Steve Rekuc

Final Report:
Alta Snowfall Data Analysis

Problem Statement

Skiers love skiing in good snow conditions with fresh snowfall, but most skiers (and certainly many of the big spenders in skiing) have to book their ski trips far in advance of a reasonable weather forecast. However, some areas may have snowfall patterns that can be studied to determine if some weeks or months of the ski season receive more snowfall than other times. Even the slight chance that the resort will fare better would provide a skier with valuable knowledge in booking a vacation. Skiers and snowboarders often see snowfall as uniform throughout the winter or they use heuristics (Japanuary; March is Vail's snowiest month); we try to be a little more data focused.

We focus our study on Alta, a ski resort in the Wasatch Mountains of Utah; Alta was a mining town before it became a ski resort in the 1930’s. Since Alta has operated for close to 80 years, there is a very long data set of daily snowfall measurements going back to 1944. This provides a lot of information for us to tease out trends or correlations for analysis and modeling.

We look at Alta snowfall as it relates to seasonal trends or other factors that could help predict snowfall far enough in advance that someone could plan a vacation. Short-term indicators of snowfall (within 14 days) are not helpful, as that is just weather forecasting. We construct a model that shows seasonal trends in snowfall that takes into other long-term weather indicators.

Data Wrangling

We acquired climate and weather data from 3 separate sources for this project. We first retrieved snowfall, precipitation, and other daily weather measurements from a National Oceanic and Atmospheric Association (NOAA) for Alta; this contains 21,809 rows of data with 22 columns from November 1944 to May 2020. We were able to quickly throw out a lot of these data columns that contained little or no information, leaving us with just 7 columns: date, precipitation, snowfall, snow depth, maximum temperature, minimum temperature, and observed temperature. We filled in a lot of the missing weather data based on what made sense for that measured value: we filled in 0 for precipitation and snowfall since those are typically 0 (most days are dry); we assumed the previous day’s snow depth to fill those 2,211 missing values.

To fill missing temperature values, I used a variety of methods. If minimum (689 days) or maximum (596 days) temperature was not recorded, I assumed that was done by those recording (manually for much of the snowfall history) because it was the same as the observed temperature, so I used the observed temperature as minimum or maximum in these cases. For cases where only the observed temperature was missing (1370 days), I used an average between the minimum and maximum temperature values (split the difference). There were still days remaining where more than 1 temperature value was missing; for these days, I used the mean temperature (minimum, maximum, or observed accordingly) on that month and day to fill in for missing temperature values. I then corrected for some temperature anomalies to maintain the correct order: Tmax > Tobs > Tmin.

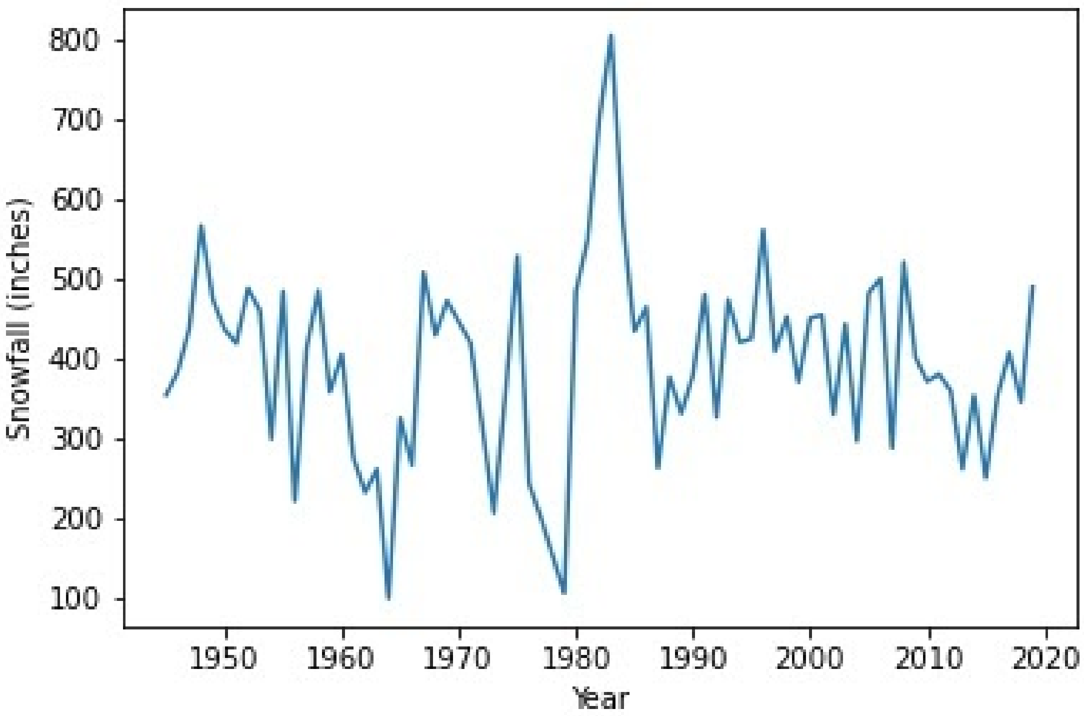
The next data set that we had the pleasure of wrangling with was the Pacific Decadal Oscillation (PDO). PDO is a very long-term measure of climate and can impact precipitation from Australia to Alaska. PDO monthly measurements go back to 1854, but since we don't have consistent snowfall data from Alta, we only take PDO since Alta had daily snowfall measurements available (November 1944). Since these are monthly measurements and change over long periods of time (decades), we marked the measurement as occurring on the 15th of each month and used linear interpolation to fill in all of the day between those measurements.

Our final data set was the Atlantic Mult-Decadal Oscillation (AMO); this is similar to the PDO but for the ocean on the opposite side of the continent. Measurements were monthly as well and linearly interpolated to fill in the missing values for each date.

We added day and month as features to the data set for quicker aggregation to finish with a dataframe that was 21,809 rows x 11 columns.

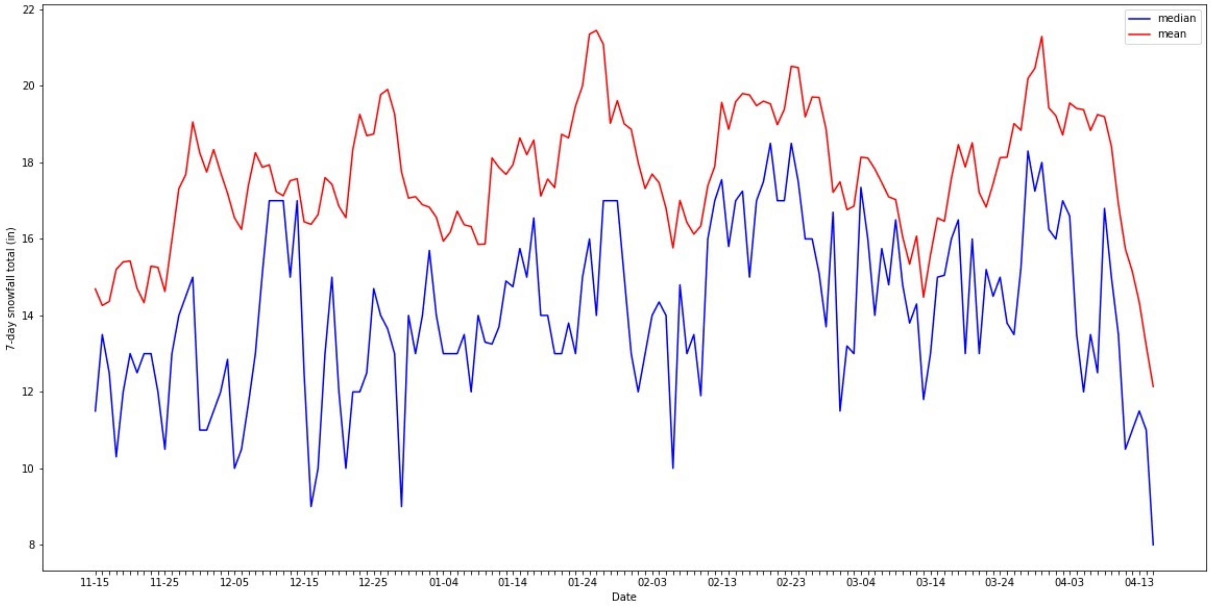
Exploratory Data Analysis

There are a lot of different aspects to snowfall that we would like to investigate. Let's begin by looking at how snowfall has changed over the years at the base of Alta.

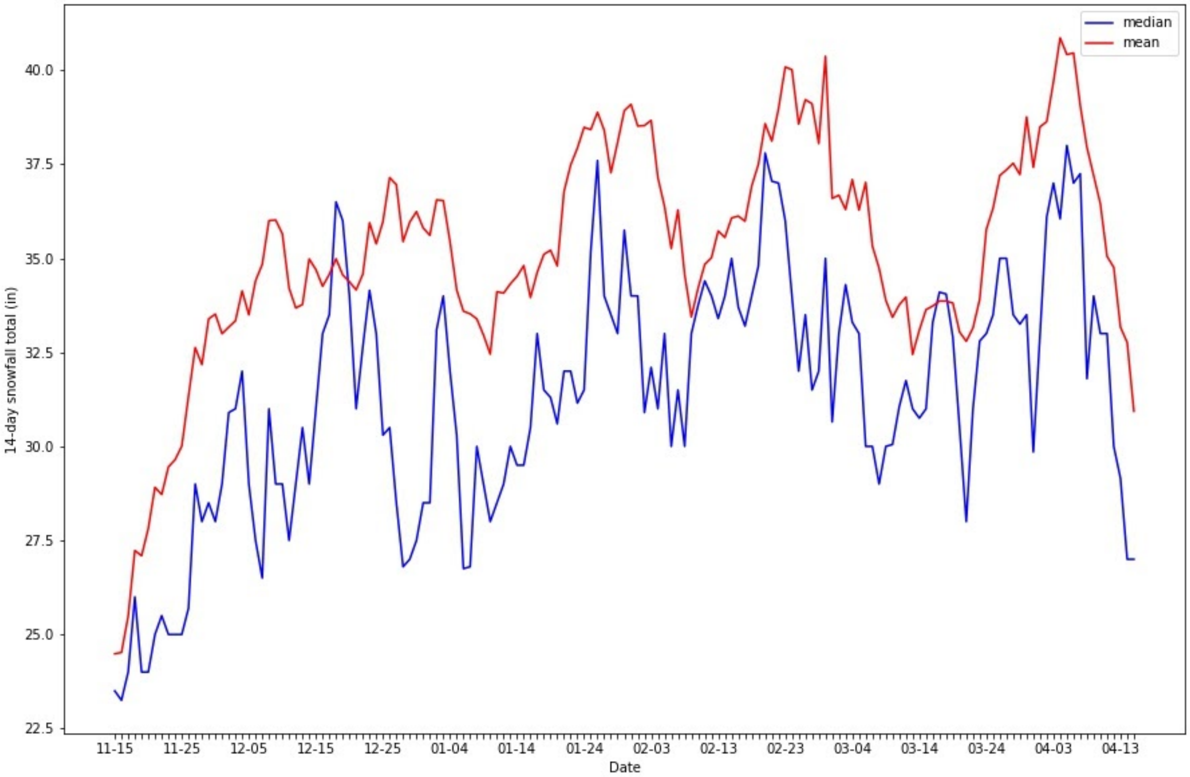


We can see that annual snowfall has a lot of variability. It has a mean of 397 inches with a standard deviation of 119 inches. There doesn't appear to be any discernible trend in snowfall over the years: it looks rather random.

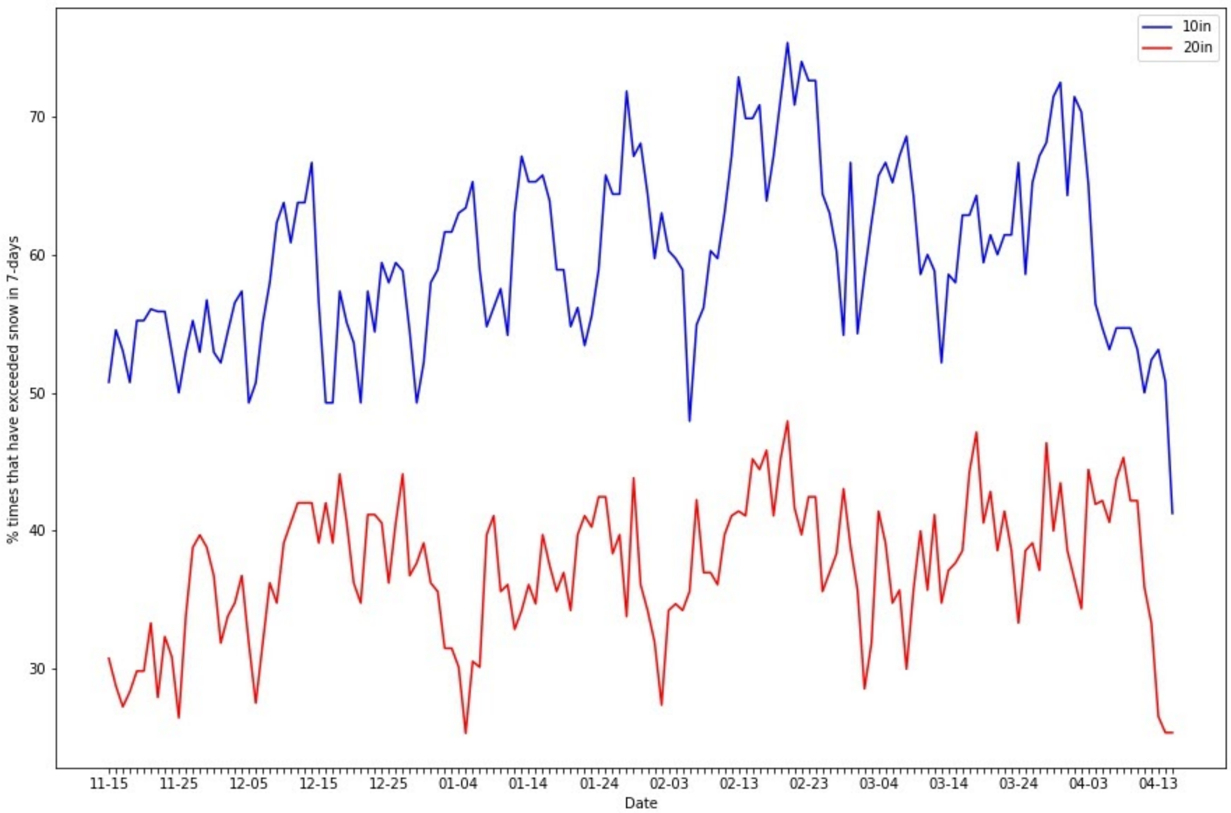
Most importantly, let's take a look at how snowfall varies throughout the ski season; since day-to-day snowfall varies significantly, we use a 7-day sum of the snowfall and then take the mean and median of those sums for each particular date.



There is still a lot of variability. Let's look at a larger window of snowfall: 14-days.

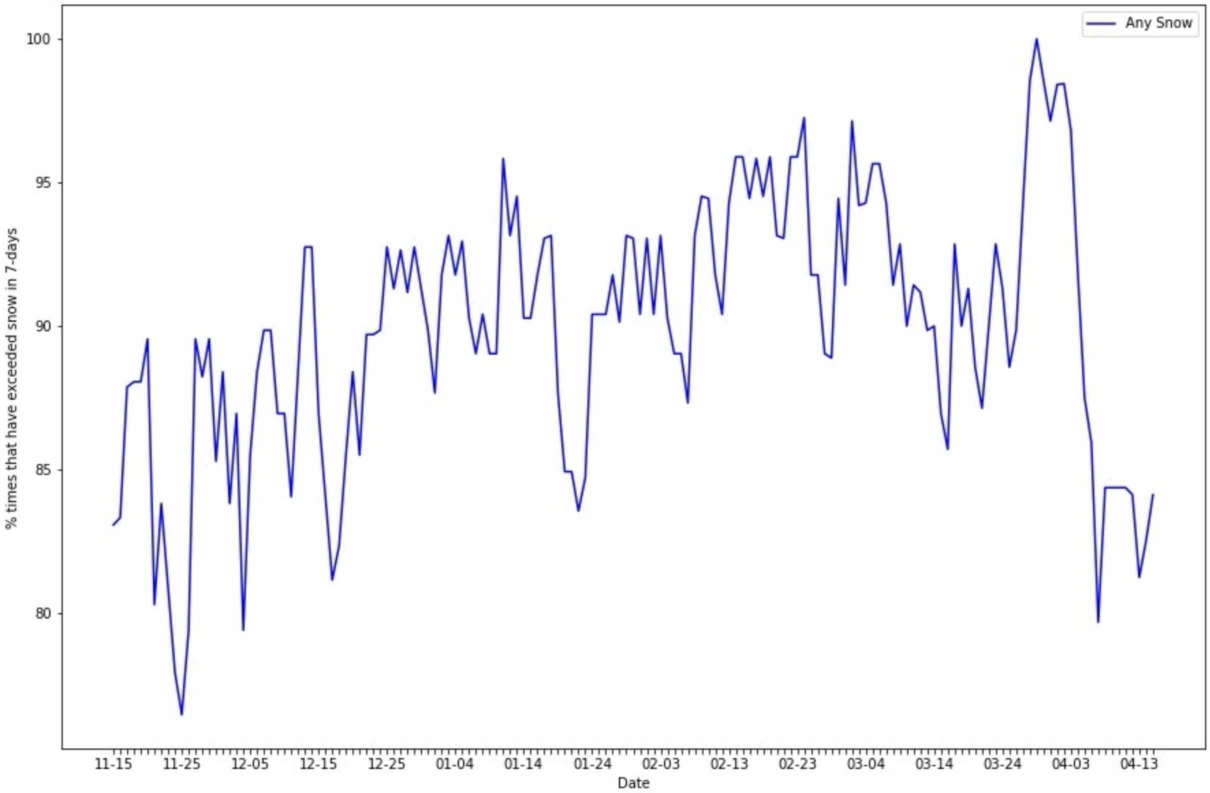


Some of the variability goes down and we see the trends of the winter more clearly: there are some days that more consistently have better snow than others. Snowfall certainly looks slightly better in late January, late February, and late March. To look at the consistency, we consider how often Alta received 10 inches or 20 inches of snowfall in a 7-day period.



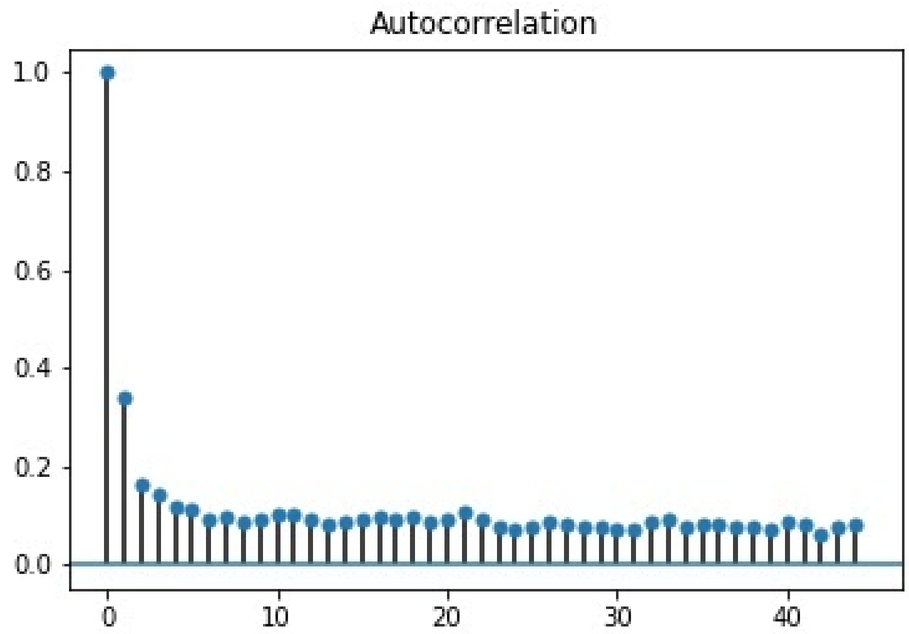
Those same times of year look promising for consistent snow: late January, mid to late February, and late March. Late January does look a little less consistent than the other prominent times.

While a snowy week is great, skiers at least would like to have some snow if we're going to a ski resort. So, we calculated the percentage of years that Alta received at least some snow during each 7-day period (ending on the listed date and rolling backward 7 days) at Alta.



The scale on the graph can be deceiving, but most winter weeks receive some snow in at least 80% of the years, but there is one stand-out date: March 29th. In every year of recorded snowfall at Alta, the week ending March 29th (March 23rd through April 29th) has seen some measurable snowfall; as little as 0.5 inches in 1966 and as much as 85 inches in 1983. In 71% of those years, Alta received 10+ inches of snow in that week; in 40% of those years, they received 20+ inches of snow. That looks like a good week to ski and these plots provide good direction in modeling the snowfall.

Since this is a time-series that we are modeling, let's look at the Autocorrelation function for snowfall to see if the previous day(s) of snowfall have any effect on the current day's snowfall:



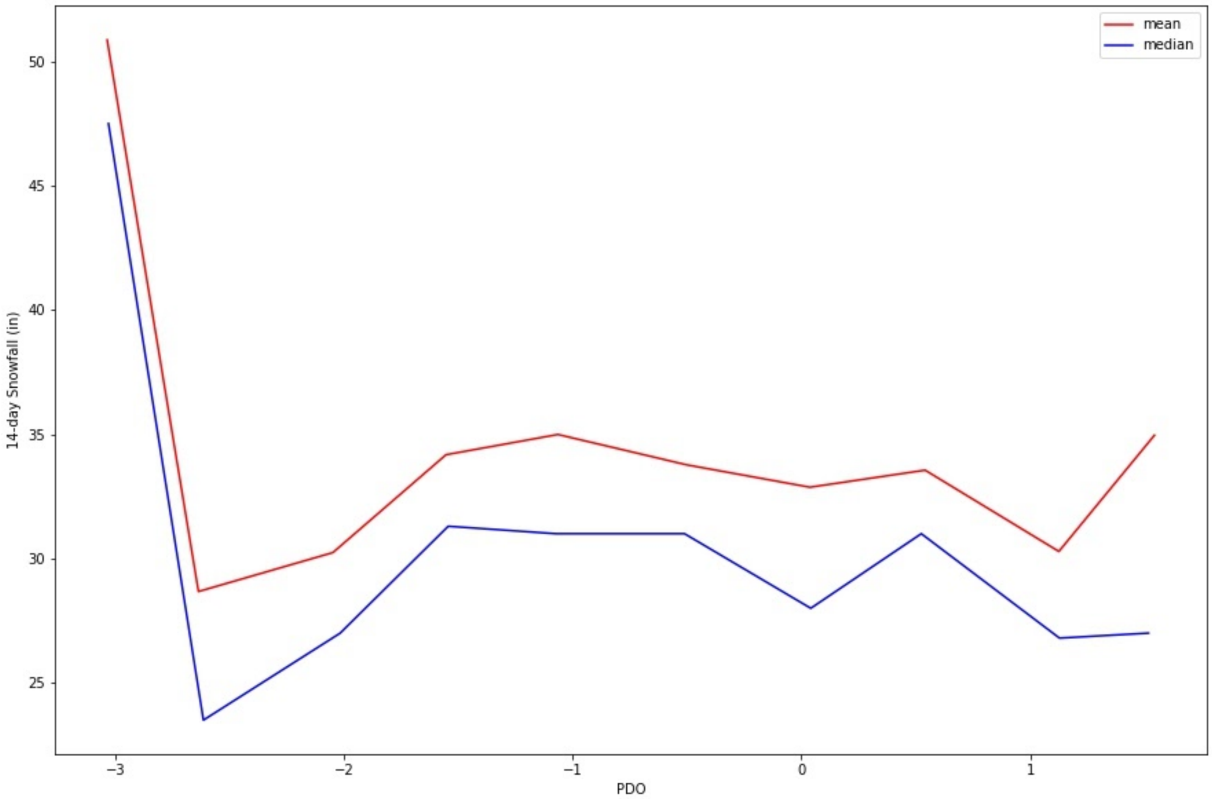
There is a slight correlation to the previous day of snowfall, but it's not much. That concludes our exploration into the time component of snowfall, but we still have to look at the other factor (features) that could potentially influence snowfall.

Test for stationary with respect to specific dates? I tested and got a good result of this being stationary once we subtract out the yearly and seasonal components - in other words, the dates are stationary.

Exploratory Data Analysis: The Other Features

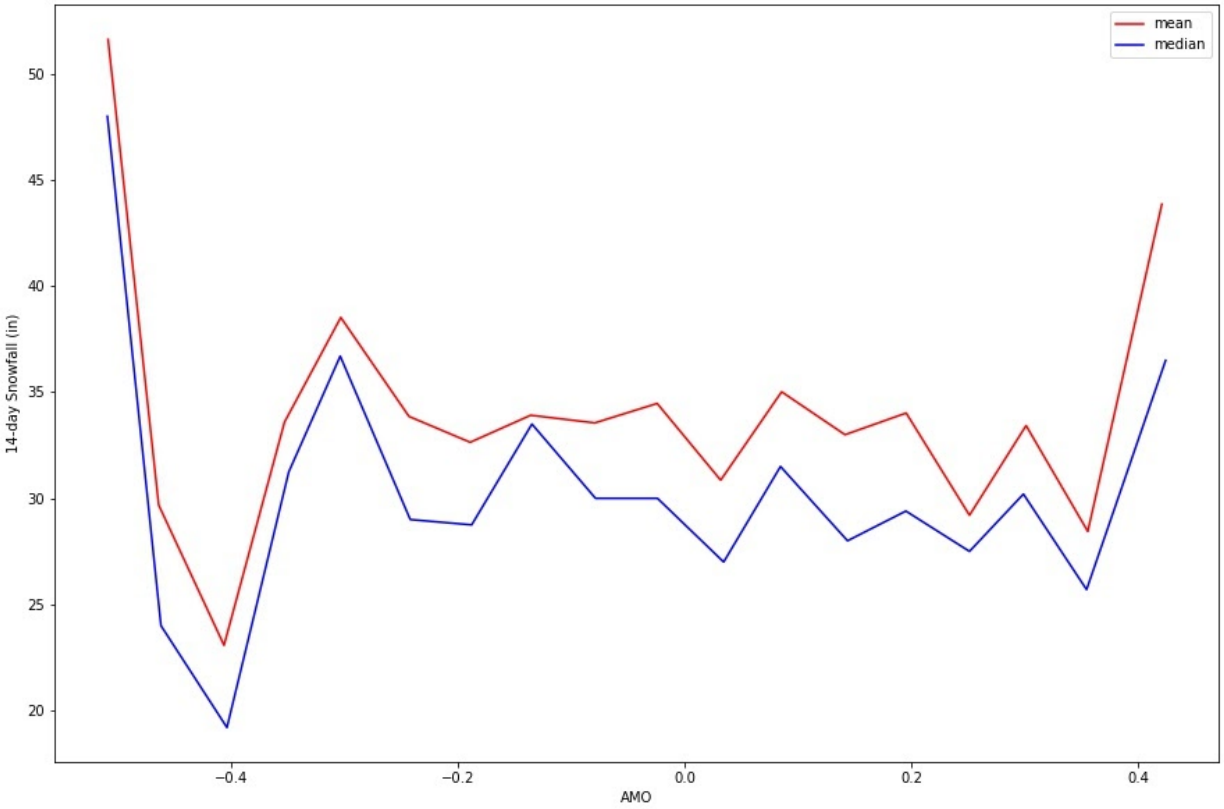
We are obviously very interested in how snowfall varies by seasonality at Alta, but the other factors that can affect snowfall in longer time-scales than the weather. We want to look at those features here.

One of those such features is Pacific Decadal Oscillation; this is an index of the ocean temperatures in the pacific relative to each other and the traditional temperature distribution in the Pacific. People noticed that this index had an impact on weather in the western United States with some areas being more heavily affected than others. Let's see if that index has any impact on the snowfall at Alta during the winter.



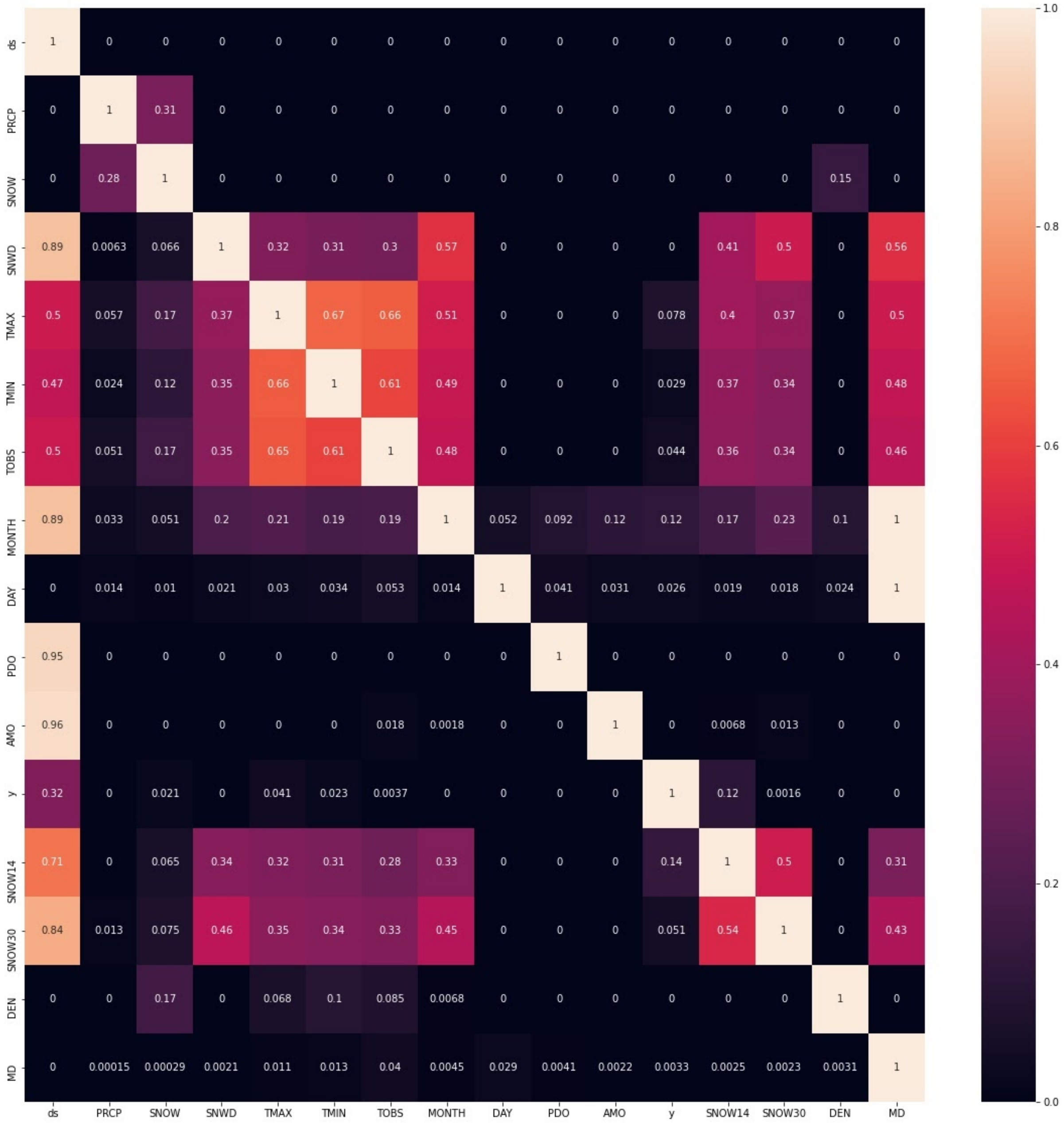
It looks like PDO has almost no effect on snowfall unless PDO goes really far negative - less than -2.9 on the index. Looking at winters of the past, this only happened in 4 months: January 1949, February 1949, March 1956, and November 2011; those were snowy times.

Let’s see if the AMO has any effect on snowfall; this is an index of ocean temperatures in the Atlantic Ocean relative to traditional temperature distribution.



Snowfall also looks to be extremely high when AMO is low or high, but not really affected in the middle range. AMO was only below -0.484 twice (April 1974 and March 1976) and AMO was only above 0.384 twice (in November 1955 and April 2010); these values may be outliers as well.

Before we move on to the next phase, we take a look at the correlation heatmap.



	ds	PRCP	SNOW	SNWD	TMAX	TMIN	TOBS	MONTH	DAY	PDO	AMO	y	SNOW14	SNOW30	DEN	MD
ds	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRCP	0	1	0.31	0	0	0	0	0	0	0	0	0	0	0	0	0
SNOW	0	0.28	1	0	0	0	0	0	0	0	0	0	0	0	0.15	0
SNWD	0.89	0.0063	0.066	1	0.32	0.31	0.3	0.57	0	0	0	0	0.41	0.5	0	0.56
TMAX	0.5	0.057	0.17	0.37	1	0.67	0.66	0.51	0	0	0	0.078	0.4	0.37	0	0.5
TMIN	0.47	0.024	0.12	0.35	0.66	1	0.61	0.49	0	0	0	0.029	0.37	0.34	0	0.48
TOBS	0.5	0.051	0.17	0.35	0.65	0.61	1	0.48	0	0	0	0.044	0.36	0.34	0	0.46
MONTH	0.89	0.033	0.051	0.2	0.21	0.19	0.19	1	0.052	0.092	0.12	0.12	0.17	0.23	0.1	1
DAY	0	0.014	0.01	0.021	0.03	0.034	0.053	0.014	1	0.041	0.031	0.026	0.019	0.018	0.024	1
PDO	0.95	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
AMO	0.96	0	0	0	0	0	0.018	0.0018	0	0	1	0	0.0068	0.013	0	0
y	0.32	0	0.021	0	0.041	0.023	0.0037	0	0	0	0	1	0.12	0.0016	0	0
SNOW14	0.71	0	0.065	0.34	0.32	0.31	0.28	0.33	0	0	0	0.14	1	0.5	0	0.31
SNOW30	0.84	0.013	0.075	0.46	0.35	0.34	0.33	0.45	0	0	0	0.051	0.54	1	0	0.43
DEN	0	0	0.17	0	0.068	0.1	0.085	0.0068	0	0	0	0	0	0	1	0
MD	0	0.00015	0.00029	0.0021	0.011	0.013	0.04	0.0045	0.029	0.0041	0.0022	0.0033	0.0025	0.0023	0.0031	1

We believe there are other factors that could have an impact on snowfall - specifically weather in the past. If there was a lot of snow in the early part of the season, could it mean there will be more snow than normal later in the season? For this reason, we included lag variables for such things as temperature and precipitation. We introduced these features for modeling, but we did not investigate those features in this exploratory data analysis; we could go into this forever. We constructed a larger heatmap, but it is not clearly legible with that many features (for a more detailed version, see the notebook). The additional features that we considered:

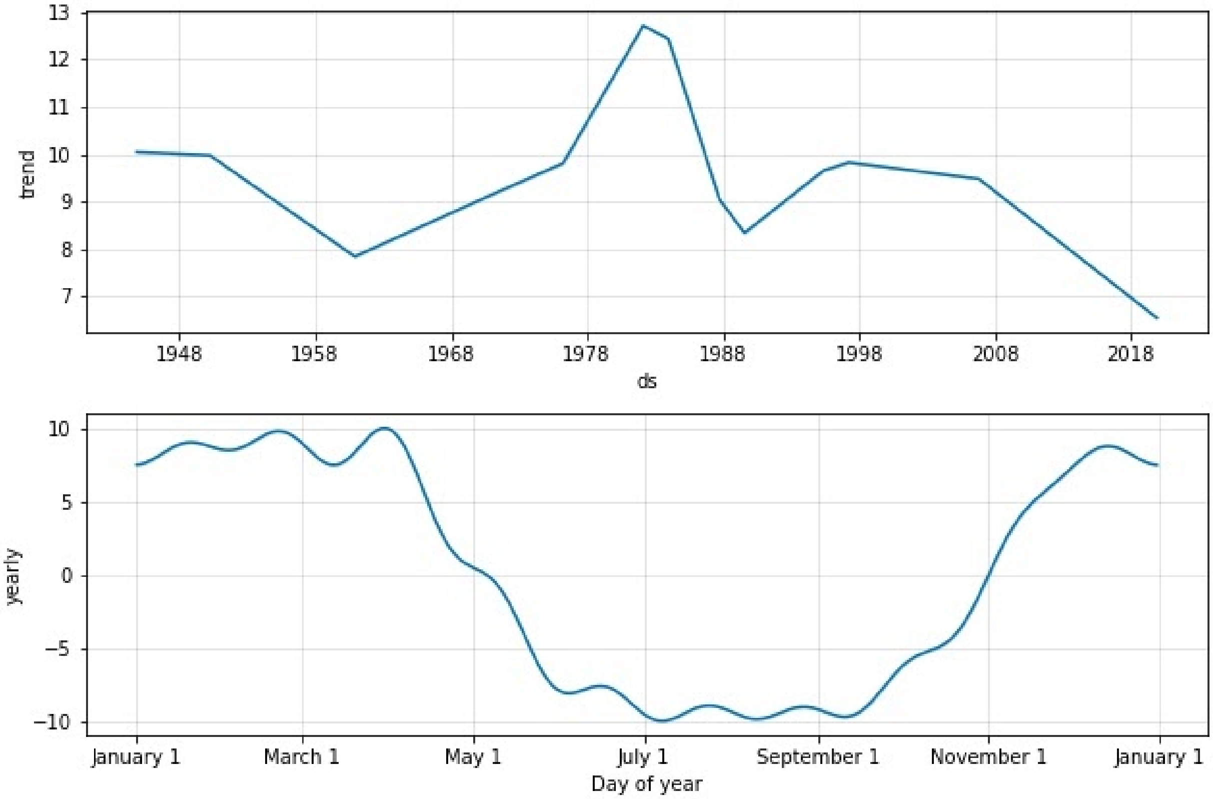
square of maximum temperature
square of minimum temperature
temperature maximum 7-day average

temperature minimum 7-day average
observed temperature 7-day average
temperature maximum 14-day average
Temperature minimum 14-day average
Temperature observed 14-day average
temperature maximum 30-day average
Temperature minimum 30-day average
Temperature observed 30-day average

We continue on to the next section with modeling the seasonality and annual snowfall trends as long with the effects of these features.

Modeling

The focus of our modeling is the seasonality of the snowfall and picking a good time to ski, so we tried some time-series modeling. We tried using MOTIF to tease out seasonality from the data, but that wasn't setup to easily account for seasonality and annual trends. Instead, Facebook's Prophet had some really good initial results for modeling 7-day snowfall total. This is time series modeling that easily breaks out the seasonal trends in data (includes lag 365) and shows the trend through that time as well.



This is a great model to visualize snowfall, as the model results are broken into 2 components that can be summed to be the predicted snowfall for a 7-day period: sum the annual component (top) with the seasonal component (bottom). There are slight but noticeable trends in the winter with the same peaks (late February and late March) and valleys (early March) that we observed in the Exploratory Data Analysis.

That's great to see the model matched our explorations. How well did it perform? We built the above model by slicing off the last ski year (November 15, 2018 - April 15, 2019) and then can test versus this model. The RMSE for that model was 14.45 inches; this is for an average wintertime snowfall of 16.7 inches; that's a very large error. When we performed the test on hold-out data for the last 3 winters, we got RMSE that was not particularly inspiring compared to the dummy model (average of 16.7 inches):

Model 1: Time Series in Prophet

Year	RMSE Model	RMSE Dummy
2016-17	16.72in	16.27in
2017-18	10.46in	11.38in
2018-19	14.45in	14.03in
Average RMSE	13.88in	13.90in

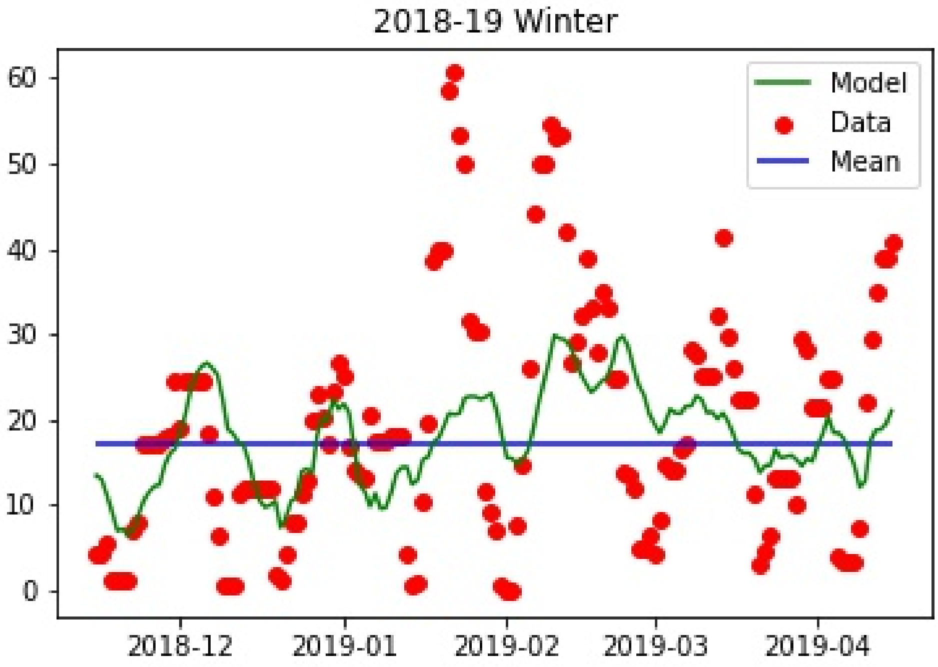
Our time series model had only a slightly better average than the dummy and the dummy model performed better in 2 of the 3 years.

Let's see if we can improve that model performance by including some regressors; in our next model, we include PDO, AMO, 7-day average maximum temperature, 7-day average minimum temperature, and 7-day average observed temperature. These results were much better:

Model 2: Time Series plus regression

Year	RMSE Model	RMSE Dummy
2016-17	13.26in	16.27in
2017-18	8.28in	11.38in
2018-19	12.11in	14.03in
Average RMSE	11.21in	13.90in

Let's see how one of those models look compared to the actual snowfall for just the last winter modeled (2018-19):



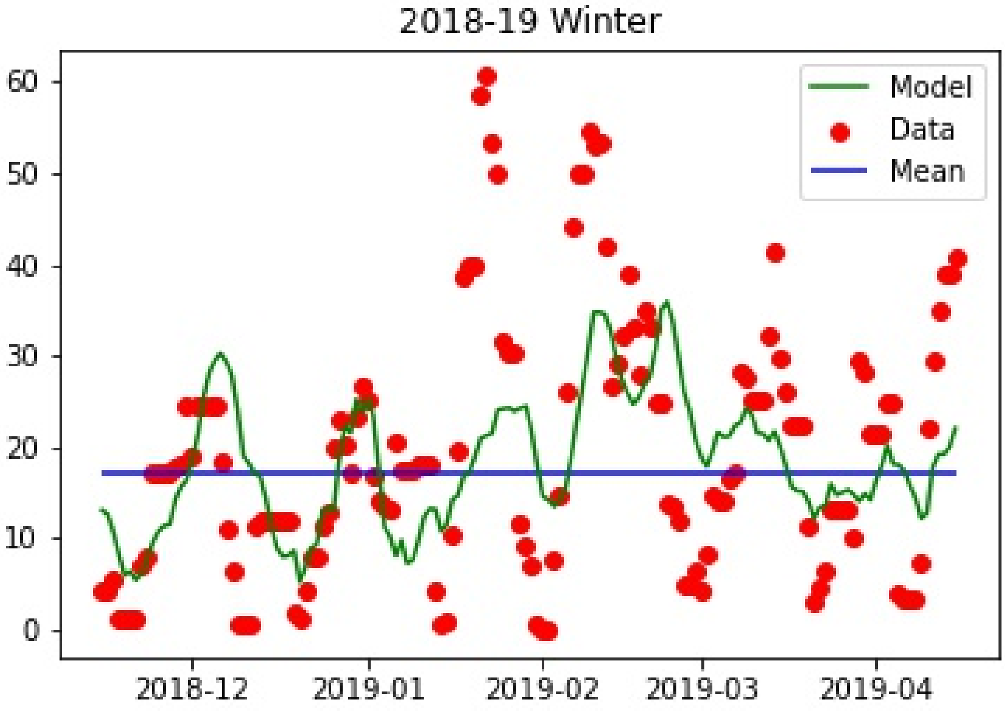
That obviously looks better than assuming the average (dummy model), but you can see the vast variability in the snowfall. The models underpredict this variability, but look much better than the dummy model, but if the models really were able to predict this variability, the models would likely overpredict a lot at the same time.

We next tried to add more variables to the regression to see if there was any improvement. Specifically, we added: precipitation lags for 30, 60, 90, and 120 days; the square of maximum temperature (7-day average); the square of both PDO and AMO. We hoped these additional variables would be good for creating more features for our model to work with. Let's see the results:

Model 3: Time Series plus regression with more variables

Year	RMSE Model	RMSE Dummy
2016-17	12.71in	16.27in
2017-18	8.46in	11.38in
2018-19	12.06in	14.03in
Average RMSE	11.07in	13.90in

That's barely any better at predicting snowfall. Let's see how that looks in the model plotted versus the actual snowfall for Winter 2018-19.



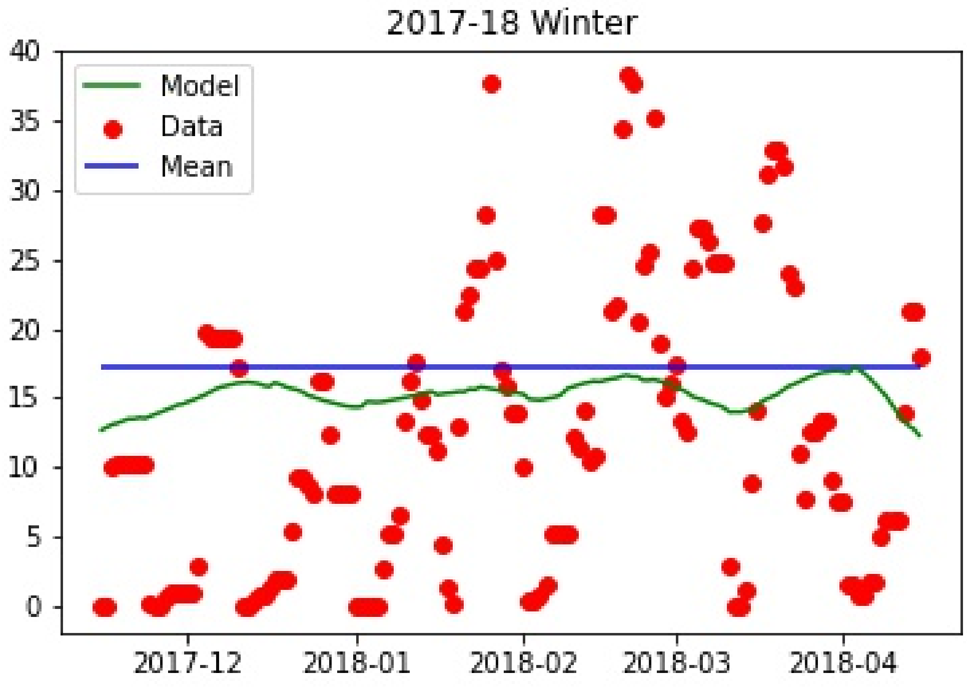
That looks extremely similar to the previous plot. Those features don't improve the model at all, but those features are actually something that we can "see" far enough in advance to make a prediction about snowfall.

In this next model, we considered only the features that we would know at least 30 days in advance; this is far enough out that a guest could plan a ski vacation and we weren't just relying on current weather. This model was built with the Prophet Time Series Model plus a regressor on PDO, AMO, and precipitation lag:

Model 4: Time Series plus regression with long-range variables

Year	RMSE Model	RMSE Dummy
2016-17	16.45in	16.27in
2017-18	10.43in	11.38in
2018-19	14.28in	14.03in
Average RMSE	13.64in	13.90in

That results is a little bit better than just the time series model and slightly better than the dummy model as well. Let's see how it looks for the year it performed the best, winter 2017-18:



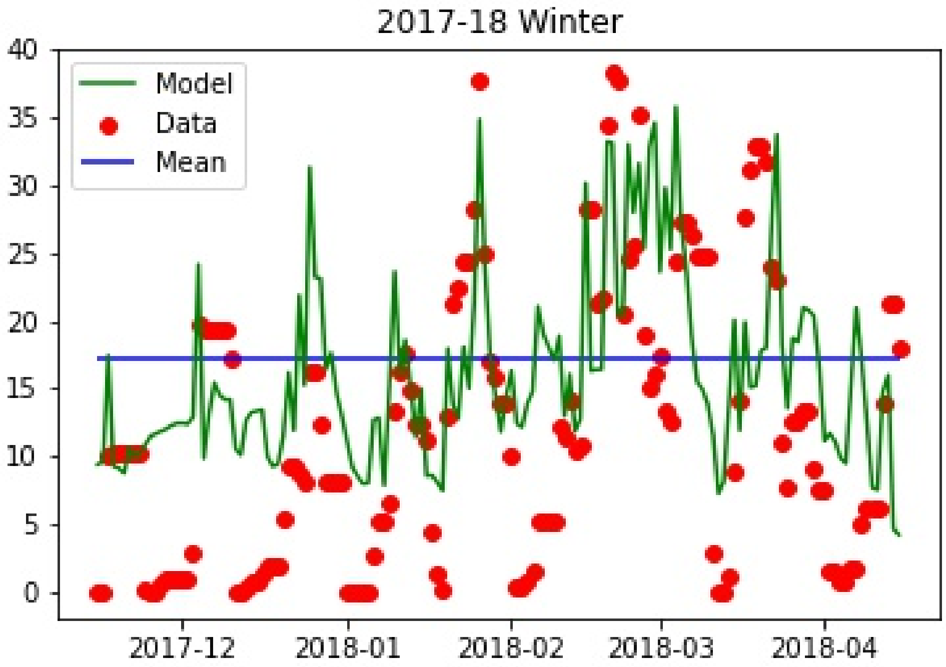
You can see the massive variability in snowfall, but it's clear that the model drops slightly below the average (dummy) at the times when it made sense. This year was a little bit less total snowfall than normal and the downward trend predicted by Prophet accounted for that.

The regressor slightly improved the resulting modeling, but what if we used something stronger instead of just a standard linear regressor, like a Random Forest Regressor? We tried running Prophet on the time series and then using a Random Forest Regressor on the residuals and the results were the best yet (only a slight margin over the Prophet plus Regressor).

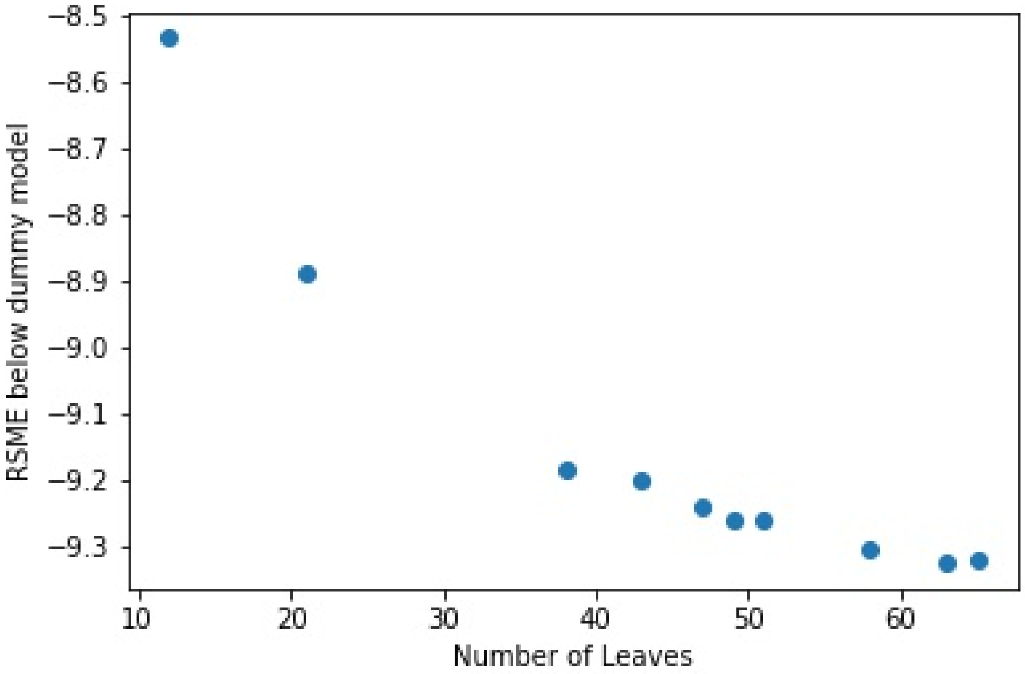
Model 5: Time Series plus Random Forest Regressor

Year	RMSE Model	RMSE Dummy
2016-17	12.10in	16.27in
2017-18	9.03in	11.38in
2018-19	11.20in	14.03in
Average RMSE	10.78in	13.90in

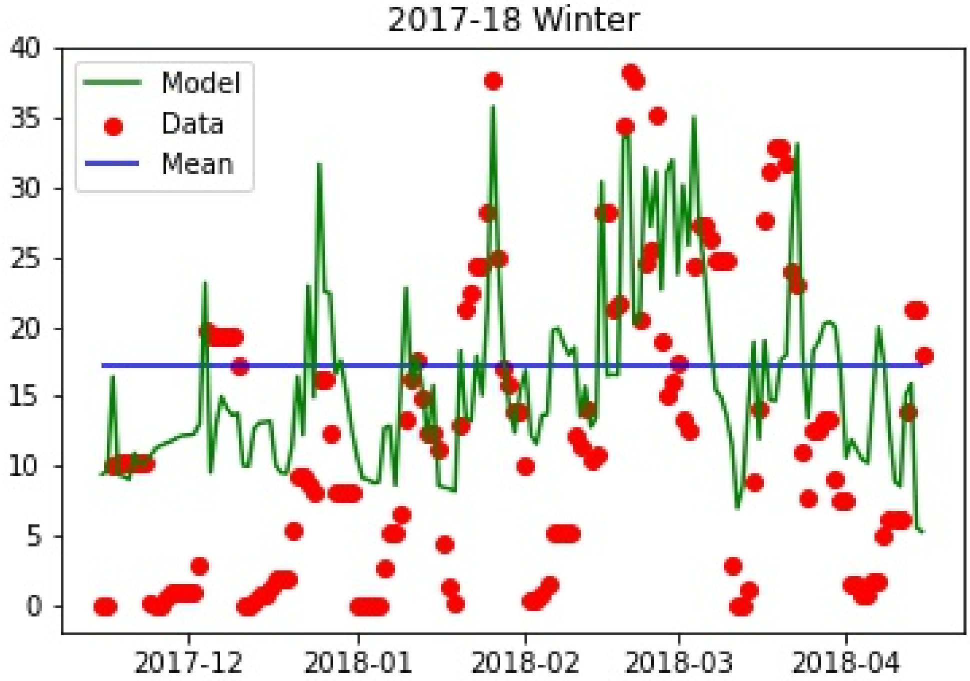
This is the best performing model so far; let's see how it looks relative to the actual values for Winter 2017-18:



That looks like it follows the resulting snowfall more closely, but it is dependent on a lot of short term variables, such as temperature. Before we eliminate those variables, let’s see if we can tune for a hyperparameter on this RF model; specifically, we tuned the maximum number of leaf nodes to avoid overfitting. Here is the plot of those nodes versus the improvement in RMSE over the dummy (average) model:

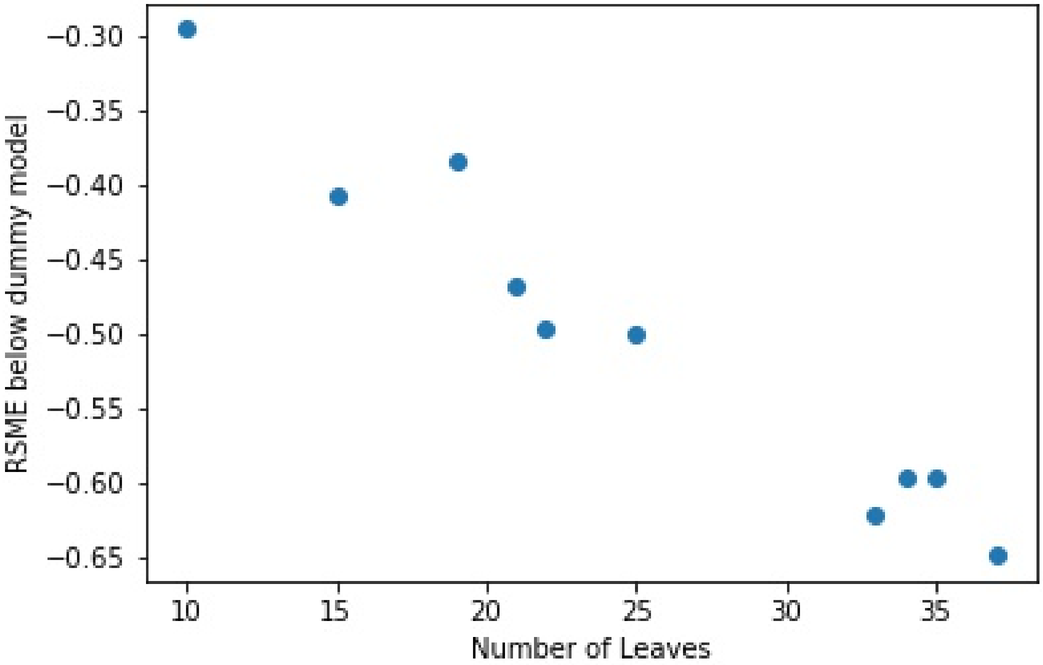


Around 39 leaves seems to be the best model or elbow; that model provides for a total 9.18 inches improvement in the RMSE over the dummy model sum for all 3 winters. Let’s see how this model performs on plots.

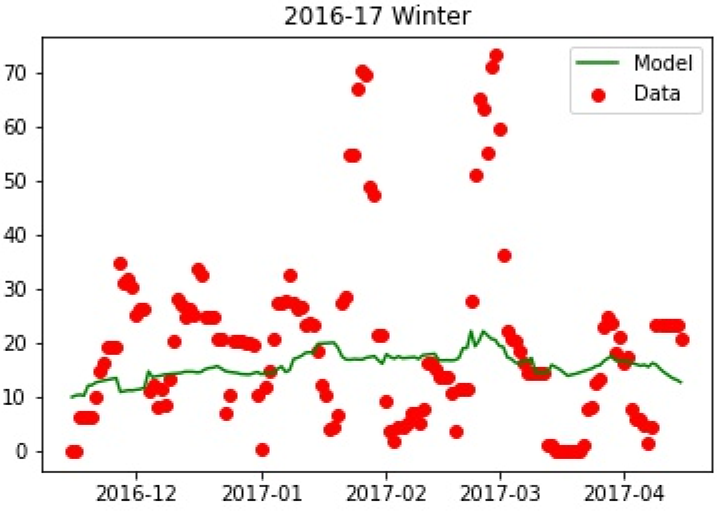


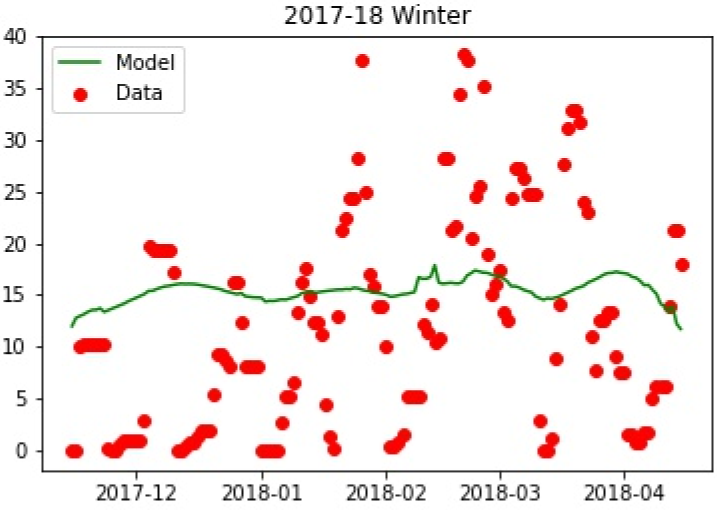
That is a better fit (possibly overfit), but does include variables that need to be known much closer to the arrival date. Let's see if we can tune a model that doesn't require recent weather data to predict snowfall.

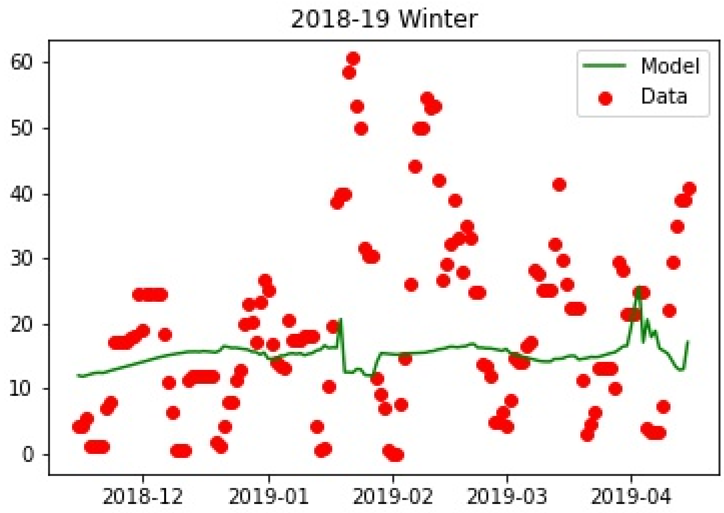
For our final model, we turn to the Prophet for time series modeling and then tack on a Random Forest Regressor on the Residuals. The features we provide to Random Forest are only features that could be measured (or predicted like PDO and AMO) over a month in advance. Using values much closer to the arrival is just cheating, as you may as well use a weather forecast at that point. Let's see if we could tune a good model out of those parameters. We plot the resulting difference in RMSE from the dummy model versus the maximum number of leaves.



Based on this plot, I would recommend a model with maximum leaves of 33. Let's build that and compare it to actual values for the 2017-18 season.


2016-17 Winter
Model
Data
0
10
20
30
40
50
60
70
2016-12
2017-01
2017-02
2017-03
2017-04


2017-18 Winter
Model
Data
0
5
10
15
20
25
30
35
40
2017-12
2018-01
2018-02
2018-03
2018-04



Those graphs are not pretty, but they are better than the dummy model that assumes the average by a slight amount.

Recommendations

The best performing model for estimating snowfall was built using Prophet to model the time series and then a Random Forest on the residuals; this is basically a boosting approach to Ensemble modeling. While this is the best performing modeling, it did include many short-term variables that would not be available until the weather forecast for that day was available - which defeats the purpose of the model (to predict snowfall beyond the typical weather forecast).

Instead, we recommend using the final model that we produced: Prophet on the time series and then using a Random Forest Regressor on the residuals. This produced the best (slightly) results using variables that could be measured more than a month in advance. Here is how that model looked for the last 3 years of snowfall:

We used this model to predict snowfall over the last 3 winter (2016-17, 2017-18, 2018-2019); when we did so, our model had an average RMSE of 0.60 inches better than the dummy model (that just assumed average snowfall for every week). We enjoyed looking at the data and modeling Alta snowfall, but after all of that work, our models were barely better than the traditional skier heuristic model for Alta: skiers assumed that each week in the winter is the same. It's good that we affirmed that starting point.

We investigated the snowfall at the base of Alta since 1944. We retrieved data from a number of different sources, explored that data from trends, and produced a number of different models. Our models were barely better than the dummy model unless we included variables that were really close to a particular day. This is the best we could produce right now, but in looking at this problem, we see far more opportunity to model this snowfall with better accuracy.

Future Research

While we explored this data and produced these models, we noted some approaches and some data that we could have included in our models that might have improved our results. We highlight some of the potential improvement and opportunity for future work:

1. **Log of snowfall in models -** In our exploratory data analysis, snowfall looked like it was exponentially distributed - with most snowfall stacked at the low end with a long tail to the right. With this kind of distribution, it would make sense for us to take the log of snowfall to convert it to something much closer to a normal distribution. This will probably lead to much better results in modeling. We jumped into the modeling without transforming the data like this and missed out on this opportunity. This is the best potential improvement in our model.
   <insert figures of exponential distribution>
2. **Substitute values for Zero Snowfall** - Our datasets included a lot of zero values for snowfall. In fact, over 80% of the measurements had no snowfall. If we take the log of snowfall values, then we need to substitute values for zero snowfall (log of zero is undefined); we might as well make those a little richer to indicate conditions where would not be able to fall or would be melting. Perhaps we would use 0.1 in for 0 snowfall but still below 32 degree; 0.01 in for 0 snowfall but above 40 degrees; 0.001 in for snowfall and above 50 degree and so on (just as examples). This adds a richness to the model that indicates that not only is it not snowing, but the conditions are adverse to any snowfall and likely are causing the snow to melt.
3. **Include Southern Ocean Oscillation Index (SOI) that indicates El Nino / La Nina** - The Southern Ocean Oscillation Index (SOI) is very similar to PDO but operates on a much shorter time scale: years and months rather than decades. This is the index that is often used to indicate an El Nino (warm Pacific) or La Nina (cold Pacific) Winter. This would have been great to include in our models, but was left out as I was not confident in the measurements I found and those measurements did not extend back to the beginning of Alta’s snowfall record (in some cases). Future models would be best to include this measured value from a reliable source.
4. **Include more lag variables and features in general -** We manually created some of our lag variables that might have an impact on our model. It would be better if we created some more lag variables (temperature, PDO, AMO, different aggregates) to determine if one of those has an impact on snowfall. With a larger variety of variables, we could just test those variables by running them through the Predictive Power Score or calculating the correlation coefficient; we would just drop the features that did not have a connection

to snowfall and model with the remaining that do (but were not too closely correlated with each other).

5. **Different Ensemble Models** - we only tried using a couple of different ensemble models: Prophet's built-in regressor and Random Forest on the residuals from Prophet. There are many other ensemble models that may combine the feature with time series for a much more accurate model (multiplying, voting, averaging, bagging, etc.). We would need to test out more of these approaches to see if the effects of features and time series to produce a more accurate model.
6. **Classification Problem** - Snowfall is sparse and an event; it might be better to try this as a classification problem that sorts between snowfall and no snowfall; or between different "bins" of snowfall totals.
7. **Look for the Opposite** - Days that do NOT have snowfall are abundant. Would it be better to build a classifier that looks for those days instead. Skiers would still gain better insight into when they should go skiing to avoid those times of no snow.
8. **Bootstrap Sampling** - Would it make sense to bootstrap sample different snowfall values for each day to build a smoother and more detailed model. We could essentially simulate 10,000 years of snowfall by resample (with replacement) the snowfall values.
9. **Inspect Other Mountains with More Seasonality** - Alta is known for its consistent snowfall and doesn't experience as much variability as other mountains or regions. While that is great for skiers, it can make modeling more difficult as we try to tease out seasonality or feature importance. If the snowfall were to vary more with PDO or AMO or seasonally then it would make our results more interesting.
10. **Larger window of snowfall accumulation** - We used a 7-day snowfall sum in our models, but that still allows for some significant variability. A larger snowfall window would likely smooth out some of that variability and provide a better snowfall prediction. The issue with a larger window: that may be longer than most people intend for a ski vacation.

All of these ideas provide potential improvements on our model, but we really like the top 3 improvements. We will revisit those at a later time. The existing model is good enough to publish for your average Alta skier; for those looking for a slight edge in snowfall, this may provide some insight, but for most it will affirm that snowfall is close to the same throughout Alta's ski season.

Thank you for reviewing this snowfall analysis. We hope you have a great ski season.