IMPROVING MOUNTAIN SNOWFALL FORECASTS IN THE SOUTHWESTERN US USING MACHINE LEARNING METHODS

by

Charles Andrew Hoopes

Copyright © Charles Andrew Hoopes 2021

A Thesis Submitted to the Faculty of the

DEPARTMENT OF HYDROLOGY AND ATMOSPHERIC SCIENCES

In Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

In the Graduate College

THE UNIVERSITY OF ARIZONA

2021

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Master's Committee, we certify that we have read the thesis prepared by: Charles Andrew Hoopes

titled: IMPROVING MOUNTAIN SNOWFALL FORECASTS IN THE SOUTHWESTERN US USING MACHINE LEARNING METHODS

and recommend that it be accepted as fulfilling the thesis requirement for the Master's Degree.

Christopher L. Castro — Date: May 26, 2021
Christopher Castro

Ali Behrangi — Date: May 26, 2021
Ali Behrangi

Hsin-I Chang — Date: May 27, 2021
Hsin-I Chang

Reza Ehsani — Date: May 27, 2021
Mohammad Reza Ehsani

Final approval and acceptance of this thesis is contingent upon the candidate's submission of the final copies of the thesis to the Graduate College.

I hereby certify that I have read this thesis prepared under my direction and recommend that it be accepted as fulfilling the Master's requirement.

Christopher L. Castro — Date: May 26, 2021
Christopher Castro
Hydrology and Atmospheric Sciences

TABLE OF CONTENTS

I. Introduction……………………………………………………………………………4

II. Data and Methodology………………………………………………………..........7

- a. Surface and Upper-Level Meteorological Data………………………………8
  1. NARR Reanalysis……………………………………………………..8
  2. ECMWF ERA5 Reanalysis……………………………………………9
  3. High-Resolution WRF-HRRR Reanalysis……………………………9
- b. Target Snow Cover Data………………………………………………………9
- c. Methodology…………………………………………………………………10
  1. Spatial Domain………………………………………………………10
  2. Data Pre-Processing…………………………………………………..11
  3. Control Method – MLR……………………………………………...12
  4. Feed-Forward Neural Network………………………………………13
  5. Verification of Performance…………………………………………..13
  6. Parameter Optimization………………………………………………14
  7. Support Vector Machines……………………………………………..15
  8. K-Nearest Neighbor Algorithms……………………………………..16

III. Results……………………………………………………………………………..17

- a. Parameter Optimization………………………………………………………17
- b. Final MLR and Network Performance……………………………………….18

c. SVM and KNN Tests……………………………………………………………19

d. Data and Machine Learning Spatial Statistics……………………………..19

e. Alternate Input Data: ERA-5and WRF-HRRR…………………………….21

IV. Discussion……………………………………………………………………….22

V. Conclusion………………………………………………………………………25

VI. Tables and Figures………………………………………………………………26

VII. References……………………………………………………………………...37

ABSTRACT

Snowfall forecasting has historically been an area of difficulty for operational meteorologists, particularly in the remote complex terrain of the Western US. Attempts at improving forecasts have been made, but skill is still poor, with snowfall routinely overpredicted. A major reason for this overprediction has been a failure to accurately predict snow-liquid ratios (SLR) ahead of major events. This research proposes, develops, and tests multiple machine learning methods for dynamic SLR prediction for the Sky Islands of southeast Arizona in the form of a simple feed-forward neural network, a support vector machine, and a k-nearest neighbor algorithm. Input parameters were chosen based on variables found by previous studies to have a regression-based relationship with SLR, with a focus on the lower-mid levels of the troposphere. These parameters were also used to construct a multiple linear regression model, and its performance was compared with the machine learning methods. Each of the machine learning methods showed significant improvement compared to the multiple linear regression. When tested on historical events, nearly 95% of the network-predicted SLR values fell within the margin of error of observed SLRs, calculated using verification data from Zeng et al (2018), with slightly higher accuracies for both the SVM and KNN algorithms. Each showed significant gain in skill compared to the multiple linear regression model. Current and future work is focusing on shifting to higher resolution data, as well as adjusting the model to look at a wider region within the mountainous Western US in order to achieve greater operational benefit.

1. Introduction

The mountainous terrain of the Western US poses many challenges to operational forecasting. Rapid changes in elevation and local climate create tight spatial gradients which are challenging for numerical weather prediction models to resolve, even at convective-permitting resolution. The specific problem of interest in this work is the prediction of snowfall in complex terrain during winter storm events, with a focus on the region of southern Arizona. The mountain ranges in this part of Arizona tend to be relatively small in spatial scale (order of 10s of km) and are locally characterized as the Sky Islands. Although snowfall prediction in such unpopulated, high mountainous areas is not necessarily concerning, from the perspective of operational short-term weather prediction, it is highly relevant for longer term hydroclimate forecasts. Mountain snowfall during the cool season is vital for water supply in the Western U.S., inclusive of southern Arizona. Winter snowpack is a principal source of streamflow in this region, and a source of water supply for agricultural and urban uses. Snowpack is also necessary to replenish groundwater, through the process of Mountain Block Recharge.

Snowfall prediction in mountain ranges like the Sky Islands remains a major challenge, in part due to errors in predicting the snow-liquid ratio (SLR). Operational meteorologists have not traditionally considered variation in SLR in the prediction of mountain snowfall. National Weather Service operational forecast practice in the Southwest U.S assumed either a 10:1 ratio or based the SLR on the local climatology computed by Baxter et al (2005). The weakness in this type of approach is the assumption that SLR during all winter storm events is constant. A more dynamic approach to SLR forecasting, as will be presented in this work, could inform better snowfall forecasting, for example through post-processing of output from the National Blend of Models (NBM) within the NWS. Current SLR predictions within the NWS in the western generally over-predict SLR, resulting in many over-forecasted snow events.

Previous work has investigated key physical parameters that are most highly statistically related to SLR within the western U.S., using multiple regression approaches. The NWS Sacramento office developed a regression algorithm for SLR in the Sierra Nevada mountains, using temperature at two different levels, valley and mountain, as predictors. Though relatively simple, this prediction method reduced SLR error and resulted in an increase in forecast skill for snow events compared to static SLR prediction methods (Van Cleave 2013). Other approaches have focused more on extracting parameters of physical interest. Alcott and Steenburgh (2010) found that mountain top level temperature and wind speed had the strongest relationships with SLR within the Wasatch Mountains in Northern Utah, with humidity, wind direction, and the time of year being secondary considerations. The variable weights in their regression were relatively larger at levels above and below the mountain crest, with the largest values at the crest (Alcott and Steenburgh 2010).

The variables provided by these mountain SLR studies provide insights about potential parameters of interest and the physical rationale for considering them in a predictive scheme. Temperature, for example, governs the type of ice crystal habits and therefore the snow density (Alcott and Steenburgh 2010). Additionally, unrimed dendrites have lower densities and higher SLR than similarly unrimed columns. Rimming of crystals greatly increases density, so melting and refreezing will drastically lower SLR, often by a factor of over 50% (Power et al 1964). Both processes are highly dependent on in-cloud temperatures, and mountainous regions are often enveloped by clouds during winter storms. Supersaturation is another factor influencing ice crystal habit, with more saturated regions favoring dendrites, and less saturated regions favoring plates. Additionally, supersaturation affects the ability of a crystal to rime. Rimming is more likely with higher moisture, due to the increased concentration of supercooled droplets, particularly if temperatures are close to freezing. Fall speed and surface area, both dependent on ice crystal type, affect the ability of a falling crystal to collide with supercooled droplets and undergo significant rimming (Power et al 1964).

Once the crystals fall below the cloud base, several other factors become important in influencing SLR. There are processes related to the temperature and humidity. Melting and sublimation can reduce the amount of snow water reaching the ground.   Provided sufficient moisture, there may be continued rimming and aggregation of falling snow. Wind mechanically removes the outer branches of dendritic crystals, making them more plate-like (Alcott and Steenburgh 2010). Wind direction may also be relevant, as speed-reducing interactions with the terrain are more pronounced with upslope flow. Elevation is related to the reduction of frictional effects with altitude, that is, closer to the mountain peaks. Lastly, the time of year and the SWE level affect irregularities posed by post-event compaction and radiation-induced melting and sublimation.

All the aforementioned studies have employed a multiple linear regression approach to the diagnosis or prediction of SLR. An alternative approach to SLR prediction, which will be explored in this work, is with the use of machine learning. Machine learning has gained rapid recognition as a computational tool to solve a variety of problems in the field of science, in part because of its ability to account for nonlinearity in statistical relationships. There are two guiding hypotheses of this work. First, machine learning is likely to improve forecast SLR in complex terrain, as compared to a regression-based approach.   If so, then a machine learning-based algorithm can thus be applied to any source of gridded numerical weather prediction model output for improved SLR prediction. Second, machine learning can yield insight into the physical processes that drive SLR, as advanced ML methods permit quantification of the relative importance of individual predictor variables.

The most relevant example of a study in the vein of machine learning applied to SLR prediction, though somewhat now dated, is Roebber et al. (2003) which used classification with a neural network rather than regression. Temperature, humidity, wind speed, and time of year were used as inputs to the classification, and it improved upon the 10:1 ratio used by NWS at the time and the use of SLR climatology. However, this study avoided the complex terrain of the Western U.S., limiting its methodological utility for operational forecasting in the Rocky Mountain region. This study investigates the use of machine learning to predict SLR in much more complex terrain, as a proof of concept in the Sky Islands of southern Arizona. Multiple machine learning techniques are evaluated, each one regression-based rather than classification-based, with results from each one compared against a multiple linear regression. The region of southern Arizona was chosen due to the low climatological SLRs in the Sky Island ranges, a known problem of operational overprediction of SLR, and because of the need for more accurate forecast characterization of snowfall to inform longer-term hydroclimate forecasting.

2. Data and Methodology

To improve snowfall predictions in the Sky Island region of SE Arizona, multiple machine learning methods were tested, utilizing data from a 10-year winter period, from 2008-09 to 2018-2019, months of November to March. Variables selected for ML input are based on recommendations in Alcott and Steenburgh (2010), as their study location in the Wasatch Mountains of Utah is most similar in environment and climatology to Arizona's Sky Islands. Variables included here are temperature at 500-850 mb, wind speed at 500-700 mb, wind direction and humidity at 850-700 mb, and the month of the year. Fixed-point surface temperatures from lower elevations were also input into the network, based on the 2013 Van Cleave NWS Sacramento Method.

2a. Surface and Vertical Meteorological Data

Input variables were taken from multiple sources. For both the surface (temperature, wind) and multi-layered parameters (upper-level temperature, wind, humidity), three different reanalysis datasets were used: The North American Regional Reanalysis (NARR), the ECMWF-produced ERA-5 reanalysis, and the University of Utah's WRF-HRRR archive. A 10 Year climatology was analyzed using NARR and ECMWF ERA-5, with additional extreme events selected from WRF-HRRR from the 2016-2017 winter onward. Multiple reanalyses were used to prove the flexibility and adaptability of each machine learning method across the varied sources of gridded model data.

1) NARR Reanalysis

The North American Regional Reanalysis (NARR), produced by the National Center for Environmental Prediction (NCEP), is a 32 km reanalysis, with data available from January 1979 to the present day. Temporal resolution is daily, with data being taken between 12z of the previous day and 12z of the current day. Specific humidity is considered rather than relative humidity because of its availability in the NARR data. Below 700 mb, data was taken every 25 mb, while it was taken every 50 mb above 700 mb. Data is obtained from a variety of sources, including radiosondes, dropsondes, aircraft, surface observations, and satellites, assimilated onto the grid and averaged temporally (Mesinger et al 2006).

2) ECMWF ERA5 Reanalysis

The ERA5 reanalysis is similar to the NARR in terms of how the data was organized and collected, although it has slightly higher horizontal resolution (28 km) and slightly poorer vertical resolution in the 700-750 mb layer (no 725 mb layer). It also has a much higher temporal resolution, with data being available hourly (Hersbach et al 2020). Unlike the NARR, relative humidity is available, but specific humidity was kept to maintain consistency for a direct comparison between the datasets. This dataset is global in scale and produced by the ECMWF, though for this study, the data were zoomed in to focus solely on the region of interest.

3) HIGH-RESOLUTION WRF-HRRR REANALYSIS

The University of Utah WRF-HRRR configuration was first described and used to measure air quality in the Salt Lake City Valley (Blaylock et al 2017). This configuration and its archive have since become available globally, and it is much higher in both horizontal and 700-500 mb vertical resolution than the previous two reanalysis datasets. However, it is also much more limited in its temporal range, with data unavailable prior to the winter of 2016-2017. Due to the limited data range and the increased computational expense, our consideration of the WRF-HRRR data in this study focused on a specific series of heavy snowfall events in the period of interest, with the threshold for the events described later in the paper. Eight suitable events were found between 2016 and 2019.

2b. Target Snow Cover Data

Snow data over the period of interest was obtained from a gridded dataset produced at the University of Arizona as described by Zeng et al (2018). This dataset contains both snow depth and SWE at a 4km resolution from 1981 to present, with a daily temporal resolution. The dataset was created by combining physical observations such as NWS COOP and

SNOTEL sites with gridded precipitation data from PRISM, coming from a combination of radar data and climate aided interpolation (CAI). SLR, as the amount of solid snow per liquid equivalent, could be calculated directly from the parameters in this dataset, by dividing the day-to-day difference in daily snow depth by the corresponding difference in SWE. SWE was directly taken as an additional input parameter, and one final parameter, elevation, was taken from the WRF Geogrid at 3 km resolution, from the author's personal configuration, centered on the Sky Islands of Southeast Arizona. The WRF used was version 3.8, an improvement on the version 3 described in Skamarock et al (2008). A more detailed view of the input parameters, their resolution, and data source can be found in table 1.

2c. Methodology

1) Spatial Domain

The focus for the initial tests of the network focused on the Sky Islands of southeast Arizona and southwest New Mexico. Boundaries for the region extended N-S from the US-Mexican border to the Mogollon Rim (~31-34 deg N) and E-W from west of the Tucson Metro Area to Silver City, NM (~112-108 deg W). The domain boundaries approximately follow the boundaries of the Tucson Local Forecast Office area of operations on the northern and eastern sides of the city, where the greater concentration of mountains means more winter snowfall events are likely, although a slight extension into western New Mexico exists due to the presence of additional sky islands. Figure 1 shows the approximate area of the domain, with terrain features visible.

2) Data Pre-Processing

Since the UA-SNOW data is only available with a daily 12z-12z temporal resolution, both the ERA-5 and WRF-HRRR temporal scales are adjusted. For both datasets, the hourly files are averaged for each day of the 10-year period, from 12z of the previous day to 12z of the day recorded by the dataset. No temporal adjustment was necessary for the NARR, as its temporal scaling matched that of the target dataset.

Due to the differing resolutions and projections of the various input data sources, spatial downscaling onto the snow data was also required, with additional scaling being necessary for the terrain data. Each grid point in the Zeng et al. dataset is bound to the closest reanalysis point, closeness being determined by the difference in latitude and longitude between the points. This scheme assumes that the larger NARR and ERA grid boxes are fully representative of each of the smaller boxes within, and vice-versa for the WRF-HRRR grid boxes, higher in resolution than the Zeng et al. grid. This assumption is likely adequate for the multi-layered inputs, which are mostly influenced by the synoptic weather pattern. However, it will not work for surface temperature because of its large dependence on elevation, and rapid change in space within mountainous ranges. Thus, surface temperatures are taken from the reanalysis at a fixed point of lower elevation, the point being taken depending on the mountain range. These points were located at the approximate locations of Tucson, Safford, and Benson, with care being taken to take points of relatively similar altitudes. These fixed points are also in line with the study of Van Cleave (2013) which used similar fixed-point valley-level temperatures to forecast SLR in the Sierra Nevadas.

Additionally, a filter was applied to eliminate marginal events, as they are highly noisy due to error sensitivity for both snow depth and SWE with smaller values. With this filter, points can only be placed into the input array if the daily snow depth increase is greater than 5 cm (~2 in). It was empirically determined that this threshold provided the best balance between network skill and having a large, robust data set for the analysis. About 20000 data points remained usable once this filter was applied for both the NARR and ERA-5 datasets.

This threshold is easily exceeded for high-elevation snow events of any significance in the Sky Islands.

To ensure robustness, only 75% of the spatiotemporal points are used to train the linear regression and the machine learning methods, with the remaining 25% withheld for testing. This is in essence a manual cross-validation of the data, containing 4 separate “folds” of data. Within this k-folding method, data is partitioned into k folds, with k-1 folds being used to train the network and the last fold being reserved specifically for validation. Data division is random, and training data does not factor into the accuracy tests, only the initially withheld testing data. Additionally, 25 runs are performed on each configuration, with statistics being taken from the mean and standard deviation of the configuration. This allows for a more representative and robust view of skill through slight variations caused by the randomized data withholding. It also allows for testing of statistical significance resulting from an adjustment of the input parameters and machine learning methods, allowing for objective comparisons to be made between each method and each input configuration.

3) Control Method - MLR

To show that the Neural Network produced a statistically significant improvement over previous methods, a multiple linear regression (MLR) model needs to be developed as a control method, using the same parameters and data division as the neural network. The regression equation would have 27 separate coefficients with the NARR data on the most optimal configuration, with the number of coefficients varying slightly for the ERA-5 and WRF-HRRR data, due to differences in vertical resolution. The equation, developed using the training data, is solved for each set of inputs on the testing data.

4) Feed Forward Neural Network

The first machine learning tests utilize a standard feed-forward neural network in the form of a multilayer perceptron, directly comparing its outputs to those of the MLR. The network structure consists of an input layer, an output layer, and several hidden layers, defined by individual weights quantifying the relationships between individual inputs and the output data, as well as their relative importance to the overall relationship. Optimal weights are determined via training of the network starting from a first guess, using a least-squares error minimization method, the training continuing until an optimal configuration of weights is found. This is accomplished via back-propagation, the comparison of network-predicted values with the true values in the output layer, in this case the Zeng et al. data. In this way, the perceptron is a trial-and-error process, albeit one that is highly dependent on pattern recognition.

For this network, the input layer consisted of 27 individual nodes, each node representing an atmospheric variable at a single level. There was only one output, the targeted SLR data calculated from the Zeng et al. dataset. Using the number of layers, the optimal number of nodes within the hidden layers can be specified. There are several possible rules of thumb for computing this number, the one used here was one plus two-thirds of the number of input layers plus output layers. For the NARR data, this gave a total of 19 major nodes within the hidden layer, as only one was used to maintain consistent performance between runs.

5) Verification of Performance

To evaluate the performance of each machine learning method, the results they produced need to be objectively compared with those of the MLR. For each method, the predicted SLRs are plotted against the SLRs calculated from the Zeng et al. dataset, assumed to be the true values for each spatiotemporal event point. To account for the known errors of the verification data, the accuracy of both methods is determined by the percentage of hits, a hit being defined as the network/regression-output SLR being within 20% of the SLR value calculated from the Zeng et al. dataset. 20% is the defined margin of error for this dataset, calculated from given error characteristics in SWE and snow depth. In this manner, accuracy is determined for each individual run, and the mean and the standard deviation of the series of runs is calculated for each method. Significance is determined by a two-tailed t-test at a 95% confidence interval.

6) Parameter Optimization

Although the regression created by Alcott and Steenburgh for the Wasatch mountains is a good starting point for determining parameters of importance, differing regional climatologies between northern Utah and southeastern Arizona may necessitate changes in the predictor variables used for an optimal network configuration. To determine the individual relationships between variables and their relative importance for SLR in the Southwest, scatter matrices are created. This shows the relationships between individual inputs and the target SLR data, allowing for variables with little to no correlation to be sidelined, provided that their removal does not significantly lower the performance of the network and/or the regression. Significance is determined by a two-tailed t-test at the 95% confidence level, using the mean and standard deviation of the R^2 values and the % correctness for both the neural network and the linear regression. Additional new input parameters can be tested in a similar way, the variables being retained if they produced a significant improvement in overall skill.

7) Support Vector Machines

After the finalization of input parameters, alternative machine learning methods such as Support Vector Machines (SVM) and K-Nearest Neighbor algorithms (KNN) are tested using the same manual 4-fold cross-validation scheme. SVM methods, although initially designed for linear use, are effective at nonlinear classification through use of kernels, which implicitly map inputs onto multidimensional feature spaces. For this problem, a radial basis function, or gaussian, kernel is found to be the most suitable. With the gaussian kernel, the mapping of inputs is used to calculate the distance between them in this theoretical future space, and the output of this function forms a matrix of support vectors, which can then look for patterns among the input variables. The space here utilizes a high number (>25) of dimensions, equivalent to the number of input layers in the base neural network, although this is inconsequential as SVM can theoretically support up to an infinite number of dimensions in its theoretical space.

Even with the kernel function, computing the optimized output vectors requires a solver algorithm, as SVM optimization is a quadratic programming problem. For this problem, sequential minimal optimization (SMO) is used, an iterative algorithm which uses and optimizes two separate Lagrange multipliers and repeats for each possible pair until training is finished, upon which the model is tested for convergence, meaning an optimal parameter fit is found through a minimization of error. The input data is standardized according to the weighted mean and standard deviation prior to training the SVM to ensure convergence, as unstandardized data may fail to converge if different inputs have different base orders of magnitude, as is the case here. This standardization also allows for better coordination with the gaussian kernel function, as the mean and the standard deviation are the core parameters of a gaussian distribution.

SVM has been used successfully in a meteorological context by several previous studies. Mellit et al (2013) tested multiple machine learning methods on short-term prediction of surface-level meteorological parameters. Out of the four methods tested, the SVM performed the best in both correlation and in accuracy for the prediction of variables such as temperature, wind, and humidity, all of which are theorized to have an influence on SLR according to Alcott and Steenburgh (2010). Although the Mellit et al study used a linear kernel function rather than a radial basis function, the fact that SVM outperformed other machine learning methods using predictors similar to the variables of interest for this study makes it an appealing method to test on the problem of SLR forecasting.

8) K-Nearest Neighbor Algorithm

The KNN algorithm, similarly to the SVM, places vectors for training onto a multidimensional feature space and can use up to a practically infinite number of dimensions, though only 27 are needed for this study. Feature extraction through a principal component analysis, done by the model as a pre-processing step, is a common way of making the size of the KNN model more manageable. After the data is ready for use, instead of using a kernel function based on the distance between vectors, the KNN looks at a single output vector and looks for k vectors which are the most like it in terms of value.

For a regression-based KNN, the predicted output is the weighted average of the outputs of the nearest neighbors, and instead of k vectors being assigned, vectors are weighted in accordance with the distance from the selected multidimensional input vector, with vectors closer to the input being weighted much more highly, similar to the methods used by Huang et al (2017), only for a regression-based, rather than a classification-based, model. This study aimed to improve precipitation forecasts through pattern recognition, and through the use of a distance-based weighting method for nearest neighbors, managed to overcome previous obstacles for KNN use caused by uneven base distributions of precipitation events. This improved the KNN algorithm to a skill level comparable to an SVM, used in the study as a control. The KNN is much more sensitive to the local data structure as the function is built in a point-by-point manner and only includes the most highly-weighted neighbors. Results will be more accurate if the network structure is assumed correctly. Although not strictly necessary, the data is once again standardized to ensure consistency with the SVM.

3. Results

3a. Parameter Optimization

The first tests of the network, with the configuration described in the methodology, showed highly accurate results, with a 91% accuracy rate and an R^2 of 0.67 on average for the network function. This was an improvement over the regression model, which had an accuracy rate of 82% and an R^2 of 0.39. Specific humidity had the best individual correlation, being particularly important at lower levels, followed by temperature, also from lower in the atmosphere and near the surface. While temperature appeared completely linear, the humidity correlations appeared to show some nonlinearity even though the linear correlations were stronger than those of the temperatures. For the wind speed and wind direction, correlations were much lower than for the temperature and humidity correlations, casting doubt on their importance to the SLR problem.

The network was re-tested without these variables, and wind speed was found to significantly affect network skill while no such effect was found for wind direction. Next, additional input variables were added and tested, of which only elevation and SWE were found to significantly improve the skill of the network. Additionally, temperature elevation was lowered to focus on the 850-500 mb range, with the surface temperatures being taken from about 900 mb, about the lowest elevation available in southeastern Arizona, and this change also improved overall skill. Thus, the finalized configuration consisted of 850-500 mb temperature, fixed-point low elevation surface temperature, 700-500 mb wind speed, 850-

700 mb specific humidity, the month of the snow event, the event SWE, and the altitude. Inputs of the multi-layer parameters were taken every 25-50 mb, resulting in a total of 27 inputs for the NARR data.

3b. Final MLR and Network Performance

After all these changes, network accuracy improved to 95% while the R^2 of its function improved to 0.79. For the regression, accuracy improved to 84% while R^2 improved to 0.44. Standard deviations were low for all accuracy parameters (<0.01 for R^2, 0.5-1% for hit rate), therefore only a small change in skill was needed to prove significance at the 95% level, and this was achieved by the optimized configuration. Figures 2 and 3, along with table 2, give a visual depiction of the results, showing the performance of both the linear regression and the neural network, along with notable selected input parameters, which scatter matrices are used to compare. These scatter matrices can be seen in figures 4 and 5.

The variables shown in figures 4 and 5, surface temperature, upper-level temperature, and specific humidity, have the highest individual correlations and therefore the most influence on SLR. Multi-layer variables were vertically averaged for conciseness in the scatter matrices, but, out of the multi-layer variables, temperature and specific humidity showed the highest correlations in the 750-800 mb range, well above the desert floor, but lower than the crests of the higher ranges. Colder temperatures and drier overall conditions favored higher SLR values, while warmer, wetter conditions tended to produce lower SLR values. This is consistent with the findings of previous studies, indicating that the parameters are being handled well by the Machine Learning methods.

3c. SVM and KNN Tests

After the optimal configuration of input parameters and the overall importance of each input variable are found, additional machine learning methods are tested on the same configuration. Two of these methods, support vector machines (SVM) and k-nearest neighbor algorithms (KNN) were done using the same procedure as the base neural network, with multiple trials and a random 75/25 data split, using each of the described input data sources. The third method, a random forest, came from a single run of the data using NARR inputs. The statistics of these methods, in addition to how they compare with the previous methods, are shown in table 3. Plots of the new methods can be seen in figure 6 for the SVM, figure 7 for the KNN, and figure 8 for the random forest, and an objective comparison between each of the methods can be seen in table 3. All these methods significantly improved on the basic neural network, and the random forest, despite its single trial, was significantly better than both SVM and KNN. Due to the random forest data being a single run, a z-test was used to determine statistical significance as opposed to the two-tailed t-test still used for the other methods. The random forest improved raw accuracy to over 98% with its singular run, while the ensemble of SVM/KNN produced an accuracy lower by 0.5-1%. The relatively low run-to-run variation (0.2-0.5% hit rate standard deviation) with these new methods caused even this small change to be statistically significant at the 95% confidence level.

3d. Data and Machine Learning Spatial Statistics

To further gauge performance of each method against the assumed true SLR values, a spatial analysis is performed, with the average SLR for each grid point over the period of study being calculated, for both the Zeng et al. dataset and each of the predictive algorithms. For each of the predictive methods, a spatial map of MAE relative to the Broxton SLR is also plotted, while a map of the number of significant events for each gridpoint, needed to compute MAE spatially, is plotted alongside the Zeng et al SLR map. These maps can be seen in figure 9 for the Zeng et al. data, figure 10 for the MLR, figure 11 for the neural network, figure 12 for the SVM, and figure 13 for the KNN. An additional plot is also shown in figure 14, showing the percent reduction in MAE for each of the three machine learning methods relative to the MLR. Figure 15 shows the elevation across the region of study, included for reference with the spatial error maps.

Some general trends are evident in these maps. First, the maps confirm generally better performance across the board of each of the three Machine Learning methods compared to the MLR, with lower MAE throughout the study area. Performance was observed to be slightly better with SVM and KNN, and both of their MAEs are generally lower than that of the feed-forward network. Performance is very good in the high mountain regions, but there are areas of poorer performance in more marginal lower-elevation regions. This reduced performance is most visible in the less-adaptable MLR, and is likely the result of fewer events and differing conditions compared to higher elevation regions. This provides further rationale for including elevation as a predictor variable, which serves to compensate somewhat in the higher-skill machine learning methods. Error value reductions for machine learning are inconsistent between points, particularly for the neural network and the SVM; this may be the result of the limited number of data points, particularly at lower elevations. The KNN does not have the same problem, with nearly complete elimination of error at lower-elevation points, error in the higher peaks remains notable, although reduced compared to the MLR, and similar to the other machine learning methods. Overall, mean SLR values are lower than the climatology of Baxter et al (2005), this may be a result of warming temperatures in the region, as warmer temperatures will produce lower SLRs, as shown in figure 4.

3e. Alternate Input Data: ERA5 and WRF-HRRR

Testing all these methods with the ERA-5 data produced slightly lower overall skill for each of the methods, but this change was not statistically significant at the 95% confidence level. On the other hand, testing the WRF-HRRR data on a small sample from major post-2016 events resulted in a significant increase in skill for all the methods; the hit rates of both additional datasets can be seen in table 3 alongside the NARR results, for each method the data was used to test. Skill increase was particularly large with the MLR, which had been far below the level of the ML methods for the prior input configurations, although its skill remained slightly below that of the other methods. The Neural Network also showed a notably high improvement, with skill improving to a slightly higher level than both the SVM and KNN networks. The higher horizontal and vertical resolution for the input variables played a major role in this improvement, as did the removal of more marginal events, which tended to produce outliers, particularly when mixing of rain was involved. Both extreme high and extreme low SLR events, present but rare on the full 10-year event time series, are absent in the reduced WRF-HRRR event series. The MLR struggled with such events due to the constraints on range imposed by its governing equation.

The WRF-HRRR improvement should be viewed cautiously, considering the relatively few input points (<5000) and the increased uniformity of the conditions between events. Still, given the accuracy of the networks built off lower resolution data that considered these additional factors and of the limited-information high resolution network, it makes sense that a high-resolution network could be trained to perform very well across a wider region, given a longer period of data availability. For both of the alternative input datasets, correlations between the variables and the relative importance of the inputs remained similar, with fixed-point surface temperature and upper-level temperature and specific humidity having the most influence, wind speed and elevation having the least (while still significantly affecting network skill) and elevation and SWE being in the middle. The small differences between them which did exist were not statistically significant.

4. Discussion

Based on the high overall accuracy of the models, machine learning is a viable strategy for improving SLR and snowfall prediction in mountainous regions, as each method considered improves on the MLR model. Performance is also consistent across multiple sources of input data, a confirmation that machine learning is capable of adapting to a wide variety of data sources. The difference in relative performance between the WRF-HRRR and the other two model sets is noteworthy. For both the NARR and the ERA-5, the SVM was the most accurate method, by a small but significant margin over the KNN algorithm. Both methods had an advantage of several percentage points over the feed-forward neural network, while the MLR was lower than each machine learning method in terms of correlation and accuracy.

For the WRF-HRRR data, the control MLR, although still lower in accuracy compared to the machine learning methods, produced much improved results. The machine learning methods also no longer showed statistically significant performance differences between them, as the feed-forward network's skill improved to a similar level as both the SVM and KNN. One possible reason may be the pre-selection of significant events eliminating outliers which negatively impacted performance in the full time series with lower resolution data. This appears to be the case for the large MLR improvement, as the MLR is limited in range and thus performs poorly for both unusually high and unusually low SLR events. However, the base neural network appears to handle these outliers just fine, with most of the WRF-HRRR improvement coming near the middle of the SLR distribution. This suggests that the feed-forward network reacts better to higher-resolution input data than either the SVM or KNN methods, which appear to handle coarser resolution better. Further tests, performed in multiple regions, will be necessary to confirm this. If this pattern continues, it may provide insight into a relationship between the quality, quantity and resolution of input data and the relative performance of different ML methods, providing information on which ML method may be best for a region depending on its climatological snowfall characteristics.

The correlations of individual parameters also warrant discussion, as there is contrast between the results found in our region of interest compared to other previous studies. Temperature remained a predictor of high importance between the studies. Specific humidity gained a much higher level of influence. Wind, seen as highly relevant by multiple prior studies, was barely significant in the mountains of southeastern Arizona. Local climatology, and its resulting influence on crystal shape and interaction with liquid water, play a major role in this discrepancy. Arizona is both warmer and drier than more northerly subranges of the Rockies examined in previous studies. The warmth is particularly significant, as it will end up resulting in more opportunities for rimming and/or mixing with rain, particularly on the lower and middle slopes of mountain ranges, where conditions tend to be more marginal for snowfall. Rimming and mixing both work to lower SLR, and this is one possible contributing factor to the relatively low climatological SLR values compared to other subranges in the Rockies. It also results in the amount of moisture, represented by specific humidity, being much more important, as the amount of available moisture plays a far greater role in determining the level of rimming at marginal temperatures, due to the reduced ability of droplets to move from a supercooled state to a frozen state. A moister atmosphere will, therefore, result in far greater rimming than a less moist one, the difference being more pronounced in a warmer region than in a colder one, where a higher proportion of droplets will have frozen completely. This also applies to marginal high-level events, where snow is confined to the highest peaks. Pressure gradients also tend to be weaker in the Southwest due to the lack of strong low-pressure systems, resulting in reduced wind speeds and therefore less variation in the wind, making such variations harder to relate to SLR changes largely induced by temperature and moisture differences. Additionally, wind's main influences come from the breaking down of dendrites, which would be less common in Arizona due to the warm temperatures and, more importantly, the lack of moisture, as dendrites require high supersaturation to form.

These points lead into the debate between a more local approach and a more national approach for implementing these ML methods. Future work will aim at widening the scope of each method to focus on a wider region of the Western US, testing performance to determine whether a larger-scale approach is viable considering the climatological differences in both SLR and influencing parameters across the Rocky Mountain region. Input parameters may need to be added or adjusted to account for differing characteristics, as wind speed, though found to be of relatively low importance by each ML method in the Southwest, was indicated as the second-most important variable by Alcott and Steenburgh's (2010) study of the Wasatch Mountains in Northern Utah.

Success of this larger-scale approach depends on these variables, and how well the ML methods handle differences in input parameter weight across the region. Should these differences be handled poorly, more localized networks will be necessary, input parameters and weights being governed on a more regional basis, like with Van Cleave's (2013) MLR, designed specifically for the NWS Sacramento office. Due to their flexible weighting systems depending on the input/output relationships, machine learning techniques would be easily customizable for this purpose.

5. Conclusion

Prediction of SLR in complex terrain remains a major challenge, even with the recent switch to a more dynamic national predicting system. Multiple Machine Learning methods were employed to improve this problem, using atmospheric and climatological variables known or thought to influence SLR as inputs. Of these inputs, temperature, from the desert floor up to 700 mb, and humidity, in the 850-700 mb range, were found to correlate most strongly with SLR, particularly in the intermediate 750-800 mb range, corresponding with the transitional region in between the desert floor and the mountain tops. Variables such as wind speed, the time of year, the elevation, and the SWE were less important but also found to play a significant role in determining SLR. The machine learning methods, when tested in the sky island region of southeastern Arizona and southwestern New Mexico, produced highly accurate results for SLR in mountainous regions, with accuracies of over 95% within the margin of error for the verification dataset. Each method showed a significant increase over a multiple linear regression of the same input variables, with results being relatively consistent across multiple datasets for the input values, barring differences in resolution and the removal of outliers. Further tests are needed over a wider area to prove the viability of such a system in operational forecasting, either as a national algorithm or as a local adjustment to current methods.

Acknowledgments.

This research was supported by the University of Arizona Hydrology and Atmospheric Science department.

TABLES

Variable	Input/Output	Layers	Resolution	Data Source
Surface Temperature	Input	Surface (point-bound)	3-32 km	NARR, ERA-5, WRF-HRRR

Upper-Level Temperature	Input	850-500 mb	3-32 km	NARR, ERA-5, WRF-HRRR
Upper-Level Wind	Input	700-500 mb	3-32 km	NARR, ERA-5, WRF-HRRR
Upper-Level Specific Humidity	Input	850-700 mb	3-32 km	NARR, ERA-5, WRF-HRRR
Month of Year	Input	N/A	N/A	Broxton et al
SWE	Input	Surface	4 km	Broxton et al
Elevation	Input	Surface	3 km	WRF Geogrid
SLR	Output	Surface	4 km	Broxton et al

Table 1: Finalized input variables for each method used, with specifications.

Variable	Surface Temperature	Specific Humidity	850-500 mb Temperature	SLR
Surface Temperature	N/A	0.47	0.48	0.25
Specific Humidity	0.47	N/A	0.34	0.31
850-500 mb temperature	0.48	0.34	N/A	0.16

Table 2: Correlation coefficients for the variables more strongly correlated with each other and with SLR. Weaker correlations not shown, but still significant to network skill.

Method	MLR	Neural Network	SVM	KNN	Random Forest
% Accuracy	NARR: 83.6 ERA5: 83.9 HRRR: 95.1	NARR: 94.6 ERA5: 94.3 HRRR: 98.7	NARR: 97.5 ERA5: 97.5 HRRR: 98.4	NARR: 97.0 ERA5: 97.0 HRRR: 98.2	NARR: 98.3
95% Significant	ERA5: No HRRR: Yes	ERA5: No HRRR: Yes	ERA5: No HRRR: Yes	ERA5: No HRRR: Yes	N/A

Table 3: Final accuracy for each method, for each input dataset with significance.

FIGURES

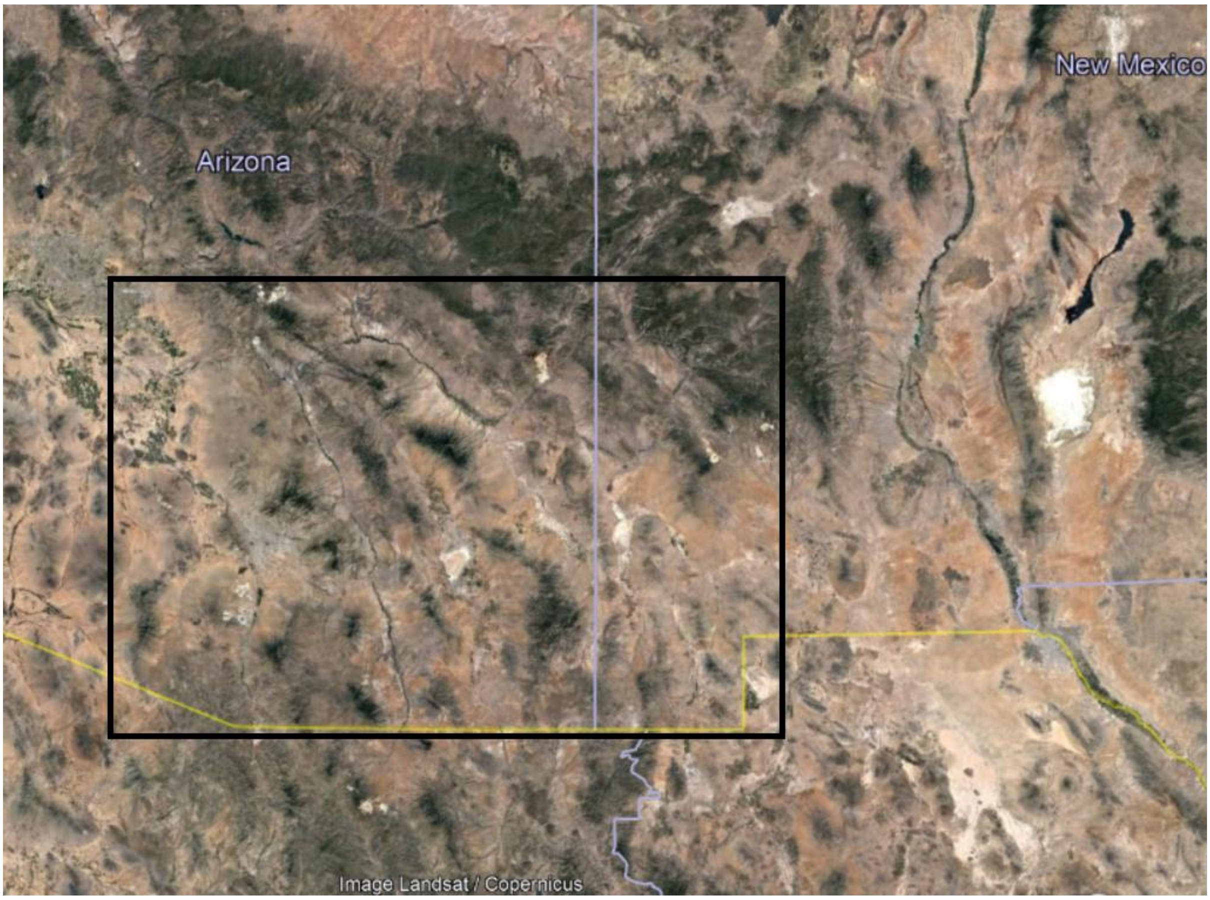

Figure 1: Approximate area of interest for MLR and machine learning simulations, bound by the black line, shown against a wider map of the surrounding area, with terrain features and political boundaries visible. Black box encompasses, approximately, 112-108 degrees west and 31-34 degrees north, southern bound adjusted slightly northward as no data available outside CONUS.

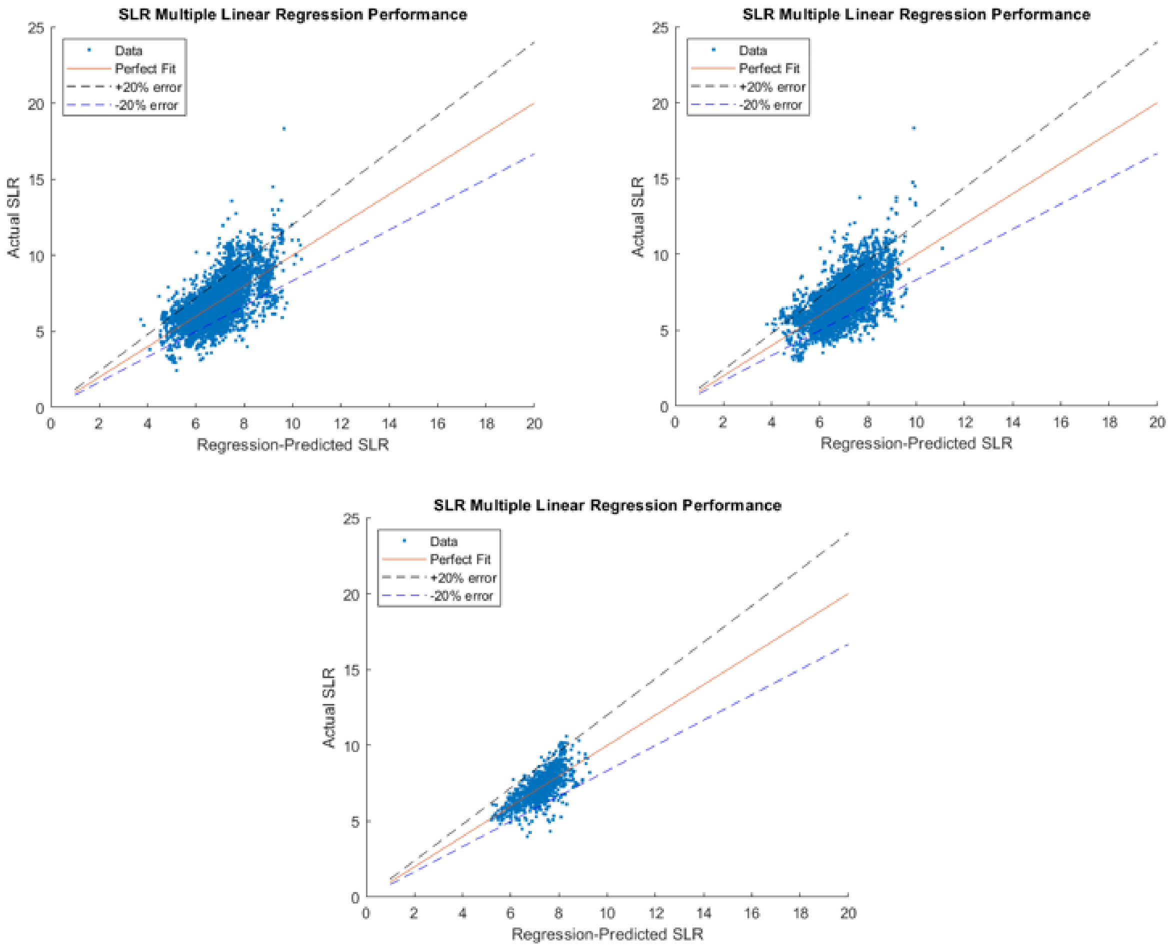

Figure 2: Results of a multiple linear regression method in predicting SLR. Points produced by the regression are compared against a line where predictions equal actual SLR, with error lines used to define the hit rate. NARR input variables are used in the top-left plot, ERA5 in the top right, ad WRF-HRRR on the bottom.

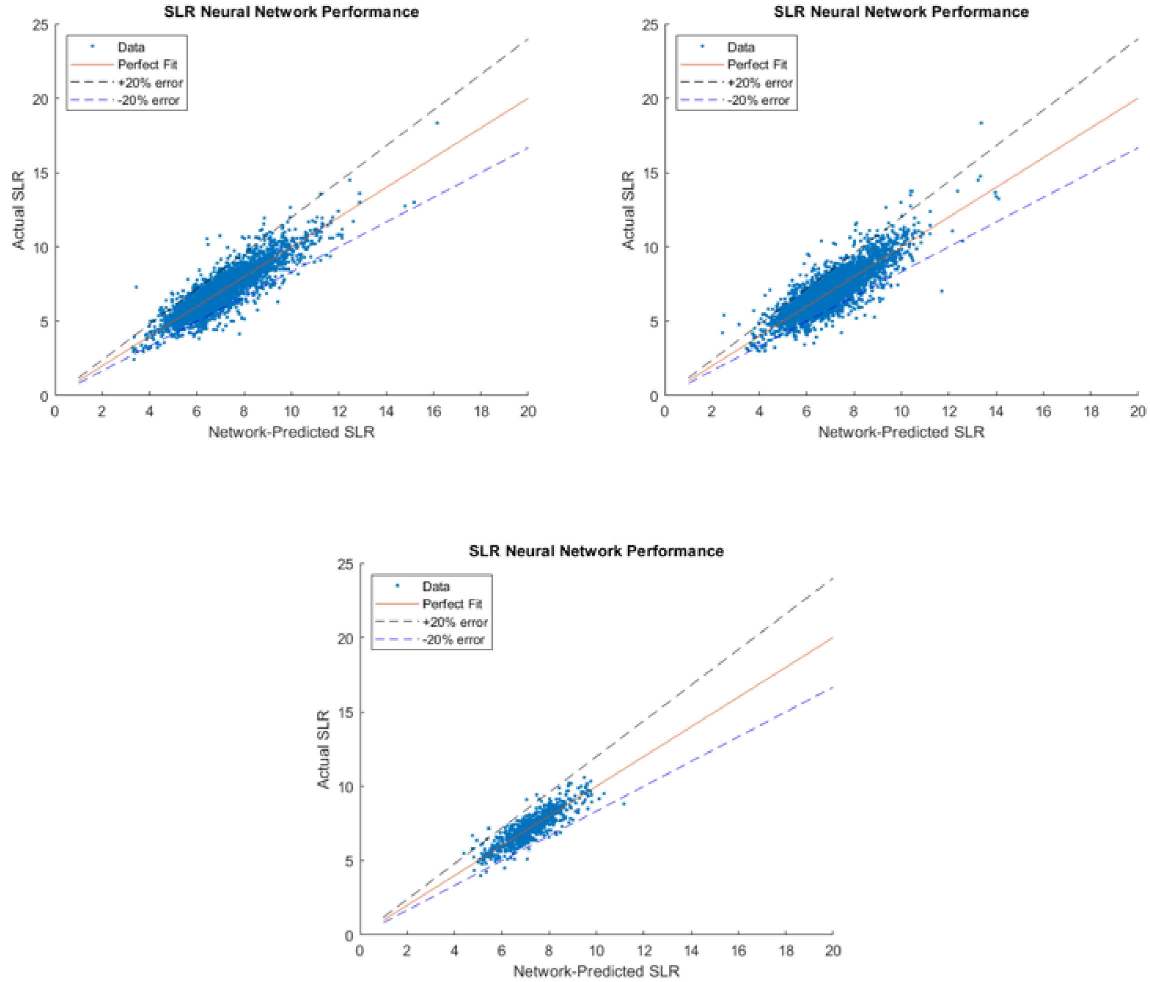

Figure 3: Same as figure 2 but with neural network results.

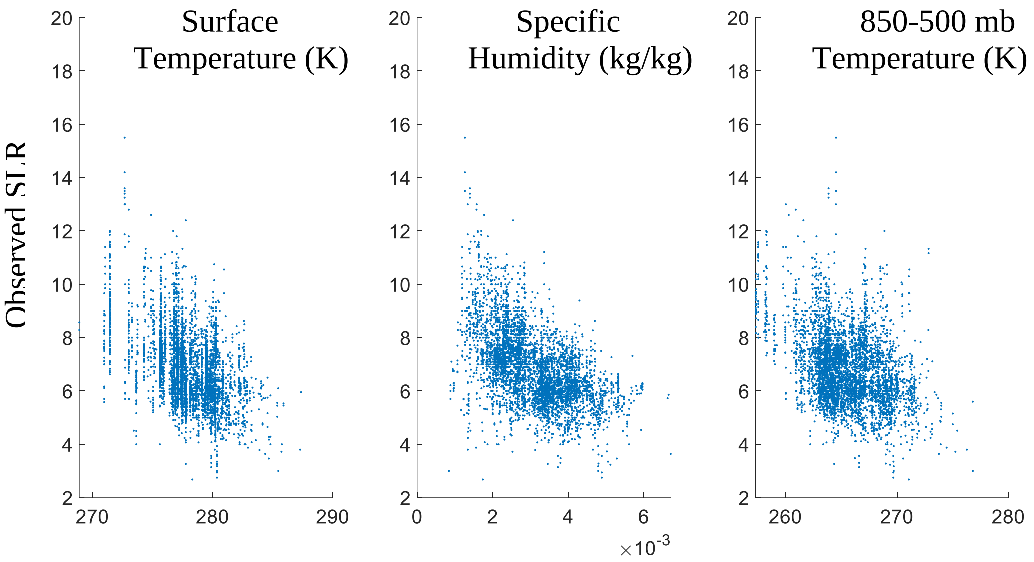

Figure 4: Scatter matrix comparing temperature and humidity parameters with the output SLR data.

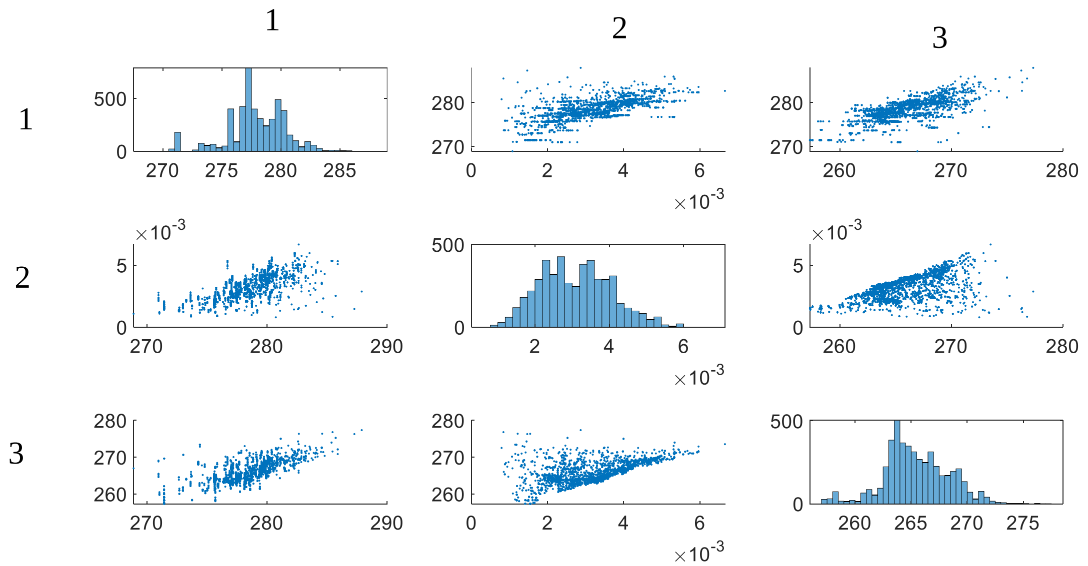

Figure 5: Scatter matrices showing the correlations of temperature and humidity parameters with each other. Row and column 1 are the fixed-point surface temperatures, in K, row and column 2 are the specific humidities in kg/kg, and row and column 3 are the 850-500 mb temperatures, in K.

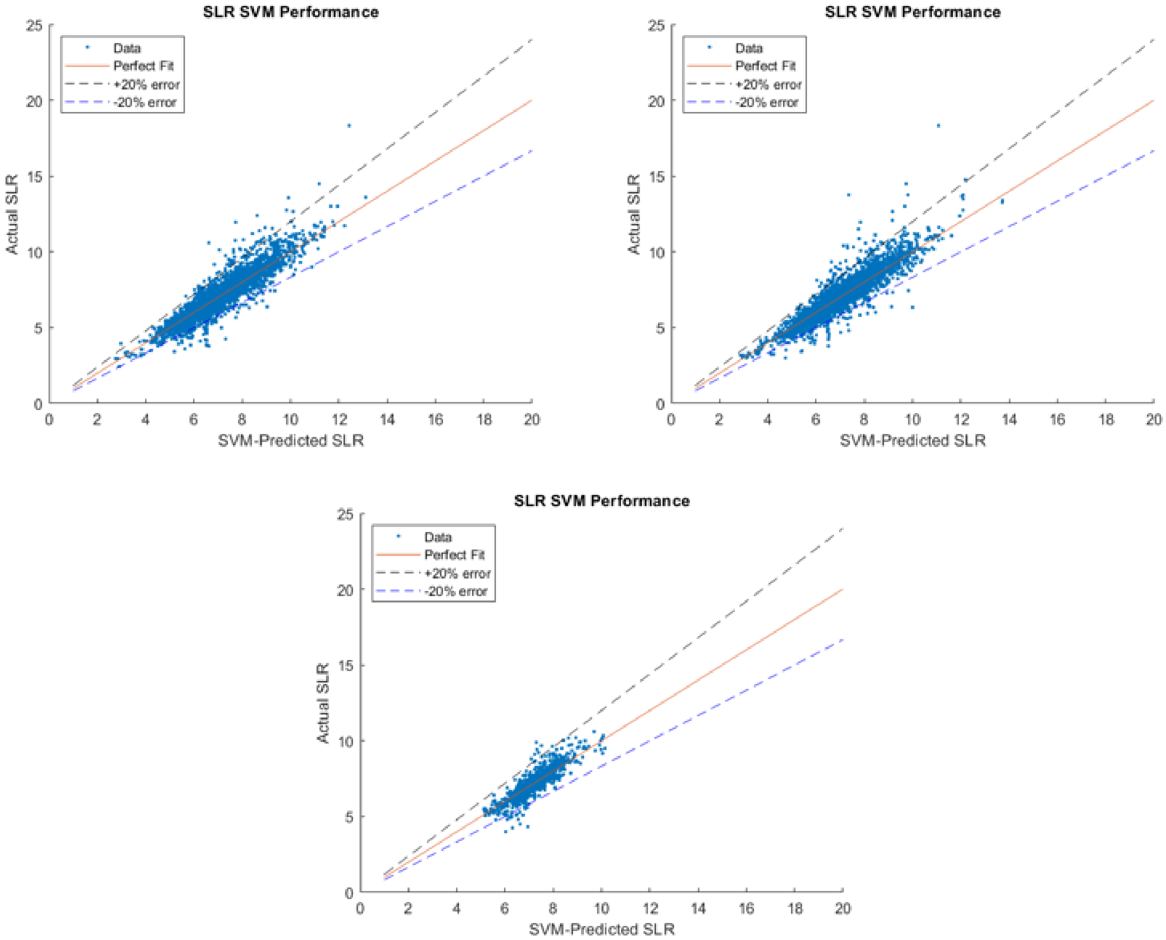

Figure 6: Same methodology as figures 2 and 3, comparing SLR predicted by SVM with actual SLR.

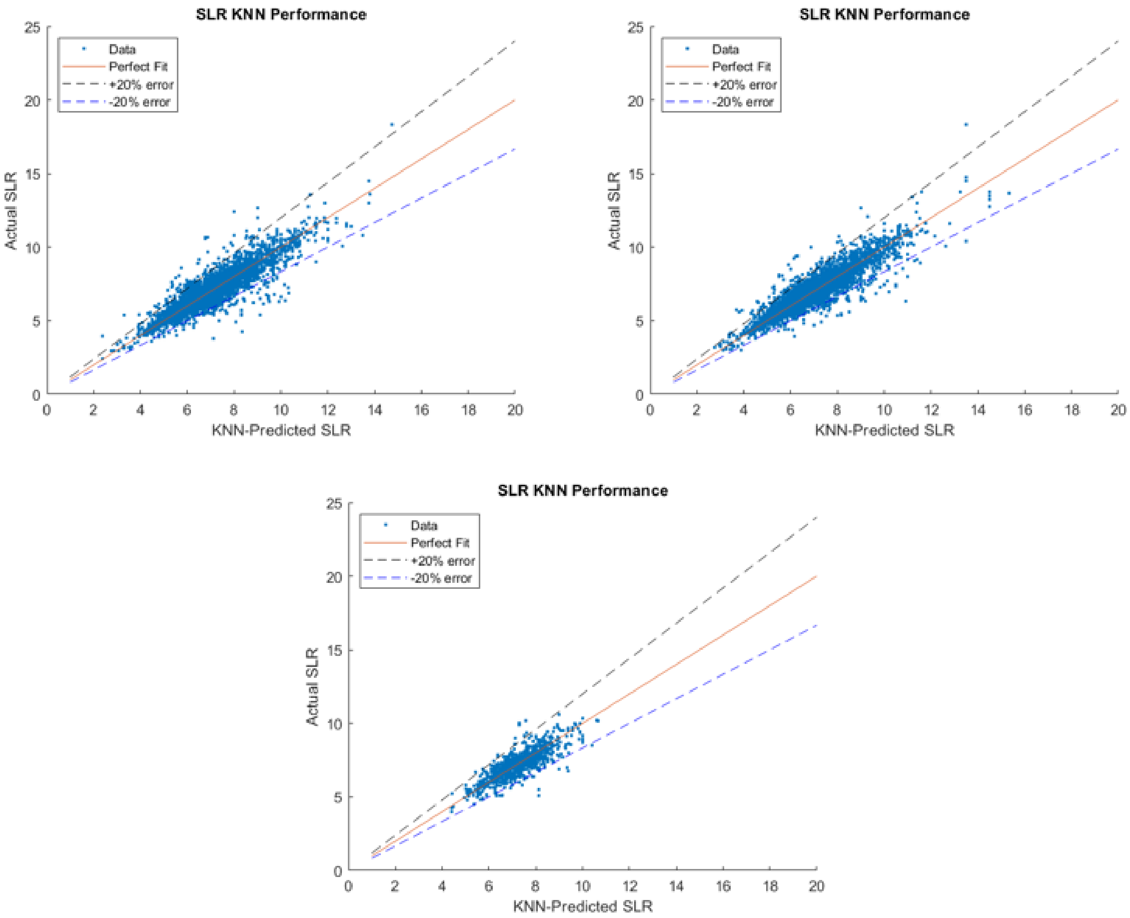

Figure 7: Comparing KNN-predicted SLR with actual SLR, same methodology as previous figures.

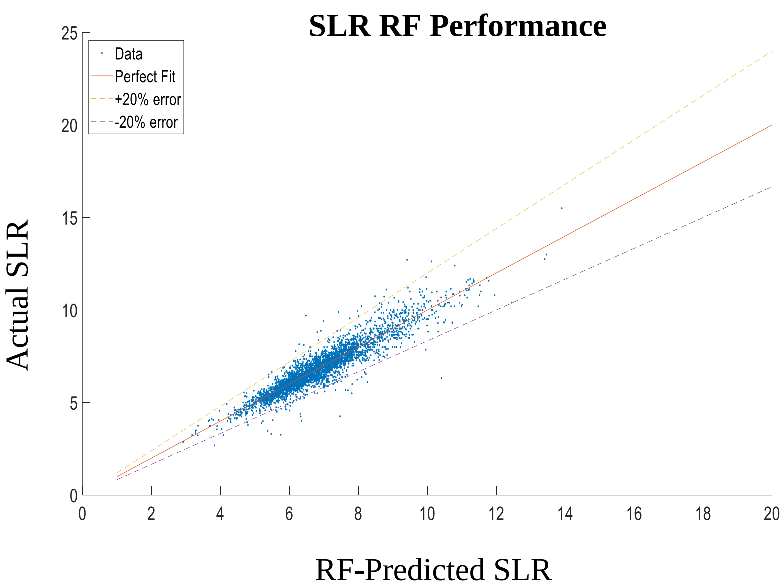

Figure 8: Comparing random forest-predicted SLR with actual SLR, same methodology as before.

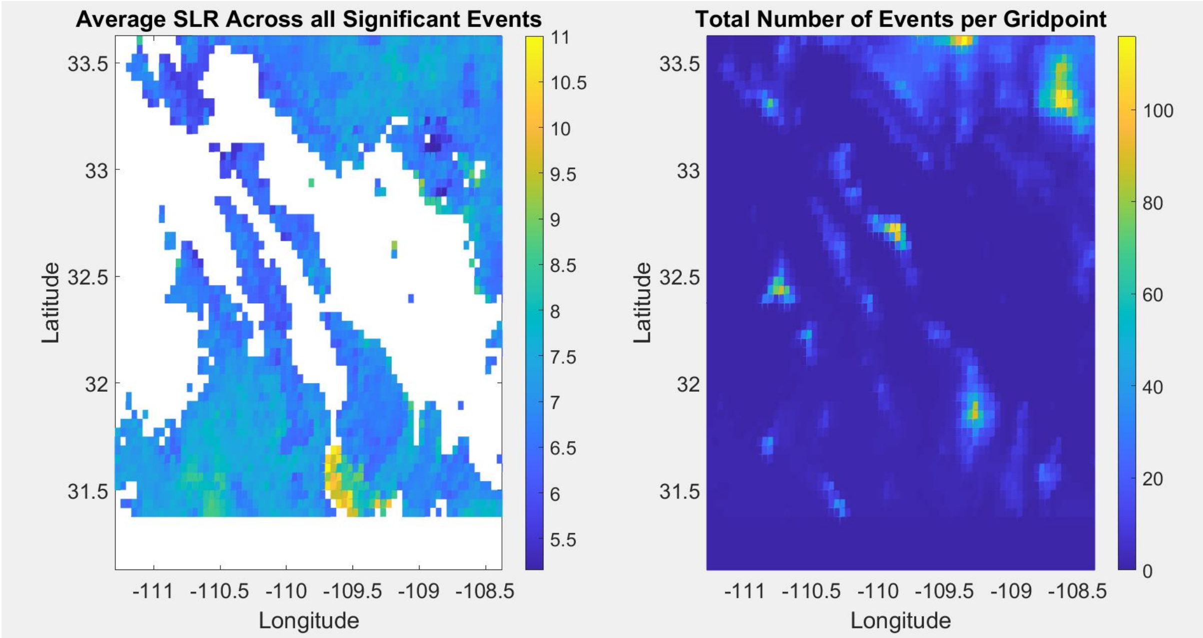

Figure 9: Maps of temporally-averaged Broxton SLR for each gridpoint and the number of events per gridpoint. Gridpoints with zero significant events over the 2008-2019 period show up as blank. Higher elevation points have more events.

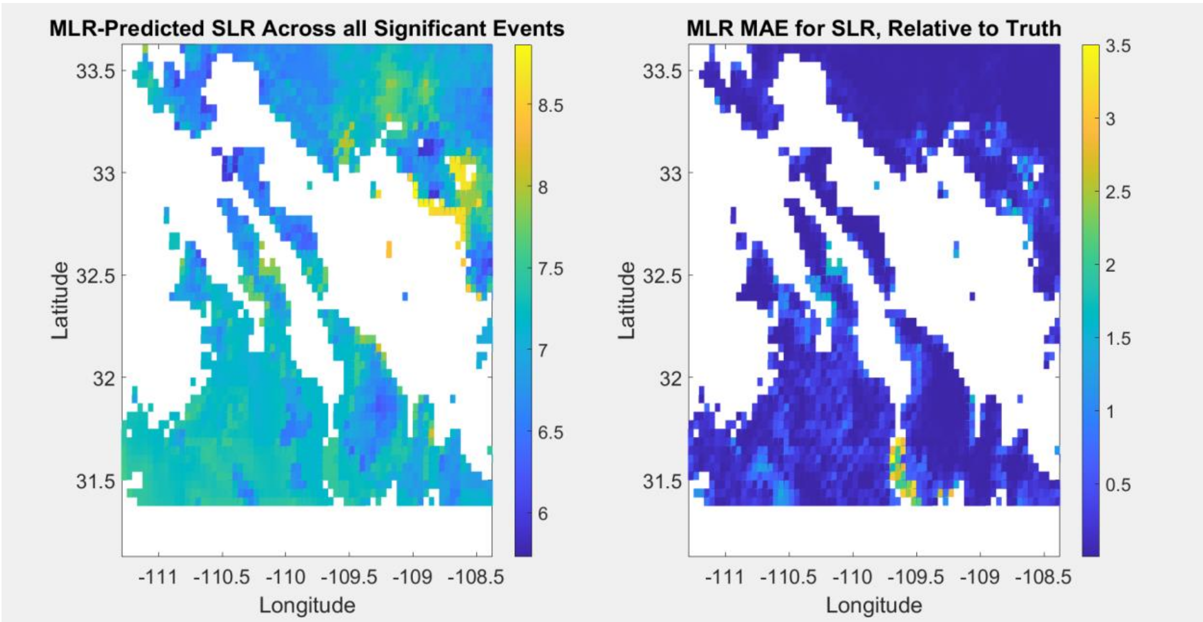

Figure 10: Map of SLR Predictions and MAE for the MLR method, using all possible 2008-2019 spatiotemporal event points.

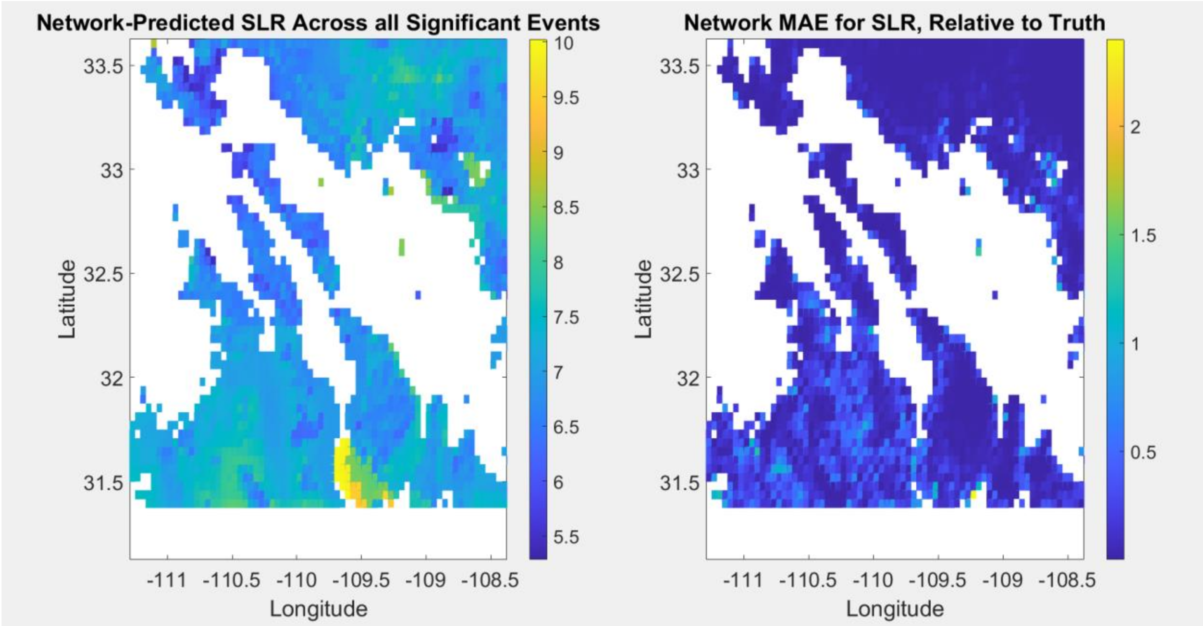

Figure 11: Map of SLR Predictions and MAE for the feed-forward network method, using all possible 2008-2019 spatiotemporal event points.

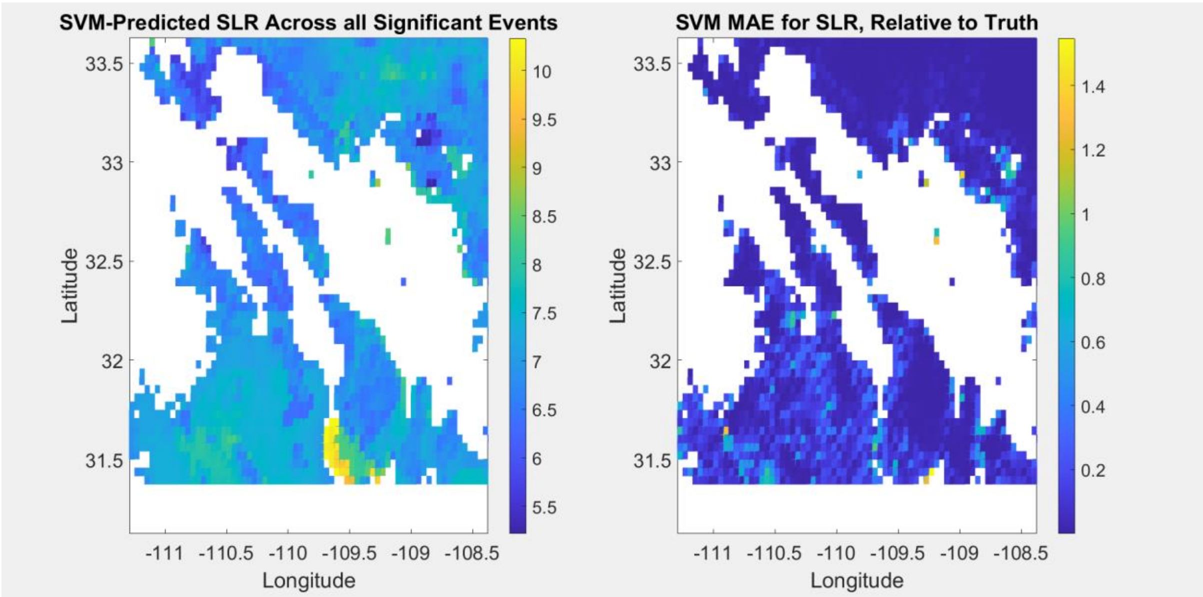

Figure 12: Map of SLR Predictions and MAE for the SVM method, using all possible 2008-2019 spatiotemporal event points.

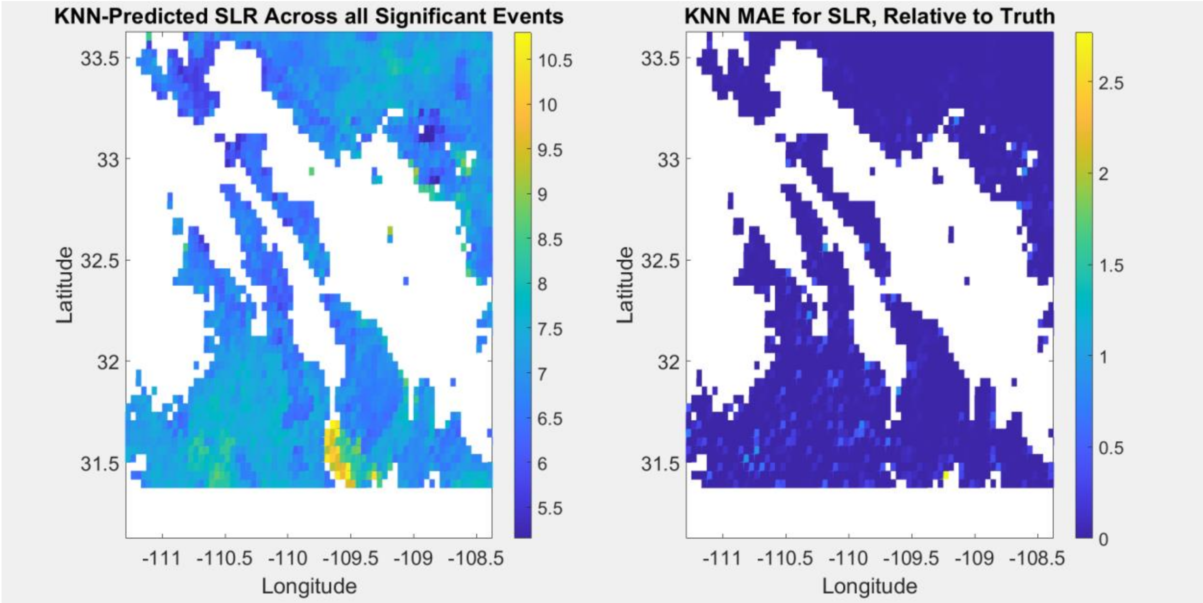

Figure 13: Map of SLR Predictions and MAE for the KNN method, using all possible 2008-2019 spatiotemporal event points.

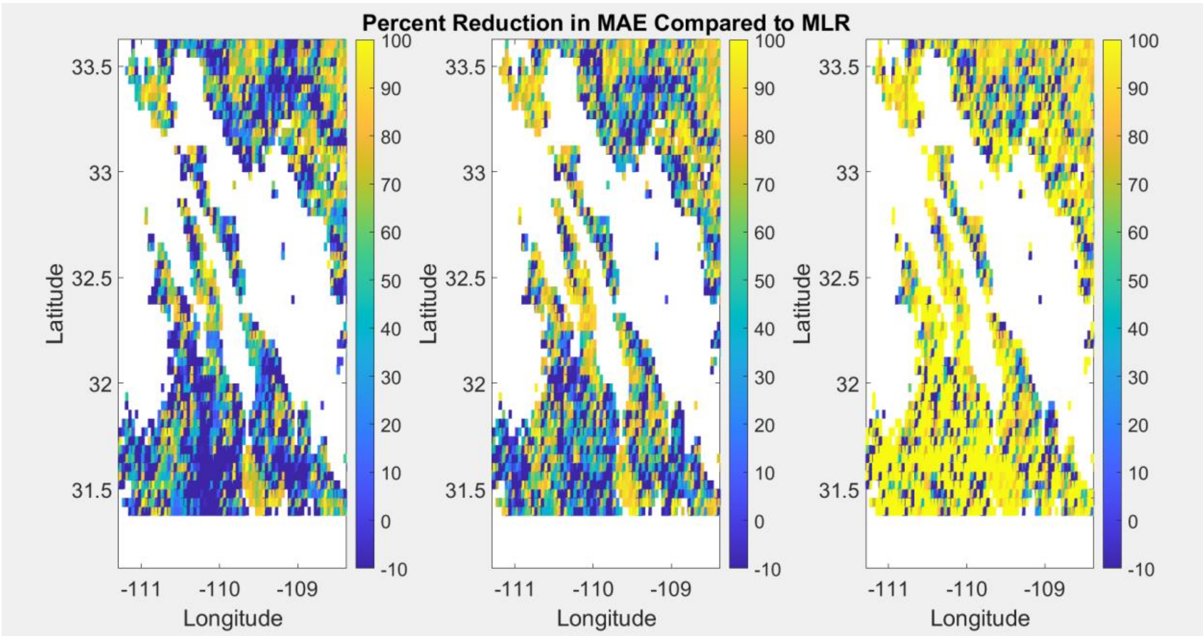

Figure 14: Percent reduction in MAE for each of the three machine learning methods relative to the MLR. From left to right: (Neural Network, SVM, KNN).

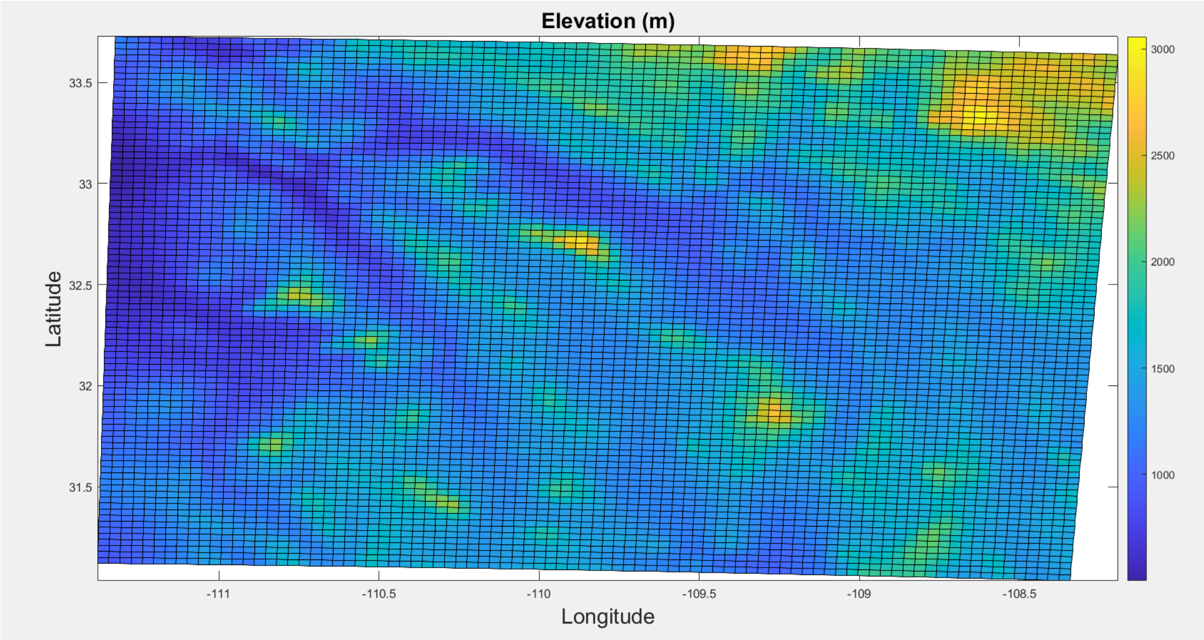

Figure 15: Map of elevation over the domain, included as reference for figures 9-14.

REFERENCES

Broxton, P., X. Zeng, and N. Dawson. 2019. *Daily 4 km Gridded SWE and Snow Depth from Assimilated In-Situ and Modeled Data over the Conterminous US, Version 1*. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5067/0GGPB220EX6A.

Zeng, X., P. Broxton, and N. Dawson. 2018. Snowpack Change From 1982 to 2016 Over Conterminous United States, *Geophysical Research Letters*, **45**, 12940-12947. https://doi.org/10.1029/2018GL079621.

Baxter, M. A., C. E. Graves, and J. T. Moore, 2005: A Climatology of Snow-to-Liquid Ratio for the Contiguous United States. *Wea. Forecasting*, **20**, 729-744, https://journals.ametsoc.org/doi/pdf/10.1175/WAF856.1.

Van Cleave, D., 2013: Snow-to-Liquid Ratios over the Northern Sierra Nevada: Climatology and a Predictive Methodology. *National Weather Service, Sacramento, California*, https://www.weather.gov/media/wrh/online_publications/TAs/TA1303.pdf.

Alcott, T. I., and W. J. Steenburgh, 2010: Snow-to-Liquid Ratio Variability and Prediction at a High-Elevation Site in Utah's Wasatch Mountains. *Wea. Forecasting*, **25**, 323-337, https://doi.org/10.1175/2009WAF2222311.1.

Power, B. A., P. W. Summers, and J. D'Avignon, 1964: Snow Crystal Forms and Rimming Effects as Related to Snowfall Density and General Storm Conditions. *J. Atmos. Sci.*, **21**, 300-305, https://doi.org/10.1175/1520-0469(1964)021<0300:SCFARE>2.0.CO;2.

Roebber, P. J., S. L. Bruening, D. M. Schultz, and J. V. Cortinas, Jr., 2003: Improving Snowfall Forecasting by Diagnosing Snow Density. *Wea. Forecasting*, **18**, 264-287, https://doi.org/10.1175/1520-0434(2003)018<0264:ISFBDS>2.0.CO;2.

Mesinger, F., and Coauthors, 2006: North American Regional Reanalysis. *Bull. Amer. Meteoro. Soc.*, **87**, 343-360, https://doi.org/10.1175/BAMS-87-3-343.

Hersbach, H., and coauthors, 2020: The ERA5 Global Reanalysis. *Quart. J. Roy. Meteor Soc.*, **146**, 1999-2049, https://doi.org/10.1002/qj.3803.

Blaylock, B. K., J. D. Horel, and E. T. Cosman, 2017: Impact of Lake Breezes on Summer Ozone Concentration in the Salt Lake Valley. *J. Appl. Meteor. Climatol.*, **56**, 353-370, https://doi.org/10.1175/JAMC-D-16-0216.1.

Skamarock, W. C., and coauthors, 2008: A Description of the Advanced Research WRF, Version 3. *NCAR Tech. Notes*, 1-125, https://opensky.ucar.edu/islandora/object/technotes%3A500/datastream/PDF/view.

Mellit, A., A. M. Pavan, M. Benghanem, 2013: Least Squares Support Vector Machine for Short-Term Prediction of Meteorological Time Series. *Theoretical and Applied Climatol.*, **111**, 297-307, https://link.springer.com/article/10.1007/s00704-012-0661-7.

Huang, M., R. Lin, S. Huang, and T Xing, 2017: A Novel Approach for Precipitation Forecast via Improved K-Nearest-Neighbor Algorithm. *Advanced Engineering Informatics*, **33**, 89-95, https://www.sciencedirect.com/science/article/pii/S1474034616304621?casa_token=pwP01SucnVgAAAAA:_nsAdseHN2Nnr3ecuSqYNBJXgdT30P8Iks_MsOF8Ku0kEZJc_zszYyIUISXn5gKSFuN0whmLyA.