Model Report for TeamUArizona's Submission to the Snowcast Showdown

Patrick Broxton[1], Xubin Zeng[2], Ali Behrangi[2], Gou Yue Niu[2]
[1]School of Natural Resources and the Environment, University of Arizona, Tucson, AZ, USA
[1]Department of Hydrology and Atmospheric Sciences, University of Arizona, Tucson, AZ, USA

1) Summary

In semiarid regions like the western United States (US), snowpacks are a critical water resource, as many river flows are dominated by snowmelt (Brown et al. 2008, Furniss 2010). Therefore, it is critical to have accurate and robust methods to measure snow water equivalent (SWE) at high spatiotemporal resolution. For decades, point measurements of SWE have been collected by National Resource Conservation Service (NRCS) Snowpack Telemetry (SNOTEL) and California Department of Water Resources snow stations across the mountains of the western US (Serreze et al., 1999), however these point measurements may be unrepresentative of snow conditions in the surrounding landscape (Molotch et al., 2006; Broxton et al., 2019). Newer remote sensing technologies (e.g. airborne LiDAR) have enabled accurate aerial measurement of snow depth (and estimation of SWE by combining them with snow density estimates) over limited regions (e.g. Baños et al., 2011; Deems et al., 2013; Painter et al., 2016), but the cost associated with these collections makes it impossible to collect these data over very large areas (e.g. the entire western US). On the other hand, satellite remote sensing is much more spatially extensive but has demonstrated limited success for measuring SWE in mountainous areas (Dozier 1989; Frei et al. 2012). Other products that map SWE based on a combination of in-situ snowpack measurements and gridded hydrometeorologic data produce more accurate estimates of SWE across many regions, including mountainous areas (Cho et al. 2020, Dawson et al. 2019). Given the importance of SWE estimation across the Western US (e.g. for water supply), it is important to leverage various information to develop the most accurate SWE analysis.

For this competition, we leverage one of these merged products, the University of Arizona (UA) SWE product (Broxton et al., 2017; Zeng et al., 2018), to train regression models to predict SWE across the western US from SNOTEL and California Data Exchange Center (CDEC) station data. Specifically, for each of ~22,000 1 km^2 grid cells across the Western US, UA SWE data are used as the target variable for training Multilinear Regression (MLR) models to predict SWE based only on real time SNOTEL and CDEC data. Although some gridded SWE products (including a recent version of UA SWE data) already produce SWE estimates in real time, these MLR models could represent a new way to leverage ground station data that may improve real-time SWE estimation in remote mountainous regions. Our SWE predictions from these MLR models are evaluated against Airborne Snow Observatory (ASO; Painter et al., 2016) LiDAR-based SWE estimates in California, Colorado, and Wyoming, as well as ~215 SNOTEL / CDEC data across the western US (Figure 1).

In addition to the MLR models to predict SWE, we also developed algorithms to fill missing SNOTEL/CDEC predictor data and to perform additional bias correction for areas that have observed SWE data in the past (e.g. from past ASO collections). Ultimately, model predictions of SWE at the target grid cells are made by averaging ensembles of MLR models which are generated by using training data from different combinations of years. Based on our tests, these ensembles performed better than individual MLR models, and MLR models produced results that were as good, or better than a few machine learning methods that we tested.

2) Model Development

2.1) Data

The UA SWE data (Broxton et al., 2016; Zeng et al., 2018), used to train the MLR models, is gridded snowpack data based on interpolation of snow amount from thousands of SNOTEL/CDEC stations (https://www.nrcs.usda.gov/wps/portal/wcc/home/) and National Weather Service Cooperative Observer stations (https://www.ncdc.noaa.gov/snow-and-ice/daily-snow/) across the conterminous United States using gridded climate data from Parameter Regression Independent Slopes Model (PRISM; Daly et al., 1997**;** https://prism.oregonstate.edu/). The UA SWE data were originally generated at 4 km resolution, but here, we use an 800 m version of these data based on 800 m PRISM data (generated by combining 800 m PRISM monthly climatology data with 4 km daily PRISM data). These 800 m UA SWE data also include corrections to account for small scale effects of terrain and forest cover variability using machine learning of lidar snow datasets (Broxton et al., 2019). They are broadly similar to 4 km UA SWE data at larger (>= 10km) scales, but have some differences at local scales in areas with complex topography or substantial forest cover.

SNOTEL/CDEC SWE data are highly accurate point measurements based on snow pillow data at NRCS SNOTEL stations and snow stations operated by the California Department of Water Resources. These data are used both during model training (as predictors and targets during training of the MLR SWE models for ~215 grid cells with SNOTEL/CDEC stations) and for model evaluation. These data are obtained directly from files that are downloaded from the competition website, but to check for data quality issues, past SNOTEL/CDEC data are also compared with data for these sites that we have in our own database (from https://www.nrcs.usda.gov/wps/portal/wcc/home/; see Section 2.2).

ASO SWE data, used to generate bias correction factors for the MLR SWE models for some of the grid cells during model development as well as for model evaluation, are based on the combination of airborne LiDAR and model-generated snow density estimates (Hedrick et al., 2018). These data are some of the best aerial snow data available over mountainous areas. ASO data used here come from data files provided during this competition, as well as from the National Snow and Ice Data Center (https://nsidc.org/data/aso/data-summaries), and from the ASO Inc. website (https://www.airbornesnowobservatories.com).

Although we tested a few other gridded data sources as target features during model training (the 4 km version of the UA SWE data, https://nsidc.org/data/nsidc-0719, and SNODAS, https://nsidc.org/data/g02158), our final solution is trained with 800 m UA SWE data because it results in the most accurate SWE models (see section 3.4). Satellite data is not used because of temporal sparseness (e.g. due to orbital paths or clouds), and the fact that our previous studies have found them to be more useful for snow cover than SWE estimation.

2.2) Feature Engineering

Our solution is set up to make SWE predictions based on unaltered SNOTEL/CDEC predictor data which are provided for real-time prediction in the evaluation stage of this competition. However, during model training, several steps are applied to past SNOTEL/CDEC SWE observations that were provided for model training/prediction for this competition. First, SNOTEL/CDEC data from the development phase of this competition were compiled to create data files compatible with the input feature list for the evaluation stage. Furthermore, based on visual inspection and comparison against data from CDEC / SNOTEL sites that we have in our database, some quality control was performed on the CDEC features. In particular, one CDEC site had unrealistically stable SWE values for long periods (Figure 2a), while others had non-zero SWE values when there was probably no snow (e.g. Figure 2b). In a few cases, there were also obvious spurious jumps, and there were sometimes differences between the provided data for CDEC stations and what we had in our database (which was obtained from the SNOTEL website (https://wcc.sc.egov.usda.gov/) (e.g. Figure 2c). We manually removed some of the data that were obviously incorrect, but we chose to only make limited alterations (e.g. we did not use data from the NRCS) because during the real-time phase, the provided SNOTEL/CDEC data (which contestants are required to use) likely have the same characteristics as those provided in the development phase files. Indeed, when we tested using the NRCS data in the model development, we observed slightly worse eventual model performance.

Like with the SNOTEL/CDEC predictor data, ground feature evaluation data (which includes some SNOTEL/CDEC data and some ASO data) provided in this competition are simply copied from the files from the development stage files into the correct rows in the validation data files for the evaluation stage. Other data, (UA SWE and additional ASO data) are queried from spatial maps. For the UA SWE data (which is already at ~1 km^2 resolution), only one UA SWE pixel is queried for each 1 km^2 feature, but for ASO data (which are based on 50 m SWE maps), all 50 m^2 pixels with the 1 km^2 feature are queried and averaged.

Similar to the SNOTEL/CDEC data, we also perform some quality control on the ASO data provided in this competition. In particular, we compared the provided features with those generated from the 50 m ASO SWE maps. In general, the match was extremely close, but in a very small number of cases, there were pixels that had zero values in the provided competition data in areas where the downloaded ASO data showed ‘no data’.

2.3) Algorithm

Our solution depends largely on two algorithms to 1) fill in missing SNOTEL/CDEC predictor data, and 2) use MLR to predict SWE for the target grid cells based on these filled data (note that MLR requires no missing predictor data). Both algorithms are conceptually similar in that surrounding SNOTEL/CDEC data are used to predict SWE for target grid cells or stations with missing data.

The gap filling algorithm for the predictor data differs, though, from traditional MLR in that it involves making separate predictions for each ground station (based on linear regression between SWE at the predicted and surrounding stations), and weighting the predictions according to their correlation strength when data is not missing:

$$SWE_{filled} = \frac{\sum w_i \times SWE_i}{\sum w_i} \quad (1)$$

where SWE_{filled} is the calculated SWE value for a particular station, SWE_i is the SWE prediction based on data from station i, determined using linear regression with data from the predicted station, and w_i is the weight given to that prediction. These weights are determined as:

$$w_i = {R^2}_i^{\,m} \quad (2)$$

where, ${R^2}_i$ is the coefficient of determination for the linear relationship between the predictor and predicted station, and the exponent m is a parameter used to enhance the weight given to high ${R^2}_i$ values. If data from one or more of the surrounding stations is missing, those stations are not used and weights for the remaining predictor stations are adjusted accordingly, allowing this methodology be used when multiple stations have missing data (a common occurrence).

For each predicted station, potential predictor stations are only used if they meet one of two criteria: the R^2 between the predictor and the predicted station is above a minimum threshold, R^2_{min}, or they are among the *n* stations with the highest R^2 values with the predicted station. This ensures that where available, only the best predictors are used, but that a certain number of stations will be available for model prediction (in case multiple stations have missing data). Note that here and elsewhere, all parameter values are adjusted manually based on maximizing cross validated (section 3.1) performance of the SWE models.

These models are applied to all stations, creating timeseries of input predictors that have no data gaps. After this step, for each station, an additional temporally varying bias correction factor is used to adjust the gap filled values based on their difference from the actual SWE values for the most recent time when actual data exists. For example, if at a given time, the actual SWE value is 4 inches, and the above procedure predicts 5 inches, then a bias correction of 4/5=0.8 would be applied to the next calculated value. If there are multiple days of missing values, this bias correction factor gradually shifts toward one. This temporally varying bias correction step minimizes discontinuities between the previous observed value and the calculated values across data gaps (see Figure 3a for an example).

Next, traditional MLR, implemented in Python's scikit-learn package (https://scikit-learn.org/stable/), is used to predict SWE at the target grid cells, based on the gap filled predictor data. These models are trained to predict either observed SWE values when the training record (which consists of weekly observations from 2013-2021) contains > 50 SWE observations, or UA SWE values when the training record contains ≤ 50 SWE observations as this ensures that enough samples are available for model training (and as explained in section 3.1, distinguishes between grid cells whose data is primarily derived from SNOTEL/CDEC stations vs. ASO data). As with the gap filling procedure above, potential predictor stations are identified based on whether the R^2 value between each potential predictor and the predicted station is above a minimum threshold, R^2_{min}, or they are among the *n* stations with the highest R^2 values with the predicted station (though note that the value of R^2_{min} and *n* in this step can be different from above). MLR models are trained individually for each grid cell. While this effectively causes each grid cell to have its own unique SWE prediction, irrespective of SWE for surrounding grid cells, the fact that most grid cells are trained using UA SWE data, which has continuous spatial fields of SWE for each day, ensures that grid cells that are next to each other behave similarly (at least to the extent that they are similar in the UA SWE data).

A final bias correction step is then performed to provide additional adjustment to SWE values for grid cells where there are multiple (>2) past observations:

$$SWE_{bc} = a \times (SWE_{MLR} - b)^c \quad (3)$$

where SWE_{bc} is the bias corrected SWE value, SWE_{MLR} is the SWE value predicted using MLR, and *a*, *b*, and *c* are coefficients for the fitted relationship. We use a nonlinear bias correction because some of the differences between SWE_{MLR} and observed SWE values are nonlinear (e.g. Figure 3c).

Although the issue never occured for the 9 years that the model was tested for, we also added code to limit excessively large weekly changes in SWE, which could theoretically happen if, for example, there are spurious observations for some of the predictor stations (e.g. like the one shown in Figure 2c, but affecting multiple stations). As such, if the modeled SWE value at a given station is greater than 10 times either the 1st or 99th percentile of all SWE changes for a station (corresponding to a very large positive or negative SWE increments), then the SWE value for that station is set to the previous week's value.

This combination of ensemble MLR prediction and bias correction, as outlined above, is selected because it gives reasonable performance for predicting SWE, yet its linear nature prevents problems with overfitting. While the MLR approach generally makes better predictions than the weighted averaging that is used for gap filling step, the weighted averaging is used for the gap filling step because it allows for missing data in the predictor data. Furthermore, the temporal bias correction of the gap filled data, as well as the fact that data gaps in many (but certainly not all) cases are fairly short, makes this gap filling step pretty robust. The MLR models, on the other hand, are chosen over several other Machine Learning methods because of their computational efficiency and favorable performance (see Section 3.4).

3) Model Robustness

3.1) Training and Testing

Training is done on a split sample such that one water year is held out for model testing, and most data from other years are used for model training. In addition, SWE predictions are generated not with a single model, but by averaging an ensemble of predictions. These ensembles are generated, for each year, by withholding from model training both the data from the predicted year, but also data from one additional year. For example, ensembles are generated for 2021 by withholding data from 2021 and 2013 to train models for ensemble 1, 2021 and 2014 for ensemble 2, and so forth. Ultimately SWE predictions for each year are generated by averaging predictions from 9 ensemble members, except for predictions in 2022, which are the average of 10 ensemble members. As shown in Section 3.2, the ensemble prediction is almost always better than prediction from each individual ensemble member.

These models are evaluated using cross validation, where each year's data is predicted from ensembles trained only with data from other years. Performance metrics include Root Mean Squared Error (RMSE), normalized RMSE (RMSE divided by the mean value), coefficient of determination (R^2), and bias. For computation of the above metrics, each occurrence of non-missing observed data in any grid cell is used as a data point. For testing purposes, observations with certain characteristics are grouped together (i.e. from a given month, from grid cells in the Sierras vs the Central Rockies vs. other areas, or from grid cells where validation data primarily comes from ASO data or SNOTEL/CDEC data). Note that grid cells are categorized as 'Sierra', 'Central Rockies', or 'Other' based on attribute values in the spatial data files provided in this competition, and they are categorized as 'ASO' or 'SNOTEL/CDEC' grid cells based on whether they have fewer / more than 50 weekly samples from 2013-2021. Fifty samples (which corresponds to ~15% of sample dates) is probably effective at distinguishing between grid cells where most data come from ASO vs. SNOTEL/CDEC because the ASO has not flown over any area more than 50 times, while a vast majority of SNOTEL/CDEC stations (or possibly all of them) would have contributed substantially more than 50 observations in the training dataset.

3.2) Overall Performance

Overall, model performance is variable from year to year (black lines in Figure 4). This does not mean, though, that the model performs dramatically differently for different years. Rather, it is mainly reflective of the fact that model evaluation occurred in different areas each year (e.g. some early years had ASO samples in only a few basins, while later years had expanded ASO coverage). Indeed, model performance for SNOTEL/CDEC grid cells is much more consistent from year to year (with a notable drop of R^2 and increase in normalized RMSE in 2015, which was a very dry year). By contrast, model performance for ASO grid cells shows considerably more year-to-year variability. The fact that earlier years have less ASO data and later years have more ASO data means that overall performance metrics are weighted toward the performance of SNOTEL/CDEC and ASO grid cells for earlier and later years, respectively

In general, error metrics are substantially better for SNOTEL/CDEC grid cells (which have R^2 values above 0.9, normalized RMSEs of 0.25-0.5, and RMSEs generally less than 4 inches) than ASO grid cells (which have R^2 values, normalized RMSEs, and RMSEs of 0.7-0.9, 0.25-1, and 4-10 inches, respectively). This may be caused by a number of factors. For one thing, models are trained using UA SWE data for the ASO grid cells, and the observed data are only used for bias correction for these locations (as these grid cells have relatively few observations). However, it could also be that it is simply harder to predict SWE for ASO grid cells based on nearby SNOTEL/CDEC data. Most SNOTEL/CDEC stations tend to be located in forest gaps and clearings in areas with seasonal snowpack (i.e. they are sited to measure substantial snowpacks), and they may be better at predicting snowpack for other similar locations (i.e. other SNOTEL/CDEC stations) and less able to predict snowpack for dissimilar locations (e.g. areas with shallow snowpacks, which make up a substantial portion of the ASO grid cells). Finally, the ASO data, itself, has more measurement uncertainty than the SNOTEL/CDEC SWE data, which may be due either the uncertainty in the LiDAR snow depth measurement from high altitude flights, or the required use of modeled snow density estimates to convert these measurements to SWE. Conversely, SNOTEL/CDEC data might not be representative of the surrounding 1 km pixel, but this should not affect model performance here because in this competition, the SWE data for most of these grid cells are essentially equivalent to the SNOTEL/CDEC station data.

Figure 4 also has several other important features. First, it shows that SWE prediction based on the ensemble mean is almost always (with occasional exceptions) better than individual ensemble members even though it is a simple average of the ensemble members. This occurs fairly consistently across individual years, and so we are fairly confident that using the ensemble mean should continue to give better results than individual models. Second, our methodology to train the MLR models with UA SWE data results in better SWE prediction than the UA SWE data itself (compare the blue thick dashed and solid lines in Figure 4a-d, which represent, respectively, the performance of UA SWE data and the MLR models trained with UA SWE data for ASO grid cells). Some of this improvement might be due to the additional bias correction for some of the ASO grid cells, but even where bias correction is not employed (when there are less than three ASO observations in the training set for a particular grid cell), the improvement is still substantial. For example, the overall RMSE between the MLR predictions and ASO observations for a set of ASO grid cells that cannot possibly have any bias correction (<= 2 observations in the historical record) is 3.06 inches, while the RMSE between UA SWE and ASO observations for the same grid cells is 6.23 inches (though we recognize that these grid cells are located over limited areas, e.g. some basins in Colorado, and so may not represent the overall relative performance of UA SWE vs the MLR models). This improvement could be due to the fact that by tying the predictions to nearby SNOTEL/CDEC stations, the MLR models are particularly effective at characterizing how the current year's SWE relates to other years in the training period.

3.3) Temporal and Spatial Differences

Figure 5 shows how our model's performance changes over the course of the season, and across different regions. Again, the performance for SNOTEL/CDEC grid cells is more consistent than for ASO grid cells due to the spatial differences in the ASO data collection discussed in Section 3.1. Model performance (in terms of normalized RMSE, R^2, and bias) for SNOTEL/CDEC grid cells (dotted lines in Figure 5) is fairly consistent across regions. RMSE for these stations is higher for Sierra Nevada than Central Rockies and other grid cells, mainly because Sierra Nevada grid cells have more snow (compare Figures 5a and 5e). Generally, our model's performance for SNOTEL/CDEC grid cells is highest in March and April (when R^2 values are above 0.9 and normalized RMSEs are below 0.3). R^2 drops and normalized RMSE goes up during the melt season, especially in June. This is unsurprising because snowpack during the melt season becomes much more variable (Deems et al., 2008; Schirmer et al., 2011). It is also possible that the relationships between the grid cells and the predictor stations change during the melt season (e.g. Egli and Jonas, 2009), though it is difficult to quantify how. During model development, we experimented with using different models for different months (or for earlier months vs later months), but were unable to improve over the methodology as described above, possibly because further subdivision resulted in the models not having enough observations for good model training.

The performance of ASO data (solid lines in Figure 5), on the other hand, varies less predictably across seasons. As with the SNOTEL/CDEC grid cells, RMSEs for Sierra Nevada ASO grid cells are higher than for Central Rockies ASO grid cells (note that at the time of writing, there are no ASO validation data outside of those two regions) because of more snow in the Sierra Nevada. However other error metrics do not necessarily show systematically better performance in one region over another. Generally, when there is a lot of ASO data (in April and May for Sierra Nevada grid cells), R^2 values for ASO grid cells are above 0.8.

3.4) Additional Tests

To ensure that our model performance is optimized, we also perform some additional tests. Specifically, we use other readily available gridded SWE datasets, the 4 km UA SWE data, and 1 km SNODAS data for model training, and we explore a few other machine learning techniques - random forest (RF), and multilayer perceptron (MLP) neural network models. Table 1 gives the overall cross validated RMSE for all grid cells, SNOTEL/CDEC grid cells, and ASO grid cells when 800 m UA SWE data is replaced with 4 km UA SWE and 1 km SNODAS data in the model training, and the MLR modelling is replaced with RF and MLP (note that for all of these tests, other aspects of the modelling, e.g. filling of predictor data, bias correction, and use of ensembles, is consistent). Using the 4km UA SWE or 1 km SNODAS data generally results in lower model performance, regardless of region. For example, RMSEs are 15-25% higher when using SNODAS data, and 7-15% higher when using 4 km UA data than when using 800m UA data. Similarly, MLR results in lower RMSEs than RF, regardless of region or pixel grouping (i.e. SNOTEL/CDEC or ASO grid cells). MLP, on the other hand, results in fairly comparable, or sometimes slightly better, model performance than MLR. However, in the end, we chose MLR because we judged the linear structure to be potentially more reliable in case the model encountered conditions that were not represented in the model training.

Finally, we tested whether further improvements could be obtained by considering additional physiographic attributes (Figure 6). In general, model errors are fairly evenly distributed, regardless of the physiographic attribute that we tested, indicating that our solution probably cannot be optimized based on these attributes. There is a slight (~10%) low bias of our model SWE compared to observed SWE (Figure 6a). Correcting this bias, though, is non-trivial, as a simple global bias adjustment resulted in a higher RMSE.

4) Considerations for Use

Overall, our model is capable of making relatively accurate SWE predictions across the western US while at the same time being simple and easy to implement. It has very low data requirements, requiring only SNOTEL/CDEC data to operate in real time and it includes an algorithm to effectively handle missing SNOTEL/CDEC predictor data, giving it a lot of robustness for real-time application. The use of MLR rather than other, more computationally expensive machine learning techniques also makes it fast. For example, using a laptop with relatively modest processing power (tested using an Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz with 8 GB of RAM), model training takes approximately 10 minutes per ensemble member to train the model for 22,000 grid cells, while model inference (running the trained model) for this same set of grid cells takes 20-30 seconds (other machine learning models were much slower). While our implementation does not include any sort of parallelization, it should also be simple to train different ensemble members simultaneously, further enhancing performance.

Our model can also be implemented across a much wider range of 1 km grid cells across the western US because the training targets come from gridded SWE data, which cover the entire region. Although we used the 800m UA SWE data to train the models, our approach here can also be used with other datasets (e.g. 1 km SNODAS, albeit with lower model performance, as we show). While some of these datasets are also available in real time (and therefore also able to do real-time SWE prediction for the target grid cells), our approach seems to lead to improvement over the original gridded data (Figure 4), both because of the use of LiDAR data-informed bias corrections, and the way that it leverages SNOTEL/CDEC observations, which provide a robust way of representing how snowpack at a particular time compares to other times during the model training period.

What is less clear is how well this methodology would perform for areas that are far from SNOTEL/CDEC stations, such as lowland areas outside of the mountains (as nearly all of the data used for model validation here come from mountainous areas that have SNOTEL/CDEC stations nearby). It is possible that in lowland areas, this approach will perform worse than the original gridded data at predicting SWE because the SNOTEL/CDEC stations are too far away. Therefore, any application of this approach requires additional research about where it should vs. should not be implemented and how to combine it with existing SWE data. An additional issue that is not addressed in our solution, but should be investigated further, is how to incorporate satellite data to improve the solution. For example, satellite data could be especially useful for detecting if an area is snow covered or not, which could help guide the discrimination of snow / no snow areas in lowland areas (as well as in the mountains in the fall and spring). In addition, c-band radar-based satellite remote sensing (e.g. that used by the Sentinel sensor) shows promise to detect snow depth in mountainous regions (Lievens et al, 2019**)**. It is challenging to incorporate satellite data though because of data gaps / confusion (e.g. due to cloud cover obscuring or getting confused with snow). Our future work could address these as well as other issues raised in this report (e.g., the negative trend in Figure 6a).

References

Baños, I. M., García, A. R., i Alavedra, J. M., i Figueras, P. O., Iglesias, J. P., i Figueras, P. M., and López, J. T., Assessment of airborne LIDAR for snowpack depth modeling, Boletín de la Sociedad Geológica Mexicana, 2011, 63, 95-107.

Broxton, P.D., Zeng, X. and Dawson, N. (2016) Why do global reanalyses and land data assimilation products underestimate snow water equivalent? Journal of Hydrometeorology 17(11), 2743-2761.

Broxton, P.D., van Leeuwen, W.J. and Biederman, J.A. (2019) Improving Snow Water Equivalent Maps With Machine Learning of Snow Survey and Lidar Measurements. Water Resources Research 55, 3739– 3757.

Brown, T.C., Hobbins, M.T. and Ramirez, J.A. (2008) Spatial Distribution of Water Supply in the Coterminous United States 1. JAWRA Journal of the American Water Resources Association 44(6), 1474-1487

Cho, E., Jacobs, J. M., & Vuyovich, C. M. (2020). The value of long-term (40 years) airborne gamma radiation SWE record for evaluating three observation-based gridded SWE data sets by seasonal snow and land cover classifications. Water Resources Research, 56, e2019WR025813. DOI: 10.1029/2019WR025813.

Daly, C., Taylor, G. and Gibson, W. (1997) The PRISM approach to mapping precipitation and temperature, pp. 20-23, Citeseer.

Dawson, N., Broxton, P., & Zeng, X. (2018). Evaluation of Remotely Sensed Snow Water Equivalent and Snow Cover Extent over the Contiguous United States, Journal of Hydrometeorology, 19(11), 1777-1791.

Deems, J. S., Painter, T. H., and Finnegan, D. C., Lidar measurement of snow depth: a review, J. Glaciol, 2013, 59, 467-479.

Deems, J. S., Fassnacht, S. R., and Elder, K. J., Interannual consistency in fractal snow depth patterns at two Colorado mountain sites, Journal of Hydrometeorology, 2008, 9, 977-988.

Schirmer, M., Wirz, V., Clifton, A., and Lehning, M., Persistence in intra-annual snow depth distribution: 1. Measurements and topographic control, Water Resources Research, 2011, 47.

Dozier, J., 1989: Spectral Signature of Alpine Snow Cover from the Landsat Thematic Mapper. Remote sens. Environ., 28, 9-22.

Egli, L., and Jonas, T., Hysteretic dynamics of seasonal snow depth distribution in the Swiss Alps, Geophysical research letters, 2009, 36.

Frei, A., M. Tedesco, S. Lee, J. Foster, D.K. Hall, R. Kelly, and D.A. Robinson, 2012: A review of global satellite-derived snow products. Advances in Space Research, 50(8), 1007-1029.

Furniss, M.J. (2010) Water, climate change, and forests: watershed stewardship for a changing climate, DIANE Publishing.

Hedrick, A. R., Marks, D., Havens, S., Robertson, M., Johnson, M., Sandusky, M., et al. (2018). Direct insertion of NASA Airborne Snow Observatory-derived snow depth time series into the iSnobal energy balance snow model. Water Resources Research, 54, 8045– 8063. https://doi.org/10.1029/2018WR023190

Lievens, H., Demuzere, M., Marshall, HP. et al. Snow depth variability in the Northern Hemisphere mountains observed from space. Nat Commun 10, 4629 (2019). https://doi.org/10.1038/s41467-019-12566-y

Molotch, N. P., and R. C. Bales (2006), SNOTEL representativeness in the Rio Grande headwaters on the basis of physiographics and remotely sensed snow cover persistence, Hydrol. Process., 20(4), 723-739, doi: 10.1002/hyp.6128.

Painter, T. H. and Coauthors (2016), The Airborne Snow Observatory: Fusion of scanning lidar, imaging spectrometer, and physically-based modeling for mapping snow water equivalent and snow albedo. Remote Sens. Environ., 184, 139–152, doi:10.1016/j.rse.2016.06.018.

Serreze, M. C., M. P. Clark, R. L. Armstrong, D. A. McGinnis, and R. S. Pulwarty (1999), Characteristics of western United States snowpack from telemetry (SNOTEL) data, Water Resour. Res., 35(7), 2145-2160, doi: 10.1029/1999WR900090.

Zeng, X., Broxton, P. and Dawson, N. (2018) Snowpack change from 1982 to 2016 over conterminous United States. Geophysical Research Letters 45(23), 12,940-912,947.

Figures and Tables

Note: Data and codes for creating the figures and tables below can be found at https://drive.google.com/file/d/17rz_0RlHTfPCj3EUQF__qnQPzlwsER2K/view?usp=sharing.

This link is not designed to be permanent but should be active until the end of the Snowcast Showdown competition.

Table 1: RMSEs for different experiments using different gridded training datasets (800 m UA SWE, 4 km UA SWE, and SNODAS data), and different machine learning techniques (MLR, RF, and MLP), considering all grid cells, SNOTEL/CDEC grid cells, and ASO grid cells. The first row in each grouping corresponds to the performance of our final solution (MLR + UA800m)

		RMSE (in)		
	Experiment	**All Grid Cells**	**SNOTEL /CDEC**	**ASO**
All Areas	**MLR (UA800m)**	5.05	2.83	6.61
	MLR (UA4km)	5.43	2.83	7.20
	MLR (SNODAS)	5.76	2.83	7.71
	RF (UA800m)	5.47	3.15	7.12
	MLP (UA800m)	5.02	2.86	6.56
Sierra Nevada	**MLR (UA800m)**	6.71	4.54	6.89
	MLR (UA4km)	7.25	4.54	7.46
	MLR (SNODAS)	7.72	4.54	7.96
	RF (UA800m)	7.33	6.03	7.45
	MLP (UA800m)	6.65	4.59	6.82
Central Rockies	**MLR (UA800m)**	3.49	2.00	5.10
	MLR (UA4km)	4.00	2.00	6.02
	MLR (SNODAS)	4.38	2.00	6.69
	RF (UA800m)	3.58	2.07	5.21
	MLP (UA800m)	3.56	1.98	5.23

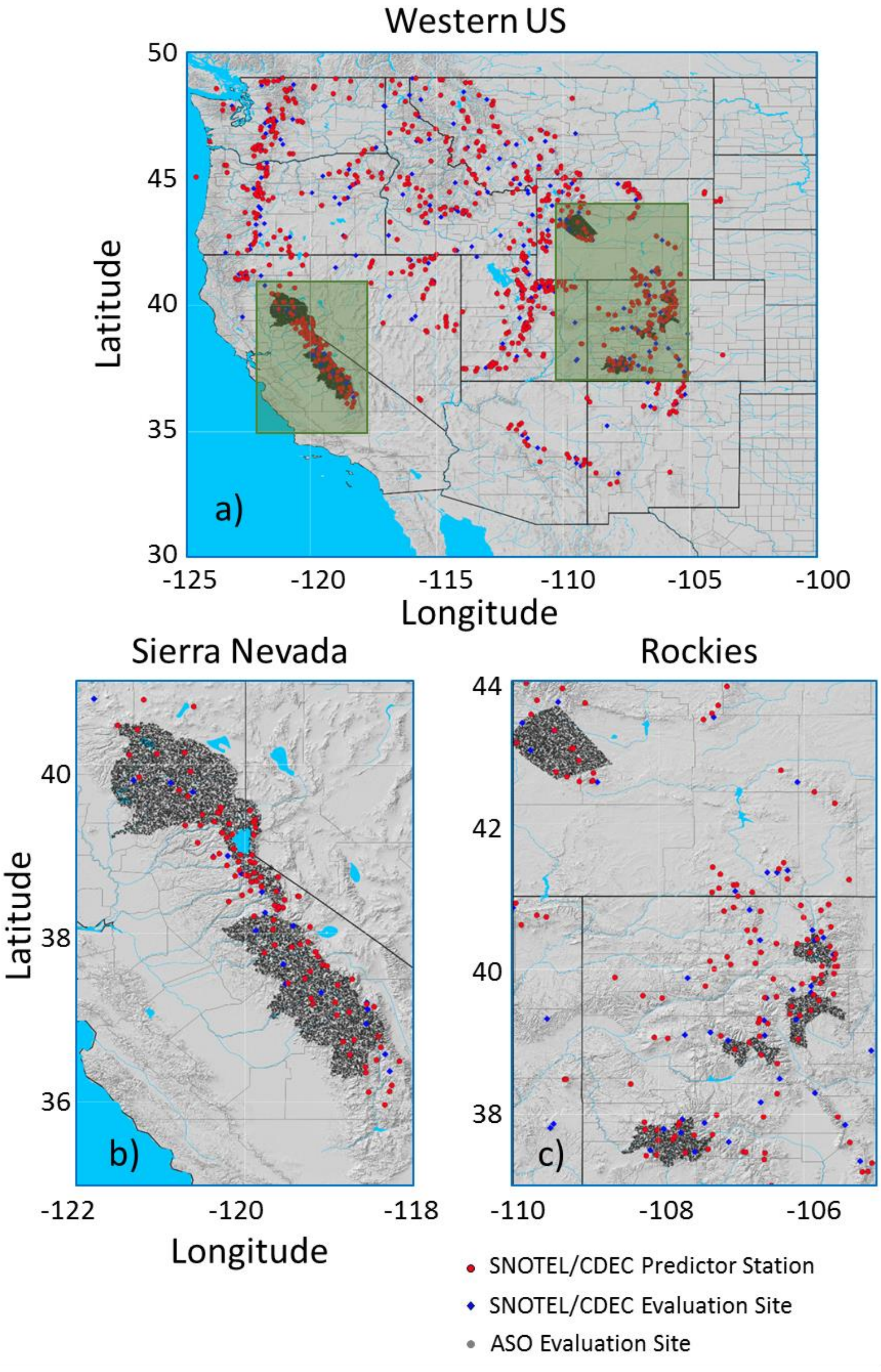

Figure 1: Locations of ground stations used as predictor variables (red dots), and grid cells where SWE is predicted (blue and dark grey dots) for the a) Western US, b) Sierra Nevada, and c) Rockies. A shaded relief map, as well as states, counties, major rivers, and other water bodies are shown in the background.

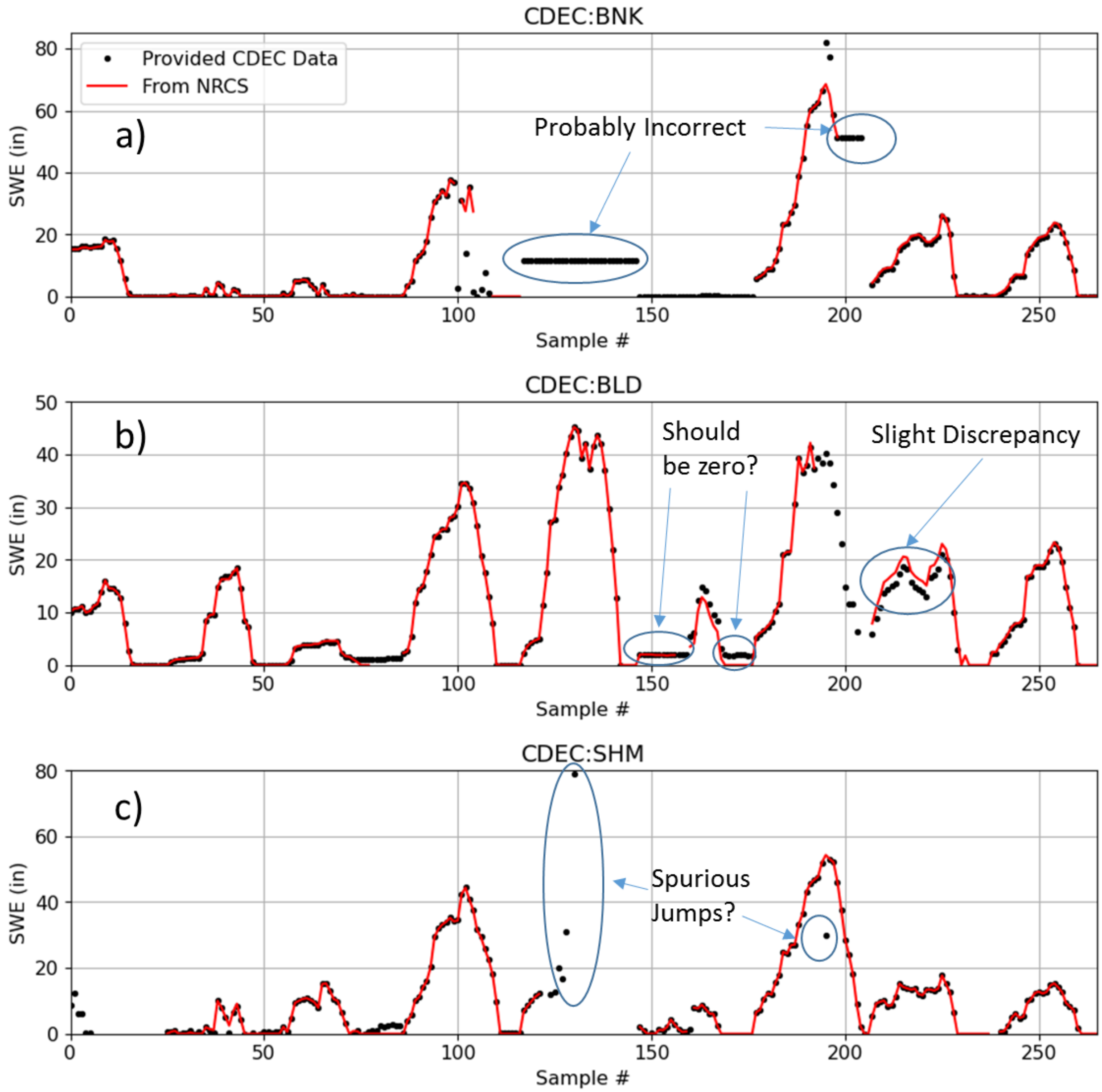

Figure 2: Examples showing possible spurious CDEC data (and some disagreement with data downloaded from NRCS). Each dot or node on the lines represent a weekly sample in the training dataset.

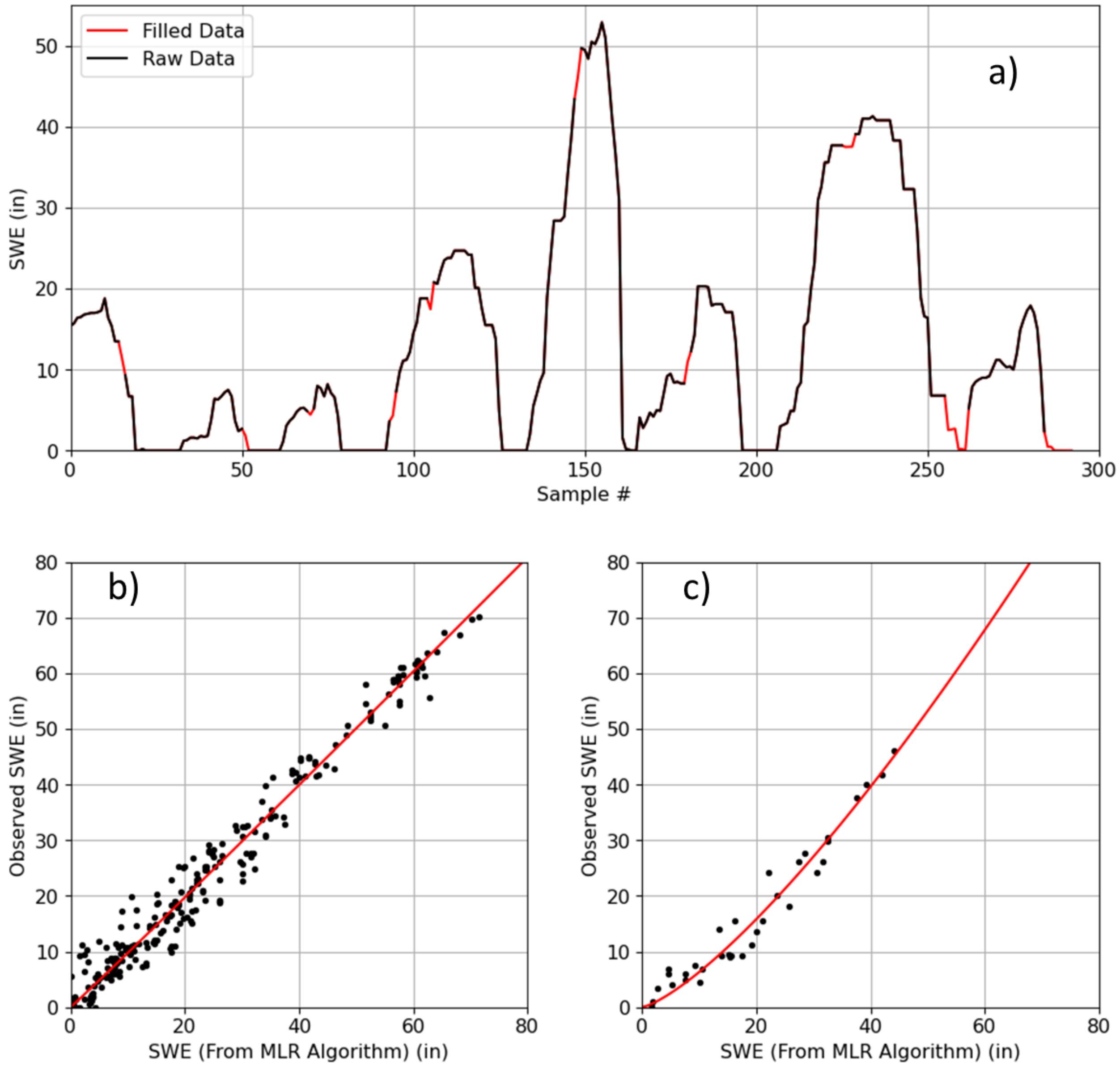

Figure 3: a) Example of a cool season weekly timeseries (extending from 2013-2020) for a CDEC site that has data gaps (black line is observed, and red line is filled). b) and c) show examples of predicted vs observed SWE, potentially requiring a bias correction factor (represented by the red lines). In b), which shows a grid cell with lots of observations (from a SNOTEL/CDEC station) the MLR prediction is already pretty good and the bias correction doesn't do much, but in c), which shows a grid cell in a basin with a relatively large amount of ASO data, a non-linear bias correction improves the SWE prediction.

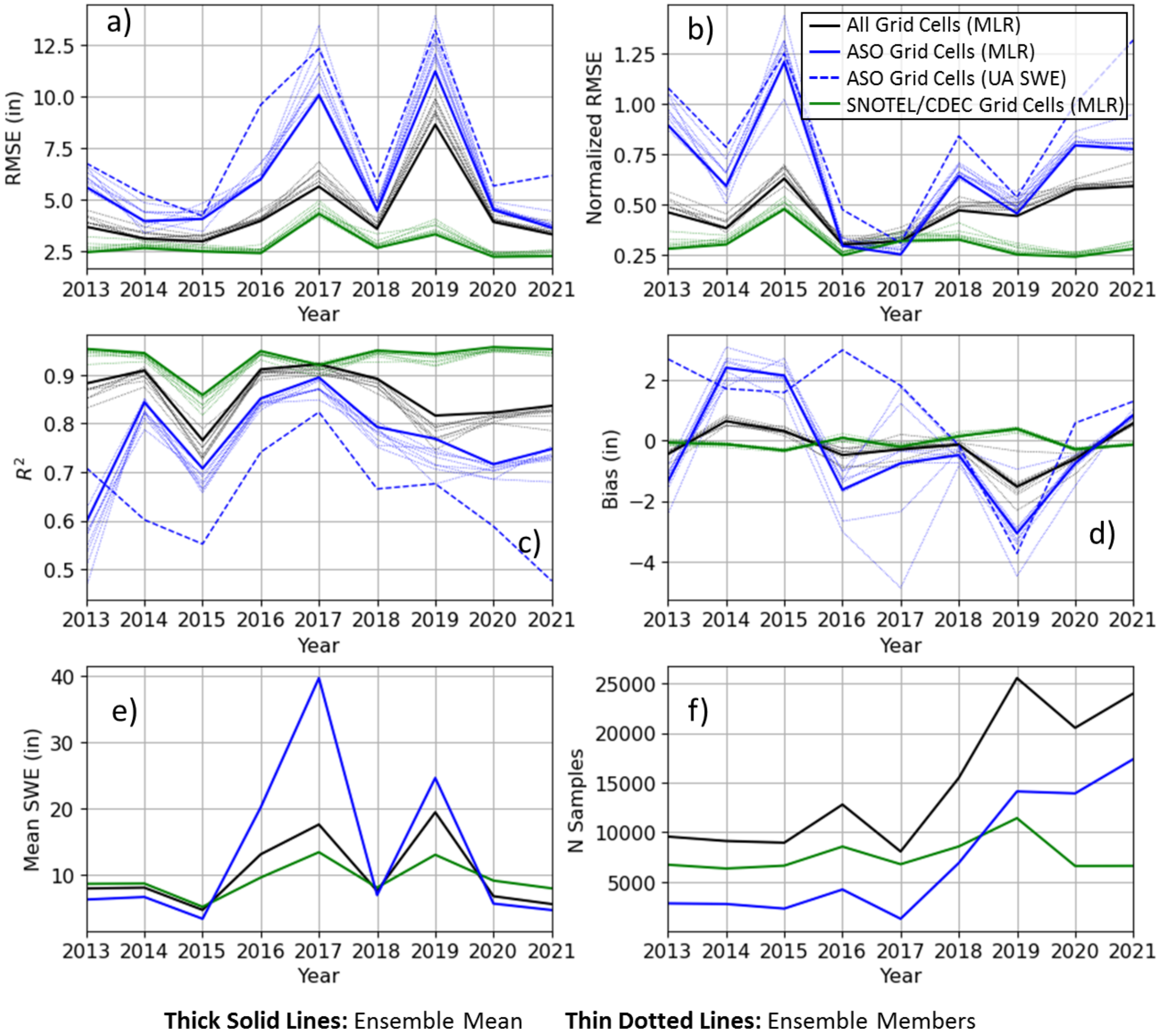

Figure 4: Yearly, cross-validated performance of the MLR models to predict SWE for validation grid cells, with different colored lines showing performance for ASO grid cells (blue), SNOTEL/CDEC grid cells (green), and all grid cells (black). For ASO grid cells, the performance of the 800 m UA SWE data are also shown (blue dashed lines). Panels a – d show performance in terms of RMSE, normalized RMSE, R^2, and bias (modeled – observed). Panel e shows the average observed SWE value for each grouping for each year, and panel f shows the number of contributing observations. As mentioned in the text, grid cells are labeled as ‘ASO’ or ‘SNOTEL/CDEC’ based on having less than or equal to 50 (ASO) vs greater than 50 (SNOTEL/CDEC) observations in the observed record (from 2013-2021) as no single watershed has had more than 50 ASO flyovers, while a vast majority of SNOTEL/CDEC stations (or possibly all of them) would have contributed more than 50 observations in the training dataset.

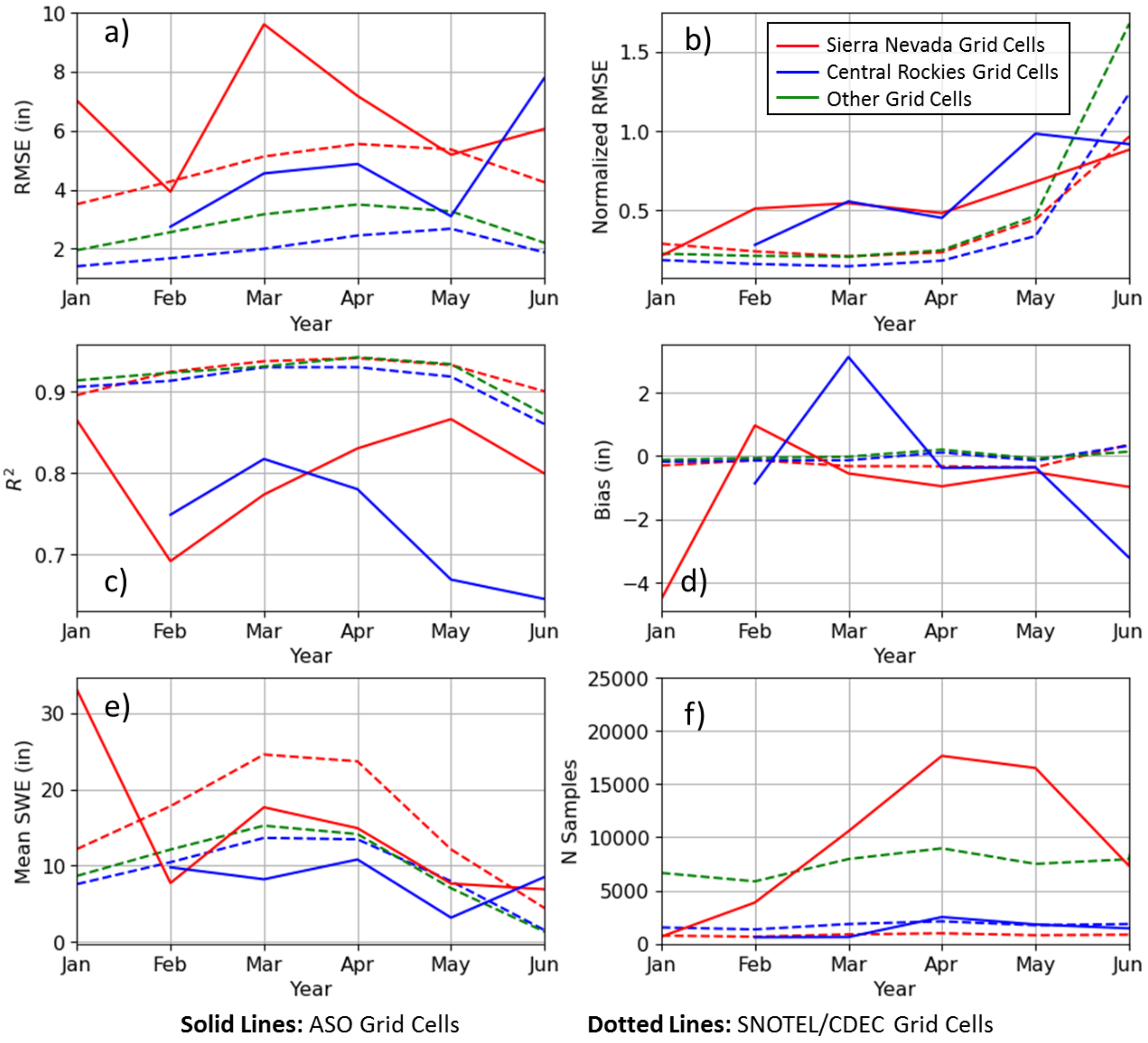

Figure 5: Average monthly, cross-validated performance of the MLR models to predict SWE for validation grid cells, with different colored lines showing performance for grid cells in the Sierra Nevada (red), Central Rockies (green), and other grid cells (black), with solid lines representing ASO grid cells, and dotted lines representing SNOTEL/CDEC grid cells (giving a total of six possible groupings, though there are no Other-ASO grid cells, so only 5 groupings are shown). As with Figure 4, panels a) – d) show performance in terms of RMSE, normalized RMSE, R^2, and bias (modeled – observed), and panels e) and f) show the average observed SWE value for each grouping for each year the number of contributing observations to each grouping. ASO and SNOTEL/CDEC grid cells are distinguished as in Figure 4, and the three regions are those in the provided spatial data files. Note that a majority of the grid cells in California belong to the Sierra Nevada grouping, while the Central Rockies grouping include most grid cells in Colorado, northern New Mexico, and extreme southern Wyoming.

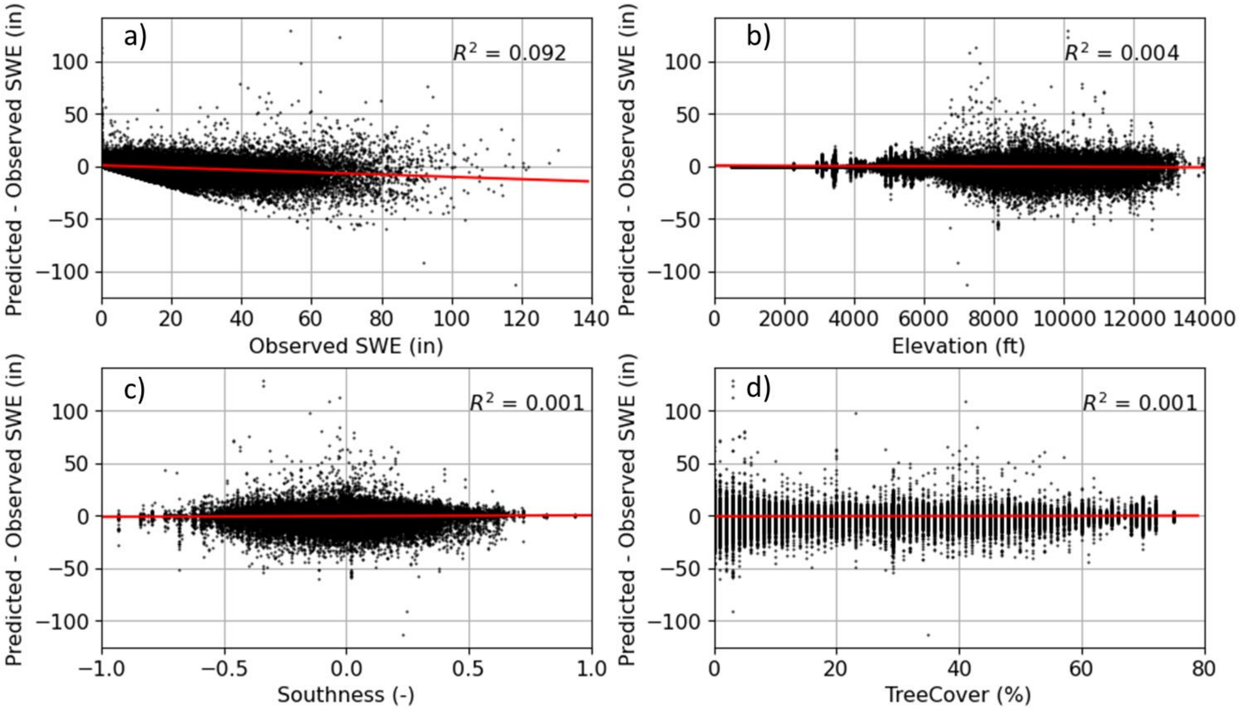

Figure 6: MLR model errors as a function of observed SWE (panel a) and a few important physiographic characteristics (panels b-d). Note that the MLR models give a slight low bias when SWE is large (represented by the decreasing trend in panel a), however errors are roughly evenly distributed according to other physiographic attributes. Note that $Southness = -\sin(slope) \times \sin(aspect)$, where positive values indicate south facing slopes, and that % Treecover is from the Global 30m Landsat Tree Canopy Version 4 dataset (https://lcluc.umd.edu/metadata/global-30m-landsat-tree-canopy-version-4).