Using observations to update a Bayesian snow model

IKT441-G: ICT Seminar 2, Data Mining

Anders Refsdal Olsen

Mikael Paavola

Nicolas Anderson

Department of Information and Communication Technology

Faculty of Engineering and Science

University of Agder, 2018

Abstract

In this paper, the relation between observations and predictions in snow accumulation is explored using a Dynamic Bayesian network (DBN). By utilizing DBN, the group was able to create a generic model for snow accumulation and further add observations in order to minimize the error rate. A trend was discovered and a regression based on the results was calculated and plotted to graphically show the discovered trend. These findings provide evidence for the existence of a correlation between the amount of observations and prediction error rate.

Keywords: Hydrology Prediction and observation, Dynamics Bayesian Model(DBN), Snow water equivalents (SWE), Snow accumulation and melt, snow modelling, measurement

Preface

Using observations to update a Bayesian snow model, is part of a collaborative student project between the University of Agder (UiA), Norway and the University of Pittsburgh (Pitt), USA for the course *IKT441 Data Mining* lectured at UiA. At UiA, 6 students have been divided into two groups, namely group 1 for Bayesian snow modeling network utilizing noisy data [1], and group 2 for using observations to update a Bayesian snow model. Group 1 includes Eivind Lindseth, Erik Mathisen, Halvor Songøygard Smørvik. While group 2 consists of Anders Refsdal Olsen, Mikael Paavola and Nicolas Anderson. Both groups worked together to replicate the model from Mathuessen and Granmos previous research [2]. When the original research was confirmed, the two groups separated and started new research on their own. The project was naturally limited, due to the domain knowledge of the research members which are ICT students with no hydrology background. The entire project period was limited to one half semester (march to May 2018). This project is suggested and supervised by Prof. Ole-Christoffer Granmo and Dr. Bernt Viggo Matheussen. We would like to thank them both for their help and and useful advice to get this project going.

Anders Refsdal Olsen, Mikael Paavola, Nicolas Anderson
Grimstad, June 1, 2018

Contents

Abstract **2**

Preface **3**

1 Introduction **1**

1.1 Background 1

1.2 Hypotheses 1

1.3 Research Question 2

1.4 Assumptions and Limitations 2

1.5 Outline 3

2 Theoretical Background **4**

2.1 Snow Water Equivalent (SWE) 4

2.2 Study Site and Observed data 4

2.3 State of the Art 4

3 Method **6**

3.1 Software Tools 6

3.1.1 GeNIe and SMILE 6

3.1.2 Angular 6

3.2 Dynamics Bayesian Network (DBN) 7

3.3 Network Configuration 7

3.4 Finding a prediction 9

3.5 Simulation score metric 9

4 Results **11**

4.1 No observations 13

4.2 One observation 14

4.3 Monthly observations 15

4.4 Weekly observations 16
4.5 Scenario comparison 17

5 Discussion and Conclusion **19**

6 Future Work **21**

Chapter 1

Introduction

1.1 Background

The understanding of uncertainty in hydrological predictions is a vital key for hydropower operations and water resources management. In addition, it helps to reduce the operating costs, failure rates, ensure the quality of the product. In hydrology prediction, there are many variables to consider, some being precipitation and air temperature. Forecasting is industry standard, however, these are not always accurate and do tend to deviate from the observed weather to some degree. To compensate for this, power companies use real-world observations and different measurement tools in order adjust their models on the weather. The problem gets more complicated in cold regions, where precipitation is accumulated as snow. To compensate for this, complex models have been created and are used today by power companies. In addition, some of these models require vast knowledge of the environment surrounding one particular reservoir.
Following up on B. Matheussen and O. Granmo's article Modeling Snow Dynamics Using a dynamic Bayesian Network, hereafter called MG [2], this project builds on their results and expands upon it. In order to do this, observations are added to the equation. Using the existing measurement system, the group tries to model and find a correlation between the amount of observations and a score of a simulation.

1.2 Hypotheses

In this project, the group is going to either prove or disprove the hypothesis below:

1. There exists a correlation between the amount of observations and the quality of the predictions.

2. It is possible to plot a correlation between the amount of observations and the quality of the predictions using regression.

The first hypothesis is chosen to explore the effect on the uncertainty of the DBN models' predictions after an observation. The second hypothesis is to show the correlation between the number of observations and the RMSD (root mean square deviation). Specifically, by analyzing the graph, one can determine the necessary number of observations to achieve a desired RMSD value.

1.3 Research Question

When predicting the snow accumulations, the cost of operations mainly consists of how many measurements that is captured. Thus, finding how many measurements that is required to maintain a given quality, is therefore interesting in order to optimize the cost of operations. The research presented in this paper, emphasizes a proposal to find the correlation of observations and prediction deviation in snow dynamics model using the dynamic Bayesian network.

1.4 Assumptions and Limitations

Assumptions

This project follows the model proposed in MG [2] and thus shares the same assumptions about the model. Furthermore, a short summary of those assumptions are as follows.

- Sublimation, horizontal mass transport, rainfall and other physical processes are ignored in this model. The model is initialized in the warm season when there is no snow on the ground.
- The group is particularly focusing on modelling snow on a mountain.
- The model uses discrete timesteps with a resolution of 1 day (24 hours).

Limitations

The project had work flow that resulted in some limitations. They are as below:

- The group members has a background in Information and Communications Technology, thus have little prior knowledge in modelling snow dynamics nor about the field of hydrology. Because of this, parts of the project time was used to understand the work done in MG [2]. [1]
- The time allowed for this project was half of the spring semester of 2018.
- There are several methods for calculating how close two plots are to one another, however in this report the group choose to only use Root Mean Square Deviation due to time limitations. The group acknowledges that there exists other methods.

1.5 Outline

This paper is divided into several chapters as below.

- **Chapter 2** - Describes the data used in the research. In addition, some domain specific background is mentioned to help understand the essence of the research.
- **Chapter 3** - Explains how the research was conducted and how mathematical and practical challenges was solved.
- **Chapter 4** - Presents the findings in the research in detail.
- **Chapter 5** - The results are discussed and a conclusion based on the results are concluded.
- **Chapter 6** - Some suggestions for future improvement are listed.

[1] Thanks to Bernt Viggo Matheussen and Ole Christoffer Granmo for spending much time helping the group to better understand the hydrology domain.

Chapter 2

Theoretical Background

As stated in Limitations (Section 1.4), MG [2] was considered the current state of the art, due to time constraints and the groups limited knowledge in the field of hydrology.

2.1 Snow Water Equivalent (SWE)

Snow Water equivalent is a common snowpack measurement. It is the amount of water found in a given snowpack, and can be thought of as the depth of water found if the snowpack were to melt instantaneously. [3]

2.2 Study Site and Observed data

The data for our work is taken from the Mt.Hood field site positioned in northern Oregon, USA (latitude: 45.32, longitude -121.72, elevation 1637 meters above the sea level, site number 651). This site is managed by National Resources Conservation Service and gathers the daily observations data namely precipitation, air temperature and Snow Water Equivalent (SWE). The data is open source and can be available online for academic purposes. This site was chosen for two reasons; first, it has almost steady winter snow pack. Second, it is situated at fairly high elevation. According to work in MG [2], the daily change in SWE would be considered as Gaussian. The data set is large enough as it involves the simulation period from 1 October 1989, until 30 September 2012. [4]

2.3 State of the Art

In MG [2], a novel snow accumulation and melt model as a dynamic Bayesian network (DBN) was developed. Uncertainty was explicitly encoded and DBN was trained by

using Monte Carlo (MC) analysis, performed with deterministic hydrology model (HM) under a broad range of reasonable parameter configurations. Then, the trained DBN was checked against field observations of precipitation. The results showed that this DBN model can be applied to reason about uncertainty, without resampling from the deterministic model. Briefly, the DBN's ability to reproduce the mean of the observations was alike to what could be achieved with the HM, but with a more realistic description of uncertainty. Without calibrating, this DBN gave good results with a correlation of 0.93 between the mean of the simulated data and observations. These outcomes signify that hybrids of classical deterministic hydrology models and DBNs may provide new solutions to estimation of uncertainty in hydrological predictions.

Ultimately the model can predict the amount of Snow Water Equivelant accumulation on a given day using data such as temperature, precipitation and knowledge of the previous days.

Chapter 3

Method

3.1 Software Tools

Several tools and techniques was used in order to obtain the results described in Chapter 4. In this chapter, the tools and its topology that was essential to the results is described in detail.

3.1.1 GeNIe and SMILE

GeNIe and SMILE are two components that helps in the creation of Bayesian network. SMILE is a set of C++ libraries that enables the creating and simulation of advanced Bayesian networks, including dynamic Bayesian Networks (DBN). GeNIe is a graphical component that allows interaction with SMILE models through a graphical interface. The two components are maintained by BayesFusion LLC [5], but originally created by Marek J. Druzdzel in his paper from 1999 about SMILE and GeNIe [6].

3.1.2 Angular

Angular [7] is a JavaScript-based; open-source; front-end web application framework. This software is usually used to create front-end web applications that interacts with a backend api. However, in this project, the group decided to create a tool that allowed for rapid plotting and in an easy way, read the results from simulations. Furthermore, a tool called Plotly [8] was used in order to create interactive plots with explanatory colors. [1]

[1] The complete Angular module can be viewed at Github [9]

3.2 Dynamics Bayesian Network (DBN)

Dynamic Bayesian networks (DBN) was utilized in this project in order to enable a Bayesian network to preserve previous observations and probabilities automatically. Therefore it is vital to have DBN that over time learns how much snow that is accumulated in the selected regions. As in MG [2], the DBN was applied in this project almost identically. More details into network configuration would be further explained in Section 3.3.

3.3 Network Configuration

In this section, the topology of the Bayesian network is described. The network contains the same nodes as in MG [2]. However, some values are different as seen in this section.

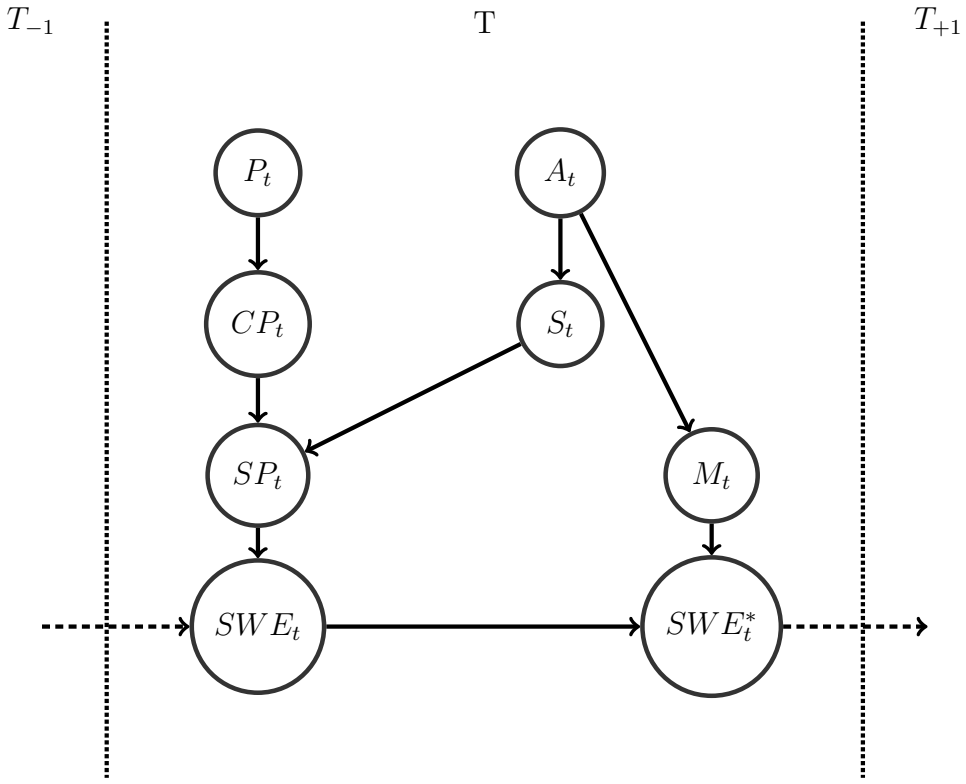

Figure 3.1: Schematics of the Bayesian Network used to model snow accumulation and melt in this report.

The Figure 3.1 was created using SWE steps of 50mm.

- P_t: This precipitation node has 7 steps, yielding the range from 0mm to 300mm. Further, not a single date involve more than 300mm of precipitation in the data set.

- CP_t: This node is the correct precipitation factor that uses a uniform distribution from 1 to 1.6 to the precipitation values. According to the group's experimentation discovery, 500mm is enough as maximum. This node has 11 steps.

- A_t: This node is air temperature and have a temperature range from 0° C to 20° C. In the model 3.1, 0° C is the freezing point and hence all temperatures in minus range counts as zero in this model. Consequently, all precipitation that counts as zero would be automatically assumed as snow. This node has a step of 2° C.

- S_t: Consists of two output values. When the temperature is 0° C or below than that, snow is anticipated.

- M_t: The melt node that shows the potential melt in mm. The CP_t of this node includes the melting probabilities. See Equation 3.3.

- SP_t: SWE accumulation node. Depending upon the precipitation and temperature, this node would add SWE from CP_t node in case of snow, otherwise nothing would be added. See Equation 3.2.

- SWE_t: The SWE node in the current time step that adds SWE from SWE_{t-1} and SP_t nodes. See Equation 3.1.

- SWE_t^*: The expected SWE node in next time step where the melt is subtracted from SWE_t. This is the final result of the node.

- SWE_{t-1}: The start value node where SWE will be maintained as it is added to the SWE_t node.

$$SWE_t = SWE_{t-1} + SP_t - M_t \tag{3.1}$$

As seen in Equation 3.1, SWE_t uses the previous observation (or initial conditions if t=1) combined with the calculated snow precipitation and melting in order to calculate the accumulated snow in the given timestep.

$$SP_t = \begin{cases} Prcp_t \cdot PCorr & if T_{airt,t} < 0.0 \\ (Prcp_t - (0.5 \cdot T_{airt,t} \cdot Prcp_t)) \cdot PCorr & if 0.0 \leq T_{airt,t} < 2.0 \\ 0.0 & if T_{airt,t} \geq 2.0 \end{cases} \tag{3.2}$$

The Snow Melt (M_t) can be computed as a linear function in terms of air temperature ($T_{airt,t}$) and the degree-day factor (C_x) as in the Equation 3.3. Here the snow availability constrains the amount of melt. In addition, the melt can only happen when the air temperature is higher than a certain threshold (T_{thres}).

$$M_t = \begin{cases} C_x \cdot T_{airt,t} & if(T_{airt,t} > T_{thres}) \wedge (SWE_{t-a} > C_x \cdot T_{airt,t}) \\ SWE_{t-1} & if(T_{airt,t} > T_{thres}) \wedge (C_x \cdot T_{airt,t} \geq SWE_{t-1}) \\ 0.0 & if T_{airt,t} \leq T_{thres} \end{cases} \tag{3.3}$$

In Equation 3.3, the amount of melt is expressed as a set of function with different conditionals.

For each node in Figure 3.1, the state space of the associated hydrological variable is described in the discrete domain, with causal relationships among the variables can be mentioned as conditional probability tables (CP_t). Elaborating the details of CP_t of the node are part of modelling and therefore Group 1 would explain about them in their report exclusively. [1]

3.4 Finding a prediction

In order to find a prediction for what the simulation shows, it is simply not enough to use the value of highest probability. Therefore, statistical expectation is used.

$$\mu_t = \sum_{x \in SWE_t} x \cdot P(SWE_t = x) \tag{3.4}$$

In Equation 3.4, the prediction (expectation) for the given time step is calculated. In the equation, x is discrete level of SWE in a given time step. With this, a predicted value is possible to obtain for each time step and it is possible to make a predicted plot across the entire data set.

3.5 Simulation score metric

Visually, it is easy to compare two series plots to one another, but a mathematical comparison is needed to be able to compare several simulations against each other. There are several methods of mathematically two plots against each other. However, the one used

in this project is called Root Mean Square Deviation (RMSD).

$$RMSD = \sqrt{\frac{\sum_{t=1}^{T}(\hat{P}_t - P_t)^2}{T}} \tag{3.5}$$

In equation 3.5, the RMSD is calculated using the predicted value that is calculated for each Timestep (T) in Figure 3.1. With such a metric, it is possible to give a simulation of predictions a score that may be used to compare with other simulations. With this score, represented as a single finite value it is possible to simply check which simulation has the lowest score.

$$BestSimulation = MIN(RMSD(Simulation_x), RMSD(Simulation_y)) \tag{3.6}$$

Equation 3.6 shows how to compare two simulations using equation 3.5, individually on two simulations and select the one with lowest value.

Chapter 4

Results

In this chapter the group presents the results from various simulations. The results contains 4 different scenarios:

- **No observations** - A scenario where no adjustments is feeded to the network. (Section 4.1)
- **One observation** - A scenario where there are only one observation feeded to the network on the first day each year. (Section 4.2)
- **Monthly observations** - A scenario where one observation is feeded at the start of each month. (Section 4.3)
- **Weekly observations** - A scenario where one observation is feeded at the start of each week. (Section 4.4)

All of the scenarios have a range of 23 years and a resolution containing days. In each of the scenarios, the entire range is examined along with the first season (first 365 days/timesteps). In the following charts in Section 4.1; 4.2; 4.3; 4.4, the X-Axis represents the timesteps (T) and the Y axis is the snow water equivalent discrete level (SWE).
In the following sections the figures are linked in such a manner.

- The Figures 4.1; 4.5; 4.9; 4.13, shows a probability chart across all of the seasons.
- The Figures 4.2; 4.6; 4.10; 4.14, shows how the predictions vs the observed value accumulation of snow during the entire simulation.
- The Figures 4.3; 4.7; 4.11; 4.15, shows a probability chart that is centered around season 1 (first 365 timesteps).

- The Figures 4.4; 4.8; 4.12; 4.16, shows how the predictions vs the observed value accumulation of snow centered around season 1 (first 365 timesteps).

In Section 4.5, the 4 scenarios are compared. In addition a regression along with the results are plotted to show a trend.

4.1 No observations

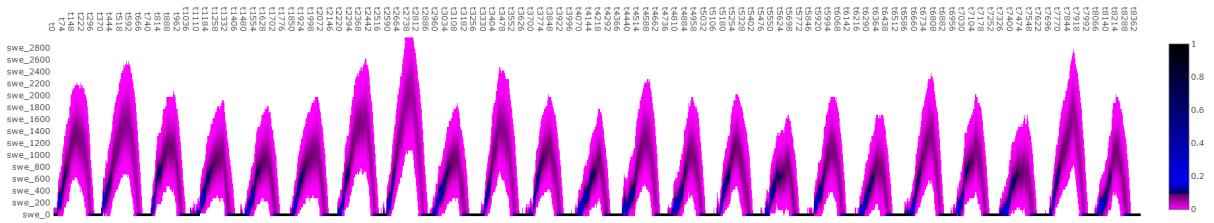

Figure 4.1: Probability distribution from all seasons without observations

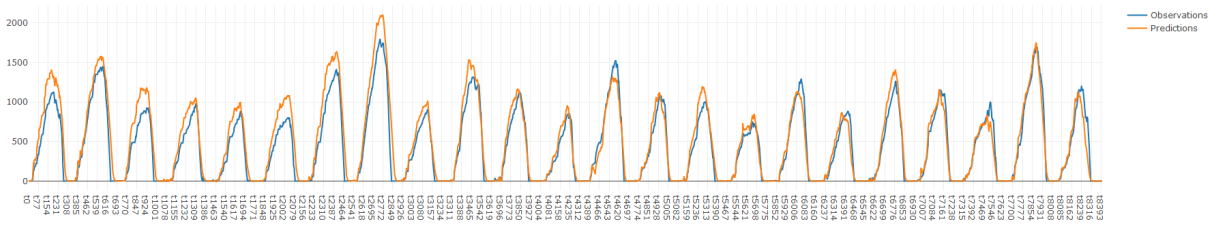

Figure 4.2: Observations vs prediction from all seasons without observations

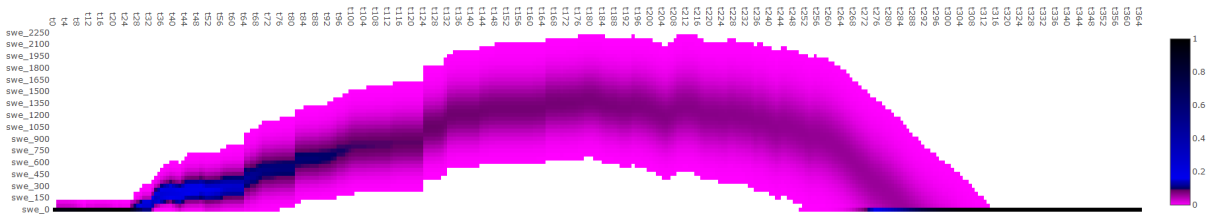

Figure 4.3: Probability distribution from season 1 without observations

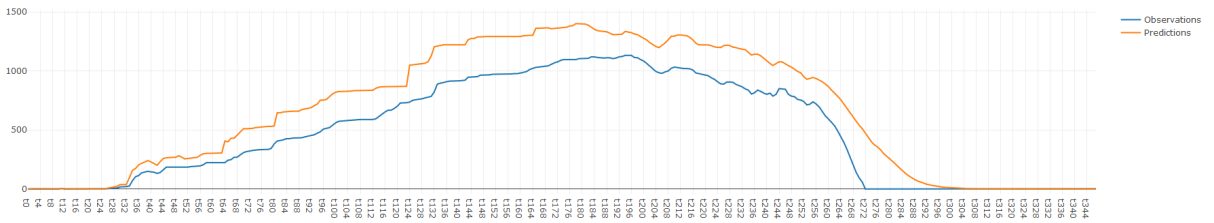

Figure 4.4: Observations vs prediction from one season without observations

4.2 One observation

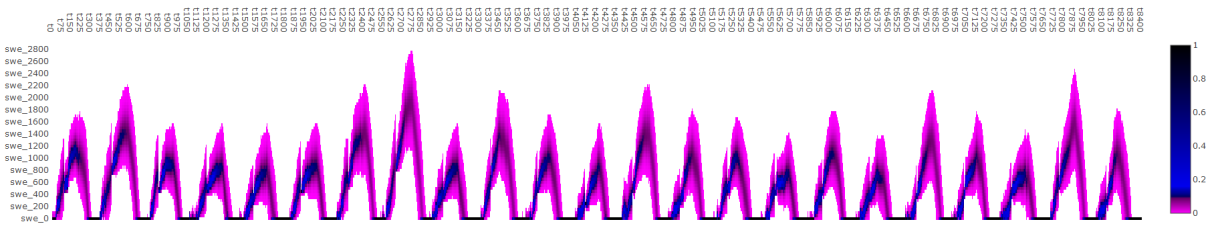

Figure 4.5: Probability distribution from all seasons with one observation each season

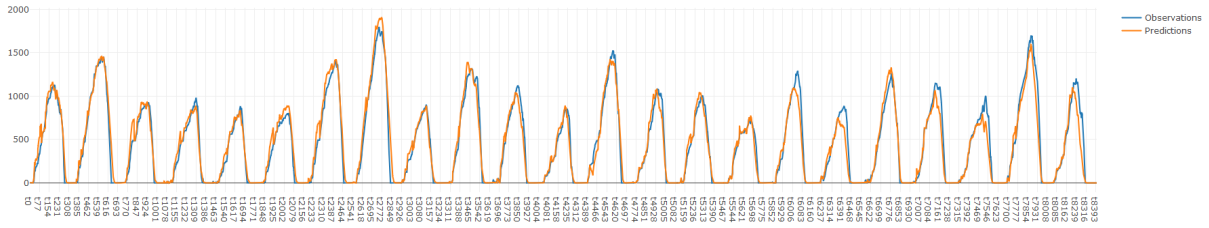

Figure 4.6: Observations vs prediction from all seasons with one observation

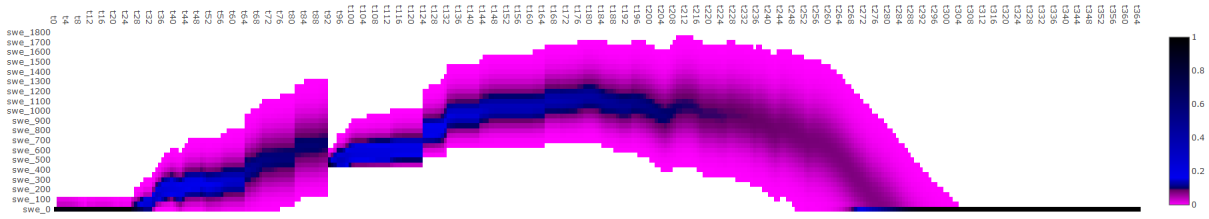

Figure 4.7: Probability distribution from season 1 with one observation

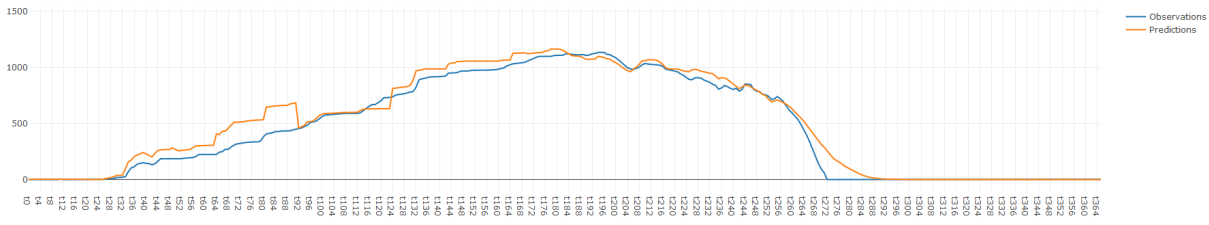

Figure 4.8: Observations vs prediction from season with one observation

4.3 Monthly observations

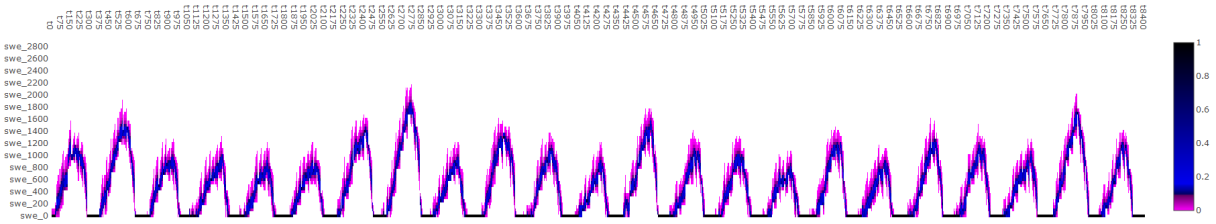

Figure 4.9: Probability distribution from all seasons with one observation each month

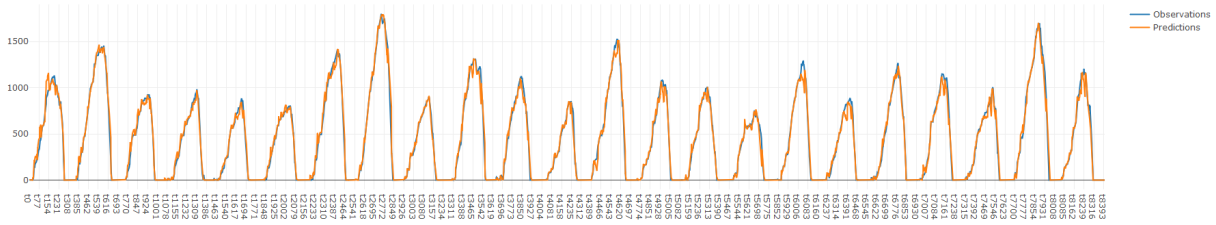

Figure 4.10: Observations vs prediction from all seasons with one observation each month

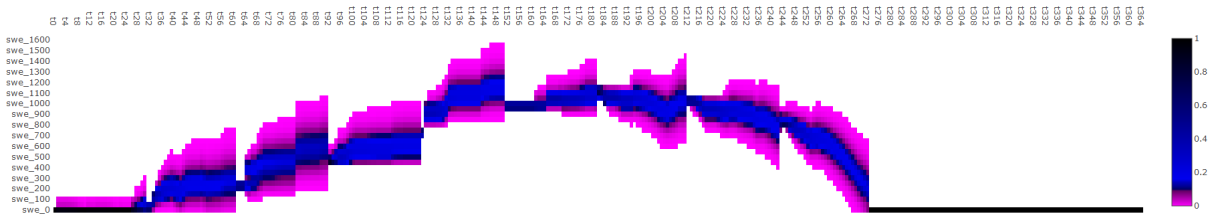

Figure 4.11: Probability distribution from season 1 with one observation each month

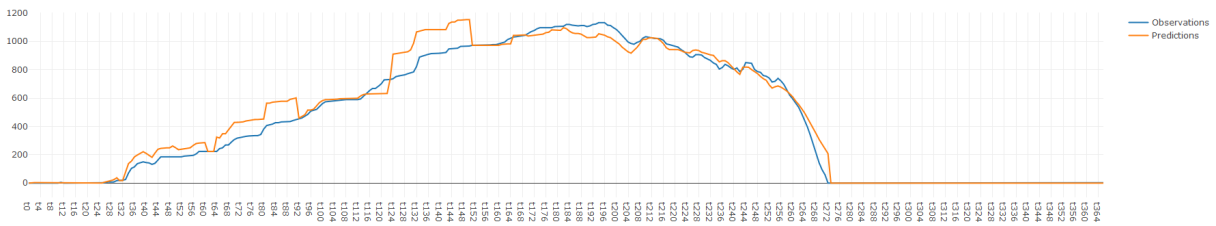

Figure 4.12: Observations vs prediction from one season with one observation each month

4.4 Weekly observations

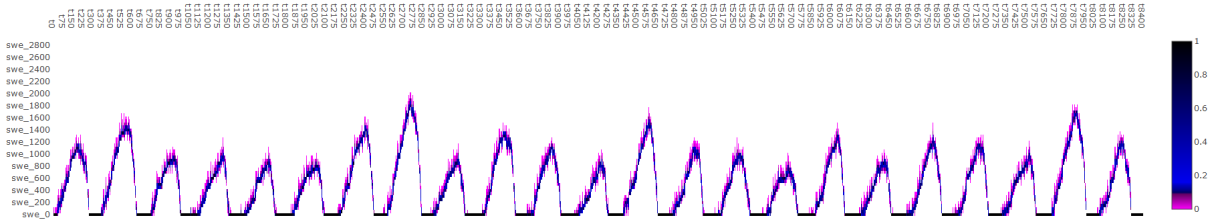

Figure 4.13: Probability distribution from all seasons with one observation each week

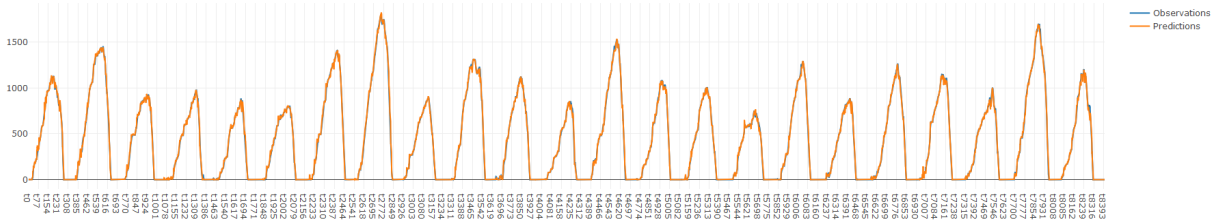

Figure 4.14: Observations vs prediction from all seasons with one observation each week

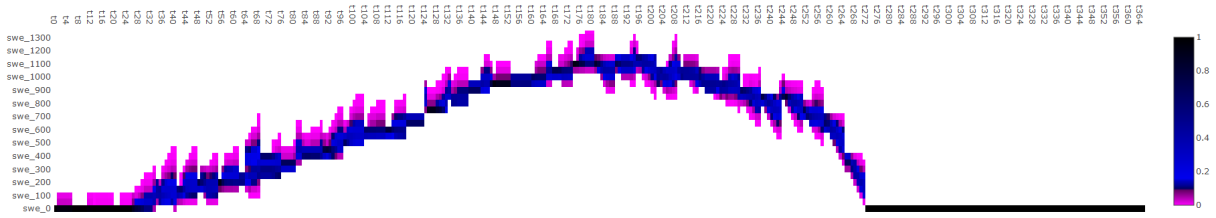

Figure 4.15: Probability distribution from season 1 with one observation each week

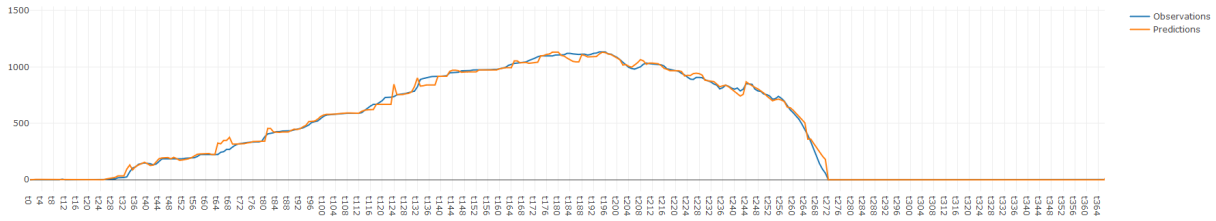

Figure 4.16: Observations vs prediction from one season with one observation each week

4.5 Scenario comparison

	Observations	RMSD
Scenario 1	0	143.86
Scenario 2	23	102.36
Scenario 3	276	51.88
Scenario 4	1196	20.78

Table 4.1: The amount of observations and the calculated Root Mean Square Deviation/Error in each scenario, during all seasons.

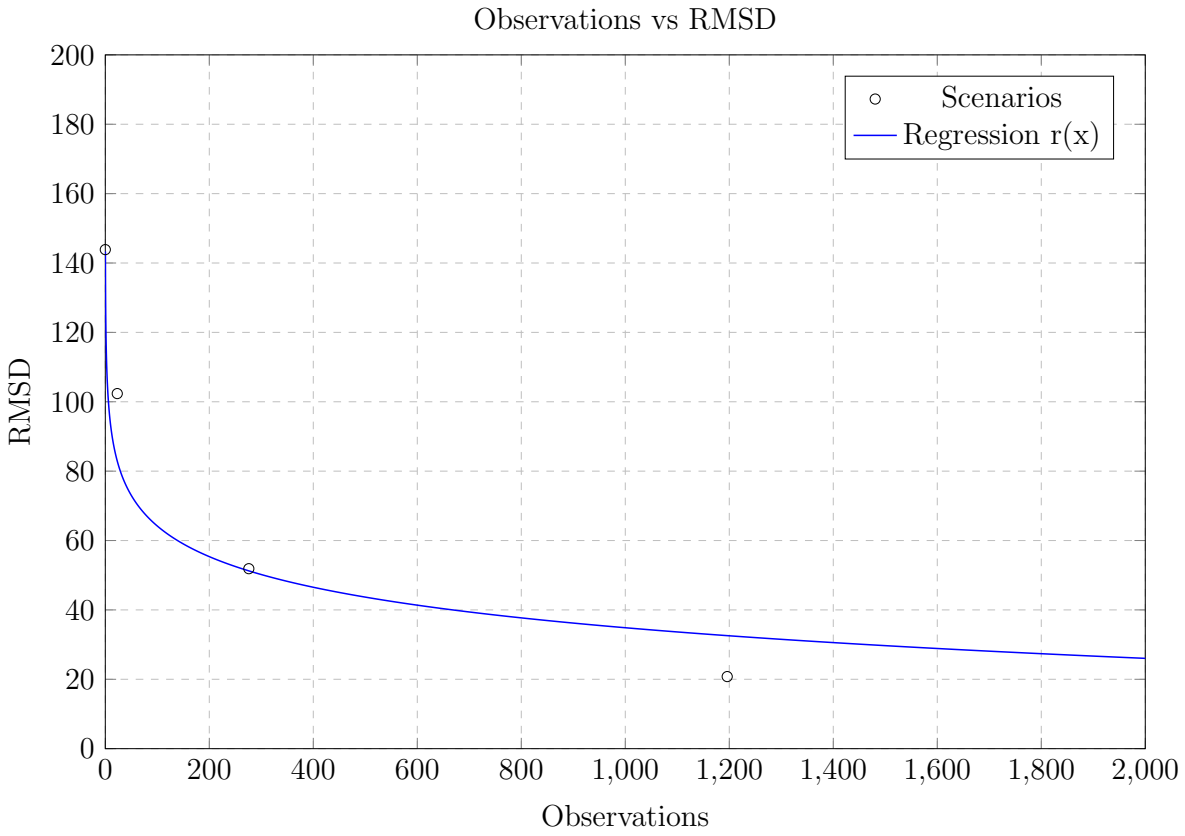

Figure 4.17: The correlation between observations and RMSD represented using a graph and a logarithmic regression function with a correlation coefficient (r) of -0.966358.

Something to note is that in order to create a logarithmic regression line of the data, a zero observation is not possible. The zero observation was therefore adjusted to 0.1 in

order to create something close the same value.

$$(122.826) + (-12.7344 \cdot ln(x)) \tag{4.1}$$

Equation 4.1 shows how the regression is plotted. The constants are precalculated using logarithmic regression.

Chapter 5

Discussion and Conclusion

The results from Chapter 4 showed a clear trend between the RMSD value and the number of observations. This can be seen visually by comparing the graphs for no observations, with those of 1 observation per year to weekly observations. This was confirmed by the decreasing RMSD value. Our results provide evidence for our hypotheses. And it was possible to create a graph that shows a correlation between the number of observations and the RMSD value. This can be used to weigh the cost of doing some number of observations against the gain in accuracy.

It is important to note that the graph was made using only 4 data points. It shows a tendency, but more tests and data points should be applied to make the graph more accurate. Very few observations are needed to get an improvement on the RMSD value. However, as more observations are added, the rate of RMSD improvement decreases. This leaves some interesting questions, like what is the sweet spot for the amount of observations? How many observations should be done before the cost of operations makes it invaluable?

The model needs only temperature and precipitation to begin predicting SWE, making it a generic model. Observations can be added to it to make it more specialized and tuned to some environment, increasing the quality, but in turn, makes the model no longer that generic.

The plot 4.17 shows that the simulation that has more observations yield the smaller RMSD value. The smaller RMSD means better quality of the predictions. This proves our first hypothesis. The regression plot itself provides significant evidence of our second hypothesis. That brings to our research question which is to find the correlation of observations and prediction deviation. Finally, the result briefly suggests that by adding observations to a model, the quality of predictions increases. In turn, that makes it easier

to plan in order to get the approximate amount of measurements to fit a desired quality.

Chapter 6

Future Work

During the work with this paper, the following future work research topics emerged as interesting points to dive further into.

- Enabling the snow dynamics to work backwards in time. This would remove the sharp cuts from the graph on an observation and make model learn ”ahead of time”. One early assumption is that this enables an even better RMSD on the predictions.
- Given more time, it would have been interesting to do more scenarios than just four. This could have improved the RMSD regression line. Maybe in a future project this could be done and maybe even some new discoveries would emerge.
- An interesting aspect of the simulation is that it’s resolution is of days. Given the existence of the data, it could be more practical for shorter simulations to use an even smaller resolution of i.e. hours.

Bibliography

[1] E. Lindseth, E. Mathisen, and H. Smørvik, *Bayesian snow modeling network utilizing noisy data*, [Report from students at University of Agder; accessed 31. May 2018], May 2018.

[2] B. V. Matheussen and O.-C. Granmo, "Modeling Snow Dynamics Using a Bayesian Network," *SpringerLink*, pp. 382–393, Jun. 2015. DOI: 10.1007/978-3-319-19066-2_37.

[3] *What is Snow Water Equivalent? | NRCS Oregon*, [Online; accessed 22. May 2018], May 2018. [Online]. Available: https://www.nrcs.usda.gov/wps/portal/nrcs/detail/or/snow/?cid=nrcs142p2_046155.

[4] *MT HOOD TEST SITE SNOTEL Data | NRCS Oregon*, [Online; accessed 14. April 2018], Apr. 2018. [Online]. Available: https://www.nrcs.usda.gov/wps/portal/nrcs/detail/or/home/?cid=nrcs142p2_046290.

[5] *BayesFusion, LLC*, [Online; accessed 21. May 2018], May 2018. [Online]. Available: https://www.bayesfusion.com.

[6] M. J. Druzdzel, "Smile: Structural modeling, inference, and learning engine and genie: A development environment for graphical decision-theoretic models," in *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence*, ser. AAAI '99/IAAI '99, Orlando, Florida, USA: American Association for Artificial Intelligence, 1999, pp. 902–903, ISBN: 0-262-51106-1. [Online]. Available: http://dl.acm.org/citation.cfm?id=315149.315504.

[7] *Angular*, [Online; accessed 4. April 2018], Apr. 2018. [Online]. Available: https://angular.io.

[8] *Plotly - Modern Visualization for the Data Era*, [Online; accessed 4. April 2018], Apr. 2018. [Online]. Available: https://plot.ly.

[9] *andersro93/School.ICT441.HydroPlotter*, [Online; accessed 21. May 2018], May 2018. [Online]. Available: `https://github.com/andersro93/School.ICT441.HydroPlotter`.