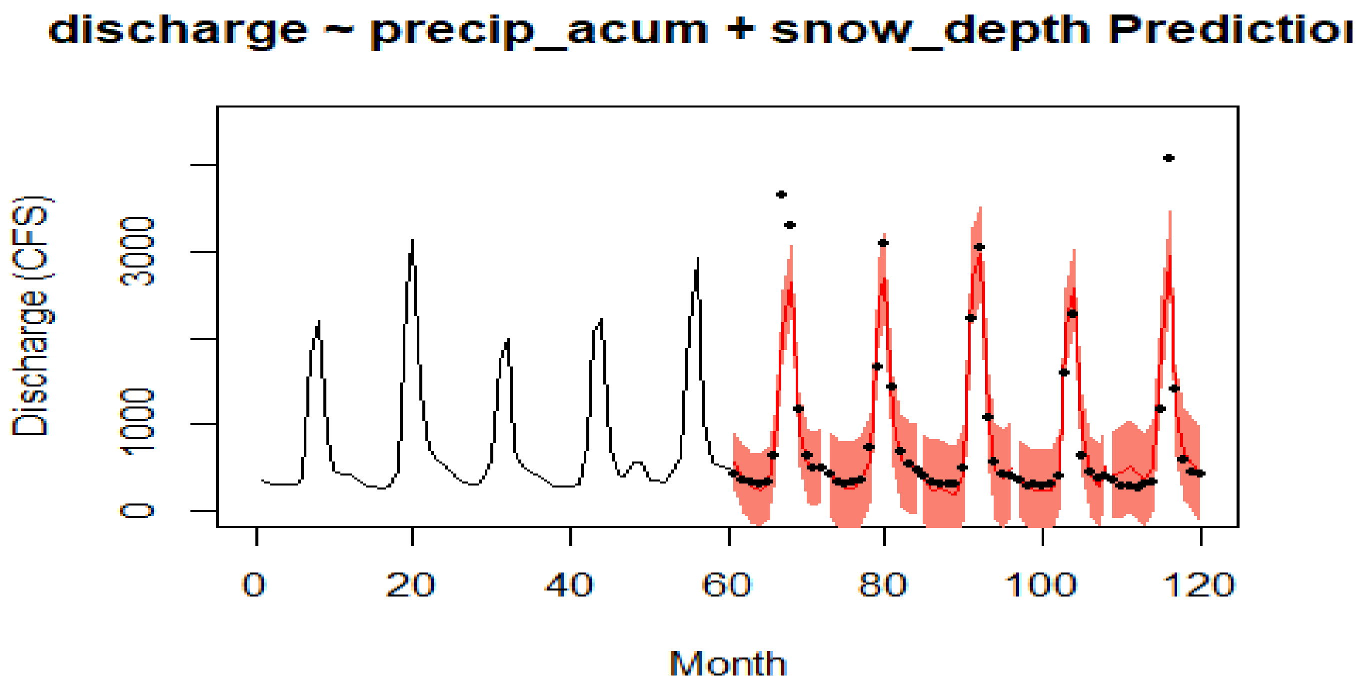
With covariates that all visually show correlation with the response, I decided to train models on all possible combinations of the covariates. Since there are only three covariates, there are only 2^3 possible combinations: one model with no covariates, three models with one covariate, three models with two covariates, and one model with all three covariates. But since the Holt-Winters model contains no covariates, I replace the no-covariate ARIMA model with the Holt-Winters (it felt redundant otherwise).

Results

After performing 5-fold cross-validation for 12 month ahead forecasts on all eight models, we have results. It turns out that all seven SARIMA models perform similarly well. The best model is the two covariate SArIMA model: discharge ~ snow_depth + precip_accum with 5-fold average RMSE 489.3. And as expected, the Holt-Winters model has the worst performance with 5-fold average RMSE 542.9. Table 1 displays results for all the models. Also interesting, the top four best models all have the same SARIMA framework: (2,0,0)(2,1,0)[12] which matches our expectations from the PACF plot in the exploratory analysis above. The predictor `precip_accum` appears to be the best covariate since it appears in all of the top four best models, with `snow_depth` being a close second as it appears in the top two best models. Figure 9 shows all the predictions for the best two models compared to the true `discharge`. The predictions are almost indistinguishable between the two models, and the RMSE are so close that we would expect them to be almost identical models. For each of these models there are only three (maybe four) true `discharge` points that lie outside the confidence intervals.

no.covariates	type	details	covariates	RMSE
2	SARIMA	(2,0,0)(2,1,0)[12]	snow_depth + precip_accum	489.3
3	SARIMA	(2,0,0)(2,1,0)[12]	snow_depth + precip_accum + temp	489.5
1	SARIMA	(2,0,0)(2,1,0)[12]	precip_accum	490.9
2	SARIMA	(2,0,0)(2,1,0)[12]	precip_accum + temp	491.1
1	SARIMA	(1,0,0)(2,1,0)[12]	temp	500.2
1	SARIMA	(1,0,0)(2,1,0)[12]	snow_depth	504.2
2	SARIMA	(0,0,1)(2,1,0)[12]	snow_depth + temp	507.1
0	Holt-Winters	beta=T, gamma=T	none	542.9

Table 1: List of all eight models and RMSE score.



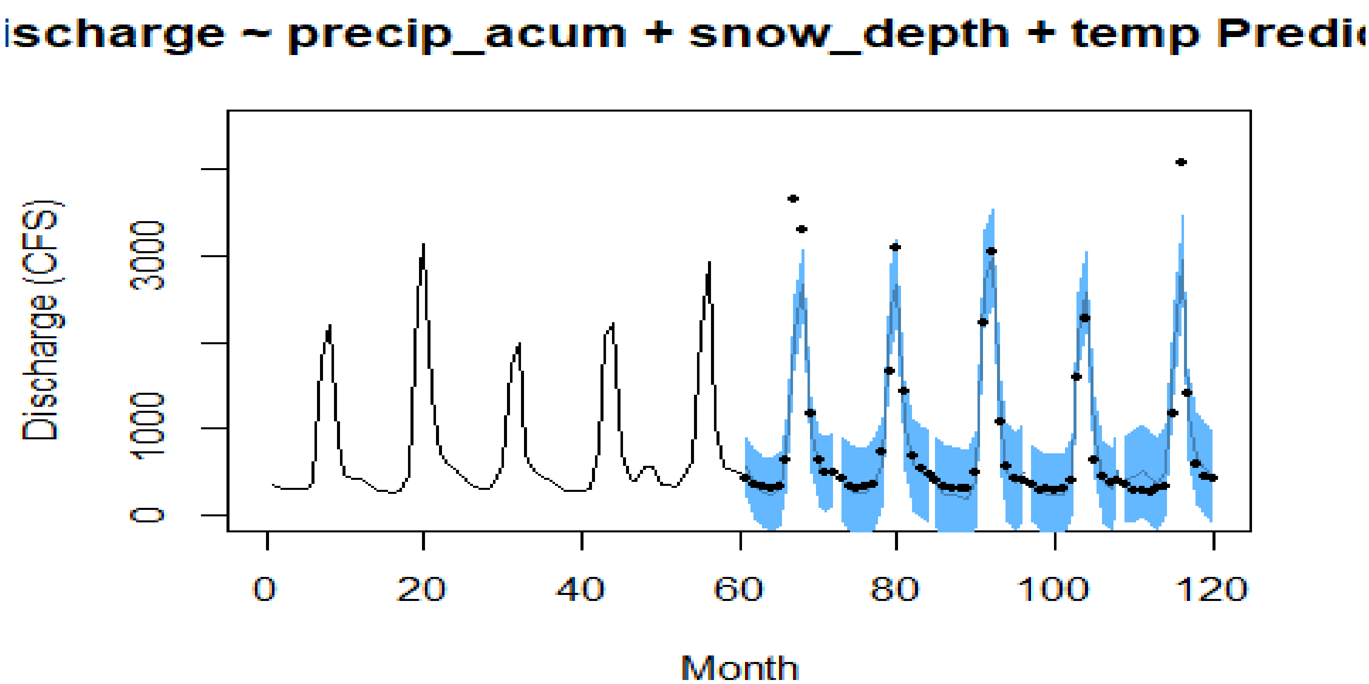


Figure 9: Plots of the best two models' predictions with shaded 95% confidence intervals and true `discharge` as black points.

As a final step in this analysis, I wanted to look at the ignorance scores of these top two performing models. To do so, I have to first fit Bayesian SARIMA models in order to obtain posterior distributions. I used the `stan_sarima` function from the `bayesforecast` package to fit the models. The MCMC process used four chains of 2,000 iterations each after a warm-up of 1,000 iterations. I assume convergence because the Rhat from these simulations are all extremely close to one (within three significant digits) for all parameters, and the Neff ranges from 1,500 to 5,500 out of 8,000 possible iterations (4 chains x 2,000 iterations each). After obtaining the posterior distribution, I used the `logs_sample` function from the `scoringRules` package to compute the ignorance scores. I used the same 5-fold cross-validation methodology to find the average ignorance score for each month in the 12 month ahead forecast. The results of the ignorance scores are in Table 2.

month	best_model	next_best_model
Nov	7.415870	7.507205
Dec	7.266354	7.367533
Jan	7.167804	7.164513
Feb	7.201289	7.180823
Mar	7.273644	7.213526
Apr	7.352401	7.388678
May	15.906762	12.404566
Jun	26.482068	25.928281
Jul	7.604127	7.440319
Aug	7.384580	7.702115
Sep	7.557044	7.884000
Oct	7.426139	7.257925

Table 2: Ignorance scores for the 5-fold cross-validated 12 month ahead predictions for the best two models

Discussion

Each of the eight models performed similarly, with the top seven models (SARIMA with different covariates) being nearly identical and the eighth (Holt-Winters) only slightly off. Analyzing this, we can say that each of the three predictors: precipitation accumulation, snow depth, and temperature did a good job at predicting water flow in the Gallatin River at the mouth of the canyon. The best predictor in this data set is precipitation accumulation, and the worst predictor in this data set, temperature, still did a pretty good job. The different sets of covariates as predictors did not create any marked difference in performance. Therefore, the best model is probably the SARIMA model with just precipitation accumulation as the only covariate since one predictor is the simplest and most interpretable. It was interesting to perform a Bayesian analysis too. I'm glad I tried (and succeeded) in fitting a Bayesian time series model and applied an ignorance score as the loss function. This is an important and more modern approach to time series analysis. I appreciate all that you have taught me. Thanks for reading!