AN ENSEMBLE APPROACH OF MAPPING SNOW WATER EQUIVALENT IN UTAH BY USING STATION OBSERVATION DENSITY

by

Logan D. Schneider

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Statistics

Approved:

Brennan Bean, Ph.D. Major Professor	Jürgen Symanzik, Ph.D. Committee Member
Yan Sun, Ph.D. Committee Member	D. Richard Cutler, Ph.D. Interim Vice Provost of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2022

ABSTRACT

AN ENSEMBLE APPROACH OF MAPPING SNOW WATER EQUIVALENT IN UTAH BY USING STATION OBSERVATION DENSITY

by

Logan D. Schneider, Master of Science

Utah State University, 2022

Major Professor: Brennan Bean, Ph.D.
Department: Mathematics and Statistics

April 1st snow water equivalent (SWE) estimates of the mountain snowpack are used by many western states for water management planning. Land-based weather stations provide accurate measures of local snowpack characteristics but, due to limited spatial coverage, struggle to characterize state or regional landscapes with consistent accuracy. In contrast, high resolution national level gridded climate products, including SWE estimates, struggle to appropriately characterize local snowpack conditions in mountainous terrain. This study investigates the accuracy and bias of SWE predictions throughout Utah and explores an ensemble approach to provide improved estimates of snowpack characteristics by using local land-based measurements to dynamically update national level products. This process shows marginal decreases in variance and provides a template that can be used to explore different weighting schemes in the future. The R package `rsnodas` provides the framework for implementing the ensemble method of combining predictions from multiple data sources, which has the advantage of improving the stability of the predictions over time and

space. Also included in this thesis is a discussion of the challenges and potential opportunities for improving currently available SWE products for operational use.

(59 pages)

PUBLIC ABSTRACT

AN ENSEMBLE APPROACH OF MAPPING SNOW WATER EQUIVALENT IN UTAH BY USING STATION OBSERVATION DENSITY

Logan D. Schneider

Mountain snowpack is an important resource for water management planning in Utah. Snow water equivalent (SWE) is the amount of water contained in a snowpack. A few organizations predict SWE throughout the United States but struggle making accurate predictions in mountainous regions. Weather stations provide accurate measurements of SWE but have limited spatial coverage that hinders the ability to make accurate estimates statewide. This thesis examines the accuracy of current models and proposes using local weather measurements to improve upon national level predictions. An R statistical software package named `rsnodas` implements this process while allowing the public access to a variety of temperature and SWE datasets. The package also provides a method for combining predictions from different data sources, which has the advantage of improving the stability of the predictions over time and space. Also included in this thesis is a discussion of the general merits of the map combination approach along with potential avenues for improvement in the future.

ACKNOWLEDGMENTS

It takes a community for someone to be successful. I would like to thank those that helped and supported me. First and foremost, I'd like to thank my academic mentor, advisor, and friend Dr. Brennan Bean. His continual guidance, input, support was invaluable to make this work possible. I'd like to thank my committee members Dr. Jürgen Symanzik and Dr. Yan Sun for dedicating time and feedback to help refine and improve this thesis.

Finally, a special thank you to my family and friends for your love, support, and patience listening to me explain this idea and concepts over and over again.

Logan Schneider

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Table Page

CHAPTER 1

INTRODUCTION

Snowmelt runoff is a major source of summer water for the western United States (Mote et al., 2005, Serreze et al., 1999). Sustainable water resource management plans rely upon accurate estimates of the water available in the springtime snowpack (Lv et al., 2019). Snow water equivalent (SWE), a measure of the water present within a snowpack, is measured manually via snow courses, where wintertime hikers take snow cores along a particular path every month, or automatically via Snowpack Telemetry (SNOTEL) stations, which use steel pillows filled with antifreeze to automatically record the weight of snow and the water content every hour. Historically, both sets of stations have been used to forecast water availability, though many of the old snow courses have been replaced by the modern SNOTEL stations.

April 1st SWE estimates are used to identify trends of snowpack conditions for effective water management planning (Bohr and Aguado, 2001). SNOTEL and Snow Course stations provide accurate measurements of local snowpack characteristics but, due to limited spatial coverage, struggle to characterize changes in the snowpack across the landscape. There are satellite-based products that aim to characterize region-level snowpack trends and provide an alternative to the land-based measurement approach by inferring the water content of the snowpack using radiation (Frei et al., 2012). Two notable snowpack data sources that make use of satellite information include the National Oceanic and Atmospheric Administration's Snow Data Assimilation System (SNODAS) and University of Arizona (UA) datasets. Both use combinations of satellite information, atmospheric modeling, and land-based weather station measurements to provide gridded estimates of SWE for the conterminous

United States at a daily resolution. The method by which the different data sources are combined is unclear, as will be explored in Chapter 2. A third data product named Daymet uses daily meteorological observations to estimate the variables of minimum and maximum temperature (Tmin and Tmax) and precipitation (PPT) which are used to infer SWE (M. Thornton et al., 2021).

Overall, each of these data sources struggle to explain the complex nature of SWE in Utah. Daymet provides SWE estimates at a high resolution (1 km) but tends to severely underestimate the true value of the end of season snowpack in the state of Utah, with April 1 SWE values about half those measured at SNOTEL stations in prelminary investigations. In contrast, the UA dataset uses a better model for deriving SWE, but does so at a lower resolution (4 km), which smooths over some of the sharp changes in SWE that are typically observed in mountainous regions. Even SNODAS, which is the best of the three gridded products in terms of accuracy and resolution, has demonstrated struggles to predict SWE in mountainous regions. For example, Clow et al. (2012) evaluated SNODAS in the Colorado Rocky Mountains and reported that only 30% of SWE variance was explained in alpine regions. These shortcomings highlight the need for local snow models that address region-specific trends not fully captured in continental scale models.

Christensen and Sain (2012) highlights the increasing use of ensembles or collections of models for estimating environmental processes. The approach in this thesis implements similar strategies by combining national-level gridded climate products with state-level snow load models generated using land based observations to produce ensemble estimates of SWE in near real time. In a recent study, Yang et al. (2022) proposed combining ground-based observations with satellite-derived snow data using a linear regression model to improve real-time SWE predictions in the Sierra Nevada mountains in California. Yang et al. (2022) combined ground snow pillow

SWE measurements, physical geographic data, physically-based historical SWE patterns, and satellite-observed daily mean fractional snow-covered area (DMFSCA) in a linear regression model. One goal of their study was to inspect the influence of satellite DMFSCA. This was determined by a comparison of linear regression models of land-based measurements with and without the information of DMSCA from satellite observations. The integration of this satellite information, in-situ measurements, and historical patterns decreased bias and increased model performance in terms of R^2 as compared to SNODAS and a National Water Model (Gochis et al., 2018) for the water years of 2013-2017.

The ensemble method described in this thesis has some similarities and differences in the approach taken by Yang et al. (2022). Yang et al. (2022) included physically-based historical SWE patterns which were not included in the current effort. This ensemble method described in this thesis utilizes SWE measurements, elevation, and climate data to create a generalized additive model (GAM) with splines on sphere (SOS) to account for possible non-linear trends rather than a linear regression model. This model differs from traditional interpolation techniques which seek to fit the input data exactly at the sampled location. Instead, the GAM approach focuses on capturing general patterns in the SWE over elevation and temperature, using a location-specific adjustment provided by the SOS step. The state-level model is then dynamically combined with the national-level SNODAS model, where increased preference is given to the state-level approach in the parts of Utah with the highest spatial density of land-based measurements. It is this dynamic combination of state and national-level models based on land-based station spatial density that is the primary novel contribution of this thesis.

This strategy of combining ground-based and remote sensing by observation density has shown improvements when compared to SNODAS estimates. Figure 1.1 dis-

plays the ensemble process. Using the station observation, a GAM is created and used to generate a gridded map of SWE estimates. A monotonic function is applied to the station density to get the maps of weights that are multiplied to the GAM map while the remaining weights are applied to the satellite estimates like SNODAS. The products are then added together to get the final SWE predictions. This methodology utilizes predictions, like SNODAS estimates, without knowing all the climate and temperature variables to fully recreate a model.

Chapter 4 discusses the results from the ensemble of SNODAS and ground-based observations which displays a decrease in standard deviation of the errors. The median of the errors of the blending technique isn't as volatile or have as large of a spread in the median bias over the years but does display an underestimation bias. This study highlights the value of implementing multiple sources and models for estimating SWE. This process and framework suggests multiple avenues of exploration with promise to provide better regional predictions of SWE in mountainous areas like Utah. The potential information gained by the use of this method leads to making more educated decisions with water management in Utah and can be implemented easily from the GitHub R package, `rsnodas`. The methodology outlined in this thesis has the flexibility to allow for and include different criteria that can be implemented. For example, another criteria that could be used is to include some measure of agreement or similarity of outputs from different models or the use of different variables. One caution with the ensemble approach is an over-aggressive regression to the mean, which can smooth over important variability in the SWE of the snowpack across the landscape.

The remainder of this thesis proceeds with Chapter 2 investigating the advantages and disadvantages of each source with information on available software to download data from SNODAS, Daymet, UA, PRISM climate maps, and SNOTEL

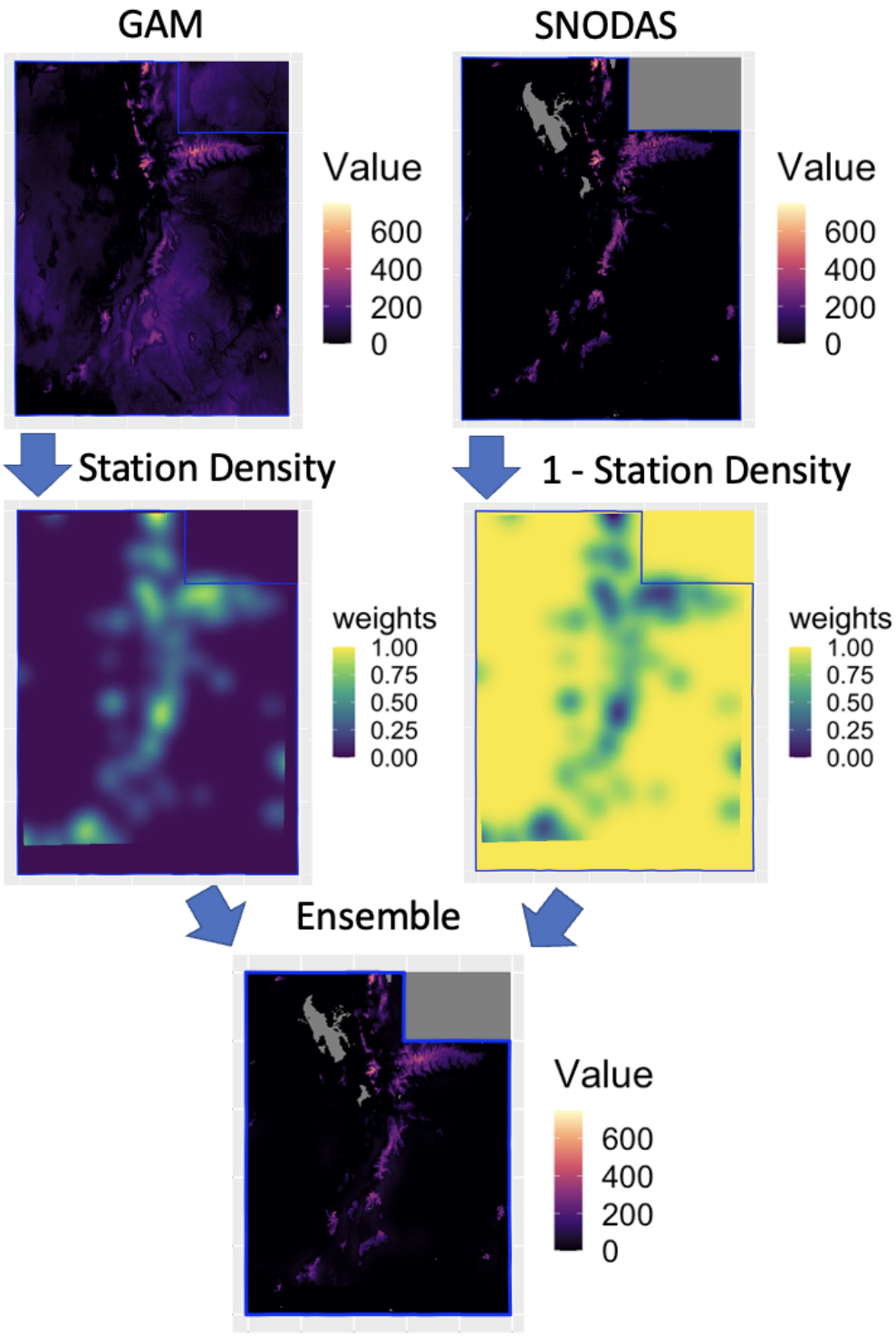


Fig. 1.1: Process of the ensemble approach. The top row maps are multiplied by the weight maps located directly below and added together to get the final map.

measurements. Chapter 3 describes the methodology of constructing a model using land-based SWE measurements that can be integrated with the national-level SWE data products. This approach uses the spatial density of the land-based weather stations as the means of blending the local and national models. In short, greater regard will be given to the local model in locations with a relatively high spatial density of land-based station measurements. The results and software contributions will be presented in Chapter 4, followed by a discussion about potential improvements, challenges, and future work in Chapter 5. Appendix A includes a vignette of the capabilities and functionality of the R package `rsnodas`. Appendix B discusses challenges from data sources, methods and models tested with preliminary results, and concluding discussing why certain models and data sources weren't used.

CHAPTER 2

Data Sources and Software

Snow Data Assimilation System (SNODAS), Parameter-elevation Regressions on Independent Slopes Model (PRISM), University of Arizona (UA), and Daymet supply data for temperature and other climate variables such as precipitation, minimum and maximum temperature, and vapor pressure. Each data source is considered for inclusion in the ensemble modeling approach described further in Chapter 3. This chapter discusses what each data source provides including their strengths and weaknesses. Table 2.1 lays out a summary of the variables and spatial resolution that each data source offers.

Table 2.1: Table of the variables available for the data sources of SNODAS, PRISM, UA, and Daymet.

Data Source	**Resolution**	**Variables available**
SNODAS	1 km	SWE, SnD, SM, SSP, SBS, PPT, SPT
PRISM	4 km*	PPT, Tmin, Tmax, TDmean, VPmin, VPDmax, Elev, SOLTOTAL
UA	4 km	SWE, SnD
Daymet	1 km	Tmin, Tmax, PPT, VP, SR, SWE

*maps available at 800 m resolution for a fee.

One product of this thesis is software that conveniently downloads the previous data types. This software is discussed further in Chapter 2.6.

2.1 Snow Data Assimilation System

SNODAS is a national data product that provides daily estimates of multiple snowpack properties at a one kilometer resolution (Shuman and Ambrose, 2003), including:

- SWE (mm)
- snow depth (SnD)
- snow melt runoff (SM)
- sublimation from the snow pack (SSP)
- sublimation of blowing snow (SBS)
- solid and liquid precipitation (PPT)
- snow pack average temperature (SPTave)

SNODAS integrates satellite, airborne platforms, and ground stations with model estimates to estimate these variables since 2004. Barrett (2003) provides further insight on the process of SNODAS utilizes, "Each day, analysts decide whether or not to use remote sensing and ground based observations to update the snow water equivalent state in the model." This use of multiple sources of data, analysts to decide on using information, and the high resolution of estimates are strengths of SNODAS and allow for providing useful SWE estimates. Although SNODAS informs readers of the data they use, they fail to provide information on how they blend the different data sources in the modelling process.

A few challenges with remote sensing SWE observations using microwaves is the limited ability to measure deep snow, snow under forest canopies, and redistributed snow (Kinar and Pomeroy, 2015, Lv et al., 2019). SNODAS SWE estimates in mountainous regions struggle explaining the distribution of SWE and researchers have tried to improve their estimates by using linear regression and other methods (Clow et al., 2012, Yang et al., 2022). SNODAS has been providing estimates since September 2003 which makes it impossible to observe long-term changes in snowpack. It will be shown that SNODAS provides the highest resolution and estimates of SWE currently available and still displays evidence of bias from time to time. Information about SNODAS daily estimates are available and steps for downloading data in near real-time are available at the following link:

https://nsidc.org/data/g02158

2.2 Parameter-elevation Regressions on Independent Slopes Model

Since 1895 to the present, the PRISM climate group has maps for the following climate elements:

- precipitation (PPT)
- minimum and maximum temperature (Tmin and Tmax)
- mean dew point (tdmean)
- minimum and maximum vapor pressure deficit (vpdmin and vpdmax)
- elevation (Elev)
- total global shortwave solar radiation on a horizontal surface (soltotal)

The PRISM climate group gathers observations from a variety of climate stations with the goal of identifying short and long term patterns in the climate variables over time (Oregon State University, 2014). The geographic and temporal resolution varies by product. PRISM provides yearly, monthly, and annual maps at a four kilometer resolution for free.

PRISM maps are used for the climate information provided to make a map of predictions from a model. The variables used in the model were PPT and Elev. Mean or average temperature were considered in the model for predicting SWE. Overall, this data source provides a lot of information and numerous opportunities for exploration. The link below provides options to download data manually and more information on available products offered by PRISM:

https://prism.oregonstate.edu/

2.3 University of Arizona

UA provides daily estimate maps of snow water equivalent and snow depth in the United States. These maps date back to October 1981 and are in four kilometer resolution. UA estimates are derived using climate variables from PRISM along with SWE and snow depth measurements from SNOTEL, Snow Course, and cooperative observer network (COOP) stations. All of these are used as input into a snow density model that provides the final, mapped estimates (Broxton et al., 2017).

When compared to the on-site measurements, the UA tends to underestimate. One possible explanation is due to the inability to characterize the sharp changes in mountainous snowpack using a four kilometer resolution grid. One potential remedy is to utilize the available climate variable maps, like elevation, at a finer resolution to downscale map estimates. These efforts to improve UA are not explored in this

thesis. More information about the UA product is made available at the following website:

https://nsidc.org/data/nsidc-0719/versions/1#

2.4 Daymet

Daymet provides daily, monthly, and annual maps at a one kilometer resolution for North America since 1980 and Puerto Rico since 1950. It includes the following variables:

- precipitation (PPT)
- SWE
- minimum and maximum temperature (Tmin and Tmax)
- vapor pressure (VP)
- shortwave radiation (SR)

Estimates are provided through statistical modeling techniques that use ground-based observations as input. It uses a simple model to estimate SWE and admits that the purpose of these estimates is to validate other Daymet outputs, rather than to use their SWE estimates directly. The authors actually encourage researchers to derive their own SWE models if that is the primary variable of interest (P. E. Thornton et al., 2021). Preliminary analysis of Daymet for April 1st 2015 confirmed that SWE predictions were too inaccurate as predictions were about half of the values recorded at SNOTEL station locations. However, Daymet provides other temperature and climate-related variables that could be used to downscale other SWE products, such as the UA data. More information about Daymet and its available data offerings is available at the following link:

Table 2.2: SNOTEL sites

State	Number of Sites
Alaska	77
Arizona	23
California	34
Colorado	116
Idaho	86
Montana	92
New Mexico	29
Nevada	56
Oregon	81
South Dakota	2
Utah	125
Washington	76
Wyoming	89

.

2.5 SNOTEL

Manual collection of snowpack data has been done since 1920's and Utah has records that date back to 1912 (Julander and Bricco, 2006). The automated versions or SNOTEL sites began to be installed in the late 1970's and early 1980's and provide information on snowpack. Table 2.2 shows there are 896 sites spread across Utah, Arizona, California, Colorado, Idaho, Montana, New Mexico, Nevada, Oregon, Washington, Wyoming, Alaska, and South Dakota Hufkens, 2020. Figure 2.1 shows the locations of SNOTEL stations in the United States excluding stations in Alaska and South Dakota.

The goal of the SNOTEL network is to "provide information for runoff volume forecasting using empirical relationships between point values of SWE, antecedent soil moisture, historical temperature and precipitation data and observed runoff." (Molotch and Bales, 2005). SNOTEL stations provide hourly and daily measurements

of the following variables:

- precipitation (PPT)
- temperature
- snow depth (SnD)
- snow water equivalent (SWE)

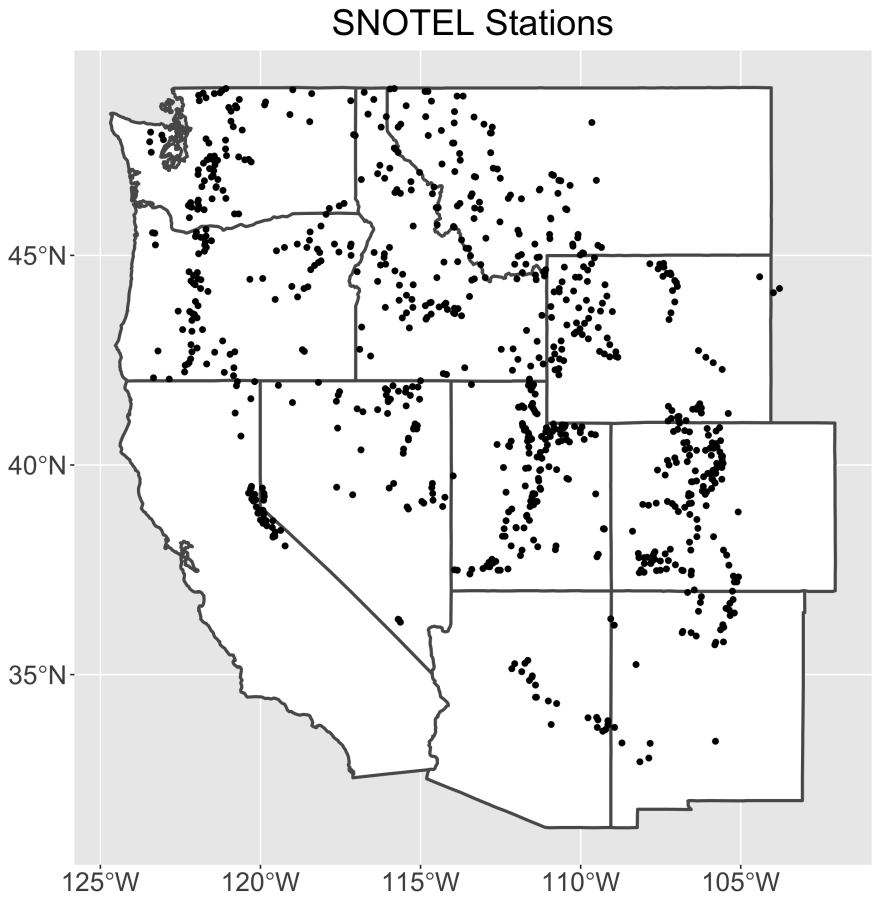


Fig. 2.1: SNOTEL sites throughout the western United States. The sites in Alaska and South Dakota aren't pictured.

SNOTEL stations estimate the rate of snowmelt by utilizing climate variable like temperature, weight, SnD, and SWE. The SnD measurements at these sites aren't

as accurate as the SWE readings. SWE measurements from these sites are used to validate gridded SWE estimates. It is important to note that the SWE values collected at SNOTEL stations may not be representative of the average value in the gridded one kilometer area. For example, a SNOTEL station covered in shade while the rest of the surrounding landscape lacks shade cover will over represent SWE. However, the SWE observation that SNOTEL stations provide are the best estimate for the "true" ground SWE values. SNOTEL data will be useful in creating GAMs or other models to predict SWE based on variable information.

2.6 Software

Each data source provides user guides on how to store, access, and download their products. Additionally, there are other websites, GitHub and Comprehensive R Archive Network (CRAN), or platforms designed to store and share software capable of performing certain tasks through functions and R packages. R is an open-source software built for to manipulate and store data, compute statistics, and generate graphics. All graphics and computations in this thesis were completed in R (R Core Team, 2021). After considerable effort of searching on GitHub and CRAN, there were multiple R packages able to facilitate direct downloading of data from Daymet and SNODAS into R. However, there were no packages or functions available to download the UA data. In order to access the UA data, the user must create an Earthdata account to manually download the data, which is provided via nc files that contain an entire calendar year of data.

The R package `daymetr` (Hufkens et al., 2018) was used for downloading daymet data. Daymetr is supported by CRAN, which generally means that the software is less error-prone and likely to be supported long term. `snowdl` (Smith, 2021) is another R package on GitHub that has the capability to download Daymet and SNODAS.

However, the robustness of the `snowdl` is in question. After downloading two specific days of data and substantial trouble shooting, one of the two downloaded days yielded an error and wasn't able to download data for the other day.

One method to download PRISM data is using the R package `prism` (Hart and Bell, 2015), available on CRAN. This package was effective and efficient in downloading PRISM data. It allows the user to set up a directory or location to store all the PRISM files and has functions to quickly plot, find, and manage the climate data files.

The following packages are capable of downloading SNODAS data:

- `snowdl` (Smith, 2021)
- `snodasr` (Marinos, 2020)
- `rwrfhydro` (National Center for Atmospheric Research, 2019)
- `NOHRSC_SNODAS` (Burakowski, 2017)

`NOHRSC_SNODAS` used only bash scripts and was not useful in downloading data in R. `rwrfhydro` has multiple functions that allowed for downloading data but are difficult to use. `snodasr` and `snowdl` both had general public licenses (GPL) and useful code. `snowdl` struggled to download data from daymet but proved rather effective with downloading SNODAS data and provided an option to download in parallel. `snodasr` was a user-friendly package with simple and effective functions that provided insight on downloading data with date-specific issues that other packages failed to address.

An R package `rsnodas` was created as part of this thesis and incorporates functionality from `snowdl` and `snodasr` while adding more options of downloading other types of data that SNODAS offers. `rsnodas` allows users to use one package to download data from SNODAS, PRISM, and SNOTEL stations. Supplemental instructions

and examples of using `rsnodas` will be demonstrated in the Appendix **??**. A development version of the R package `rsnodas` is available on GitHub and can be found at https://github.com/lschneider93/rsnodas

In contrast to the gridded climate products, the land-based SNOTEL measurements can also be downloaded using an R package, `snotelr` (Hufkens, 2020). This package is user-friendly and provides examples of how to download information for locations. Although easy to use, one problem was comparing the downloaded data from `snotelr` to another function available in the created GitHub package `rsnodas`. `snotelr` reports values one position after the decimal point while `rsnodas` reports 2 values after the decimal point. This difference between these values is due to rounding.

CHAPTER 3

METHODOLOGY

Ensemble modeling is a process that includes averaging the output from independently created models to estimate an outcome of interest. Ensemble models aggregate predictions from each base model in hopes of reducing the generalization error of the final estimate (Kotu and Deshpande, 2014). The general idea behind the ensemble process is the belief that each individual model is unbiased, but collectively the average knowledge of the models can result in the reduction of noise, variance, and bias. Consolidating multiple model outputs effectively requires identifying which model(s) perform best in a given situation. A common ensemble approach is taking the average of the outputs or adding weights to each output based on some criteria like distance or similarity. The ensemble approach in this thesis involves combining a map of estimates from a GAM using ground-based measurements of the snow pack with gridded SWE products from SNODAS. The predicted maps are weighted based on the relative spatial density of the the SNOTEL measurement locations. This means more weight will be given to the GAM outputs in areas with a higher density of land-based stations. The dynamic ensemble approach outlined in this thesis creates a framework for future, more complex, spatial ensemble approaches that combine larger sets of gridded and land-based model estimates.

3.1 Generalized Additive Models

This dynamic ensemble approach involves building a model based on in-situ measures and converting them to spatial gridded predictions. This thesis uses a generalized additive model (GAM) with splines on the sphere (SOS) to create these maps. A

GAM can account for non-linear effects but uses a framework similar to linear regression. The method replaces the coefficients (β_i) of a linear regression with smoothing functions (s_i). Each smoothing function is "fit using a scatterplot smoother (e.g., a cubic smoothing spline or kernel smoother), and provide an algorithm for simultaneously estimating all smoothing function" (Hastie et al., 2009). These functions allow GAMs to fit a large variety of data and can be used in conjunction with regularization to avoid overfitting. This is often accomplished using some form of penalized regression and leave one out cross-validation which helps generalize the model to account for new data points. (S. Wood, 2017)

One thing to note is that GAMs do not model interaction terms and aren't as easy to interpret as linear regression models because there is no parametric model summary. However, GAMs can provide understanding about the relationships between response and explanatory variables when compared to other machine learning algorithms like random forest or support-vector machines.

Let $\boldsymbol{u_\alpha}$ represent $\alpha = 1, \ldots n$ measurement locations (either a pair of lat/long values for geographical coordinates or a pair of x/y values for Cartesian coordinates) and let $\boldsymbol{u^*}$ represent an arbitrary location of interest. Further, let $\boldsymbol{X}$ represent an $n \times p$ matrix of explanatory variables (not including geographical location) and x_1 represents the first column of $\boldsymbol{X}$. Finally, let $\hat{z}_l(\boldsymbol{u^*})$ denote the estimate of the variable of interest from the model at the location of interest using the land-based observations. The predictions from the GAM model can then be represented mathematically as

$$\hat{z}_l(\boldsymbol{u^*}|\boldsymbol{X}(\boldsymbol{u_\alpha})) = s_0 + s_1(x_1) + s_2(x_2) + \ldots + s_k(x_p). \tag{3.1}$$

Spatial modeling assumes that there is predictive power in the location of the observations beyond that captured in traditional explanatory variables. A geographic smoothing spline with position (s_{sos}), using the longitude and latitude, will be added

to the equation 3.1. This additional term, s_{sos}, models spatial patterns on a spherical surface using a "splines on the sphere" approach (Wahba, 1981; S. N. Wood, 2003). The end result will be an estimation of variable at a location using position and climate variables in the represented as

$$\hat{z}_l(\boldsymbol{u}^*|\boldsymbol{X}(\boldsymbol{u}_{\boldsymbol{\alpha}})) = s_0 + s_1(x_1) + s_2(x_2) + \ldots + s_k(x_k) + s_{sos}(\boldsymbol{u}^*) + \epsilon_i. \tag{3.2}$$

Other mapping approaches, including random forest and linear regression, were used in predicting SWE. The random forests highlighted variables of importance such as Elev, PPT, but failed to provide smoothly varying characterizations of the landscape. Linear regression worked well at some locations but struggled to have the flexibility to describe the climate patterns throughout the state of Utah. GAMs proved to explain some SWE distribution patterns while providing less errors in cross-validation. The generalized additive model implemented in this thesis included the variables of latitude and longitude, elevation, slope, annual precipitation, and aspect or orientation.

3.2 Ensemble Methodology

After determining the satellite predicts and ground based model(s) that will be used in interpolation. The second step of dynamic ensemble approach is to determine the optimal weights for combining satellite and land model predictions. The previous subsection describes how to get the estimate of the land-based observations, $\hat{z}_l(\boldsymbol{u}^*)$. There are other satellite estimates $\hat{z}_s(\boldsymbol{u}^*)$ provided by other services such as SNODAS, UA, Daymet, or any combination of the three. The ground or land-based model predictions are weighted, based on observation density ($f(\rho)$), and combined

with other gridded predictions. This method can be used over any spatially reference area ($\mathcal{R}$) in which there are multiple model estimates and land-based stations taking measurements. Let $\hat{z}(\boldsymbol{u}^*)$ represent the estimate of the variable of interest and with the subscripts of c, l, and s. Each subscript represents the following: the combined estimate, the land-based model estimate ($\hat{z}_l(\boldsymbol{u}^*)$), and the other satellite source estimate ($\hat{z}_s(\boldsymbol{u}^*)$). Below is an equation of how to calculate the combined estimate of the variable of interest at an arbitrary location.

$$\hat{z}_c(\boldsymbol{u}^*) = f(\rho(\boldsymbol{u}^*)) \cdot \hat{z}_l(\boldsymbol{u}^*|\boldsymbol{X}(\boldsymbol{u}_\alpha)) + (1 - f(\rho(\boldsymbol{u}^*))) \cdot \hat{z}_s(\boldsymbol{u}^*) \tag{3.3}$$

One way to allow for other multiple satellite source predictions like UA and Daymet is by expanding and adding models, m, and weights, λ_i to each of those satellite estimates. This is represented mathematically as

$$\hat{z}_c(\boldsymbol{u}^*) = f(\rho(\boldsymbol{u}^*)) \cdot \hat{z}_l(\boldsymbol{u}^*|\boldsymbol{X}(\boldsymbol{u}_\alpha)) + (1 - f(\rho(\boldsymbol{u}^*))) \cdot \sum_{i=1}^{m} \lambda_i \hat{z}_{s,i}(\boldsymbol{u}^*) \tag{3.4}$$

$$\text{where } 0 \leq f(\rho(u)) \leq 1,\ 0 \leq \lambda_i \leq 1, \text{ and } \textstyle\sum_{i=1}^{m} \lambda_i = 1.$$

The function, $f(\rho(\boldsymbol{u}^*))$, applied to the location's point density ($\rho(\boldsymbol{u}^*)$ must be monotonically increasing so as to give less weight to land based model predictions as station density decreases. The output, a value in the interval of [0, 1], determines the weight or percentage of the estimate that will be given to the land based model. The remaining percentage, $1 - f(\rho(\boldsymbol{u}^*))$, is going to be given to the predictions of the other satellite source model, $\hat{z}_s(\boldsymbol{u}^*)$.

One challenge with this method is that the weighting is almost completely arbitrary. This allows for exploration of nearly endless and different weighting schemes

and functions. However, there should a process to incorporate data-driven rules to determine the weighting scheme and or function applied to the density. The data-drive rules could use information from previous year differences or a measure of similarity between model outputs.

CHAPTER 4

RESULTS

The process of this ensemble approach was used to predict the April 1st SWE estimates from 2004 to 2021. This approach involved combining the SNODAS and GAM model predictions. The theory of combining SNODAS and UA was explored but smaller errors were obtained when UA was excluded. Different weight schemes could implement UA SWE predictions and Daymet climate variables in the future. Figure 4.1 compares the SWE estimates of SNODAS and UA in 2015.

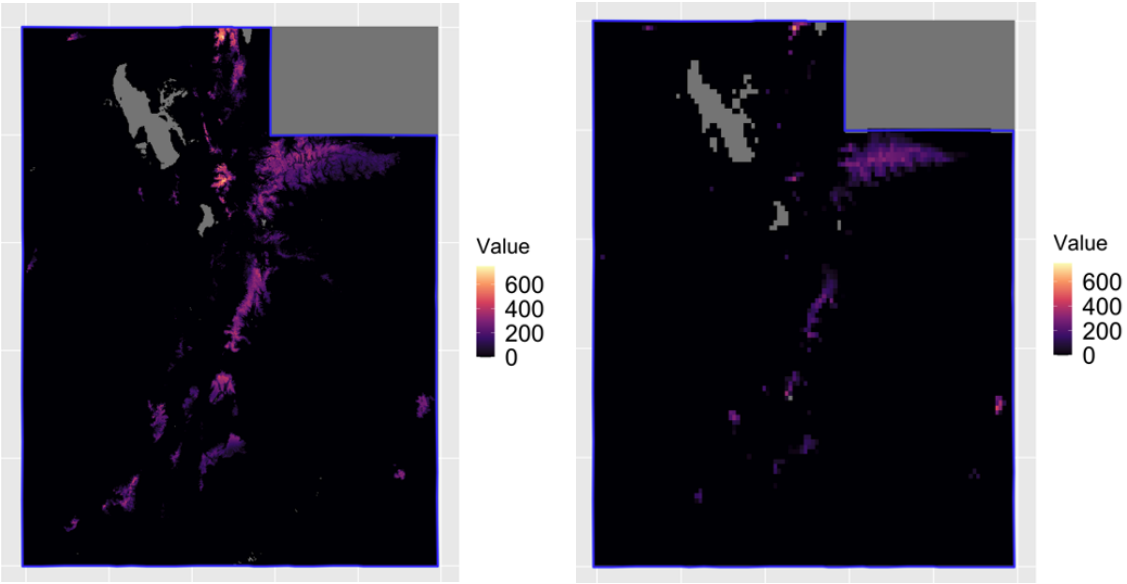


Fig. 4.1: (Left to right:) SNODAS and UA Gridded Products of SWE predictions in Utah for 2015.

A GAM based on SNOTEL stations and using PRISM climate variables produces a map of predictions throughout Utah. This model uses the variables, annual precipitation, elevation, slope, aspect and location. A density map is calculated based

on station observation locations. A simple function utilizes the density to return an output between 0 and 1 that determines the weight given to the GAM model. The function takes the density value at a location and divides by the max density value. Effectively, 100% is the max weight the ground-based model is allowed.

Before a comparison of each model, it is crucial to understand how errors are calculated. In-situ or ground-based measurements are the primary way to validate satellite-derived maps, such as SNODAS and UA, but site specific measurement conditions may make it such that the land-based measurement is not be representative of the surrounding landscape. The result from subtracting the model predicted value from the ground based value measures the error. A positive residual or error denotes an under-prediction and a negative error denotes an over-prediction. In order to improve the accuracy of SWE predictions, this ensemble process explores using different schemes and compares the residuals with the three models (SNODAS, GAM, and the Ensemble). The max weight given to the GAM is 100% and the blended model predictions will shift percentages between the GAM and SNODAS model by giving more or less depending on station density.

As previously discussed, this weighting approach is intuitive, but arbitrary. There isn't a real technique or criteria to establish the weights given to each model. One way to try to incorporate data into deciding the weights would be incorporating previous year(s) differences.

4.1 Cross-validated Results

10 fold cross-validation validates the results from the ensemble approach. This process involves using 90% of the data to predict on the remaining ten percent and repeats this procedure ten times. This type of cross validation provides an accurate representation of the bias and variance, even with a small dataset, that is present in

the model. The figures below illustrate the bias and variance of each model.

It is important to recognize that the variance of the errors decreases with time for all models. GAM has the highest variance and SNODAS and the ensemble does stays right with SNODAS. The big difference is comparing SNODAS and the combined SNODAS and GAM model in 2014 and 2022. Figure **??** shows SNODAS displayed the highest variance in the errors and the ensemble method had the lowest variance for both 2014 and 2022. This suggests that the ensemble approach is more stable and less prone to having extreme difference in the errors between years.

A similar story is reflected in the medians of the errors. The SNODAS and GAM models have years when they are unbiased but the median of the errors fluctuates between higher and lower values. The stability gained from the ensemble methodology is promising, but does display a slight underestimation bias. Although the ensemble does generally underpredict, it is important to note that these predictions are estimating for a 1km area, and it seems as though SNOTEL stations tend to generally receive higher accumulations of snow than the 1km squares that contain them. These are at two different spatial resolutions and it is difficult to decide, without more SWE measurements, how accurate the ensemble model preforms.

The spread of the errors for each models varies from year to year but the ensemble tends to exhibit similar outliers and variance as SNODAS. The GAM interquartile range of the residuals always contains zero, indicating less severe bias, but displays the highest variance of the models. The residual distributions in 2015 and 2020 of each model are depicted in Figure **??**. In 2015, the errors from the GAM model look approximately normal and unbiased because the boxplot is centered around zero and roughly symmetric. It is important to note that the GAM model residuals display a much larger interquartile range. The distribution of the errors from SNODAS has the smallest interquartile range but looks to be biased since zero isn't contained in

the interquartile range. The ensemble model's has qualities of both the GAM and SNODAS error distributions seeing that the interquartile range contains zero but covers a larger range of values compared to SNODAS. Each model displays a similar number of outliers. The residual distributions in 2020 shows that SNODAS performed the best because it isn't bias and exhibits the smallest variance but exhibits two extreme outliers. The Ensemble method has a slightly larger interquartile range and has outliers slightly less extreme compared to SNODAS. The GAM residuals look to be approximately normal and non-biased but displays the largest interquartile range and has two extreme outliers

Ultimately, this ensemble approach displays marginal improvements of variance and bias. This evidence shows that the ensemble models are a promising approach for improving SWE predictions in Utah, but more work needs to be done to determine the best function for assigning weights based on station density. This work demonstrates that a combination of models can develop better predictions overall.

CHAPTER 5
CONCLUSIONS

5.1 Challenges

A common challenge with new strategies for analyzing spatial data is scalability. The process of creating a grid of estimates at a one km resolution or finer resolution is computationally expensive. However, this issue can be mitigated by focusing on improving predictions for smaller regions, and deferring fully to national products for continental scale analyses. Another issue to consider is the potential loss of valuable information when averaging all the predictions in the ensemble process. This is especially true when individual models are superior to the other models in the ensemble, making the user better off to use a single product rather than an average of many. One potential fix is to incorporate another criteria of previous year accuracy comparisons. This could be something to help weigh the model and help determine regions that models generally under or over-predict.

Another challenge is that there are between 80 and 130 land-based observations throughout Utah in a given day. This is a relatively small amount of data for predicting over all of Utah. The scale at which snow changes is finer than current available measurements and the process of validating the results requires leaving out a percentage of these observations. The removal of 10% of the observations won't make a major difference in predictability, but in practice the ensemble model will include every in-situ observation and hopefully improve the accuracy and lower variance. However, the general lack of ground-based observations of SWE makes it difficult to effectively validate the land-based models we produce, let alone the ensemble models.

Another ongoing challenge is determining best practices for handling NA's or missing values in mapped values. The current ensemble approach propagates missing values included in any one of the ensemble predictions. For example, UA doesn't provide an estimate for Burt Miller Ranch. Any ensemble that incorporates UA won't yield any estimates at that location. One option is to change the NA's to 0's but often the NA's reflect a body of water like the Great Salt Lake. In the case of an NA, that location could be checked on another satellite resources to verify if it is NA or not.

5.2 Future Work

There is a lot of exploration available with this methodology because this is a process of combining the outputs of different models. The first thing is to try to implement more complex weighting schemes of more than two SWE mapped estimates. This framework allows for the dynamic weighting to be based on criteria besides station density, such as the level of agreement of model predictions of SNODAS and UA or other data sources. For example, instead of implementing the GAM prediction by the station or observation density, Christensen and Sain (2012) suggested that the weights could be determined by some measure of covariance between the mapped values. Additional opportunities for improvement include the use of additional climate variables like to help overcome the sparse spatial distribution of weather stations measuring SWE directly. There are multiple ways to do interpolation and the performance of other types of models and utilizing other variables could be compared. Some variables of interest would be vapor pressure, max temperature from C.1, vegetation index, and using information from historical trends. Future work might also incorporate the agreement and disagreement of the already established models. For example, if the SNODAS and UA have a big disagreement, then we could give more

weight to ground-based models. This could be coupled with historical difference fields at that location. Other studies utilized previous year differences in their model and that strategy could be implemented in this approach by including previous years difference.

5.3 Reflection

Droughts are becoming an increasing concern throughout the US and especially in Utah. The ability to accurately estimate the amount of water from the mountainous snowpacks is crucial for summer water management. The current models and methods of predicting SWE have challenges that have been identified in this thesis. Each of these sources showed evidence of bias and high errors due to the complexity of predicting SWE. This thesis has shown that the ensembling model approach displayed a decrease in standard deviation of the errors and promise in mitigating the weaknesses of individual data sources, while still preserving their strengths.

The software available to download and access the data sources involved multiple R packages. One key contribution of this thesis was the amalgamation of multiple sources of code to one package that allows users to download, store, access and explore the ensemble methodology. This approach outlines and highlights the value of combining remote and land observations to better estimate SWE. This dynamic ensembling approach can be extended to other spatial variables that has a larger groups of individual models. Another benefit is the framework of this method has the flexibility to try different weighting schemes and the possibility to try criteria for the dynamic weighting. The R package, `rsnodas`, discussed in this thesis provides a framework of combining gridded and point data based on spatial density. The functions in `rsnodas` allow users the ability to download climate data from a variety of sources, calculate spatial density, and create a map of predictions. This allows

for exploration and optimizing the data sources' information to make better SWE predictions.

5.4 Software Acknowledgements

The ensemble approach in this thesis attempts to improve current products' SWE predictions in Utah by utilizing SNOTEL station data. This is accomplished by understanding the current software available for downloading the data. All computations and figures in this thesis were completed in R (R Core Team, 2021) using the following spatial statistics packages: sf (Pebesma, 2018), stars (Pebesma, 2021), terra (Hijmans, 2021), and spatstat (Baddeley and Turner, 2005). The package dplyr (Wickham et al., 2022), was used for data manipulation and wrangling, and ggplot2 (Wickham, 2016) was used for data visualizations.

A portion of this thesis involves the software contributions on GitHub at the following https://github.com/lschneider93/rsnodas. The R package, `rsnodas`, allows users to download data from SNODAS, Daymet, and PRISM maps and variables. Functions from the `rsnodas` package are used to create density maps of observations and generating a map of estimates given data points. `rsnodas` has the capabilities to read in the SNOTEL station data and extract the SWE information. Most of the figures presented in this thesis incorporated functions from this package in order to obtain results.

Bibliography

Baddeley, A., & Turner, R. (2005). spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software*, *12*(6), 1–42. https://doi.org/10.18637/jss.v012.i06

Barrett, A. P. (2003). *National operational hydrologic remote sensing center snow data assimilation system (SNODAS) products at NSIDC*. National Snow; Ice Data Center, Cooperative Institute for Research.

Bohr, G., & Aguado, E. (2001). Use of April 1 swe measurements as estimates of peak seasonal snowpack and cold-season precipitation. *Water Resources Research - WATER RESOUR RES*, *37*, 51–60. https://doi.org/10.1029/2000WR900256

Broxton, P. D., Zeng, X., & Dawson, N. (2017). The impact of a low bias in snow water equivalent initialization on CFS seasonal forecasts. *Journal of Climate*, *30*(21), 8657–8671.

Burakowski. (2017). Nohrsc basnodas [data retrieved from GitHub, https://github.com/eaburakowski/NOHRSC_SNODAS].

Christensen, W. F., & Sain, S. R. (2012). Latent variable modeling for integrating output from multiple climate models. *Mathematical Geosciences*, *44*(4), 395–410. https://doi.org/https://doi.org/10.1007/s11004-011-9321-1

Clow, D. W., Nanus, L., Verdin, K. L., & Schmidt, J. (2012). Evaluation of SNODAS snow depth and snow water equivalent estimates for the Colorado Rocky Mountains, USA. *Hydrological Processes*, *26*(17), 2583–2591. https://doi.org/https://doi.org/10.1002/hyp.9385

Frei, A., Tedesco, M., Lee, S., Foster, J., Hall, D. K., Kelly, R., & Robinson, D. A. (2012). A review of global satellite-derived snow products [Oceanography,

Cryosphere and Freshwater Flux to the Ocean]. *Advances in Space Research*, *50*(8), 1007–1029. https://doi.org/https://doi.org/10.1016/j.asr.2011.12.021

Gochis, D., Barlage, M., Dugger, A., FitzGerald, K., Karsten, L., McAllister, M., McCreight, J., Mills, J., RafieeiNasab, A., Read, L., et al. (2018). The wrf-hydro modeling system technical description,(version 5.0). *NCAR Technical Note*, *107*.

Hart, E. M., & Bell, K. (2015). *PRISM: Download data from the oregon prism project* [R package version 0.0.6]. https://doi.org/10.5281/zenodo.33663

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). Springer.

Hijmans, R. J. (2021). *terra: spatial data analysis* [R package version 1.4-22]. https://CRAN.R-project.org/package=terra

Hufkens, K. (2020). *Snotelr: Calculate and visualize 'snotel' snow data and seasonality* [R package version 1.0.4]. https://CRAN.R-project.org/package=snotelr

Hufkens, K., Basler, D., Milliman, T., Melaas, E. K., & Richardson, A. D. (2018). An integrated phenology modelling framework in r: Modelling vegetation phenology with phenor. *Methods in Ecology & Evolution*, *9*, 1–10. https://doi.org/10.1111/2041-210X.12970

Julander, R. P., & Bricco, M. (2006). An examination of external influences imbedded in the historical snow data of Utah. *All US Government Documents (Utah Regional Depository)*, 116. https://doi.org/https://digitalcommons.usu.edu/govdocs/116

Kinar, N., & Pomeroy, J. (2015). Measurement of the physical properties of the snowpack. *Reviews of Geophysics*, *53*(2), 481–544. https://doi.org/https://doi.org/10.1002/2015RG000481

Kotu, V., & Deshpande, B. (2014). *Predictive analytics and data mining: Concepts and practice with rapidminer*. Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/C2014-0-00329-2

Lv, Z., Pomeroy, J. W., & Fang, X. (2019). Evaluation of snodas snow water equivalent in western Canada and assimilation into a cold region hydrological model. *Water Resources Research*, *55*(12), 11166–11187. https://doi.org/https://doi.org/10.1029/2019WR025333

Marinos. (2020). snodasr [data retrieved from GitHub, https://github.com/marinosr/SNODASR].

Molotch, N. P., & Bales, R. C. (2005). Scaling snow observations from the point to the grid element: Implications for observation network design. *Water Resources Research*, *41*(11).

Mote, P. W., Hamlet, A. F., Clark, M. P., & Lettenmaier, D. P. (2005). Declining mountain snowpack in western North America*. *Bulletin of the American Meteorological Society*, *86*(1), 39–50. https://doi.org/10.1175/BAMS-86-1-39

National Center for Atmospheric Research. (2019). Rwrfhydro [data retrieved from GitHub, https://github.com/NCAR/rwrfhydro].

Oregon State University. (2014). Prism climate group [data retrieved from PRISM, https://prism.oregonstate.edu/].

Pebesma, E. (2018). Simple features for R: Standardized support for spatial vector data. *The R Journal*, *10*(1), 439–446. https://doi.org/10.32614/RJ-2018-009

Pebesma, E. (2021). *stars: spatiotemporal arrays, raster and vector data cubes* [R package version 0.5-5]. https://CRAN.R-project.org/package=stars

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Serreze, M. C., Clark, M. P., Armstrong, R. L., McGinnis, D. A., & Pulwarty, R. S. (1999). Characteristics of the western United States snowpack from snowpack telemetry (SNOTEL) data. *Water Resources Research*, *35*(7), 2145–2160. https://doi.org/https://doi.org/10.1029/1999WR900090

Shuman, C. S., & Ambrose, R. F. (2003). A comparison of remote sensing and ground-based methods for monitoring wetland restoration success. *Restoration Ecology*, *11*(3), 325–333. https://doi.org/https://doi.org/10.1046/j.1526-100X.2003.00182.x

Smith. (2021). Snowdl [data retrieved from GitHub, https://github.com/bsmity13/snowdl].

Thornton, M., Shrestha, R., Thornton, P., Kao, S., Wei, Y., & Wilson, B. (2021). Daymet v4 daily data for the previous month. https://doi.org/10.3334/ORNLDAAC/1904

Thornton, P. E., Shrestha, R., Thornton, M., Kao, S.-C., Wei, Y., & Wilson, B. E. (2021). Gridded daily weather data for North America with comprehensive uncertainty quantification. *Scientific Data*, *8*(1), 1–17. https://doi.org/https://doi.org/10.1038/s41597-021-00973-0

Wahba, G. (1981). Spline interpolation and smoothing on the sphere. *SIAM Journal on Scientific and Statistical Computing*, *2*(1), 5–16. https://doi.org/https://doi.org/10.1137/0902002

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., & Müller, K. (2022). *Dplyr: A grammar of data manipulation* [R package version 1.0.8]. https://CRAN.R-project.org/package=dplyr

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(1), 95–114. https://doi.org/https://doi.org/10.1111/1467-9868.00374

Wood, S. (2017). *Generalized additive models: An introduction with r* (2nd ed.). Chapman; Hall/CRC.

Yang, K., Musselman, K. N., Rittger, K., Margulis, S. A., Painter, T. H., & Molotch, N. P. (2022). Combining ground-based and remotely sensed snow data in a linear regression model for real-time estimation of snow water equivalent. *Advances in Water Resources*, *160*, 104075. https://doi.org/https://doi.org/10.1016/j.advwatres.2021.104075

APPENDICES

APPENDIX A

Package Vignette

This vignette helps users understand the capabilities of the R package, `rsnodas`. This is located on GitHub at https://github.com/lschneider93/rsnodas and allows users to read and see sample code that displays the basic functionality of the package. The software has the capability for users to utilize their own models like random forests, support vector machines, linear models and more to create gridded predictions. Figure A.1 is a raster that used a random forest model to create the predictions. The code used to create this figure is commented out code on page six of the vignette.

The software and functions in the R package `rsnodas` has the flexibility for users to utilize multiple types of models to create gridded predictions.

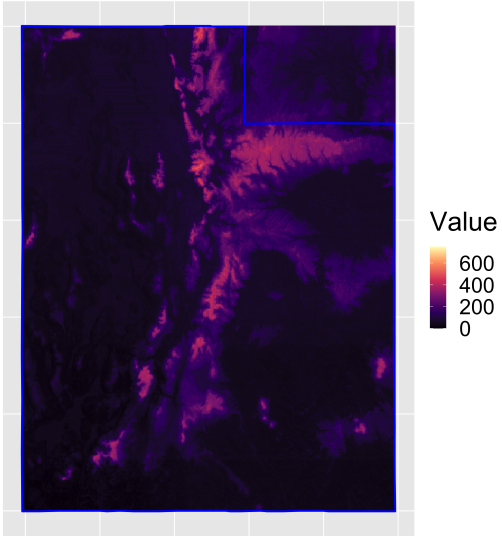


Fig. A.1: A gridded product created from a random forest model for 2015.

rsnodas vignette

Logan Schneider

July 12, 2022

Contents

1 Introduction

This rsnodas package allows users to access, clean, visualize, and analyze data from the following sources:

- Snow Data Assimulation System (SNODAS)
- University of Arizona (UA)
- Parameter-elevation Regressions on Independent Slopes Model (PRISM)
- Historical Climatology Network daily (GHCND) which provides access to Snow Telemetry (SNOTEL) stations

SNODAS, UA, and PRISM provide data in the form of gridded rasters for the contiguous United States. A summary of the variables and resolution assoicated with each data source is provided in the table below. GHCND provides in-situ measurements, spatial point data, with different temporal scales. This package focuses on downloading the daily maps and measurements from SNODAS, PRISM, and GHCND. UA data currently can't be downloaded in this package and must be downloaded manually after creating and signing into an Earthdata account. Please refer to their links for more information about each data product.

The following table lists the data products and spatial resolution of each data source. Note that like PRISM allows users to download data at the 4km resolution for free but charge a fee for their maps at a 800m when it is their daily, monthly, or yealy data. Their 30 year normal (i.e. average) maps are free at the 800m resolution and are currently from 1991-2020. These 30 year maps change from year to year and are free at the 800m resolution.

Data Source	**Resolution**	**Variables available**
SNODAS	1 km	Snow water equivalent (SWE), Snow Depth (SnD) Snow Melt runoff (SM), Precipitation (Ppt), Sublimation from Snow pack (SSP) Sublimation of blowing snow (SBS), Snow pack Temperature (SPT)
University of Arizona (UA)	4 km	SWE, SnD
PRISM	4 km 800 m*	PPT, Elevation (Elev), mean dew point (TDmean) Min/Max vapor pressure deficit (VPmin and VPDmax) Min/Max Temperature (Tmin and Tmax) total global shortwave solar radiation (SolTol)
Daymet	1 km	PPT, SWE, Shortwave radiation (SR) Tmin, Tmax, Water vapor pressure (VP)

This vignette uses functions from `tidyverse`, `ggplot2`, `sf`, and the `stars` package and will be loaded now.

```
library(tidyverse)
library(sf)
library(stars)
library(ggplot2)
```

1.1 Installation

rsnodas is available on GitHub and can be installed with devtools:

```
# install.packages("devtools")
library(devtools)
install_github("lschneider93/rsnodas")
```

The vignette will first proceed with a demonstration of how to download SNODAS, SNOTEL, and PRISM data. After downloading all of those sources of data, the vignette will show the process of creating a raster of predictions and calculating station density. The blending of SNODAS and a Generalized Additive model can then be explored.

2 Downloading SNODAS, PRISM, and SNOTEL data

2.1 SNODAS

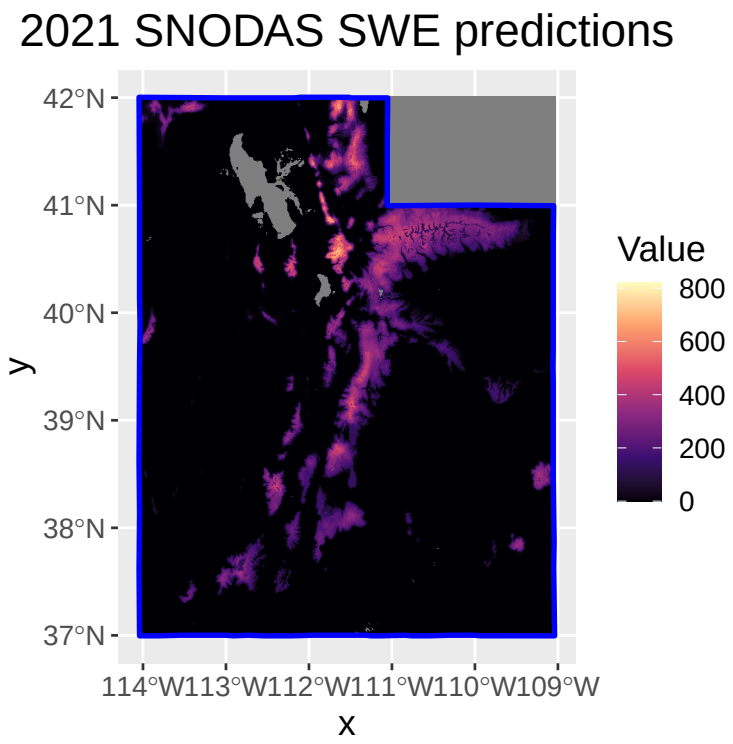
The function `download_snodas` allows us to choose the variables we want to download. Note that the dates inputted in this function need to be in 'YYYY-MM-DD' or 'YYYY-M-D' format. The function `format_date` allows users to easily create character vectors with dates in the needed format. `format_date` function can be applied inside of the `download_snodas` for mass downloading. The following example shows how to download snow water equivalent for the masked area of the US into a folder called 'snodas_data' in our working directory.

```
# Download Snodas SWE data for April 1st in 2021 and 2022
snodas_2021 <- download_snodas(dates = format_dates(day = 1,
                                                    month = 4,
                                                    year = 2021),
                masked = TRUE,
                overwrite = TRUE,
                remove_zip = TRUE,
                data_saved = c('swe'),
                out_dir = paste0(getwd(), "/snodas_data"),
                GTiff = FALSE) #
```

After downloading SNODAS, you can crop it to the area or state of interest. These next portion of code gets a shape file of Utah from the `maps` package, crops the SNODAS map to the state of Utah, and creates a visual by using `ggplot2`.

```
# Get an shape of utah from the maps package
ut_map <- maps::map("state", plot = FALSE, fill = TRUE) %>%
  sf::st_as_sf() %>%
  dplyr::filter(ID == "utah") %>%
  sf::st_transform(crs = sf::st_crs(snodas_2021[[1]]))

# Crop the maps to just the state of utah and name the values to be "Value"
snodas_ut_2021 <- sf::st_crop(snodas_2021[[1]], ut_map)
names(snodas_ut_2021) <- "Value"

# Plot SNODAS April 1st 2015 SWE map of Utah with blue outline
ggplot() +
```

```
stars::geom_stars(data = snodas_ut_2021) +
geom_sf(data = ut_map, fill = "NA", size = 1, color = "blue") +
ggtitle("2021 SNODAS SWE predictions") +
scale_fill_viridis_c(option = "A") +
theme(plot.title = element_text(hjust = 0.5, size = 18),
      text = element_text(size = 14))
```



2.2 Download PRISM climate data

The PRISM climate group has been provided climate maps since 1895 for the Precipitation (PPT), Minimum and maximum temperature (Tmin and Tmax), Minimum and maximum vapor pressure deficit (vpdmin and vpdmax), mean dew point (tdmean), Elevation (Elev), and total global shortwave solar radiation.

Most of this data from PRISM can be downloaded by utilizing the R package, `prism`, and more examples and information is available at this link. This package has three separate functions to download daily, monthly, and yearly data and store them in different file paths to keep them organized.

In comparison, we've created the `download_prism` function to seamlessly download daily, monthly, and yearly PRISM climate data and stores all the data files in one directory. This will be optimal for decreasing time and effort later to create a model for prediction. The alternatives are to manually or by use `prism` functions to download PRISM data and move all PRISM files into one directory. Currently, `download_prism` is limited to only download daily, monthly, or yearly data and can't be used to download PRISM's 30 year normal maps.

The follow example will download monthly precipitation from January 2021 until March 2021 and store in the folder called 'prism_data' in your working directory.

```
# Example of downloading 4km monthly precipitation maps from Jan. 2021 - March 2021
download_prism(sp_res = "4km", data = "ppt", t_res = "monthly",
               start_date = as.Date("2021-01-01"), end_date = as.Date("2021-03-15"),
               out_dir = paste0(getwd(), "/prism"))
```

```
## [1] "Downloading month 1 of 3"
## [1] "Downloading month 2 of 3"
## [1] "Downloading month 3 of 3"
```

Note that all PRISM files need to be in the same location in order to use the function `gam_to_df` in a later section.

2.3 Download SNOTEL

The `data-raw` folder contains an script titled `DATASET` that shows how to download the all SNOTEL data by using `download_all_ghcnd_stations`. Note that there is more than 30 gigabytes of data and this will take time to download. The functions `download_all_ghcnd_stations`, `get_station_data`, and `get_state_data` come from a currently private `snowload2` package (authors Jadon Wagstaff and Brennan Bean) and is replicated here with the permission from the authors. `get_station_data` and `get_state_data` are for the purpose of sifting through all the data to get the stations in a specific state. These functions were used to create the dataset `april_1_snotel_data` which is included as a dataset internal to the `rsnodas` package

```
# Code to download all stations in your working directory. A file path could
#   have been used instead of ".".
### This code downloads 30 GB of data and will take hour to download
download_all_ghcnd_stations(directory = ".")
```

```
# april 1st data for SNOTEL stations in Utah and subset to just April 1st, 2014.
snotel_ut <- rsnodas::april_1_snotel_data
snotel_ut_2021 <- snotel_ut[snotel_ut$DATE == "2021-04-01", ]
```

3 GAM to Data frame to Raster

This section will use the point data provided by SNOTEL sites and climate variable maps from PRISM to create estimates. The process of creating gridded raster estimates will be demonstrated in a three-step process. The first step is to create a data frame with the Longitude, Latitude, and extract PRISM climate variable information for every grid cell in the area of interest. This is accomplished by using the `gam_to_df` function. This function only looks in one directory and it is vital that all PRISM data files are in one directory.

This example shows creating a data frame with the Longitude, Latitude, and monthly precipitation values from March 2021 in Utah.

```
# creation of a data frame with all the PRISM variables of precipitation
gam_2021 <- gam_to_df(model_data = snotel_ut_2021,
                      raster_template = snodas_ut_2021,
                      path_to_prism = paste0(getwd(), "/prism"),
                      model_x <- "ppt_2021_03",
                      model_y <- "VALUE",
                      coords = c("LONGITUDE", "LATITUDE"))
head(gam_2021, 5)
```

```
##   LONGITUDE LATITUDE ppt_2021_03
## 1 -114.0458 42.00417     18.2794
## 2 -114.0375 42.00417     18.5215
## 3 -114.0292 42.00417     18.5215
## 4 -114.0208 42.00417     18.5215
## 5 -114.0125 42.00417     18.5215
```

The second step is to build a model with the SNOTEL stations data. Any model that can use the `predict` function will work here. A commmon model used for spatial data is a Generalized Additive Model (GAM). This is because GAMs account for non-linear effects and can use splines on the sphere (SOS) which account for the spherical shape of the earth. GAMs were calculated by using the function `gam` from the package `mgcv` and more information is available at this link.

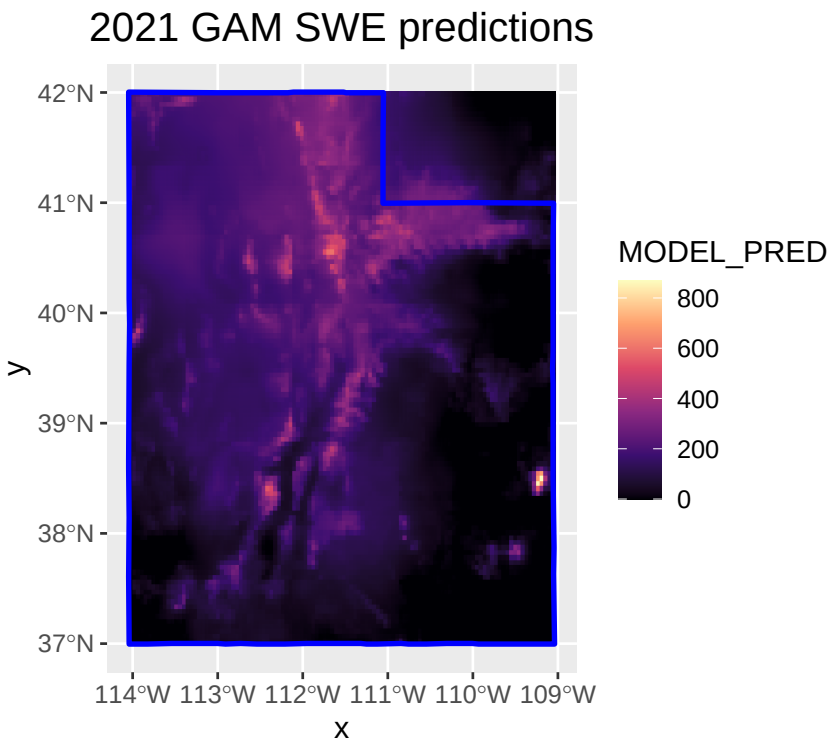
The following code creates a GAM with the SNOTEL station information using March 2021 PRISM monthly precipitation.

```
# This allows users to explore using different models with the monthly precipitation 2021.
model <- mgcv::gam(data = snotel_ut_2021,
                   VALUE ~ s(LONGITUDE, LATITUDE, bs = "sos", k = 25) +
                  s(ppt_2021_03),
                method = "REML")

# This is another example of a random forest model that uses elevation, slope and precipitation.
# NOTE: In order to run this You MUST download elevation and store it in the same folder.
# rf_model <- randomForest::randomForest(data = train.data,
#                              VALUE ~ LONGITUDE + LATITUDE + ppt_normal_annual +
#                              elevation + slope, method = "REML")
```

The final step predict the output with the variables throughout the area of interest and create a gridded output of predictions. This is accomplished by using the `df_to_raster` that takes the previously created model and data frame to create a raster.

Below is an example of creating a raster predicting snow water equivalent (SWE) using the previous model using elevation, slope and precipitation from March and the data frame with all the points in the area of Utah.

```
# After creating a model, we can make predictions of SWE with the information available.
gam_rast <- df_to_raster(model = model,
                         data_frame = gam_2021,
                         raster_template = snodas_ut_2021)

ggplot() +
  stars::geom_stars(data = gam_rast) +
  geom_sf(data = ut_map, fill = "NA", size = 1, color = "blue") +
  ggtitle("2021 GAM SWE predictions") +
  scale_fill_viridis_c(option = "A") +
  theme(plot.title = element_text(hjust = 0.5, size = 16),
        text = element_text(size = 12))
```



3.1 Station Density

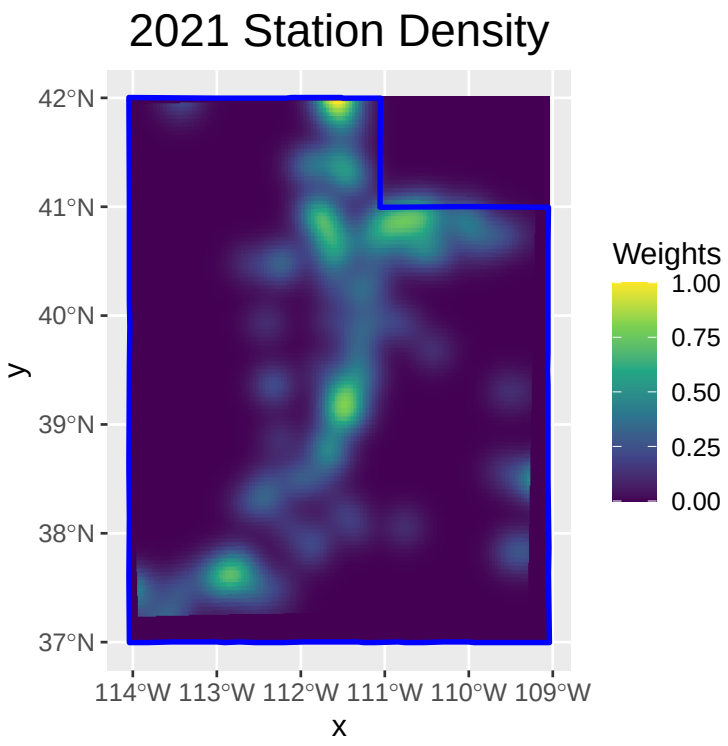
We have created a map of SWE estimates throughout Utah using point data and climate information. There are already products, like SNODAS, that estimate SWE throughout the US. SNODAS currently provides the best SWE estimates but is known to struggle predicting in mountainous regions. This map of SWE estimates created using SNOTEL stations has the potential to provide better information in mountain regions. The final map generated by blending SNODAS and the SNOTEL map would not rely solely on one prediction but on SNOTEL information and SNODAS. The goal of blending these maps is to use local information from SNOTEL sites to improve the global predictions that SNODAS provides. The combining of these maps should reduce the variance of the error of the predictions. In order to blend these maps together, we need to blend based on some criteria. This vignette will blend these two maps based on station density.

We are going to blend and weight each map by observation or station density. This uses the `points_to_density_stars` function to produce a raster of weights. The `sigma` argument will influence the kernel density calculation to include a larger or smaller distance. Changing the `max_weight` argument gives more weight to the Generalized Additive model or the model from the in-situ measurements. The max weight needs to be within the range of 0-1. This weight is a percentage and won't allow negative percentages or percentages over 100.

This creates a density map of all the SNOTEL stations in 2022.

```
dens_2021 <- points_to_density_stars(sp_points = snotel_ut_2021,
                                     coords = c("LONGITUDE", "LATITUDE"),
                                     raster_template = snodas_ut_2021,
                                     sigma = 15000,
                                     max_weight = 1,
                                     flat_crs = "+proj=utm + zone=12 + datum=WGS84")
```

```
names(dens_2021) <- "Weights"

ggplot() +
  stars::geom_stars(data = dens_2021) +
  geom_sf(data = ut_map, fill = "NA", size = 1, color = "blue") +
  ggtitle("2021 Station Density") +
  scale_fill_viridis_c(option = "D") +
  theme(plot.title = element_text(hjust = 0.5, size = 18),
        text = element_text(size = 12))
```



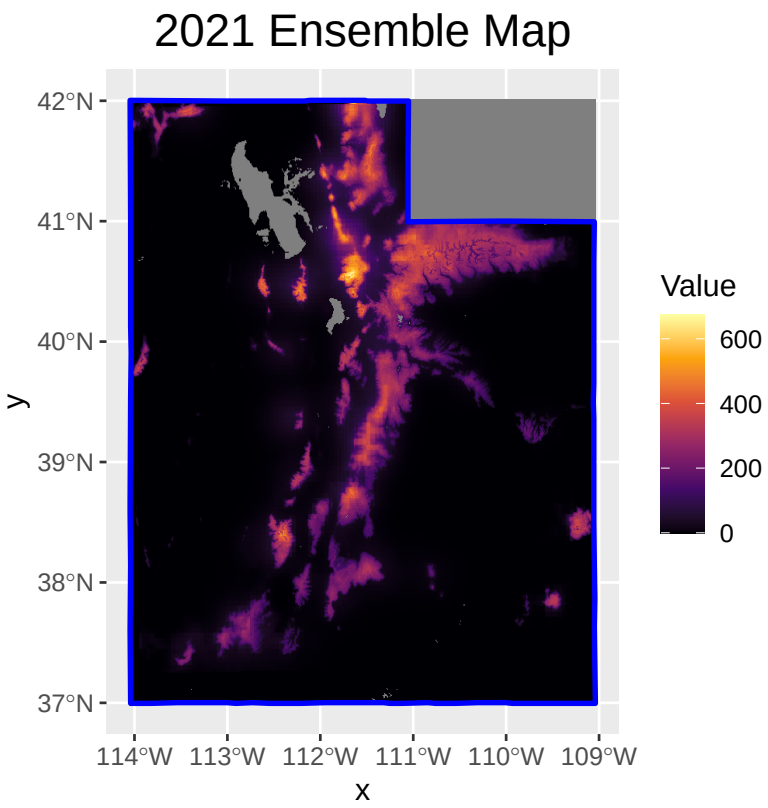
It is important to note that the station density follows the Wasatch Mountains. The data collected from SNOTEL stations are important to predicting how much water is contained in the Snowpacks for flood forecasting and water management. Therefore, it is important that we understand how much water is in the mountain snowpacks.

3.2 Ensemble SNODAS and GAM predictions

Lastly, the `blend_raster` function will blend the rasters together with the weights associated. This function could be used to apply weight to any two raster and can be applied to multiple types of spatial problems. It isn't necessary to use the GAM for the land-based predictions if that doesn't apply to your problem. Currently, this function only allows the blending of two maps together. We are going to use all the previous maps to create a final prediction or estimate of SWE.

This is blending the SNODAS and GAM model based on the SNOTEL station density.

```
comb_map <- blend_raster(raster_sate = snodas_ut_2021,
                         raster_land = gam_rast,
                         weights = dens_2021)

names(comb_map) <- "Value"
ggplot() +
  stars::geom_stars(data = comb_map) +
  geom_sf(data = ut_map, fill = "NA", size = 1, color = "blue") +
  ggtitle("2021 Ensemble Map") +
  scale_fill_viridis_c(option = "B") +
  theme(plot.title = element_text(hjust = 0.5, size = 18),
        text = element_text(size = 12))
```



4 Conclusion

This package allows users to download, store, and access multiple types of information available from SNODAS, PRISM, and SNOTEL. These data sources can explore multiple types of modeling techniques for spatial data with the objective to provide local improvements to currently availabe national gridded climate and snow products.

APPENDIX B

Data Limitations

There were multiple data sources introduced in the thesis like SNODAS, University of Arizona (UA), and Daymet. However, SNODAS was the only data source used in the ensemble process. This appendix gives an explanation why the other data sources, UA and Daymet, were excluded in the combination of SNODAS and the generalized additive model.

SNODAS and Daymet provide gridded products at a one kilometer resolution while the University of Arizona (UA) provides estimates at a four-kilometer resolution. This difference in resolution translates to UA providing less information or estimates than Daymet or SNODAS. Additionally, UA estimates are over a larger area and struggle to capture the sharp changes of SWE in the mountain snowpack. The averaging of estimates over a larger area results, in general, with an under prediction of SWE when compared to the in-situ recorded measurements. This under estimation is shown in Figure B.1 for UA. Daymet has the same resolution as SNODAS but utilizes a simple model to create the SWE estimates that fail to characterize the mountain snowpack in Utah. Figure B.1 supports the claim that the model Daymet uses struggles to explain the variability of SWE because of the high variability in the standard deviations of the error combined with a high median error value.

Figure B.1 displays the standard deviation and median of the errors for SNODAS, UA, and Daymet. The median value of errors in UA is smaller than Daymet for every year after 2004 except 2007, 2013, and 2015 and stays in a relatively small range of values that is comparable to SNODAS. UA median value of the errors is always positive and hints to UA being biased and under predicting SWE. This is

sufficient evidence to not include UA in combining with SNODAS. The combination of SNODAS and UA yielded higher variance and more biases than just using SNODAS. After determining not to include UA, the years of data for 2021 and 2022 weren't downloaded and that's why 2021 and 2022 aren't displayed in Figure B.1. Daymet exhibits an even higher median value of the errors and has a large spread of values. The standard deviations of the errors for Daymet are generally the highest out of the three models. The high median value and large standard deviations of the errors were enough to not include Daymet in the ensemble model. SNODAS displays a median of the errors that is almost centered at zero throughout the years since 2004 and displays the smallest standard deviation of the errors.

Although Daymet and UA weren't included in this ensemble approach, they might be able to provide some insight at certain locations. The need to explore historical trends of each data source at each location could be used in the future. Daymet also provided temperature and other climate variables that could be used in the modelling process.

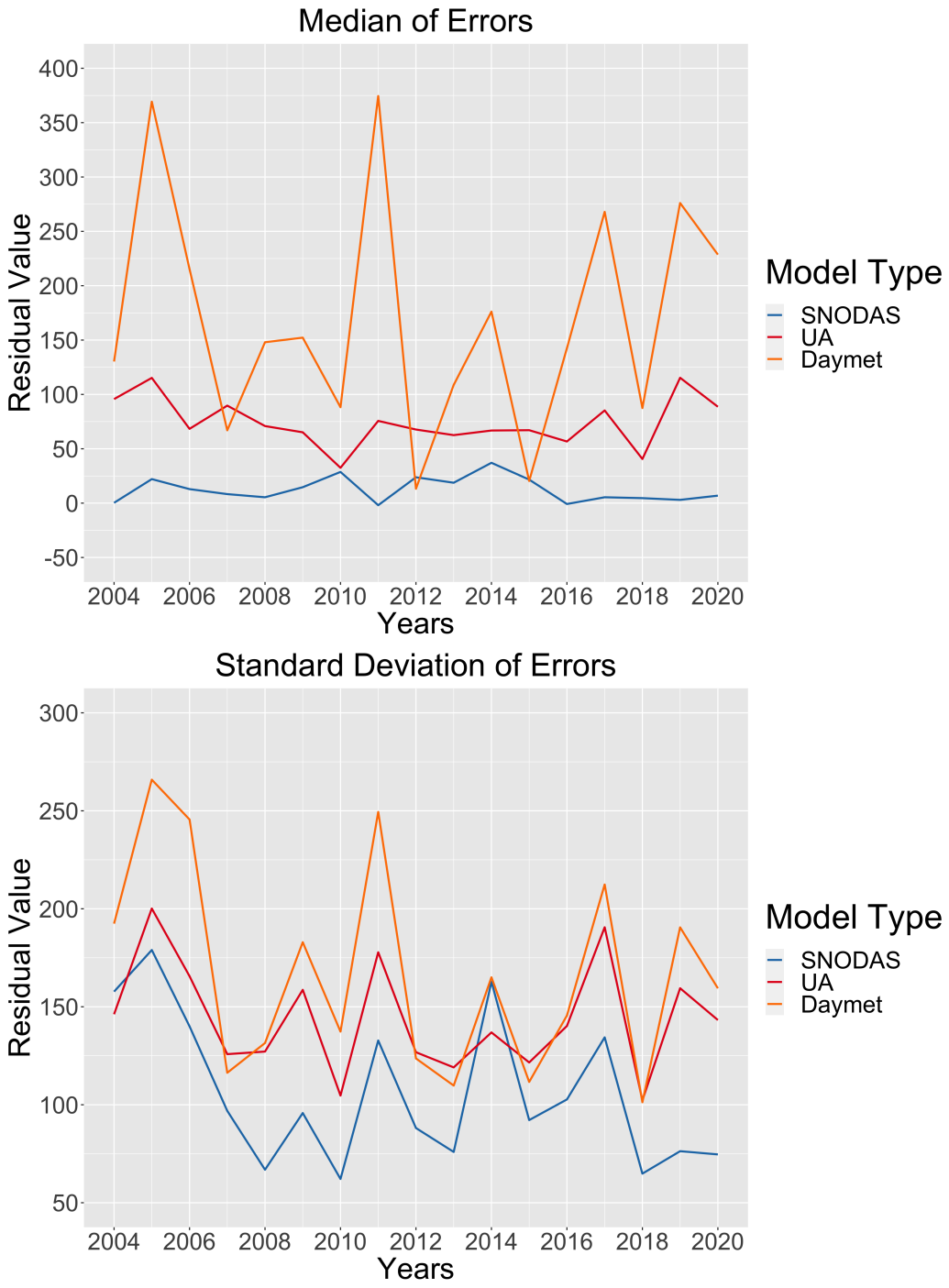


Fig. B.1: (Top to Bottom) These graphics display the Median and Standard Deviations of the errors of each data source (SNODAS, University of Arizona, and Daymet) from 2004 to 2020.

APPENDIX C

Model Limitations

A few different machine learning algorithms were explored before deciding on using a Generalized Additive Model (GAM). Random forests, support vector machines, and linear models, and spatial variograms were explored. One big advantage of random forests is the capability to explore the variable importance plot. This feature is often used to identify key variables in predicting the variable of interest or SWE in this case. Support Vector machines can provide more insight by looking at the partial dependence plots. These two techniques were utilized and shown in Figure C.1 and supported variables like annual precipitation, and elevation that are commonly used in modelling snow-related variables. The partial dependence plot was used to look at the general trend of a variable for each year. Figure C.1 shows the partial dependence curve since 2004 and each year follows a similar trend for each year. This consistent trend over the years helped us incorporate annual precipitation in predicting SWE.

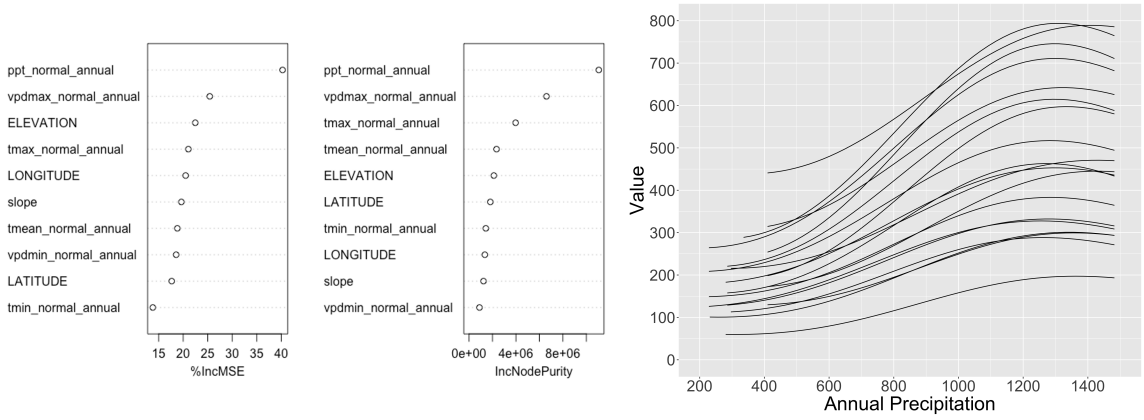


Fig. C.1: (Left:) The variable importance plot of the random forest for the year 2015. (Right:) A partial dependence plot of Annual precipitation. Each line represents a partial dependence curve for each year from 2004 to 2022.

Each year's random forest produced a similar variable importance plot with annual precipitation, elevation, vapor pressure, and max temperature were important variables but often in a different order. Linear models were also used but weren't as capable in capturing the complexity that of SWE in Utah. SNODAS provides the most accurate predictive estimates with the smallest variance. It struggles to explain the variability of SWE in Utah. Figure C.2 displays the variogram of SNODAS errors and the standard deviations of the errors from a ten-fold cross validation of a random forest and GAM gridded product.

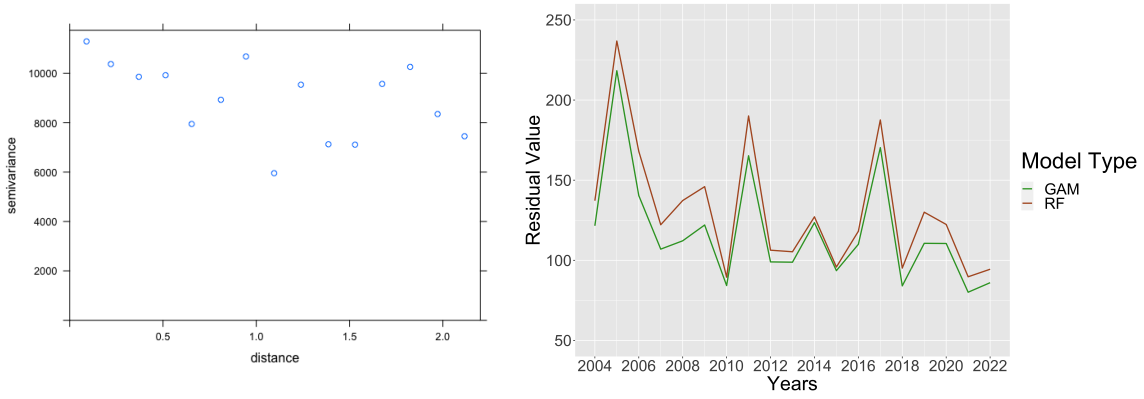


Fig. C.2: (Left to Right) The variogram of the errors of SNODAS in Utah and the standard deviations of the errors of the ten-fold cross validation from the GAM and random forest models since 2004.

Lastly, spatial auto correlation of SNODAS errors were explored used a variogram and the variogram is decreasing. This means that the errors of SNODAS are random and provides more evidence to utilize SNODAS in the ensemble process. The standard deviations of the cross-fold validation of the random forest are slightly larger than the GAM. The GAM was used because of its general acceptance to model climate and spatial variables and capacity to use splines on sphere. However, future work could entail exploring other random forest models because of their performance.